

MARKERS FOR BREAST CANCER

CROSS REFERENCE TO RELATED APPLICATIONS

5 [0001] The subject application claims priority to and benefit of USSN 60/740,971 MARKERS FOR BREAST CANCER by Cox et al., filed November 29, 2005. This application also claims priority to and benefit of USSN 60/781,483 MARKERS FOR BREAST CANCER by Cox et al., filed March 10, 2006. Each of these prior applications are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

10 [0002] Breast cancer, like other common cancers, shows familial clustering. Numerous epidemiological studies have demonstrated that, overall, the disease is approximately twice as common in first degree relatives of breast cancer patients¹. Family studies, and particularly twin studies, suggest that most if not all of this clustering has a
15 genetic basis^{2,3}. For example, Peto and Mack³ estimated that the risk of breast cancer in the MZ twin of an affected woman was approximately four-fold greater than the risk to a sister of a case.

[0003] Several breast cancer susceptibility genes have already been identified, most importantly *BRCA1* and *BRCA2*. Mutations in these genes confer a high risk of breast
20 cancer (of the order of 65% and 45%, respectively, by age 70)⁴. Mutation screening of population-based series of breast cancer cases has shown that only about 15% of the familial risk of breast cancer can be explained by mutations in these genes^{5,6}. The other known breast cancer susceptibility genes (*TP53*, *PTEN*, *ATM*, *CHEK2*) make only small contributions to the familial risk (because the predisposing mutations are rare and/or confer
25 only small risks). In total therefore, the known breast cancer susceptibility genes have been estimated to account for no more than 20% of the familial risk⁷.

[0004] Genetic variation in risk may result from rare highly-penetrant mutations (such as those in *BRCA1* and *BRCA2*) or from variants conferring more moderate risks. Several lines evidence suggest strongly that high penetrance mutations are not major
30 contributors to the residual familial risk of breast cancer. Firstly, mutation screening of multiple case families has found that the large majority of cases with a very strong family history (for example four or more affected relatives) harbor mutations in *BRCA1* or

*BRCA2*⁸. Secondly, despite extensive efforts over the past nine years, genetic linkage studies have not identified any further linked loci^{9,10}. Thirdly, segregation analyses of large series of breast cancer families have found, after adjusting for *BRCA1* and *BRCA2*, no evidence for a further major dominant breast cancer susceptibility allele^{11,12}. In the largest such analysis, Antoniou et al.¹³ found that the most parsimonious model for breast cancer was a polygenic model, equivalent to a large number of loci of small effect combining multiplicatively.

[0005] While the above analyses suggest that several low penetrance breast cancer susceptibility genes remain to be detected, the precise number of such genes is unknown.

10 Moreover, in the prior art, it is unclear whether such susceptibility alleles are common or rare in the population. The subject application focuses on alleles that are relatively common (frequencies greater than 5%) and identification of such loci is performed herein on a genome-wide basis.

SUMMARY OF THE INVENTION

15 [0006] The invention includes the identification of polymorphic loci that are correlated with breast cancer phenotypes, such as susceptibility to breast cancer. Figures 1 and 2 provides descriptions of the phenotypic loci. Figure 1 provides descriptions of preferred phenotypic loci. Accordingly, this invention provides previously unknown correlations between various polymorphisms and breast cancer susceptibility phenotypes.

20 The detection of these polymorphisms (or loci linked thereto), accordingly, provides robust and precise methods and systems for identifying patients that are at risk for breast cancer. In addition, the identification of these polymorphisms provides high-throughput systems and methods for identifying modulators of breast cancer.

[0007] Therefore, in one aspect, the invention provides methods of identifying a breast cancer phenotype for an organism or biological sample derived therefrom. The method includes detecting, in the organism or biological sample, a polymorphism or a locus closely linked thereto, the polymorphism being selected from a polymorphism of Figure 1, wherein the polymorphism is associated with a breast cancer phenotype. The methods further include correlating the polymorphism or locus to the phenotype.

30 [0008] The organism is typically a mammal, and is preferably a human patient, most typically a human female patient (although breast cancer does occur in men, and the

associations noted herein may be applicable to male patients as well). Similarly, the biological sample is typically derived from a mammal, e.g., a human patient, e.g., following appropriate informed consent practices. The methods can be used to detect breast cancer markers in samples taken from human patients, or can be used to detect markers in
5 biological samples (e.g., cells, including primary and cultured cells) derived therefrom.

[0009] The polymorphisms can be detected by any available method, including amplification, hybridization to a probe or array, or the like. In one specific embodiment, Detection includes amplifying the polymorphism, linked locus or a sequence associated therewith (e.g., flanking sequences, transcribed sequences or the like) and detecting the
10 resulting amplicon. For example, in one embodiment, amplifying includes a) admixing an amplification primer or amplification primer pair with a nucleic acid template isolated from the organism or biological sample. The primer or primer pair can be complementary or partially complementary to a region proximal to or including the polymorphism or linked locus, and are capable of initiating nucleic acid polymerization by a polymerase on the
15 nucleic acid template. The primer or primer pair is extended in a DNA polymerization reaction comprising a polymerase and the template nucleic acid to generate the amplicon. In certain aspects, the amplicon is optionally detected by a process that includes hybridizing the amplicon to an array, digesting the amplicon with a restriction enzyme, or real-time PCR analysis. Optionally, the amplicon can be fully or partially sequenced, e.g., by
20 hybridization. Typically, amplification can include performing a polymerase chain reaction (PCR), reverse transcriptase PCR (RT-PCR), or ligase chain reaction (LCR) using nucleic acid isolated from the organism or biological sample as a template in the PCR, RT-PCR, or LCR. Other technologies can be substituted for amplification, e.g., use of branched DNA (bDNA) probes.

[0010] In typical embodiments, the polymorphism or linked locus can include a SNP. Example alleles include those described in Figures 1 and/or 2. Relevant polymorphisms can be those, e.g., of Figure 1 (most preferred) or Figure 2. Preferred polymorphisms include SNPs selected from the group of SNPs described by SNP identification numbers consisting of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013,
30 SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858,

SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418. These identification numbers are Perlegen SNP identification numbers (Perlegen Sciences, Inc. in Mountain View, CA), which are publicly available and can be viewed with considerable associated information at Perlegen(dot)com, by using the company's available genome browser at genome(dot)perlegen(dot)com/browser/index(dot)html. Wild card characters (e.g., "*" symbols) can be added at the beginning of the SNP_ID to identify pertinent information for all alleles of the SNP, e.g., following the complete instructions provided. This database also links to the NCBI genomic database, thereby providing considerable additional information for the relevant genes and polymorphisms. These SNPs include SNPs associated with, e.g., the following genes: FGFR2, A2BP1, TNRC9, H19, FSTL5, LSP1, LOC388927, UNQ9391, HCN1, LOC441192, TNRC9, NR3C2, KIAA0826, FLJ31033, AACS, FRMD4A and SEC31L2 (*See also*, Figure 1). Polymorphisms linked to these genes are, accordingly, also preferred SNPs that can be associated with breast cancer polymorphisms.

[0011] Optionally, and, in certain embodiments, preferably, the method includes detecting polymorphisms in more than one such gene (e.g., in certain convenient applications, several polymorphisms can be detected simultaneously for a single patient to more completely determine or assign the relevant phenotype). Thus, in one aspect, the invention includes detecting a plurality of polymorphisms or linked loci in a plurality of said genes. This can include, e.g., detecting at least one polymorphism for each of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710 and SNP ID 84302, and/or polymorphism in FGFR2, A2BP1, TNRC9, H19, and FSTL5. Similarly, the method can include detecting at least one polymorphism for each of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418, or at least one polymorphism in each of: FGFR2, A2BP1, TNRC9, H19, FSTL5, LSP1, LOC388927, UNQ9391, HCN1, LOC441192, TNRC9, NR3C2, KIAA0826, FLJ31033, AACS,

FRMD4A and SEC31L2. In general, any combination of these or any other polymorphism/gene/locus in the figures herein can be detected, and all such combinations are optionally a feature of the invention, whether listed expressly or not. Probes or primers of the invention useful in detecting the polymorphisms herein can include a nucleotide sequence of a polymorphism of Figures 1 and/or 2, a flanking sequence thereof, or a complementary nucleic acid thereof, or a transcribed product thereof (e.g., a nRNA or mRNA form produced from a genomic sequence, e.g., by transcription or splicing). Polymorphisms can also be detected in a polypeptide sequence, e.g., for any polypeptide sequence transcribed from a given allelic form of a nucleic acid.

10 [0012] In general, any polymorphism that is linked to a QTL can be used as a marker for the QTL. Thus, markers linked to a given polymorphism of the Figures can be used as proxy markers for the given polymorphism. In general, the closer the linkage, the better the marker will be for a QTL/ polymorphism. Thus, desirably, the linked locus can be a closely linked locus that is about 5 cM or less (and, optionally, 1cM or less) from the polymorphism.

[0013] The methods optionally include correlating the polymorphism or linked locus to the breast cancer phenotype by referencing a look up table that comprises correlation information for alleles of the polymorphism or linked locus and the breast cancer phenotype. Databases that are used for this correlation can be heuristic, or otherwise capable of refining correlations based on information obtained by correlating marker-trait information.

[0014] Related compositions are a feature of the invention, e.g., a composition comprising a plurality of marker probes or amplification primers that detect or amplify a plurality of polymorphisms associated with a breast cancer phenotype, e.g., as described herein. The primers/probes can be array based, or free in solution.

[0015] In an additional aspect, methods of identifying a modulator of a breast cancer phenotype are also provided. The methods include contacting a potential modulator to a gene or gene product, e.g., wherein the gene or gene product comprises or is closely linked to a polymorphism described herein (e.g., in figures 1 and/or 2). An effect of the potential modulator on the gene or gene product is detected, thereby identifying whether the potential modulator modulates the phenotype.

[0016] The gene or gene product optionally includes a particular allele of a polymorphism selected from those listed herein, but modulators can also be tested on other alleles to identify modulators that modulate alleles specifically or non-specifically. The effects that can be tested for include any of: (a) increased or decreased expression of the gene or gene product in the presence of the modulator; (b) increased or decreased activity of the gene product in the presence of the modulator; and, (c) an altered expression pattern of the gene or gene product in the presence of the modulator.

[0017] A kit for treatment of a breast cancer phenotype can include a modulator identified by the method and instructions for administering the modulator to a patient to treat the phenotype.

[0018] In addition to the methods noted above, kits and systems for practicing the methods are also a feature of the invention. For example, a system for identifying a breast cancer phenotype for an organism or biological sample derived therefrom are one feature of the invention. The system includes, e.g., a set of marker probes or primers configured to detect at least one allele of one or more polymorphism or linked locus, e.g., where the polymorphism is any polymorphism noted herein, e.g., in Figures 1 or 2. The system optionally additionally includes a detector that is configured to detect one or more signal outputs from the set of marker probes or primers, or an amplicon produced from the set of marker probes or primers, thereby identifying the presence or absence of the allele. System instructions (e.g., software embodied in a computer of the system) that correlate the presence or absence of the allele with a predicted phenotype are typically included as components of the system.

[0019] Systems for screening modulators are also a feature of the invention. The systems can include, e.g., genes linked to a polymorphism herein, or an encoded expression products of the gene. The systems will typically include a detector that measures increased or decreased expression of the gene or gene product in the presence of the modulator; increased or decreased activity of the gene product in the presence of the modulator; or an altered expression pattern of the gene or gene product in the presence of the modulator. The systems can also include fluid handling elements for mixing and aliquotting modulator and/or the gene or product, mixing them, performing laboratory operations (e.g., purification, synthesis, cell culture, etc.). System instructions for recording modulator

effects and, optionally, for selecting modulators are also an optional feature of these systems.

[0020] Kits for performing any of the methods herein are another feature of the invention. Such kits can include probes or amplicons for detecting any polymorphism
5 herein, appropriate packaging materials, and instructions for practicing the methods.

[0021] The polymorphisms and genes, and corresponding marker probes, amplicons or primers described above can be embodied in any system herein, either in the form of physical nucleic acids or polypeptides, or in the form of system instructions that include sequence information for the nucleic acids and polypeptides. For example, the system can
10 include primers or amplicons corresponding to (or that amplify a portion of) a gene or polymorphism described herein, such as SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID
15 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418, and/or FGFR2, A2BP1, TNRC9, H19, FSTL5, LSP1, LOC388927, UNQ9391, HCN1, LOC441192, TNRC9, NR3C2, KIAA0826, FLJ31033, AACS, FRMD4A and SEC31L2. As in the methods above, the set of marker
20 probes or primers optionally detects a plurality of polymorphisms in a plurality of said genes or genetic loci. Thus, for example, the set of marker probes or primers detects at least one polymorphism in each of these polymorphisms or genes, or any other polymorphism, gene or locus in the Figures herein. Any such probe or primer can include a nucleotide sequence of any such polymorphism or gene, or a complementary nucleic acid thereof, or a
25 transcribed product thereof (e.g., a nRNA or mRNA form produced from a genomic sequence, e.g., by transcription or splicing).

[0022] Many alternate variants are embodiments of the invention. For example, the detector typically detects one or more light emission that is indicative of the presence or absence of the allele. The instructions typically comprise at least one look-up table that
30 includes a correlation between the presence or absence of the allele and the phenotype. The system optionally comprises a sample for testing, e.g., a genomic DNA, amplified genomic

DNA, cDNA, amplified cDNA, RNA, or amplified RNA. The sample can be from or derived from a mammal such as a human patient.

[0023] All features of the methods, kits and systems can be used together in combination. For example, systems for detecting modulators can be used for practicing methods of modulator detection. Systems for identifying correlations between breast cancer phenotypes and polymorphisms can be used for practicing the methods herein. Kits can be used for practicing the methods herein. Thus, described features of the systems, methods and kits can be applied to the different systems, methods and kits herein.

BRIEF DESCRIPTION OF THE FIGURES

[0024] Figure 1 is a table of preferred polymorphisms, genes and related information for polymorphisms associated with breast cancer.

[0025] Figure 2 is a table of preferred polymorphisms, genes and related information for polymorphisms associated with breast cancer.

DEFINITIONS

[0026] It is to be understood that this invention is not limited to particular embodiments, which can, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting. As used in this specification and the appended claims, terms in the singular and the singular forms “a,” “an” and “the,” for example, optionally include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to “a probe” optionally includes a plurality of probe molecules; similarly, depending on the context, use of the term “a nucleic acid” optionally includes, as a practical matter, many copies of that nucleic acid molecule. Letter designations for genes or proteins can refer to the gene form and/or the protein form, depending on context. One of skill is fully able to relate the nucleic acid and amino acid forms of the relevant biological molecules by reference to the sequences herein, known sequences and the genetic code.

[0027] Unless otherwise indicated, nucleic acids are written left to right in a 5' to 3' orientation. Numeric ranges recited within the specification are inclusive of the numbers defining the range and include each integer or any non-integer fraction within the defined range. Unless defined otherwise, all technical and scientific terms used herein have the

same meaning as commonly understood by one of ordinary skill in the art to which the invention pertains. Although any methods and materials similar or equivalent to those described herein can be used in the practice for testing of the present invention, the preferred materials and methods are described herein. In describing and claiming the present invention, the following terminology will be used in accordance with the definitions set out below.

[0028] A “phenotype” is a trait or collection of traits that is/are observable in an individual or population. The trait can be quantitative (a quantitative trait, or QTL) or qualitative. For example, susceptibility to breast cancer is a phenotype that can be monitored according to the methods, compositions, kits and systems herein.

[0029] A “breast cancer susceptibility phenotype” is a phenotype that displays a predisposition towards developing breast cancer in an individual. A phenotype that displays a predisposition for breast cancer, can, for example, show a higher likelihood that the cancer will develop in an individual with the phenotype than in members of a relevant general population under a given set of environmental conditions (diet, physical activity regime, geographic location, etc.).

[0030] A “polymorphism” is a locus that is variable; that is, within a population, the nucleotide sequence at a polymorphism has more than one version or allele. The term “allele” refers to one of two or more different nucleotide sequences that occur or are encoded at a specific locus, or two or more different polypeptide sequences encoded by such a locus. For example, a first allele can occur on one chromosome, while a second allele occurs on a second homologous chromosome, e.g., as occurs for different chromosomes of a heterozygous individual, or between different homozygous or heterozygous individuals in a population. One example of a polymorphism is a “single nucleotide polymorphism” (SNP), which is a polymorphism at a single nucleotide position in a genome (the nucleotide at the specified position varies between individuals or populations).

[0031] An allele “positively” correlates with a trait when it is linked to it and when presence of the allele is an indicator that the trait or trait form will occur in an individual comprising the allele. An allele negatively correlates with a trait when it is linked to it and

when presence of the allele is an indicator that a trait or trait form will not occur in an individual comprising the allele.

[0032] A marker polymorphism or allele is “correlated” or “associated” with a specified phenotype (breast cancer susceptibility, etc.) when it can be statistically linked (positively or negatively) to the phenotype. That is, the specified polymorphism occurs more commonly in a case population (e.g., breast cancer patients) than in a control population (e.g., individuals that do not have breast cancer). This correlation is often inferred as being causal in nature, but it need not be—simple genetic linkage to (association with) a locus for a trait that underlies the phenotype is sufficient for correlation/ association to occur.

[0033] A “favorable allele” is an allele at a particular locus that positively correlates with a desirable phenotype, e.g., resistance to breast cancer, e.g., an allele that negatively correlates with predisposition to breast cancer. A favorable allele of a linked marker is a marker allele that segregates with the favorable allele. A favorable allelic form of a chromosome segment is a chromosome segment that includes a nucleotide sequence that positively correlates with the desired phenotype, or that negatively correlates with the unfavorable phenotype at one or more genetic loci physically located on the chromosome segment.

[0034] An “unfavorable allele” is an allele at a particular locus that negatively correlates with a desirable phenotype, or that correlates positively with an undesirable phenotype, e.g., positive correlation to breast cancer susceptibility. An unfavorable allele of a linked marker is a marker allele that segregates with the unfavorable allele. An unfavorable allelic form of a chromosome segment is a chromosome segment that includes a nucleotide sequence that negatively correlates with the desired phenotype, or positively correlates with the undesirable phenotype at one or more genetic loci physically located on the chromosome segment.

[0035] “Allele frequency” refers to the frequency (proportion or percentage) at which an allele is present at a locus within an individual, within a line, or within a population of lines. For example, for an allele “A,” diploid individuals of genotype “AA,” “Aa,” or “aa” have allele frequencies of 1.0, 0.5, or 0.0, respectively. One can estimate the allele frequency within a line or population (e.g., cases or controls) by averaging the allele

frequencies of a sample of individuals from that line or population. Similarly, one can calculate the allele frequency within a population of lines by averaging the allele frequencies of lines that make up the population.

[0036] An individual is “homozygous” if the individual has only one type of allele at a given locus (e.g., a diploid individual has a copy of the same allele at a locus for each of two homologous chromosomes). An individual is “heterozygous” if more than one allele type is present at a given locus (e.g., a diploid individual with one copy each of two different alleles). The term “homogeneity” indicates that members of a group have the same genotype at one or more specific loci. In contrast, the term “heterogeneity” is used to indicate that individuals within the group differ in genotype at one or more specific loci.

[0037] A “locus” is a chromosomal position or region. For example, a polymorphic locus is a position or region where a polymorphic nucleic acid, trait determinant, gene or marker is located. In a further example, a “gene locus” is a specific chromosome location (region) in the genome of a species where a specific gene can be found. Similarly, the term “quantitative trait locus” or “QTL” refers to a locus with at least two alleles that differentially affect the expression or alter the variation of a quantitative or continuous phenotypic trait in at least one genetic background, e.g., in at least one population or progeny.

[0038] A “marker,” “molecular marker” or “marker nucleic acid” refers to a nucleotide sequence or encoded product thereof (e.g., a protein) used as a point of reference when identifying a locus or a linked locus. A marker can be derived from genomic nucleotide sequence or from expressed nucleotide sequences (e.g., from an RNA, nRNA, mRNA, a cDNA, etc.), or from an encoded polypeptide. The term also refers to nucleic acid sequences complementary to or flanking the marker sequences, such as nucleic acids used as probes or primer pairs capable of amplifying the marker sequence. A “marker probe” is a nucleic acid sequence or molecule that can be used to identify the presence of a marker locus, e.g., a nucleic acid probe that is complementary to a marker locus sequence. Nucleic acids are “complementary” when they specifically hybridize in solution, e.g., according to Watson-Crick base pairing rules. A “marker locus” is a locus that can be used to track the presence of a second linked locus, e.g., a linked or correlated locus that encodes or contributes to the population variation of a phenotypic trait. For example, a marker locus can be used to monitor segregation of alleles at a locus, such as a QTL, that are

genetically or physically linked to the marker locus. Thus, a “marker allele,” alternatively an “allele of a marker locus” is one of a plurality of polymorphic nucleotide sequences found at a marker locus in a population that is polymorphic for the marker locus. In one aspect, the present invention provides marker loci correlating with a phenotype of interest, e.g., breast cancer susceptibility/ resistance. Each of the identified markers is expected to be in close physical and genetic proximity (resulting in physical and/or genetic linkage) to a genetic element, e.g., a QTL, that contributes to the relevant phenotype. Markers corresponding to genetic polymorphisms between members of a population can be detected by methods well-established in the art. These include, e.g., PCR-based sequence specific amplification methods, detection of restriction fragment length polymorphisms (RFLP), detection of isozyme markers, detection of allele specific hybridization (ASH), detection of single nucleotide extension, detection of amplified variable sequences of the genome, detection of self-sustained sequence replication, detection of simple sequence repeats (SSRs), detection of single nucleotide polymorphisms (SNPs), or detection of amplified fragment length polymorphisms (AFLPs).

[0039] A “genetic map” is a description of genetic linkage (or association) relationships among loci on one or more chromosomes (or linkage groups) within a given species, generally depicted in a diagrammatic or tabular form. “Mapping” is the process of defining the linkage relationships of loci through the use of genetic markers, populations segregating for the markers, and standard genetic principles of recombination frequency. A “map location” is an assigned location on a genetic map relative to linked genetic markers where a specified marker can be found within a given species. The term “chromosome segment” or designates a contiguous linear span of genomic DNA that resides on a single chromosome. Similarly, a “haplotype” is a set of genetic loci found in the heritable material of an individual or population (the set can be a contiguous or non-contiguous). In the context of the present invention genetic elements such as one or more alleles herein and one or more linked marker alleles can be located within a chromosome segment and are also, accordingly, genetically linked, a specified genetic recombination distance of less than or equal to 20 centimorgan (cM) or less, e.g., 15 cM or less, often 10 cM or less, e.g., about 9, 8, 7, 6, 5, 4, 3, 2, 1, 0.75, 0.5, 0.25, or 0.1 CM or less. That is, two closely linked genetic elements within a single chromosome segment undergo recombination during meiosis with each other at a frequency of less than or equal to about 20%, e.g., about 19%, 18%, 17%,

16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.75%, 0.5%, 0.25%, or 0.1% or less. Once a correlation/ association between a phenotype (e.g., breast cancer predisposition) and a polymorphic locus is identified, e.g., by comparison of a statistical frequency of the locus in cases and controls, any polymorphism that is linked to the associated locus can be used as a proxy marker for the correlated locus.

[0040] A “genetic recombination frequency” is the frequency of a recombination event between two genetic loci. Recombination frequency can be observed by following the segregation of markers and/or traits during meiosis. In the context of this invention, a marker locus is “associated with” another marker locus or some other locus (for example, a breast cancer susceptibility locus), when the relevant loci are part of the same linkage group, due to physical chromosomal association, and are in linkage disequilibrium. This occurs when the marker locus and a linked locus are found together in progeny more frequently than if the loci segregate randomly. Similarly, a marker locus can also be associated with a trait, e.g., a marker locus can be “associated with” a given trait (breast cancer resistance or susceptibility) when the marker locus is in linkage disequilibrium with the trait (this can be detected, e.g., when the marker is found more commonly in case versus control populations). The term “linkage disequilibrium” refers to a non-random segregation of genetic loci or traits (or both). In either case, linkage disequilibrium implies that the relevant loci are within sufficient physical proximity along a length of a chromosome so that they segregate together with greater than random frequency (in the case of co-segregating traits, the loci that underlie the traits are in sufficient proximity to each other). Linked loci co-segregate more than 50% of the time, e.g., from about 51% to about 100% of the time. Advantageously, the two loci are located in close proximity such that recombination between homologous chromosome pairs does not occur between the two loci during meiosis with high frequency, e.g., such that closely linked loci co-segregate at least about 80% of the time, more preferably at least about 85% of the time, still more preferably at least 90% of the time, e.g., 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.5%, 99.75%, or 99.90% or more of the time.

[0041] The phrase “closely linked,” in the present application, means that recombination between two linked loci (e.g., a SNP such as one identified in Figures 1 or 2 (e.g., that is correlated to a breast cancer phenotype by comparison of case and control populations) and a second linked polymorphism) occurs with a frequency of equal to or less

than about 20%. Put another way, the closely (or “tightly”) linked loci co-segregate at least 80% of the time. Marker loci are especially useful in the present invention when they are closely linked to target loci (e.g., QTL for breast cancer, or, alternatively, simply other breast cancer marker loci). The more closely a marker is linked to a target locus, the better an indicator for the target locus that the marker is. Thus, in one embodiment, tightly linked loci such as a marker locus and a second locus display an inter-locus recombination frequency of about 20% or less, e.g., 15% or less, e.g., 10% or less, preferably about 9% or less, still more preferably about 8% or less, yet more preferably about 7% or less, still more preferably about 6% or less, yet more preferably about 5% or less, still more preferably about 4% or less, yet more preferably about 3% or less, and still more preferably about 2% or less. In highly preferred embodiments, the relevant loci (e.g., a marker locus and a target locus such as a QTL) display a recombination frequency of about 1% or less, e.g., about 0.75% or less, more preferably about 0.5% or less, or yet more preferably about 0.25% or less, or still more preferably about 0.1% or less. Two loci that are localized to the same chromosome, and at such a distance that recombination between the two loci occurs at a frequency of less than about 20%, e.g., 15%, more preferably 10% (e.g., about 9 %, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.75%, 0.5%, 0.25%, 0.1% or less) are also said to be “proximal to” each other. When referring to the relationship between two linked genetic elements, such as a genetic element contributing to a trait and a proximal marker, “coupling” phase linkage indicates the state where the “favorable” allele at the trait locus is physically associated on the same chromosome strand as the “favorable” allele of the respective linked marker locus. In coupling phase, both favorable alleles are inherited together by progeny that inherit that chromosome strand. In “repulsion” phase linkage, the “favorable” allele at the locus of interest (e.g., a QTL for breast cancer susceptibility) is physically associated on the same chromosome strand as an “unfavorable” allele at the proximal marker locus, and the two “favorable” alleles are not inherited together (i.e., the two loci are “out of phase” with each other).

[0042] The term “amplifying” in the context of nucleic acid amplification is any process whereby additional copies of a selected nucleic acid (or a transcribed form thereof) are produced. Typical amplification methods include various polymerase based replication methods, including the polymerase chain reaction (PCR), ligase mediated methods such as the ligase chain reaction (LCR) and RNA polymerase based amplification (e.g., by

transcription) methods. An “amplicon” is an amplified nucleic acid, e.g., a nucleic acid that is produced by amplifying a template nucleic acid by any available amplification method (e.g., PCR, LCR, transcription, or the like).

[0043] A “genomic nucleic acid” is a nucleic acid that corresponds in sequence to a heritable nucleic acid in a cell. Common examples include nuclear genomic DNA and amplicons thereof. A genomic nucleic acid is, in some cases, different from a spliced RNA, or a corresponding cDNA, in that the spliced RNA or cDNA is processed, e.g., by the splicing machinery, to remove introns. Genomic nucleic acids optionally comprise non-transcribed (e.g., chromosome structural sequences, promoter regions, enhancer regions, etc.) and/or non-translated sequences (e.g., introns), whereas spliced RNA/cDNA typically do not have introns. A “template genomic nucleic acid” is a genomic nucleic acid that serves as a template in an amplification reaction (e.g., a polymerase based amplification reaction such as PCR, a ligase mediated amplification reaction such as LCR, a transcription reaction, or the like).

[0044] An “exogenous nucleic acid” is a nucleic acid that is not native to a specified system (e.g., a germplasm, cell, individual, etc.), with respect to sequence, genomic position, or both. As used herein, the terms “exogenous” or “heterologous” as applied to polynucleotides or polypeptides typically refers to molecules that have been artificially supplied to a biological system (e.g., a cell, an individual, etc.) and are not native to that particular biological system. The terms can indicate that the relevant material originated from a source other than a naturally occurring source, or can refer to molecules having a non-natural configuration, genetic location or arrangement of parts.

[0045] The term “introduced” when referring to translocating a heterologous or exogenous nucleic acid into a cell refers to the incorporation of the nucleic acid into the cell using any methodology. The term encompasses such nucleic acid introduction methods as “transfection,” “transformation” and “transduction.”

[0046] As used herein, the term “vector” is used in reference to polynucleotides or other molecules that transfer nucleic acid segment(s) into a cell. The term “vehicle” is sometimes used interchangeably with “vector.” A vector optionally comprises parts which mediate vector maintenance and enable its intended use (e.g., sequences necessary for replication, genes imparting drug or antibiotic resistance, a multiple cloning site, operably

linked promoter/enhancer elements which enable the expression of a cloned gene, etc.). Vectors are often derived from plasmids, bacteriophages, or plant or animal viruses. A “cloning vector” or “shuttle vector” or “subcloning vector” contains operably linked parts that facilitate subcloning steps (e.g., a multiple cloning site containing multiple restriction endonuclease sites).

[0047] The term “expression vector” as used herein refers to a vector comprising operably linked polynucleotide sequences that facilitate expression of a coding sequence in a particular host organism (e.g., a bacterial expression vector or a mammalian cell expression vector). Polynucleotide sequences that facilitate expression in prokaryotes typically include, e.g., a promoter, an operator (optional), and a ribosome binding site, often along with other sequences. Eukaryotic cells can use promoters, enhancers, termination and polyadenylation signals and other sequences that are generally different from those used by prokaryotes. In one optional embodiment, a gene corresponding to a loci herein is cloned into an expression vector and expressed, with the gene product(s) to be used in the methods and systems herein for modulator identification.

[0048] A specified nucleic acid is “derived from” a given nucleic acid when it is constructed using the given nucleic acid’s sequence, or when the specified nucleic acid is constructed using the given nucleic acid.

[0049] A “gene” is one or more sequence(s) of nucleotides in a genome that together encode one or more expressed molecule, e.g., an RNA, or polypeptide. The gene can include coding sequences that are transcribed into RNA which may then be translated into a polypeptide sequence, and can include associated structural or regulatory sequences that aid in replication or expression of the gene. Genes of interest in the present invention include those that include or are closely linked to the loci of Figures 1 and/or 2.

[[0050] A “genotype” is the genetic constitution of an individual (or group of individuals) at one or more genetic loci. Genotype is defined by the allele(s) of one or more known loci of the individual, typically, the compilation of alleles inherited from its parents. A “haplotype” is the genotype of an individual at a plurality of genetic loci on a single DNA strand. Typically, the genetic loci described by a haplotype are physically and genetically linked, i.e., on the same chromosome strand.

[0051] A “set” of markers or probes refers to a collection or group of markers or probes, or the data derived therefrom, used for a common purpose, e.g., identifying an individual with a specified phenotype (e.g., breast cancer resistance or susceptibility). Frequently, data corresponding to the markers or probes, or derived from their use, is stored
5 in an electronic medium. While each of the members of a set possess utility with respect to the specified purpose, individual markers selected from the set as well as subsets including some, but not all of the markers, are also effective in achieving the specified purpose.

[0052] A “look up table” is a table that correlates one form of data to another, or one or more forms of data with a predicted outcome to which the data is relevant. For example,
10 a look up table can include a correlation between allele data and a predicted trait that an individual comprising one or more given alleles is likely to display. These tables can be, and typically are, multidimensional, e.g., taking multiple alleles into account simultaneously, and, optionally, taking other factors into account as well, such as genetic background, e.g., in making a trait prediction.

[0053] A “computer readable medium” is an information storage media that can be accessed by a computer using an available or custom interface. Examples include memory (e.g., ROM or RAM, flash memory, etc.), optical storage media (e.g., CD-ROM), magnetic storage media (computer hard drives, floppy disks, etc.), punch cards, and many others that are commercially available. Information can be transmitted between a system of interest
15 and the computer, or to or from the computer to or from the computer readable medium for storage or access of stored information. This transmission can be an electrical transmission, or can be made by other available methods, such as an IR link, a wireless connection, or the like.

[0054] “System instructions” are instruction sets that can be partially or fully
25 executed by the system. Typically, the instruction sets are present as system software.

[0055] A “translation product” is a product (typically a polypeptide) produced as a result of the translation of a nucleic acid. A “transcription product” is a product (e.g., an RNA, optionally including mRNA, or, e.g., a catalytic or biologically active RNA) produced as a result of transcription of a nucleic acid (e.g., a DNA).

[0056] An “array” is an assemblage of elements. The assemblage can be spatially ordered (a “patterned array”) or disordered (a “randomly patterned” array). The array can
30

form or comprise one or more functional elements (e.g., a probe region on a microarray) or it can be non-functional.

[0057] As used herein, the term “SNP” or “single nucleotide polymorphism” refers to a genetic variation between individuals; *e.g.*, a single nitrogenous base position in the DNA of organisms that is variable. As used herein, “SNPs” is the plural of SNP. Of course, when one refers to DNA herein, such reference may include derivatives of the DNA such as amplicons, RNA transcripts thereof, etc.

DETAILED DESCRIPTION

OVERVIEW

10 [0058] The invention includes new correlations between the polymorphisms of Figures 1 and 2 (and genes that include or are proximal to the polymorphisms) and breast cancer predisposition. Certain alleles in, and linked to, these genes or gene products are predictive of the likelihood that an individual possessing the relevant alleles will develop breast cancer. Accordingly, detection of these alleles, by any available method, can be used
15 for diagnostic purposes such as early detection of susceptibility to breast cancer, prognosis for patients that present with breast cancer and in assisting in diagnosis, *e.g.*, where current criteria are insufficient for a definitive diagnosis.

[0059] The identification that the polymorphisms, genes or gene products of Figures 1 or 2, are correlated with breast cancer phenotypes also provides a platform for screening
20 potential modulators of breast cancer disorders. Modulators of the activity of any genes or encoded proteins corresponding to the polymorphisms of Figures 1 and 2, are expected to have an effect on breast cancer. Thus, methods of screening, systems for screening and the like, are features of the invention. Modulators identified by these screening approaches are also a feature of the invention.

25 [0060] Kits for the diagnosis and treatment of breast cancer, *e.g.*, comprising probes to identify relevant alleles, packaging materials, and instructions for correlating detection of relevant alleles to breast cancer are also a feature of the invention. These kits can also include modulators of breast cancer and/or instructions for treating patients using conventional methods.

METHODS OF IDENTIFYING BREAST CANCER PREDISPOSITION

[0061] As noted, the invention provides the discovery that certain genes or other loci of Figure 1 and 2, are linked to breast cancer phenotypes. Thus, by detecting markers (e.g., the SNPs in Figures 1 or 2 or loci closely linked thereto) that correlate, positively or negatively, with the relevant phenotypes, it can be determined whether an individual or population is likely to be comprise these phenotypes. This provides enhanced early detection options to identify patients that are at risk for breast cancer, making it possible, in some cases, to prevent actual development of cancer, e.g., by taking early preventative action (e.g., any existing therapy, including prophylactic surgery, diet, exercise, available medications, etc.). In addition, use of the various markers herein also adds certainty to existing diagnostic techniques for identifying whether a patient is suffering from a particular form of breast cancer. Furthermore, knowledge of whether there is a molecular basis for the disease can also assist in determining patient prognosis, e.g., by providing an indication of how likely it is that a patient can respond to conventional therapy for breast cancer. Disease treatment can also be targeted based on what type of molecular disorder the patient displays.

[0062] Detection methods for detecting relevant alleles can include any available method, e.g., amplification technologies. For example, detection can include amplifying the polymorphism or a sequence associated therewith and detecting the resulting amplicon.

This can include admixing an amplification primer or amplification primer pair with a nucleic acid template isolated from the organism or biological sample (e.g., comprising the SNP or other polymorphism), e.g., where the primer or primer pair is complementary or partially complementary to at least a portion of the gene or tightly linked polymorphism, or to a sequence proximal thereto. The primer is typically capable of initiating nucleic acid polymerization by a polymerase on the nucleic acid template. The primer or primer pair is extended, e.g., in a DNA polymerization reaction (PCR, RT-PCR, etc.) comprising a polymerase and the template nucleic acid to generate the amplicon. The amplicon is detected by any available detection process, e.g., sequencing, hybridizing the amplicon to an array (or affixing the amplicon to an array and hybridizing probes to it), digesting the amplicon with a restriction enzyme (e.g., RFLP), real-time PCR analysis, single nucleotide extension, allele-specific hybridization, or the like.

[0063] The correlation between a detected polymorphism and a trait can be performed by any method that can identify a relationship between an allele and a phenotype. Most typically, these methods involve referencing a look up table that comprises correlations between alleles of the polymorphism and the phenotype. The table can include data for multiple allele-phenotype relationships and can take account of additive or other higher order effects of multiple allele-phenotype relationships, e.g., through the use of statistical tools such as principle component analysis, heuristic algorithms, etc.

[0064] Within the context of these methods, the following discussion first focuses on how markers and alleles are linked and how this phenomenon can be used in the context of methods for identifying breast cancer phenotypes, and then focuses on marker detection methods. Additional sections below discuss data analysis.

Markers, Linkage And Alleles

[0065] In traditional linkage (or association) analysis, no direct knowledge of the physical relationship of genes on a chromosome is required. Mendel's first law is that factors of pairs of characters are segregated, meaning that alleles of a diploid trait separate into two gametes and then into different offspring. Classical linkage analysis can be thought of as a statistical description of the relative frequencies of cosegregation of different traits. Linkage analysis is the well characterized descriptive framework of how traits are grouped together based upon the frequency with which they segregate together. That is, if two non-allelic traits are inherited together with a greater than random frequency, they are said to be "linked." The frequency with which the traits are inherited together is the primary measure of how tightly the traits are linked, i.e., traits which are inherited together with a higher frequency are more closely linked than traits which are inherited together with lower (but still above random) frequency. Traits are linked because the genes which underlie the traits reside near one another on the same chromosome. The further apart on a chromosome the genes reside, the less likely they are to segregate together, because homologous chromosomes recombine during meiosis. Thus, the further apart on a chromosome the genes reside, the more likely it is that there will be a recombination event during meiosis that will result in two genes segregating separately into progeny.

[0066] A common measure of linkage (or association) is the frequency with which traits cosegregate. This can be expressed as a percentage of cosegregation (recombination frequency) or, also commonly, in centiMorgans (cM), which are actually a reciprocal unit of

recombination frequency. The cM is named after the pioneering geneticist Thomas Hunt Morgan and is a unit of measure of genetic recombination frequency. One cM is equal to a 1% chance that a trait at one genetic locus will be separated from a trait at another locus due to recombination in a single generation (meaning the traits segregate together 99% of the time). Because chromosomal distance is approximately proportional to the frequency of recombination events between traits, there is an approximate physical distance that correlates with recombination frequency. For example, in humans, 1 cM correlates, on average, to about 1 million base pairs (1Mbp).

[0067] Marker loci are themselves traits and can be assessed according to standard linkage analysis by tracking the marker loci during segregation. Thus, in the context of the present invention, one cM is equal to a 1% chance that a marker locus will be separated from another locus (which can be any other trait, e.g., another marker locus, or another trait locus that encodes a QTL for breast cancer), due to recombination in a single generation. The markers herein, e.g., those listed in Figure 1 and 2, can correlate with breast cancer. This means that the markers comprise or are sufficiently proximal to a QTL for breast cancer that they can be used as a predictor for the trait itself. This is extremely useful in the context of disease diagnosis.

[0068] The polymorphisms of Figures 1 and 2 have been identified as being more prevalent in case (breast cancer patient) versus control populations. Any marker that is linked to a trait locus of interest (e.g., in the present case, a QTL or identified linked marker locus for breast cancer, e.g., as in Figures 1 and 2) can be used as a marker for that trait. Thus, in addition to the markers noted in Figures 1 and 2, other markers closely linked to the markers itemized in these Figures can also usefully predict the presence of the marker alleles indicated in the figures (and, thus, the relevant phenotypic trait). Such linked markers are particularly useful when they are sufficiently proximal to a given locus so that they display a low recombination frequency with the given locus. In the present invention, such closely linked markers are a feature of the invention. Closely linked loci display a recombination frequency with a given marker of about 20% or less (the given marker is within 20cM of the given marker). Put another way, closely linked loci co-segregate at least 80% of the time. More preferably, the recombination frequency is 10% or less, e.g., 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.5%, 0.25%, or 0.1% or less. In one typical class of embodiments, closely linked loci are within 5 cM or less of each other.

[0069] As one of skill in the art will recognize, recombination frequencies (and, as a result, map positions) can vary depending on the map used (and the markers that are on the map). Additional markers that are closely linked to (e.g., within about 20 cM, or more preferably within about 10 cM, or still more preferably within 5cM of) the markers identified in Figures 1 and 2 may readily be used for identification of QTL for breast cancer predisposition.

[0070] Marker loci are especially useful in the present invention when they are closely linked to target loci (e.g., QTL for breast cancer phenotypes, or, alternatively, simply other marker loci that are, themselves linked to such QTL) that they are being used as markers for. The more closely a marker is linked to a target locus that encodes or affects a phenotypic trait, the better an indicator for the target locus that the marker is (due to the reduced cross-over frequency between the target locus and the marker). Thus, in one embodiment, closely linked loci such as a marker locus and a second locus (e.g., a given marker locus of Figures 1 and 2 and an additional second locus) display an inter-locus cross-over frequency of about 20% or less, e.g., 15% or less, preferably 10% or less, more preferably about 9% or less, still more preferably about 8% or less, yet more preferably about 7% or less, still more preferably about 6% or less, yet more preferably about 5% or less, still more preferably about 4% or less, yet more preferably about 3% or less, and still more preferably about 2% or less. In highly preferred embodiments, the relevant loci (e.g., a marker locus and a target locus such as a QTL) display a recombination a frequency of about 1% or less, e.g., about 0.75% or less, more preferably about 0.5% or less, or yet more preferably about 0.25% or 0.1% or less. Thus, the loci are about 20cM, 19 cM, 18 cM, 17 cM, 16 cM, 15 cM, 14 cM, 13 cM, 12 cM, 11 cM, 10 cM, 9 cM, 8 cM, 7 cM, 6 cM, 5 cM, 4 cM, 3cM, 2cM, 1cM, 0.75 cM, 0.5 cM, 0.25 cM, 0 or .1 cM or less apart. Put another way, two loci that are localized to the same chromosome, and at such a distance that recombination between the two loci occurs at a frequency of less than 20% (e.g., about 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9 %, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, 0.75%, 0.5%, 0.25%, 0.1% or less) are said to be "proximal to" each other. In one aspect, linked markers are within 100 kb (which correlates in humans to about 0.1 cM, depending on local recombination rate), e.g., 50 kb, or even 20 kb or less of each other.

[0071] When referring to the relationship between two genetic elements, such as a genetic element contributing to breast cancer, and a proximal marker, "coupling" phase

linkage indicates the state where the “favorable” allele at the locus is physically associated on the same chromosome strand as the “favorable” allele of the respective linked marker locus. In coupling phase, both favorable alleles are inherited together by progeny that inherit that chromosome strand. In “repulsion” phase linkage, the “favorable” allele at the locus of interest (e.g., a QTL for breast cancer) is physically linked with an “unfavorable” allele at the proximal marker locus, and the two “favorable” alleles are not inherited together (i.e., the two loci are “out of phase” with each other).

[0072] In addition to tracking SNP and other polymorphisms in the genome, and in corresponding expressed nucleic acids and polypeptides, expression level differences between individuals or populations for the gene products of Figures 1 and 2, in either mRNA or protein form, can also correlate to breast cancer. Accordingly, markers of the invention can include any of, e.g.: genomic loci, transcribed nucleic acids, spliced nucleic acids, expressed proteins, levels of transcribed nucleic acids, levels of spliced nucleic acids, and levels of expressed proteins.

Marker Amplification Strategies

[0073] Amplification primers for amplifying markers (e.g., marker loci) and suitable probes to detect such markers or to genotype a sample with respect to multiple marker alleles, are a feature of the invention. In Figures 1 and 2, specific loci for amplification are provided, along with known flanking sequences in the design of such primers. For example, primer selection for long-range PCR is described in USSN 10/042,406, filed Jan. 9, 2002 and USSN 10/236,480, filed Sep. 5, 2002; for short-range PCR, USSN 10/341,832, filed Jan. 14, 2003 provides guidance with respect to primer selection. Also, there are publicly available programs such as “Oligo” available for primer design. With such available primer selection and design software, the publicly available human genome sequence and the polymorphism locations as provided in Figures 1 and 2, one of skill can construct primers to amplify the SNPs of the present invention. Further, it will be appreciated that the precise probe to be used for detection of a nucleic acid comprising a SNP (e.g., an amplicon comprising the SNP) can vary, e.g., any probe that can identify the region of a marker amplicon to be detected can be used in conjunction with the present invention. Further, the configuration of the detection probes can, of course, vary. Thus, the invention is not limited to the sequences recited herein.

[0074] Indeed, it will be appreciated that amplification is not a requirement for marker detection—for example, one can directly detect unamplified genomic DNA simply by performing a Southern blot on a sample of genomic DNA. Procedures for performing Southern blotting, standard amplification (PCR, LCR, or the like) and many other nucleic acid detection methods are well established and are taught, e.g., in Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2000 (“Sambrook”); Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2002) (“Ausubel”) and PCR Protocols A Guide to Methods and Applications (Innis *et al.* eds) Academic Press Inc. San Diego, CA (1990) (Innis).

[0075] Separate detection probes can also be omitted in amplification/detection methods, e.g., by performing a real time amplification reaction that detects product formation by modification of the relevant amplification primer upon incorporation into a product, incorporation of labeled nucleotides into an amplicon, or by monitoring changes in molecular rotation properties of amplicons as compared to unamplified precursors (e.g., by fluorescence polarization).

[0076] Typically, molecular markers are detected by any established method available in the art, including, without limitation, allele specific hybridization (ASH), detection of single nucleotide extension, array hybridization (optionally including ASH), or other methods for detecting single nucleotide polymorphisms (SNPs), amplified fragment length polymorphism (AFLP) detection, amplified variable sequence detection, randomly amplified polymorphic DNA (RAPD) detection, restriction fragment length polymorphism (RFLP) detection, self-sustained sequence replication detection, simple sequence repeat (SSR) detection, single-strand conformation polymorphisms (SSCP) detection, isozyme marker detection, northern analysis (where expression levels are used as markers), quantitative amplification of mRNA or cDNA, or the like. While the exemplary markers provided in the Figures are SNP markers, any of the aforementioned marker types can be employed in the context of the invention to identify linked loci that correlate with a breast cancer phenotype.

Example Techniques For Marker Detection

[0077] The invention provides molecular markers that comprise or are linked to QTL for breast cancer phenotypes. The markers find use in disease predisposition diagnosis, prognosis, treatment, etc. It is not intended that the invention be limited to any particular method for the detection of these markers.

[0078] Markers corresponding to genetic polymorphisms between members of a population can be detected by numerous methods well-established in the art (e.g., PCR-based sequence specific amplification, restriction fragment length polymorphisms (RFLPs), isozyme markers, northern analysis, allele specific hybridization (ASH), array based hybridization, amplified variable sequences of the genome, self-sustained sequence replication, simple sequence repeat (SSR), single nucleotide polymorphism (SNP), random amplified polymorphic DNA ("RAPD") or amplified fragment length polymorphisms (AFLP). In one additional embodiment, the presence or absence of a molecular marker is determined simply through nucleotide sequencing of the polymorphic marker region. Any of these methods are readily adapted to high throughput analysis.

[0079] Some techniques for detecting genetic markers utilize hybridization of a probe nucleic acid to nucleic acids corresponding to the genetic marker (e.g., amplified nucleic acids produced using genomic DNA as a template). Hybridization formats, including, but not limited to: solution phase, solid phase, mixed phase, or *in situ* hybridization assays are useful for allele detection. An extensive guide to the hybridization of nucleic acids is found in Tijssen (1993) Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes Elsevier, New York, as well as in Sambrook, Berger and Ausubel.

[0080] For example, markers that comprise restriction fragment length polymorphisms (RFLP) are detected, e.g., by hybridizing a probe which is typically a sub-fragment (or a synthetic oligonucleotide corresponding to a sub-fragment) of the nucleic acid to be detected to restriction digested genomic DNA. The restriction enzyme is selected to provide restriction fragments of at least two alternative (or polymorphic) lengths in different individuals or populations. Determining one or more restriction enzyme that produces informative fragments for each allele of a marker is a simple procedure, well known in the art. After separation by length in an appropriate matrix (e.g., agarose or polyacrylamide) and transfer to a membrane (e.g., nitrocellulose, nylon, etc.), the labeled

probe is hybridized under conditions which result in equilibrium binding of the probe to the target followed by removal of excess probe by washing.

[0081] Nucleic acid probes to the marker loci can be cloned and/or synthesized.

Any suitable label can be used with a probe of the invention. Detectable labels suitable for

5 use with nucleic acid probes include, for example, any composition detectable by spectroscopic, radioisotopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels include biotin for staining with labeled streptavidin conjugate, magnetic beads, fluorescent dyes, radiolabels, enzymes, and colorimetric labels. Other labels include ligands which bind to antibodies labeled with
10 fluorophores, chemiluminescent agents, and enzymes. A probe can also constitute radiolabelled PCR primers that are used to generate a radiolabelled amplicon. Labeling strategies for labeling nucleic acids and corresponding detection strategies can be found, e.g., in Haugland (2003) Handbook of Fluorescent Probes and Research Chemicals Ninth Edition by Molecular Probes, Inc. (Eugene OR). Additional details regarding marker
15 detection strategies are found below.

Amplification-based Detection Methods

[0082] PCR, RT-PCR and LCR are in particularly broad use as amplification and amplification-detection methods for amplifying nucleic acids of interest (e.g., those comprising marker loci), facilitating detection of the nucleic acids of interest. Details

20 regarding the use of these and other amplification methods can be found in any of a variety of standard texts, including, e.g., Sambrook, Ausubel, and Berger. Many available biology texts also have extended discussions regarding PCR and related amplification methods. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using
25 reverse transcriptase and a polymerase ("Reverse Transcription-PCR, or "RT-PCR"). *See also*, Ausubel, Sambrook and Berger, *above*. These methods can also be used to quantitatively amplify mRNA or corresponding cDNA, providing an indication of expression levels of mRNA that correspond to a gene product corresponding to the genes or gene products of Figures 1 and 2 in an individual. Differences in expression levels for these
30 genes between individuals, families, lines and/or populations can be used as markers for breast cancer phenotypes.

Real Time Amplification/ Detection Methods

[0083] In one aspect, real time PCR or LCR is performed on the amplification mixtures described herein, e.g., using molecular beacons or TaqMan™ probes. A molecular beacon (MB) is an oligonucleotide or PNA which, under appropriate hybridization conditions, self-hybridizes to form a stem and loop structure. The MB has a label and a quencher at the termini of the oligonucleotide or PNA; thus, under conditions that permit intra-molecular hybridization, the label is typically quenched (or at least altered in its fluorescence) by the quencher. Under conditions where the MB does not display intra-molecular hybridization (e.g., when bound to a target nucleic acid, e.g., to a region of an amplicon during amplification), the MB label is unquenched. Details regarding standard methods of making and using MBs are well established in the literature and MBs are available from a number of commercial reagent sources. *See also*, e.g., Leone et al. (1995) "Molecular beacon probes combined with amplification by NASBA enable homogenous real-time detection of RNA." Nucleic Acids Res. 26:2150-2155; Tyagi and Kramer (1996) "Molecular beacons: probes that fluoresce upon hybridization" Nature Biotechnology 14:303-308; Blok and Kramer (1997) "Amplifiable hybridization probes containing a molecular switch" Mol Cell Probes 11:187-194; Hsuih et al. (1997) "Novel, ligation-dependent PCR assay for detection of hepatitis C in serum" J Clin Microbiol 34:501-507; Kostrikis et al. (1998) "Molecular beacons: spectral genotyping of human alleles" Science 279:1228-1229; Sokol et al. (1998) "Real time detection of DNA:RNA hybridization in living cells" Proc. Natl. Acad. Sci. U.S.A. 95:11538-11543; Tyagi et al. (1998) "Multicolor molecular beacons for allele discrimination" Nature Biotechnology 16:49-53; Bonnet et al. (1999) "Thermodynamic basis of the chemical specificity of structured DNA probes" Proc. Natl. Acad. Sci. U.S.A. 96:6171-6176; Fang et al. (1999) "Designing a novel molecular beacon for surface-immobilized DNA hybridization studies" J. Am. Chem. Soc. 121:2921-2922; Marras et al. (1999) "Multiplex detection of single-nucleotide variation using molecular beacons" Genet. Anal. Biomol. Eng. 14:151-156; and Vet et al. (1999) "Multiplex detection of four pathogenic retroviruses using molecular beacons" Proc. Natl. Acad. Sci. U.S.A. 96:6394-6399. Additional details regarding MB construction and use is found in the patent literature, e.g., USP 5,925,517 (July 20, 1999) to Tyagi et al. entitled "Detectably labeled dual conformation oligonucleotide probes, assays and kits;" USP 6,150,097 to Tyagi et al (November 21, 2000) entitled "Nucleic acid detection probes having non-FRET fluorescence quenching and kits and assays including such probes" and

USP 6,037,130 to Tyagi et al (March 14, 2000), entitled "Wavelength-shifting probes and primers and their use in assays and kits."

[0084] PCR detection and quantification using dual-labeled fluorogenic oligonucleotide probes, commonly referred to as "TaqMan™" probes, can also be performed according to the present invention. These probes are composed of short (e.g., 20-25 base) oligodeoxynucleotides that are labeled with two different fluorescent dyes. On the 5' terminus of each probe is a reporter dye, and on the 3' terminus of each probe a quenching dye is found. The oligonucleotide probe sequence is complementary to an internal target sequence present in a PCR amplicon. When the probe is intact, energy transfer occurs between the two fluorophores and emission from the reporter is quenched by the quencher by FRET. During the extension phase of PCR, the probe is cleaved by 5' nuclease activity of the polymerase used in the reaction, thereby releasing the reporter from the oligonucleotide-quencher and producing an increase in reporter emission intensity. Accordingly, TaqMan™ probes are oligonucleotides that have a label and a quencher, where the label is released during amplification by the exonuclease action of the polymerase used in amplification. This provides a real time measure of amplification during synthesis. A variety of TaqMan™ reagents are commercially available, e.g., from Applied Biosystems (Division Headquarters in Foster City, CA) as well as from a variety of specialty vendors such as Biosearch Technologies (e.g., black hole quencher probes). Further details regarding dual-label probe strategies can be found, e.g., in WO92/02638.

[0085] Other similar methods include e.g. fluorescence resonance energy transfer between two adjacently hybridized probes, e.g., using the "LightCycler®" format described in U.S. 6,174,670.

Array- Based Marker Detection

[0086] Array- based detection can be performed using commercially available arrays, e.g., from Affymetrix (Santa Clara, CA) or other manufacturers. Reviews regarding the operation of nucleic acid arrays include Sapolsky et al. (1999) "High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays." Genetic Analysis: Biomolecular Engineering 14:187-192; Lockhart (1998) "Mutant yeast on drugs" Nature Medicine 4:1235-1236; Fodor (1997) "Genes, Chips and the Human Genome." FASEB Journal 11:A879; Fodor (1997) "Massively Parallel Genomics." Science 277: 393-395; and Chee et al. (1996) "Accessing Genetic Information with High-Density DNA

Arrays.” Science 274:610-614. Array based detection is a preferred method for identification markers of the invention in samples, due to the inherently high-throughput nature of array based detection.

[0087] A variety of probe arrays have been described in the literature and can be

5 used in the context of the present invention for detection of markers that can be correlated to the phenotypes noted herein. For example, DNA probe array chips or larger DNA probe array wafers (from which individual chips would otherwise be obtained by breaking up the wafer) are used in one embodiment of the invention. DNA probe array wafers generally comprise glass wafers on which high density arrays of DNA probes (short segments of
10 DNA) have been placed. Each of these wafers can hold, for example, approximately 60 million DNA probes that are used to recognize longer sample DNA sequences (e.g., from individuals or populations, e.g., that comprise markers of interest). The recognition of sample DNA by the set of DNA probes on the glass wafer takes place through DNA hybridization. When a DNA sample hybridizes with an array of DNA probes, the sample
15 binds to those probes that are complementary to the sample DNA sequence. By evaluating to which probes the sample DNA for an individual hybridizes more strongly, it is possible to determine whether a known sequence of nucleic acid is present or not in the sample, thereby determining whether a marker found in the nucleic acid is present. One can also use this approach to perform ASH, by controlling the hybridization conditions to permit single
20 nucleotide discrimination, e.g., for SNP identification and for genotyping a sample for one or more SNPs. Arrays provide one convenient embodiment for detecting multiple polymorphic markers simultaneously (or in series). For example, breast cancer susceptibility detection arrays can be constructed in which any or all of the polymorphisms noted herein (or polymorphisms linked thereto) are detected simultaneously to assign a
25 breast cancer susceptibility phenotype. Of course, any detection technology (PCR, LCR, real-time PCR, etc.) can similarly be used, e.g., with multiplex amplification/ detection reactions, or simply by running several separate reactions, e.g., simultaneously or in series.

[0088] The use of DNA probe arrays to obtain allele information typically involves the following general steps: design and manufacture of DNA probe arrays, preparation of
30 the sample, hybridization of sample DNA to the array, detection of hybridization events and data analysis to determine sequence. Preferred wafers are manufactured using a process

adapted from semiconductor manufacturing to achieve cost effectiveness and high quality, and are available, e.g., from Affymetrix, Inc of Santa Clara, California.

[0089] For example, probe arrays can be manufactured by light-directed chemical synthesis processes, which combine solid-phase chemical synthesis with photolithographic fabrication techniques as employed in the semiconductor industry. Using a series of photolithographic masks to define chip exposure sites, followed by specific chemical synthesis steps, the process constructs high-density arrays of oligonucleotides, with each probe in a predefined position in the array. Multiple probe arrays can be synthesized simultaneously on a large glass wafer. This parallel process enhances reproducibility and helps achieve economies of scale.

[0090] Once fabricated, DNA probe arrays can be used to obtain data regarding presence and/or expression levels for markers of interest. The DNA samples may be tagged with biotin and/or a fluorescent reporter group by standard biochemical methods. The labeled samples are incubated with an array, and segments of the samples bind, or hybridize, with complementary sequences on the array. The array can be washed and/or stained to produce a hybridization pattern. The array is then scanned and the patterns of hybridization are detected by emission of light from the fluorescent reporter groups. Additional details regarding these procedures are found in the examples below. Because the identity and position of each probe on the array is known, the nature of the DNA sequences in the sample applied to the array can be determined. When these arrays are used for genotyping experiments, they can be referred to as genotyping arrays. As already noted, the genotype of any or all of the polymorphisms noted herein, e.g., in Figure 1 and/or Figure 2 can be detected, e.g., to assign a breast cancer predisposition phenotype.

[0091] The nucleic acid sample to be analyzed is isolated, amplified and, typically, labeled with biotin and/or a fluorescent reporter group. The labeled nucleic acid sample is then incubated with the array using a fluidics station and hybridization oven. The array can be washed and or stained or counter-stained, as appropriate to the detection method. After hybridization, washing and staining, the array is inserted into a scanner, where patterns of hybridization are detected. The hybridization data are collected as light emitted from the fluorescent reporter groups already incorporated into the labeled nucleic acid, which is now bound to the probe array. Probes that most clearly match the labeled nucleic acid produce stronger signals than those that have mismatches. Since the sequence and position of each

probe on the array are known, by complementarity, the identity of the nucleic acid sample applied to the probe array can be identified.

[0092] In one embodiment, two DNA samples may be differentially labeled and hybridized with a single set of the designed genotyping arrays. In this way two sets of data can be obtained from the same physical arrays. Labels that can be used include, but are not limited to, cychrome, fluorescein, or biotin (later stained with phycoerythrin-streptavidin after hybridization). Two-color labeling is described in U.S. Patent No. 6,342,355, incorporated herein by reference in its entirety. Each array may be scanned such that the signal from both labels is detected simultaneously, or may be scanned twice to detect each signal separately.

[0093] Intensity data is collected by the scanner for all the markers for each of the individuals that are tested for presence of the marker. The measured intensities are a measure indicative of the amount of a particular marker present in the sample for a given individual (expression level and/or number of copies of the allele present in an individual, depending on whether genomic or expressed nucleic acids are analyzed). This can be used to determine whether the individual is homozygous or heterozygous for the marker of interest. The intensity data is processed to provide corresponding marker information for the various intensities.

Additional Details Regarding Amplified Variable Sequences, SSR, AFLP
ASH, SNPs and Isozyme Markers

[0094] Amplified variable sequences refer to amplified sequences of the genome which exhibit high nucleic acid residue variability between members of the same species. All organisms have variable genomic sequences and each organism (with the exception of a clone, e.g., a cloned cell) has a different set of variable sequences. Once identified, the presence of specific variable sequence can be used to predict phenotypic traits. Preferably, DNA from the genome serves as a template for amplification with primers that flank a variable sequence of DNA. The variable sequence is amplified and then sequenced.

[0095] Alternatively, self-sustained sequence replication can be used to identify genetic markers. Self-sustained sequence replication refers to a method of nucleic acid amplification using target nucleic acid sequences which are replicated exponentially, *in vitro*, under substantially isothermal conditions by using three enzymatic activities involved in retroviral replication: (1) reverse transcriptase, (2) Rnase H, and (3) a DNA-dependent

RNA polymerase (Guatelli et al. (1990) Proc Natl Acad Sci USA 87:1874). By mimicking the retroviral strategy of RNA replication by means of cDNA intermediates, this reaction accumulates cDNA and RNA copies of the original target.

[0096] Amplified fragment length polymorphisms (AFLP) can also be used as genetic markers (Vos et al. (1995) Nucl Acids Res 23:4407). The phrase “amplified fragment length polymorphism” refers to selected restriction fragments which are amplified before or after cleavage by a restriction endonuclease. The amplification step allows easier detection of specific restriction fragments. AFLP allows the detection large numbers of polymorphic markers and has been used for genetic mapping (Becker et al. (1995) Mol Gen Genet 249:65; and Meksem et al. (1995) Mol Gen Genet 249:74).

[0097] Allele-specific hybridization (ASH) can be used to identify the genetic markers of the invention. ASH technology is based on the stable annealing of a short, single-stranded, oligonucleotide probe to a completely complementary single-strand target nucleic acid. Detection may be accomplished via an isotopic or non-isotopic label attached to the probe.

[0098] For each polymorphism, two or more different ASH probes are designed to have identical DNA sequences except at the polymorphic nucleotides. Each probe will have exact homology with one allele sequence so that the range of probes can distinguish all the known alternative allele sequences. Each probe is hybridized to the target DNA. With appropriate probe design and hybridization conditions, a single-base mismatch between the probe and target DNA will prevent hybridization. In this manner, only one of the alternative probes will hybridize to a target sample that is homozygous or homogenous for an allele. Samples that are heterozygous or heterogeneous for two alleles will hybridize to both of two alternative probes.

[0099] ASH markers are used as dominant markers where the presence or absence of only one allele is determined from hybridization or lack of hybridization by only one probe. The alternative allele may be inferred from the lack of hybridization. ASH probe and target molecules are optionally RNA or DNA; the target molecules are any length of nucleotides beyond the sequence that is complementary to the probe; the probe is designed to hybridize with either strand of a DNA target; the probe ranges in size to conform to variously stringent hybridization conditions, etc.

[0100] PCR allows the target sequence for ASH to be amplified from low concentrations of nucleic acid in relatively small volumes. Otherwise, the target sequence from genomic DNA is digested with a restriction endonuclease and size separated by gel electrophoresis. Hybridizations typically occur with the target sequence bound to the surface of a membrane or, as described in U.S. Patent 5,468,613, the ASH probe sequence may be bound to a membrane.

[0101] In one embodiment, ASH data are typically obtained by amplifying nucleic acid fragments (amplicons) from genomic DNA using PCR, transferring the amplicon target DNA to a membrane in a dot-blot format, hybridizing a labeled oligonucleotide probe to the amplicon target, and observing the hybridization dots by autoradiography.

[0102] Single nucleotide polymorphisms (SNP) are markers that consist of a shared sequence differentiated on the basis of a single nucleotide. Typically, this distinction is detected by differential migration patterns of an amplicon comprising the SNP on e.g., an acrylamide gel. However, alternative modes of detection, such as hybridization, e.g., ASH, or RFLP analysis or array based detection are also appropriate.

[0103] Isozyme markers can be employed as genetic markers, e.g., to track isozyme markers linked to the markers herein. Isozymes are multiple forms of enzymes that differ from one another in their amino acid, and therefore their nucleic acid sequences. Some isozymes are multimeric enzymes contain slightly different subunits. Other isozymes are either multimeric or monomeric but have been cleaved from the proenzyme at different sites in the amino acid sequence. Isozymes can be characterized and analyzed at the protein level, or alternatively, isozymes which differ at the nucleic acid level can be determined. In such cases any of the nucleic acid based methods described herein can be used to analyze isozyme markers.

25 Additional Details Regarding Nucleic Acid Amplification

[0104] As noted, nucleic acid amplification techniques such as PCR and LCR are well known in the art and can be applied to the present invention to amplify and/or detect nucleic acids of interest, such as nucleic acids comprising marker loci. Examples of techniques sufficient to direct persons of skill through such *in vitro* methods, including the polymerase chain reaction (PCR), the ligase chain reaction (LCR), Q β -replicase amplification and other RNA polymerase mediated techniques (e.g., NASBA), are found in

the references noted above, e.g., Innis, Sambrook, Ausubel, and Berger. Additional details are found in Mullis et al. (1987) U.S. Patent No.4,683,202; Arnheim & Levinson (October 1, 1990) C&EN 36-47; The Journal Of NIH Research (1991) 3, 81-94; (Kwoh et al. (1989) Proc. Natl. Acad. Sci. USA 86, 1173; Guatelli et al. (1990) Proc. Natl. Acad. Sci. USA 87, 1874; Lomell et al. (1989) J. Clin. Chem 35, 1826; Landegren et al., (1988) Science 241, 1077-1080; Van Brunt (1990) Biotechnology 8, 291-294; Wu and Wallace, (1989) Gene 4, 560; Barringer et al. (1990) Gene 89, 117, and Sooknanan and Malek (1995) Biotechnology 13: 563-564. Improved methods of amplifying large nucleic acids by PCR, which is useful in the context of positional cloning of genes linked to the polymorphisms herein (Figures 1 and/ or 2), are further summarized in Cheng et al. (1994) Nature 369: 684, and the references therein, in which PCR amplicons of up to 40 kb are generated. Methods for long-range PCR are disclosed, for example, in U.S. patent application no. 10/042,406, filed January 9, 2002, entitled "Algorithms for Selection of Primer Pairs"; U.S. patent application no. 10/236,480, filed September 9, 2002, entitled "Methods for Amplification of Nucleic Acids"; and U.S. patent no. 6,740,510, issued May 25, 2004, entitled "Methods for Amplification of Nucleic Acids". USSN 10/341,832 (filed 1/14/ 03) also provides details regarding primer picking methods for performing short range PCR.

Detection of protein expression products

[0105] Proteins such as those encoded by the genes noted in Figures 1 and/or 2 are encoded by nucleic acids, including those comprising markers that are correlated to the phenotypes of interest herein. For a description of the basic paradigm of molecular biology, including the expression (transcription and/or translation) of DNA into RNA into protein, *see*, Alberts et al. (2002) Molecular Biology of the Cell, 4th Edition Taylor and Francis, Inc., ISBN: 0815332181 ("Alberts"), and Lodish et al. (1999) Molecular Cell Biology, 4th Edition W H Freeman & Co, ISBN: 071673706X ("Lodish"). Accordingly, proteins corresponding to the genes in Figures 1 and/or 2 can be detected as markers, e.g., by detecting different protein isotypes between individuals or populations, or by detecting a differential presence, absence or expression level of such a protein of interest (e.g., a gene product of Figures 1 and/or 2).

[0106] A variety of protein detection methods are known and can be used to distinguish markers. In addition to the various references noted supra, a variety of protein manipulation and detection methods are well known in the art, including, e.g., those set

forth in R. Scopes, Protein Purification, Springer-Verlag, N.Y. (1982); Deutscher, Methods in Enzymology Vol. 182: Guide to Protein Purification, Academic Press, Inc. N.Y. (1990); Sandana (1997) Bioseparation of Proteins, Academic Press, Inc.; Bollag et al. (1996) Protein Methods, 2nd Edition Wiley-Liss, NY; Walker (1996) The Protein Protocols Handbook Humana Press, NJ, Harris and Angal (1990) Protein Purification Applications: A Practical Approach IRL Press at Oxford, Oxford, England; Harris and Angal Protein Purification Methods: A Practical Approach IRL Press at Oxford, Oxford, England; Scopes (1993) Protein Purification: Principles and Practice 3rd Edition Springer Verlag, NY; Janson and Ryden (1998) Protein Purification: Principles, High Resolution Methods and Applications, Second Edition Wiley-VCH, NY; and Walker (1998) Protein Protocols on CD-ROM Humana Press, NJ; and the references cited therein. Additional details regarding protein purification and detection methods can be found in Satinder Ahuja ed., Handbook of Bioseparations, Academic Press (2000).

[0107] "Proteomic" detection methods, which detect many proteins simultaneously have been described. These can include various multidimensional electrophoresis methods (e.g., 2-d gel electrophoresis), mass spectrometry based methods (e.g., SELDI, MALDI, electrospray, *etc.*), or surface plasmon resonance methods. For example, in MALDI, a sample is usually mixed with an appropriate matrix, placed on the surface of a probe and examined by laser desorption/ionization. The technique of MALDI is well known in the art. See, e.g., U.S. patent 5,045,694 (Beavis et al.), U.S. patent 5,202,561 (Gleissmann et al.), and U.S. Patent 6,111,251 (Hillenkamp). Similarly, for SELDI, a first aliquot is contacted with a solid support-bound (e.g., substrate-bound) adsorbent. A substrate is typically a probe (e.g., a biochip) that can be positioned in an interrogatable relationship with a gas phase ion spectrometer. SELDI is also a well known technique, and has been applied to diagnostic proteomics. See, e.g. Issaq et al. (2003) "SELDI-TOF MS for Diagnostic Proteomics" Analytical Chemistry 75:149A-155A.

[0108] In general, the above methods can be used to detect different forms (alleles) of proteins and/or can be used to detect different expression levels of the proteins (which can be due to allelic differences) between individuals, families, lines, populations, etc. Differences in expression levels, when controlled for environmental factors, can be indicative of different alleles at a QTL for the gene of interest, even if the encoded differentially expressed proteins are themselves identical. This occurs, for example, where

there are multiple allelic forms of a gene in non-coding regions, e.g., regions such as promoters or enhancers that control gene expression. Thus, detection of differential expression levels can be used as a method of detecting allelic differences.

[0109] In other aspect of the present invention, a gene comprising, in linkage disequilibrium with, or under the control of a nucleic acid associated with a breast cancer phenotype may exhibit differential allelic expression. "Differential allelic expression" as used herein refers to both qualitative and quantitative differences in the allelic expression of multiple alleles of a single gene present in a cell. As such, a gene displaying differential allelic expression may have one allele expressed at a different time or level as compared to a second allele in the same cell/tissue. For example, an allele associated with a breast cancer phenotype may be expressed at a higher or lower level than an allele that is not associated with the breast cancer phenotype, even though both are alleles of the same gene and are present in the same cell/tissue. Differential allelic expression and analysis methods are disclosed in detail in U.S. patent application no. 10/438,184, filed May 13, 2003 and U.S. patent application no. 10/845,316, filed May 12, 2004, both of which are entitled "Allele-specific expression patterns." Detection of a differential allelic expression pattern of one or more nucleic acids, or fragments, derivatives, polymorphisms, variants or complements thereof, associated with a breast cancer phenotype is a prognostic and diagnostic for susceptibility/resistance to breast cancer; likewise, detection of a differential allelic expression pattern of one or more nucleic acids, or fragments, derivatives, polymorphisms, variants or complements thereof, associated with a breast cancer phenotype is prognostic for and diagnostic of breast cancer and breast cancer treatment outcomes.

Additional Details regarding Types of Markers Appropriate for Screening

[0110] The biological markers that are screened for correlation to the phenotypes herein can be any of those types of markers that can be detected by screening, e.g., genetic markers such as allelic variants of a genetic locus (e.g., as in SNPs), expression markers (e.g., presence or quantity of mRNAs and/or proteins), and/or the like.

[0111] The nucleic acid of interest to be amplified, transcribed, translated and/or detected in the methods of the invention can be essentially any nucleic acid, though nucleic acids derived from human sources are especially relevant to the detection of markers associated with disease diagnosis and clinical applications. The sequences for many nucleic acids and amino acids (from which nucleic acid sequences can be derived via reverse

translation) are available, including for the genes/proteins of Figures 1 and/or 2. Common sequence repositories for known nucleic acids include GenBank® EMBL, DDBJ and the NCBI. Other repositories can easily be identified by searching the internet. The nucleic acid to be amplified, transcribed, translated and/or detected can be an RNA (e.g., where
5 amplification includes RT-PCR or LCR, the Van-Gelder Eberwine reaction or Ribo-SPIA) or DNA (e.g., amplified DNA, cDNA or genomic DNA), or even any analogue thereof (e.g., for detection of synthetic nucleic acids or analogues thereof, e.g., where the sample of interest includes or is used to derive or synthesize artificial nucleic acids). Any variation in a nucleic acid sequence or expression level between individuals or populations can be
10 detected as a marker, e.g., a mutation, a polymorphism, a single nucleotide polymorphism (SNP), an allele, an isotype, expression of an RNA or protein, etc. One can detect variation in sequence, expression levels or gene copy numbers as markers that can be correlated to a breast cancer phenotype.

[0112] For example, the methods of the invention are useful in screening samples
15 derived from patients for a marker nucleic acid of interest, e.g., from bodily fluids (blood, saliva, urine etc.), biopsy, tissue, and/or waste from the patient. Thus, tissue biopsies, stool, sputum, saliva, blood, lymph, tears, sweat, urine, vaginal secretions, ejaculatory fluid or the like can easily be screened for nucleic acids by the methods of the invention, as can essentially any tissue of interest that contains the appropriate nucleic acids. These samples
20 are typically taken, following informed consent, from a patient by standard medical laboratory methods.

[0113] Prior to amplification and/or detection of a nucleic acid comprising a marker, the nucleic acid is optionally purified from the samples by any available method, e.g., those taught in Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in
25 Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2001 ("Sambrook"); and/or Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through
30 2002) ("Ausubel"). A plethora of kits are also commercially available for the purification of nucleic acids from cells or other samples (*see*, e.g., EasyPrep™, FlexiPrep™, both from Pharmacia Biotech; StrataClean™, from Stratagene; and, QIAprep™ from Qiagen).

Alternately, samples can simply be directly subjected to amplification or detection, e.g., following aliquotting and/or dilution.

[0114] Examples of markers can include polymorphisms, single nucleotide polymorphisms, presence of one or more nucleic acids in a sample, absence of one or more nucleic acids in a sample, presence of one or more genomic DNA sequences, absence of one or more genomic DNA sequences, presence of one or more mRNAs, absence of one or more mRNAs, expression levels of one or more mRNAs, presence of one or more proteins, expression levels of one or more proteins, and/or data derived from any of the preceding or combinations thereof. Essentially any number of markers can be detected, using available methods, e.g., using array technologies that provide high density, high throughput marker mapping. Thus, at least about 10, 100, 1,000, 10,000, or even 100,000 or more genetic markers can be tested, simultaneously or in a serial fashion (or combination thereof), for correlation to a relevant phenotype, in the first and/or second population. Combinations of markers can also be desirably tested, e.g., to identify genetic combinations or combinations of expression patterns in populations that are correlated to the phenotype.

[0115] As noted, the biological marker to be detected can be any detectable biological component. Commonly detected markers include genetic markers (e.g., DNA sequence markers present in genomic DNA or expression products thereof) and expression markers (which can reflect genetically coded factors, environmental factors, or both).

Where the markers are expression markers, the methods can include determining a first expression profile for a first individual or population (e.g., of one or more expressed markers, e.g., a set of expressed markers) and comparing the first expression profile to a second expression profile for the second individual or population. In this example, correlating expression marker(s) to a particular phenotype can include correlating the first or second expression profile to the phenotype of interest.

Probe/Primer Synthesis Methods

[0116] In general, synthetic methods for making oligonucleotides, including probes, primers, molecular beacons, PNAs, LNAs (locked nucleic acids), etc., are well known. For example, oligonucleotides can be synthesized chemically according to the solid phase phosphoramidite triester method described by Beaucage and Caruthers (1981), Tetrahedron Letts., 22(20):1859-1862, e.g., using a commercially available automated synthesizer, e.g., as described in Needham-VanDevanter et al. (1984) Nucleic Acids Res., 12:6159-6168.

Oligonucleotides, including modified oligonucleotides can also be ordered from a variety of commercial sources known to persons of skill. There are many commercial providers of oligo synthesis services, and thus this is a broadly accessible technology. Any nucleic acid can be custom ordered from any of a variety of commercial sources, such as The Midland
5 Certified Reagent Company (mcrc@oligos.com), The Great American Gene Company (www.genco.com), ExpressGen Inc. (www.expressgen.com), Operon Technologies Inc. (Alameda, CA) and many others. Similarly, PNAs can be custom ordered from any of a variety of sources, such as PeptidoGenic (pkim@ccnet.com), HTI Bio-products, inc. (htibio.com), BMA Biomedicals Ltd (U.K.), Bio-Synthesis, Inc., and many others.

In Silico Marker Detection

10 [0117] In some embodiments, *in silico* methods can be used to detect the marker loci of interest. For example, the sequence of a nucleic acid comprising the marker locus of interest can be stored in a computer. The desired marker locus sequence or its homolog can be identified using an appropriate nucleic acid search algorithm as provided by, for
15 example, in such readily available programs as BLAST, or even simple word processors. The entire human genome has been sequenced and, thus, sequence information can be used to identify marker regions, flanking nucleic acids, etc.

Amplification Primers For Marker Detection

[0118] In some preferred embodiments, the molecular markers of the invention are
20 detected using a suitable PCR-based detection method, where the size or sequence of the PCR amplicon is indicative of the absence or presence of the marker (e.g., a particular marker allele). In these types of methods, PCR primers are hybridized to the conserved regions flanking the polymorphic marker region.

[0119] Suitable primers to be used with the invention can be designed using any
25 suitable method. It is not intended that the invention be limited to any particular primer or primer pair. For example, primers can be designed using any suitable software program, such as LASERGENE[®], e.g., taking account of publicly available sequence information. Flanking sequences for the polymorphisms identified herein are publicly available; accordingly, suitable amplification primers can be constructed based on well understood
30 base-pairing rules. The sequence of any amplicon can be detected as has already been discussed above, e.g., by hybridization, array hybridization, PCR, real-time PCR, LCR, or the like.

[0120] In some embodiments, the primers of the invention are radiolabelled, or labeled by any suitable means (e.g., using a non-radioactive fluorescent tag), to allow for rapid visualization of differently sized amplicons following an amplification reaction without any additional labeling step or visualization step. In some embodiments, the primers are not labeled, and the amplicons are visualized following their size resolution, e.g., following agarose or acrylamide gel electrophoresis. In some embodiments, ethidium bromide staining of the PCR amplicons following size resolution allows visualization of the different size amplicons.

[0121] It is not intended that the primers of the invention be limited to generating an amplicon of any particular size. For example, the primers used to amplify the marker loci and alleles herein are not limited to amplifying the entire region of the relevant locus, or any subregion thereof. The primers can generate an amplicon of any suitable length for detection. In some embodiments, marker amplification produces an amplicon at least 20 nucleotides in length, or alternatively, at least 50 nucleotides in length, or alternatively, at least 100 nucleotides in length, or alternatively, at least 200 nucleotides in length. Amplicons of any size can be detected using the various technologies described herein. Differences in base composition or size can be detected by conventional methods such as electrophoresis.

Detection of Markers For Positional Cloning

[0122] In some embodiments, a nucleic acid probe is used to detect a nucleic acid that comprises a marker sequence. In addition to their role in determining breast cancer phenotypes, such probes can also be used, for example, in positional cloning to isolate nucleotide sequences linked to the marker nucleotide sequence. It is not intended that the nucleic acid probes of the invention be limited to any particular size. In some embodiments, nucleic acid probe is at least 20 nucleotides in length, or alternatively, at least 50 nucleotides in length, or alternatively, at least 100 nucleotides in length, or alternatively, at least 200 nucleotides in length.

[0123] A hybridized probe is detected using, e.g., autoradiography, fluorography or other similar detection techniques depending on the label to be detected. Examples of specific hybridization protocols are widely available in the art, *see*, e.g., Berger, Sambrook, and Ausubel, *all herein*.

Generation Of Transgenic Cells

[0124] The present invention also provides cells which are transformed with nucleic acids corresponding to QTL identified according to the invention. For example, such nucleic acids include chromosome intervals (*e.g.*, genomic fragments), genes, ORFs and/or cDNAs that encode genes that correspond or are linked to QTL for breast cancer phenotypes. Additionally, the invention provides for the production of polypeptides that influence breast cancer phenotypes. This is useful, *e.g.*, to influence breast cancers, and for the generation of transgenic cells. These cells provide commercially useful cell lines having defined genes that influence the relevant phenotype, thereby providing a platform for screening potential modulators of phenotype, as well as basic research into the mechanism of action for each of the genes of interest. In addition, gene therapy can be used to introduce desirable genes into individuals or populations thereof. Such gene therapies may be used to provide a treatment for a disorder exhibited by an individual, or may be used as a preventative measure to prevent the development of such a disorder in an individual at risk.

[0125] General texts which describe molecular biological techniques for the cloning and manipulation of nucleic acids and production of encoded polypeptides include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (3rd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 2001 ("Sambrook") and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2004 or later) ("Ausubel"). These texts describe mutagenesis, the use of vectors, promoters and many other relevant topics related to, *e.g.*, the generation of clones that comprise nucleic acids of interest, *e.g.*, genes, marker loci, marker probes, QTL that segregate with marker loci, etc.

[0126] Host cells are genetically engineered (*e.g.*, transduced, transfected, transformed, etc.) with the vectors of this invention (*e.g.*, vectors, such as expression vectors which comprise a gene or an ORF derived from or related to a QTL) which can be, for example, a cloning vector, a shuttle vector or an expression vector. Such vectors are, for example, in the form of a plasmid, a phagemid, an agrobacterium, a virus, a naked polynucleotide (linear or circular), or a conjugated polynucleotide. Vectors can be introduced into bacteria, especially for the purpose of propagation and expansion.

Additional details regarding nucleic acid introduction methods are found in Sambrook, Berger and Ausubel, *infra*. The method of introducing a nucleic acid of the present invention into a host cell is not critical to the instant invention, and it is not intended that the invention be limited to any particular method for introducing exogenous genetic material into a host cell. Thus, any suitable method, e.g., including but not limited to the methods provided herein, which provides for effective introduction of a nucleic acid into a cell or protoplast can be employed and finds use with the invention.

[0127] The engineered host cells can be cultured in conventional nutrient media modified as appropriate for such activities as, for example, activating promoters or selecting transformants. In addition to Sambrook, Berger and Ausubel, *all infra*, Atlas and Parks (eds) The Handbook of Microbiological Media (1993) CRC Press, Boca Raton, FL and available commercial literature such as the Life Science Research Cell Culture Catalogue (2004) from Sigma- Aldrich, Inc (St Louis, MO) ("Sigma-LSRCCC") provide additional details.

CORRELATING MARKERS TO PHENOTYPES

[0128] One aspect of the invention is a description of correlations between polymorphisms noted in Figures 1 and/or 2 and breast cancer phenotypes. An understanding of these correlations can be used in the present invention to correlate information regarding a set of polymorphisms that an individual or sample is determined to possess and a phenotype that they are likely to display. Further, higher order correlations that account for combinations of alleles in one or more different genes can also be assessed for correlations to phenotype.

[0129] These correlations can be performed by any method that can identify a relationship between an allele and a phenotype, or a combination of alleles and a combination of phenotypes. For example, alleles in genes or loci in Figures 1 and/or 2 can be correlated with one or more breast cancer phenotypes. Most typically, these methods involve referencing a look up table that comprises correlations between alleles of the polymorphism and the phenotype. The table can include data for multiple allele-phenotype relationships and can take account of additive or other higher order effects of multiple allele-phenotype relationships, e.g., through the use of statistical tools such as principle component analysis, heuristic algorithms, etc.

[0130] Correlation of a marker to a phenotype optionally includes performing one or more statistical tests for correlation. Many statistical tests are known, and most are computer-implemented for ease of analysis. A variety of statistical methods of determining associations/correlations between phenotypic traits and biological markers are known and can be applied to the present invention. For an introduction to the topic, *see*, Hartl (1981) A Primer of Population Genetics Washington University, Saint Louis Sinauer Associates, Inc. Sunderland, MA ISBN: 0-087893-271-2. A variety of appropriate statistical models are described in Lynch and Walsh (1998) Genetics and Analysis of Quantitative Traits, Sinauer Associates, Inc. Sunderland MA ISBN 0-87893-481-2. These models can, for example, provide for correlations between genotypic and phenotypic values, characterize the influence of a locus on a phenotype, sort out the relationship between environment and genotype, determine dominance or penetrance of genes, determine maternal and other epigenetic effects, determine principle components in an analysis (via principle component analysis, or "PCA"), and the like. The references cited in these texts provides considerable further detail on statistical models for correlating markers and phenotype.

[0131] In addition to standard statistical methods for determining correlation, other methods that determine correlations by pattern recognition and training, such as the use of genetic algorithms, can be used to determine correlations between markers and phenotypes. This is particularly useful when identifying higher order correlations between multiple alleles and multiple phenotypes. To illustrate, neural network approaches can be coupled to genetic algorithm-type programming for heuristic development of a structure-function data space model that determines correlations between genetic information and phenotypic outcomes. For example, NNUGA (Neural Network Using Genetic Algorithms) is an available program (e.g., on the world wide web at cs.bgu.ac.il/~omri/NNUGA which couples neural networks and genetic algorithms. An introduction to neural networks can be found, e.g., in Kevin Gurney, An Introduction to Neural Networks, UCL Press (1999) and on the world wide web at shef.ac.uk/psychology/gurney/notes/index.html. Additional useful neural network references include those noted above in regard to genetic algorithms and, e.g., Bishop, Neural Networks for Pattern Recognition, Oxford University Press (1995), and Ripley et al., Pattern Recognition and Neural Networks, Cambridge University Press (1995). Two tables showing exemplary data sets including certain statistical analyses are shown in Figures 1 and/or 2.

[0132] Additional references that are useful in understanding data analysis applications for using and establishing correlations, principle components of an analysis, neural network modeling and the like, include, e.g., Hinchliffe, Modeling Molecular Structures, John Wiley and Sons (1996), Gibas and Jambeck, Bioinformatics Computer Skills, O'Reilly (2001), Pevzner, Computational Molecular Biology and Algorithmic Approach, The MIT Press (2000), Durbin et al., Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press (1998), and Rashidi and Buehler, Bioinformatic Basics: Applications in Biological Science and Medicine, CRC Press LLC (2000).

10 [0133] In any case, essentially any statistical test can be applied in a computer implemented model, by standard programming methods, or using any of a variety of "off the shelf" software packages that perform such statistical analyses, including, for example, those noted above and those that are commercially available, e.g., from Partek Incorporated (St. Peters, Missouri; www.partek.com), e.g., that provide software for pattern recognition
 15 (e.g., which provide Partek Pro 2000 Pattern Recognition Software) which can be applied to genetic algorithms for multivariate data analysis, interactive visualization, variable selection, neural network & statistical modeling, etc. Relationships can be analyzed, e.g., by Principal Components Analysis (PCA) mapped mapped scatterplots and biplots, Multi-Dimensional Scaling (MDS) Multi-Dimensional Scaling (MDS) mapped scatterplots, star
 20 plots, etc. Available software for performing correlation analysis includes SAS, R and MathLab.

[0134] The marker(s), whether polymorphisms or expression patterns, can be used for any of a variety of genetic analyses. For example, once markers have been identified, as in the present case, they can be used in a number of different assays for association studies.
 25 For example, probes can be designed for microarrays that interrogate these markers. Other exemplary assays include, e.g., the Taqman assays and molecular beacon assays described *supra*, as well as conventional PCR and/or sequencing techniques.

[0135] Additional details regarding association studies can be found in 10/106,097, filed March 26, 2002, entitled "Methods for Genomic Analysis;" 10/042,819, filed January
 30 7, 2002, entitled "Genetic Analysis Systems and Methods;" 10/286,417, filed October 31, 2002, entitled "Methods for Genomic Analysis;" 10/768,788, filed January 30, 2004, entitled "Apparatus and Methods for Analyzing and Characterizing Nucleic Acid

Sequences;"10/447,685, filed May 28, 2003, entitled "Liver Related Disease Compositions and Methods;" 10/970,761, filed October 20, 2004, entitled "Improved Analysis Methods and Apparatus for Individual Genotyping" (methods for individual genotyping); 10/956,224, filed September 30, 2004, entitled "Methods for Genetic Analysis."

5 **[0136]** In some embodiments, the marker data is used to perform association studies to show correlations between markers and phenotypes. This can be accomplished by determining marker characteristics in individuals with the phenotype of interest (i.e., individuals or populations displaying the phenotype of interest) and comparing the allele frequency or other characteristics (expression levels, etc.) of the markers in these
10 individuals to the allele frequency or other characteristics in a control group of individuals. Such marker determinations can be conducted on a genome-wide basis, or can be focused on specific regions of the genome (e.g., haplotype blocks of interest). In one embodiment, markers that are linked to the genes or loci in Figures 1 and/or 2 are assessed for correlation to one or more specific breast cancer predisposition phenotypes.

15 **[0137]** In addition to the other embodiments of the methods of the present invention disclosed herein, the methods additionally allow for the "dissection" of a phenotype. That is, a particular phenotypes can (and typically do) result from two or more different genetic bases. For example, a susceptibility phenotype in one individual may be the result of a "defect" (or simply a particular allele—"defect" with respect to a susceptibility phenotype is
20 context dependent, e.g., whether the phenotype is desirable or undesirable in the individual in a given environment) in a gene for in Figures 1 and/or 2, while the same basic phenotype in a different individual may be the result of multiple "defects" in multiple genes in Figures 1 and/or 2. Thus, scanning a plurality of markers (e.g., as in genome or haplotype block scanning) allows for the dissection of varying genetic bases for similar (or graduated)
25 phenotypes.

[0138] As described in the previous paragraph, one method of conducting association studies is to compare the allele frequency (or expression level) of markers in individuals with a phenotype of interest ("case group") to the allele frequency in a control group of individuals. In one method, informative SNPs are used to make the SNP
30 haplotype pattern comparison (an "informative SNP" is genetic SNP marker such as a SNP or subset (more than one) of SNPs in a genome or haplotype block that tends to distinguish one SNP or genome or haplotype pattern from other SNPs, genomes or haplotype patterns).

The approach of using informative SNPs has an advantage over other whole genome scanning or genotyping methods known in the art, for instead of reading all 3 billion bases of each individual's genome—or even reading the 3-4 million common SNPs that may be found—only informative SNPs from a sample population need to be detected. Reading these particular, informative SNPs provides sufficient information to allow statistically accurate association data to be extracted from specific experimental populations, as described above.

[0139] Thus, in an embodiment of one method of determining genetic associations, the allele frequency of informative SNPs is determined for genomes of a control population that do not display the phenotype. The allele frequency of informative SNPs is also determined for genomes of a population that do display the phenotype. The informative SNP allele frequencies are compared. Allele frequency comparisons can be made, for example, by determining the allele frequency (number of instances of a particular allele in a population divided by the total number of alleles) at each informative SNP location in each population and comparing these allele frequencies. The informative SNPs displaying a difference between the allele frequency of occurrence in the control versus case populations/groups are selected for analysis. Once informative SNPs are selected, the SNP haplotype block(s) that contain the informative SNPs are identified, which in turn identifies a genomic region of interest that is correlated with the phenotype. The genomic regions can be analyzed by genetic or any biological methods known in the art e.g., for use as drug discovery targets or as diagnostic markers.

SYSTEMS FOR IDENTIFYING BREAST CANCER PHENOTYPES

[0140] Systems for performing the above correlations are also a feature of the invention. Typically, the system will include system instructions that correlate the presence or absence of an allele (whether detected directly or, e.g., through expression levels) with a predicted phenotype. The system instructions can compare detected information as to allele sequence or expression level with a database that includes correlations between the alleles and the relevant phenotypes. As noted above, this database can be multidimensional, thereby including higher-order relationships between combinations of alleles and the relevant phenotypes. These relationships can be stored in any number of look-up tables, e.g., taking the form of spreadsheets (e.g., Excel™ spreadsheets) or databases such as an Access™, SQL™, Oracle™, Paradox™, or similar database. The system includes

provisions for inputting sample-specific information regarding allele detection information, e.g., through an automated or user interface and for comparing that information to the look up tables.

[0141] Optionally, the system instructions can also include software that accepts
5 diagnostic information associated with any detected allele information, e.g., a diagnosis that a subject with the relevant allele has a particular phenotype. This software can be heuristic in nature, using such inputted associations to improve the accuracy of the look up tables and/ or interpretation of the look up tables by the system. A variety of such approaches, including neural networks, Markov modeling, and other statistical analysis are described
10 above.

[0142] The invention provides data acquisition modules for detecting one or more detectable genetic marker(s) (e.g., one or more array comprising one or more biomolecular probes, detectors, fluid handlers, or the like). The biomolecular probes of such a data acquisition module can include any that are appropriate for detecting the biological marker,
15 e.g., oligonucleotide probes, proteins, aptamers, antibodies, etc. These can include sample handlers (e.g., fluid handlers), robotics, microfluidic systems, nucleic acid or protein purification modules, arrays (e.g., nucleic acid arrays), detectors, thermocyclers or combinations thereof, e.g., for acquiring samples, diluting or aliquoting samples, purifying marker materials (e.g., nucleic acids or proteins), amplifying marker nucleic acids, detecting
20 amplified marker nucleic acids, and the like.

[0143] For example, automated devices that can be incorporated into the systems herein have been used to assess a variety of biological phenomena, including, e.g., expression levels of genes in response to selected stimuli (Service (1998) "Microchips Arrays Put DNA on the Spot" Science 282:396-399), high throughput DNA genotyping
25 (Zhang et al. (1999) "Automated and Integrated System for High-Throughput DNA Genotyping Directly from Blood" Anal. Chem. 71:1138-1145) and many others. Similarly, integrated systems for performing mixing experiments, DNA amplification, DNA sequencing and the like are also available. See, e.g., Service (1998) "Coming Soon: the Pocket DNA Sequencer" Science 282: 399-401. A variety of automated system
30 components are available, e.g., from Caliper Technologies (Hopkinton, MA), which utilize various Zymate systems, which typically include, e.g., robotics and fluid handling modules. Similarly, the common ORCA® robot, which is used in a variety of laboratory systems,

e.g., for microtiter tray manipulation, is also commercially available, e.g., from Beckman Coulter, Inc. (Fullerton, CA). Similarly, commercially available microfluidic systems that can be used as system components in the present invention include those from Agilent technologies and the Caliper Technologies. Furthermore, the patent and technical literature
5 includes numerous examples of microfluidic systems, including those that can interface directly with microwell plates for automated fluid handling.

[0144] Any of a variety of liquid handling and/or array configurations can be used in the systems herein. One common format for use in the systems herein is a microtiter plate, in which the array or liquid handler includes a microtiter tray. Such trays are commercially
10 available and can be ordered in a variety of well sizes and numbers of wells per tray, as well as with any of a variety of functionalized surfaces for binding of assay or array components. Common trays include the ubiquitous 96 well plate, with 384 and 1536 well plates also in common use. Samples can be processed in such trays, with all of the processing steps being performed in the trays. Samples can also be processed in microfluidic apparatus, or
15 combinations of microtiter and microfluidic apparatus.

[0145] In addition to liquid phase arrays, components can be stored in or analyzed on solid phase arrays. These arrays fix materials in a spatially accessible pattern (e.g., a grid of rows and columns) onto a solid substrate such as a membrane (e.g., nylon or nitrocellulose), a polymer or ceramic surface, a glass or modified silica surface, a metal
20 surface, or the like. Components can be accessed, e.g., by hybridization, by local rehydration (e.g., using a pipette or other fluid handling element) and fluidic transfer, or by scraping the array or cutting out sites of interest on the array.

[0146] The system can also include detection apparatus that is used to detect allele information, using any of the approached noted herein. For example, a detector configured
25 to detect real-time PCR products (e.g., a light detector, such as a fluorescence detector) or an array reader can be incorporated into the system. For example, the detector can be configured to detect a light emission from a hybridization or amplification reaction comprising an allele of interest, wherein the light emission is indicative of the presence or absence of the allele. Optionally, an operable linkage between the detector and a computer
30 that comprises the system instructions noted above is provided, allowing for automatic input of detected allele-specific information to the computer, which can, e.g., store the database

information and/or execute the system instructions to compare the detected allele specific information to the look up table.

[0147] Probes that are used to generate information detected by the detector can also be incorporated within the system, along with any other hardware or software for using the probes to detect the amplicon. These can include thermocycler elements (e.g., for performing PCR or LCR amplification of the allele to be detected by the probes), arrays upon which the probes are arrayed and/or hybridized, or the like. The fluid handling elements noted above for processing samples, can be used for moving sample materials (e.g., template nucleic acids and/or proteins to be detected) primers, probes, amplicons, or the like into contact with one another. For example, the system can include a set of marker probes or primers configured to detect at least one allele of one or more genes or linked loci associated with a phenotype, where the gene encodes a polymorphism in Figures 1 and/or 2. The detector module is configured to detect one or more signal outputs from the set of marker probes or primers, or an amplicon produced from the set of marker probes or primers, thereby identifying the presence or absence of the allele.

[0148] The sample to be analyzed is optionally part of the system, or can be considered separate from it. The sample optionally includes e.g., genomic DNA, amplified genomic DNA, cDNA, amplified cDNA, RNA, amplified RNA, proteins, etc., as noted herein. In one aspect, the sample is derived from a mammal such as a human patient.

[0149] Optionally, system components for interfacing with a user are provided. For example, the systems can include a user viewable display for viewing an output of computer-implemented system instructions, user input devices (e.g., keyboards or pointing devices such as a mouse) for inputting user commands and activating the system, etc. Typically, the system of interest includes a computer, wherein the various computer-implemented system instructions are embodied in computer software, e.g., stored on computer readable media.

[0150] Standard desktop applications such as word processing software (e.g., Microsoft Word™ or Corel WordPerfect™) and database software (e.g., spreadsheet software such as Microsoft Excel™, Corel Quattro Pro™, or database programs such as Microsoft Access™ or Sequel™, Oracle™, Paradox™) can be adapted to the present invention by inputting a character string corresponding to an allele herein, or an association

between an allele and a phenotype. For example, the systems can include software having the appropriate character string information, e.g., used in conjunction with a user interface (e.g., a GUI in a standard operating system such as a Windows, Macintosh or LINUX system) to manipulate strings of characters. Specialized sequence alignment programs such as BLAST can also be incorporated into the systems of the invention for alignment of nucleic acids or proteins (or corresponding character strings) e.g., for identifying and relating multiple alleles.

[0151] As noted, systems can include a computer with an appropriate database and an allele sequence or correlation of the invention. Software for aligning sequences, as well as data sets entered into the software system comprising any of the sequences herein can be a feature of the invention. The computer can be, e.g., a PC (Intel x86 or Pentium chip-compatible DOS™, OS2™, WINDOWS™, WINDOWS NT™, WINDOWS95™, WINDOWS98™, WINDOWS2000, WINDOWSME, or LINUX based machine, a MACINTOSH™, Power PC, or a UNIX based (e.g., SUN™ work station or LINUX based machine) or other commercially common computer which is known to one of skill. Software for entering and aligning or otherwise manipulating sequences is available, e.g., BLASTP and BLASTN, or can easily be constructed by one of skill using a standard programming language such as Visualbasic, Fortran, Basic, Java, or the like.

METHODS OF IDENTIFYING MODULATORS

[0152] In addition to providing various diagnostic and prognostic markers for identifying breast cancer predisposition, etc., the invention also provides methods of identifying modulators of breast cancer phenotypes. In the methods, a potential modulator is contacted to a relevant protein corresponding to a loci in Figures 1 and/or 2, or to a nucleic acid that encodes such a protein. An effect of the potential modulator on the gene or gene product is detected, thereby identifying whether the potential modulator modulates the underlying molecular basis for the phenotype.

[0153] In addition, the methods can include, e.g., administering one or more putative modulator to an individual that displays a relevant phenotype and determining whether the putative modulator modulates the phenotype in the individual, e.g., in the context of a clinical trial or treatment. This, in turn, determines whether the putative modulator is clinically useful.

[0154] The gene or gene product that is contacted by the modulator can include any allelic form noted herein. Allelic forms, whether genes or proteins, that positively correlate to undesirable phenotypes are preferred targets for modulator screening.

5 [0155] Effects of interest that can be screened for include: (a) increased or decreased expression of a gene or gene product in Figures 1 and/or 2 in the presence of the modulator; (b) a change in the timing or location of expression; (c) or a change in localization of the proteins corresponding to loci of Figures 1 and/or 2 in the presence of the modulator.

[0156] The precise format of the modulator screen will, of course, vary, depending on the effect(s) being detected and the equipment available. Northern analysis, quantitative
10 RT-PCR and/or array-based detection formats can be used to distinguish expression levels of genes noted above. Protein expression levels can also be detected using available methods, such as western blotting, ELISA analysis, antibody hybridization, BIAcore, or the like. Any of these methods can be used to distinguish changes in expression levels that result from a potential modulator.

15 [0157] Accordingly, one may screen for potential modulators for activity or expression. For example, potential modulators (small molecules, organic molecules, inorganic molecules, proteins, hormones, transcription factors, or the like) can be contacted to a cell comprising an allele of interest and an effect on activity or expression (or both) of a gene or protein corresponding to a loci in Figures 1 and/or 2 can be detected, e.g., via
20 northern analysis or quantitative (optionally real time) RT-PCR, before and after application of potential expression modulators. Similarly, promoter regions of the various genes (e.g., generally sequences in the region of the start site of transcription, e.g., within 5 kb of the start site, e.g., 1 kb, or less e.g., within 500 bp or 250 bp or 100 bp of the start site) can be coupled to reporter constructs (CAT, beta-galactosidase, luciferase or any other available
25 reporter) and can be similarly be tested for expression activity modulation by the potential modulator. In either case, the assays can be performed in a high-throughput fashion, e.g., using automated fluid handling and/or detection systems, in serial or parallel fashion. Similarly, activity modulators can be tested by contacting a potential modulator to an appropriate cell using any of the activity detection methods herein, regardless of whether
30 the activity that is detected is the result of activity modulation, expression modulation or both.

[0158] Biosensors for detecting modulator activity detection are also a feature of the invention. These include devices or systems that comprise a gene or gene product corresponding to a loci of Figures 1 and/or 2 coupled to a readout that measures or displays one or more activity of the gene or product. Thus, any of the above described assay components can be configured as a biosensor by operably coupling the appropriate assay components to a readout. The readout can be optical (e.g., to detect cell markers or cell survival) electrical (e.g., coupled to a FET, a BIAcore, or any of a variety of others), spectrographic, or the like, and can optionally include a user-viewable display (e.g., a CRT or optical viewing station). The biosensor can be coupled to robotics or other automation, e.g., microfluidic systems, that direct contact of the putative modulators to the proteins of the invention, e.g., for automated high-throughput analysis of putative modulator activity. A large variety of automated systems that can be adapted to use with the biosensors of the invention are commercially available. For example, automated systems have been made to assess a variety of biological phenomena, including, e.g., expression levels of genes in response to selected stimuli (Service (1998) "Microchips Arrays Put DNA on the Spot" Science 282:396-399). Laboratory systems can also perform, e.g., repetitive fluid handling operations (e.g., pipetting) for transferring material to or from reagent storage systems that comprise arrays, such as microtiter trays or other chip trays, which are used as basic container elements for a variety of automated laboratory methods. Similarly, the systems manipulate, e.g., microtiter trays and control a variety of environmental conditions such as temperature, exposure to light or air, and the like. Many such automated systems are commercially available and are described herein, including those described above. These include various Zymate systems, ORCA® robots, microfluidic devices, etc. For example, the LabMicrofluidic device® high throughput screening system (HTS) by Caliper Technologies, Mountain View, CA can be adapted for use in the present invention to screen for modulator activity.

[0159] In general, methods and sensors for detecting protein expression level and activity are available, including those taught in the various references above, including R. Scopes, Protein Purification, Springer-Verlag, N.Y. (1982); Deutscher, Methods in Enzymology Vol. 182: Guide to Protein Purification, Academic Press, Inc. N.Y. (1990); Sandana (1997) Bioseparation of Proteins, Academic Press, Inc.; Bollag et al. (1996) Protein Methods, 2nd Edition Wiley-Liss, NY; Walker (1996) The Protein Protocols

Handbook Humana Press, NJ, Harris and Angal (1990) Protein Purification Applications: A Practical Approach IRL Press at Oxford, Oxford, England; Harris and Angal Protein Purification Methods: A Practical Approach IRL Press at Oxford, Oxford, England; Scopes (1993) Protein Purification: Principles and Practice 3rd Edition Springer Verlag, NY; Janson and Ryden (1998) Protein Purification: Principles, High Resolution Methods and Applications, Second Edition Wiley-VCH, NY; and Walker (1998) Protein Protocols on CD-ROM Humana Press, NJ; and Satinder Ahuja ed., Handbook of Bioseparations, Academic Press (2000). "Proteomic" detection methods, which detect many proteins simultaneously have been described and are also noted above, including various multidimensional electrophoresis methods (e.g., 2-d gel electrophoresis), mass spectrometry based methods (e.g., SELDI, MALDI, electrospray, *etc.*), or surface plasmon resonance methods. These can also be used to track protein activity and/or expression level.

[0160] Similarly, nucleic acid expression levels (e.g., mRNA) can be detected using any available method, including northern analysis, quantitative RT-PCR, or the like.

References sufficient to guide one of skill through these methods are readily available, including Ausubel, Sambrook and Berger.

[0161] Potential modulator libraries to be screened for effects on expression and/or activity are available. These libraries can be random, or can be targeted.

[0162] Targeted libraries include those designed using any form of a rational design technique that selects scaffolds or building blocks to generate combinatorial libraries.

These techniques include a number of methods for the design and combinatorial synthesis of target-focused libraries, including morphing with bioisosteric transformations, analysis of target-specific privileged structures, and the like. In general, where information regarding structure genes or gene products of Figures 1 and/or 2, is available, likely binding partners can be designed, e.g., using flexible docking approaches, or the like. Similarly, random libraries exist for a variety of basic chemical scaffolds. In either case, many thousands of scaffolds and building blocks for chemical libraries are available, including those with polypeptide, nucleic acid, carbohydrate, and other backbones. Commercially available libraries and library design services include those offered by Chemical Diversity (San Diego, CA), Affymetrix (Santa Clara, CA), Sigma (St. Louis MO), ChemBridge Research Laboratories (San Diego, CA), TimTec (Newark, DE), Nuevolution A/S (Copenhagen, Denmark) and many others.

[0163] Kits for treatment of breast cancer phenotypes can include a modulator identified as noted above and instructions for administering the compound to a patient to treat breast cancer.

REGULATING GENE EXPRESSION OF GENES

5 [0164] Gene expression (e.g., transcription and/or translation) of any gene linked to a polymorphism in Figures 1 and 2 can be regulated using any of a variety of techniques known in the art. For example, gene expression can be inhibited using an antisense nucleic acid or an interfering RNA. Inhibition of expression in particular cell-types can be used for further studying the in vitro or in vivo role of these genes, and/or as a mechanism for
10 treating a condition caused by overexpression of a linked gene, and/or for treating a dominant effect caused by a particular allele of such a gene. Gene expression modulators are one class of modulators provided by the present invention, e.g., modulators applied to modulate a breast cancer phenotype.

[0165] For example, use of antisense nucleic acids is well known in the art. An
15 antisense nucleic acid has a region of complementarity to a target nucleic acid, e.g., a target gene, mRNA, or cDNA. Typically, a nucleic acid comprising a nucleotide sequence in a complementary, antisense orientation with respect to a coding (sense) sequence of an endogenous gene is introduced into a cell. The antisense nucleic acid can be RNA, DNA, a PNA or any other appropriate molecule. A duplex can form between the antisense
20 sequence and its complementary sense sequence, resulting in inactivation of the gene. The antisense nucleic acid can inhibit gene expression by forming a duplex with an RNA transcribed from the gene, by forming a triplex with duplex DNA, etc. An antisense nucleic acid can be produced, e.g., for any gene whose coding sequence is known or can be determined by a number of well-established techniques (e.g., chemical synthesis of an
25 antisense RNA or oligonucleotide (optionally including modified nucleotides and/or linkages that increase resistance to degradation or improve cellular uptake) or in vitro transcription). Antisense nucleic acids and their use are described, e.g., in USP 6,242,258 to Haselton and Alexander (June 5, 2001) entitled "Methods for the selective regulation of DNA and RNA transcription and translation by photoactivation"; USP 6,500,615; USP
30 6,498,035; USP 6,395,544; USP 5,563,050; E. Schuch et al (1991) Symp Soc. Exp Biol 45:117-127; de Lange et al., (1995) Curr Top Microbiol Immunol 197:57-75; Hamilton et

al. (1995) Curr Top Microbiol Immunol 197:77-89; Finnegan et al., (1996) Proc Natl Acad Sci USA 93:8449-8454; Uhlmann and A. Pepan (1990), Chem. Rev. 90:543; P. D. Cook (1991), Anti-Cancer Drug Design 6:585; J. Goodchild, Bioconjugate Chem. 1 (1990) 165; and, S. L. Beaucage and R. P. Iyer (1993), Tetrahedron 49:6123; and F. Eckstein, Ed.

5 (1991), Oligonucleotides and Analogues--A Practical Approach, IRL Press.

[0166] Gene expression can also be inhibited by RNA silencing or interference.

“RNA silencing” refers to any mechanism through which the presence of a single-stranded or, typically, a double-stranded RNA in a cell results in inhibition of expression of a target gene comprising a sequence identical or nearly identical to that of the RNA, including, but
10 not limited to, RNA interference, repression of translation of a target mRNA transcribed from the target gene without alteration of the mRNA's stability, and transcriptional silencing (e.g., histone acetylation and heterochromatin formation leading to inhibition of transcription of the target mRNA).

[0167] The term “RNA interference” (“RNAi,” sometimes called RNA-mediated
15 interference, post-transcriptional gene silencing, or quelling) refers to a phenomenon in which the presence of RNA, typically double-stranded RNA, in a cell results in inhibition of expression of a gene comprising a sequence identical, or nearly identical, to that of the double-stranded RNA. The double-stranded RNA responsible for inducing RNAi is called an “interfering RNA.” Expression of the gene is inhibited by the mechanism of RNAi as
20 described below, in which the presence of the interfering RNA results in degradation of mRNA transcribed from the gene and thus in decreased levels of the mRNA and any encoded protein.

[0168] The mechanism of RNAi has been and is being extensively investigated in a number of eukaryotic organisms and cell types. See, for example, the following reviews:
25 McManus and Sharp (2002) “Gene silencing in mammals by small interfering RNAs” Nature Reviews Genetics 3:737-747; Hutvagner and Zamore (2002) “RNAi: Nature abhors a double strand” Curr Opin Genet & Dev 200:225-232; Hannon (2002) “RNA interference” Nature 418:244-251; Agami (2002) “RNAi and related mechanisms and their potential use for therapy” Curr Opin Chem Biol 6:829-834; Tuschl and Borkhardt (2002) “Small
30 interfering RNAs: A revolutionary tool for the analysis of gene function and gene therapy” Molecular Interventions 2:158-167; Nishikura (2001) “A short primer on RNAi: RNA-directed RNA polymerase acts as a key catalyst” Cell 107:415-418; and Zamore (2001)

“RNA interference: Listening to the sound of silence” *Nature Structural Biology* 8:746-750. RNAi is also described in the patent literature; see, e.g., CA 2359180 by Kreutzer and Limmer entitled “Method and medicament for inhibiting the expression of a given gene”; WO 01/68836 by Beach et al. entitled “Methods and compositions for RNA interference”;
5 WO 01/70949 by Graham et al. entitled “Genetic silencing”; and WO 01/75164 by Tuschl et al. entitled “RNA sequence-specific mediators of RNA interference.”

[0169] In brief, double-stranded RNA introduced into a cell (e.g., into the cytoplasm) is processed, for example by an RNase III-like enzyme called Dicer, into shorter double-stranded fragments called small interfering RNAs (siRNAs, also called short
10 interfering RNAs). The length and nature of the siRNAs produced is dependent on the species of the cell, although typically siRNAs are 21-25 nucleotides long (e.g., an siRNA may have a 19 base pair duplex portion with two nucleotide 3' overhangs at each end). Similar siRNAs can be produced in vitro (e.g., by chemical synthesis or in vitro transcription) and introduced into the cell to induce RNAi. The siRNA becomes associated
15 with an RNA-induced silencing complex (RISC). Separation of the sense and antisense strands of the siRNA, and interaction of the siRNA antisense strand with its target mRNA through complementary base-pairing interactions, optionally occurs. Finally, the mRNA is cleaved and degraded.

[0170] Expression of a target gene in a cell can thus be specifically inhibited by
20 introducing an appropriately chosen double-stranded RNA into the cell. Guidelines for design of suitable interfering RNAs are known to those of skill in the art. For example, interfering RNAs are typically designed against exon sequences, rather than introns or untranslated regions. Characteristics of high efficiency interfering RNAs may vary by cell type. For example, although siRNAs may require 3' overhangs and 5' phosphates for most
25 efficient induction of RNAi in *Drosophila* cells, in mammalian cells blunt ended siRNAs and/or RNAs lacking 5' phosphates can induce RNAi as effectively as siRNAs with 3' overhangs and/or 5' phosphates (see, e.g., Czauderna et al. (2003) “Structural variations and stabilizing modifications of synthetic siRNAs in mammalian cells” *Nucl Acids Res* 31:2705-2716). As another example, since double-stranded RNAs greater than 30-80 base
30 pairs long activate the antiviral interferon response in mammalian cells and result in non-specific silencing, interfering RNAs for use in mammalian cells are typically less than 30 base pairs (for example, Caplen et al. (2001) “Specific inhibition of gene expression by

small double-stranded RNAs in invertebrate and vertebrate systems” Proc. Natl. Acad. Sci. USA 98:9742-9747, Elbashir et al. (2001) “Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells” Nature 411:494-498 and Elbashir et al. (2002)

“Analysis of gene function in somatic mammalian cells using small interfering RNAs”

5 Methods 26:199-213 describe the use of 21 nucleotide siRNAs to specifically inhibit gene expression in mammalian cell lines, and Kim et al. (2005) “Synthetic dsRNA Dicer substrates enhance RNAi potency and efficacy” Nature Biotechnology 23:222-226

describes use of 25-30 nucleotide duplexes). The sense and antisense strands of a siRNA are typically, but not necessarily, completely complementary to each other over the double-
10 stranded region of the siRNA (excluding any overhangs). The antisense strand is typically completely complementary to the target mRNA over the same region, although some nucleotide substitutions can be tolerated (e.g., a one or two nucleotide mismatch between the antisense strand and the mRNA can still result in RNAi, although at reduced efficiency).

The ends of the double-stranded region are typically more tolerant to substitution than the

15 middle; for example, as little as 15 bp (base pairs) of complementarity between the antisense strand and the target mRNA in the context of a 21 mer with a 19 bp double-stranded region has been shown to result in a functional siRNA (see, e.g., Czauderna et al. (2003) “Structural variations and stabilizing modifications of synthetic siRNAs in mammalian cells” Nucl Acids Res 31:2705-2716). Any overhangs can but need not be

20 complementary to the target mRNA; for example, TT (two 2'-deoxythymidines) overhangs are frequently used to reduce synthesis costs.

[0171] Although double-stranded RNAs (e.g., double-stranded siRNAs) were initially thought to be required to initiate RNAi, several recent reports indicate that the antisense strand of such siRNAs is sufficient to initiate RNAi. Single-stranded antisense

25 siRNAs can initiate RNAi through the same pathway as double-stranded siRNAs (as evidenced, for example, by the appearance of specific mRNA endonucleolytic cleavage fragments). As for double-stranded interfering RNAs, characteristics of high-efficiency single-stranded siRNAs may vary by cell type (e.g., a 5' phosphate may be required on the antisense strand for efficient induction of RNAi in some cell types, while a free 5' hydroxyl
30 is sufficient in other cell types capable of phosphorylating the hydroxyl). See, e.g.,

Martinez et al. (2002) “Single-stranded antisense siRNAs guide target RNA cleavage in RNAi” Cell 110:563-574; Amarzguoui et al. (2003) “Tolerance for mutations and chemical

modifications in a siRNA" Nucl. Acids Res. 31:589-595; Holen et al. (2003) "Similar behavior of single-strand and double-strand siRNAs suggests that they act through a common RNAi pathway" Nucl. Acids Res. 31:2401-2407; and Schwarz et al. (2002) Mol. Cell 10:537-548.

5 [0172] Due to differences in efficiency between siRNAs corresponding to different regions of a given target mRNA, several siRNAs are typically designed and tested against the target mRNA to determine which siRNA is most effective. Interfering RNAs can also be produced as small hairpin RNAs (shRNAs, also called short hairpin RNAs), which are processed in the cell into siRNA-like molecules that initiate RNAi (see, e.g., Siolas et al.
10 (2005) "Synthetic shRNAs as potent RNAi triggers" Nature Biotechnology 23:227 - 231).

[0173] The presence of RNA, particularly double-stranded RNA, in a cell can result in inhibition of expression of a gene comprising a sequence identical or nearly identical to that of the RNA through mechanisms other than RNAi. For example, double-stranded RNAs that are partially complementary to a target mRNA can repress translation of the
15 mRNA without affecting its stability. As another example, double-stranded RNAs can induce histone methylation and heterochromatin formation, leading to transcriptional silencing of a gene comprising a sequence identical or nearly identical to that of the RNA (see, e.g., Schramke and Allshire (2003) "Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing" Science 301:1069-1074; Kawasaki and Taira
20 (2004) "Induction of DNA methylation and gene silencing by short interfering RNAs in human cells" Nature 431:211-217; and Morris et al. (2004) "Small interfering RNA-induced transcriptional gene silencing in human cells" Science 305:1289-1292).

[0174] Short RNAs called microRNAs (miRNAs) have been identified in a variety of species. Typically, these endogenous RNAs are each transcribed as a long RNA and then
25 processed to a pre-miRNA of approximately 60-75 nucleotides that forms an imperfect hairpin (stem-loop) structure. The pre-miRNA is typically then cleaved, e.g., by Dicer, to form the mature miRNA. Mature miRNAs are typically approximately 21-25 nucleotides in length, but can vary, e.g., from about 14 to about 25 or more nucleotides. Some, though not all, miRNAs have been shown to inhibit translation of mRNAs bearing partially
30 complementary sequences. Such miRNAs contain one or more internal mismatches to the corresponding mRNA that are predicted to result in a bulge in the center of the duplex formed by the binding of the miRNA antisense strand to the mRNA. The miRNA typically

forms approximately 14-17 Watson-Crick base pairs with the mRNA; additional wobble base pairs can also be formed. In addition, short synthetic double-stranded RNAs (e.g., similar to siRNAs) containing central mismatches to the corresponding mRNA have been shown to repress translation (but not initiate degradation) of the mRNA. See, for example,

5 Zeng et al. (2003) "MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms" *Proc. Natl. Acad. Sci. USA* 100:9779-9784; Doench et al. (2003) "siRNAs can function as miRNAs" *Genes & Dev.* 17:438-442; Bartel and Bartel (2003) "MicroRNAs: At the root of plant development?" *Plant Physiology* 132:709-717; Schwarz and Zamore (2002) "Why do miRNAs live in the miRNP?" *Genes & Dev.* 16:1025-1031;

10 Tang et al. (2003) "A biochemical framework for RNA silencing in plants" *Genes & Dev.* 17:49-63; Meister et al. (2004) "Sequence-specific inhibition of microRNA- and siRNA-induced RNA silencing" *RNA* 10:544-550; Nelson et al. (2003) "The microRNA world: Small is mighty" *Trends Biochem. Sci.* 28:534-540; Scacheri et al. (2004) "Short interfering RNAs can induce unexpected and divergent changes in the levels of untargeted proteins in

15 mammalian cells" *Proc. Natl. Acad. Sci. USA* 101:1892-1897; Sempere et al. (2004) "Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation" *Genome Biology* 5:R13; Dykxhoorn et al. (2003) "Killing the messenger: Short RNAs that silence gene expression" *Nature Reviews Molec. and Cell Biol.* 4:457-467; McManus (2003)

20 "MicroRNAs and cancer" *Semin Cancer Biol.* 13:253-288; and Stark et al. (2003) "Identification of *Drosophila* microRNA targets" *PLoS Biol.* 1:E60.

[0175] The cellular machinery involved in translational repression of mRNAs by partially complementary RNAs (e.g., certain miRNAs) appears to partially overlap that involved in RNAi, although, as noted, translation of the mRNAs, not their stability, is

25 affected and the mRNAs are typically not degraded.

[0176] The location and/or size of the bulge(s) formed when the antisense strand of the RNA binds the mRNA can affect the ability of the RNA to repress translation of the mRNA. Similarly, location and/or size of any bulges within the RNA itself can also affect efficiency of translational repression. See, e.g., the references above. Typically,

30 translational repression is most effective when the antisense strand of the RNA is complementary to the 3' untranslated region (3' UTR) of the mRNA. Multiple repeats, e.g., tandem repeats, of the sequence complementary to the antisense strand of the RNA can also

provide more effective translational repression; for example, some mRNAs that are translationally repressed by endogenous miRNAs contain 7-8 repeats of the miRNA binding sequence at their 3' UTRs. It is worth noting that translational repression appears to be more dependent on concentration of the RNA than RNA interference does; translational repression is thought to involve binding of a single mRNA by each repressing RNA, while RNAi is thought to involve cleavage of multiple copies of the mRNA by a single siRNA-RISC complex.

[0177] Guidance for design of a suitable RNA to repress translation of a given target mRNA can be found in the literature (e.g., the references above and Doench and Sharp (2004) "Specificity of microRNA target selection in translational repression" *Genes & Dev.* 18:504-511; Rehmsmeier et al. (2004) "Fast and effective prediction of microRNA/target duplexes" *RNA* 10:1507-1517; Robins et al. (2005) "Incorporating structure to predict microRNA targets" *Proc Natl Acad Sci* 102:4006-4009; and Mattick and Makunin (2005) "Small regulatory RNAs in mammals" *Hum. Mol. Genet.* 14:R121 - R132, among many others) and herein. However, due to differences in efficiency of translational repression between RNAs of different structure (e.g., bulge size, sequence, and/or location) and RNAs corresponding to different regions of the target mRNA, several RNAs are optionally designed and tested against the target mRNA to determine which is most effective at repressing translation of the target mRNA.

ANTIBODIES TO GENE PRODUCTS

[0178] An additional class of modulators are antibodies that bind to products of genes linked to the loci herein. The antibodies can be utilized for detecting and/or purifying the gene products e.g., in situ, to monitor the gene product. Antibodies can also be used to block function of gene products, in vivo, in situ or in vitro. As used herein, the term "antibody" includes, but is not limited to, polyclonal antibodies, monoclonal antibodies, humanized or chimeric antibodies and biologically functional antibody fragments, which are those fragments sufficient for binding of the antibody fragment to the protein.

[0179] For the production of antibodies to a relevant gene product, any of a variety of host animals may be immunized by injection with the polypeptide, or a portion thereof. Such host animals may include, but are not limited to, rabbits, mice and rats, to name but a few. Various adjuvants may be used to enhance the immunological response, depending on

the host species, including, but not limited to, Freund's (complete and incomplete), mineral gels such as aluminum hydroxide, surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, keyhole limpet hemocyanin, dinitrophenol, and potentially useful human adjuvants such as BCG (*bacille Calmette-Guerin*) and

5 *Corynebacterium parvum*.

[0180] Polyclonal antibodies are heterogeneous populations of antibody molecules derived from the sera of animals immunized with an antigen, such as target gene product, or an antigenic functional derivative thereof. For the production of polyclonal antibodies, host animals, such as those described above, may be immunized by injection with the encoded
10 protein, or a portion thereof, supplemented with adjuvants as also described above.

[0181] Monoclonal antibodies (mAbs), which are homogeneous populations of antibodies to a particular antigen, may be obtained by any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma technique of Kohler and Milstein (*Nature* 256:495-497, 1975;
15 and U.S. Patent No. 4,376,110), the human B-cell hybridoma technique (Kosbor et al., *Immunology Today* 4:72, 1983; Cole et al., *Proc. Nat'l. Acad. Sci. USA* 80:2026-2030, 1983), and the EBV-hybridoma technique (Cole et al., *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, Inc., pp. 77-96, 1985). Such antibodies may be of any immunoglobulin class, including IgG, IgM, IgE, IgA, IgD, and any subclass thereof. The
20 hybridoma producing the mAb of this invention may be cultivated *in vitro* or *in vivo*. Production of high titers of mAbs *in vivo* makes this the presently preferred method of production.

[0182] In addition, techniques developed for the production of "chimeric antibodies" (Morrison et al., *Proc. Nat'l. Acad. Sci. USA* 81:6851-6855, 1984; Neuberger et al., *Nature* 312:604-608, 1984; Takeda et al., *Nature* 314:452-454, 1985) by splicing the
25 genes from a mouse antibody molecule of appropriate antigen specificity, together with genes from a human antibody molecule of appropriate biological activity, can be used. A chimeric antibody is a molecule in which different portions are derived from different animal species, such as those having a variable or hypervariable region derived from a
30 murine mAb and a human immunoglobulin constant region. Similarly, humanized antibodies can also be produced using available techniques.

[0183] Alternatively, techniques described for the production of single-chain antibodies (U.S. Patent No. 4,946,778; Bird, *Science* 242:423-426, 1988; Huston et al., *Proc. Nat'l. Acad. Sci. USA* 85:5879-5883, 1988; and Ward et al., *Nature* 334:544-546, 1989) can be adapted to produce differentially expressed gene-single chain antibodies.

5 Single chain antibodies are formed by linking the heavy and light chain fragments of the Fv region via an amino acid bridge, resulting in a single-chain polypeptide.

[0184] In one aspect, techniques useful for the production of "humanized antibodies" can be adapted to produce antibodies to the proteins, fragments or derivatives thereof. Such techniques are disclosed in U.S. Patent Nos. 5,932,448; 5,693,762;
10 5,693,761; 5,585,089; 5,530,101; 5,569,825; 5,625,126; 5,633,425; 5,789,650; 5,661,016; and 5,770,429.

[0185] Antibody fragments which recognize specific epitopes may be generated by known techniques. For example, such fragments include, but are not limited to, the F(ab')₂ fragments, which can be produced by pepsin digestion of the antibody molecule, and the
15 Fab fragments, which can be generated by reducing the disulfide bridges of the F(ab')₂ fragments. Alternatively, Fab expression libraries may be constructed (Huse et al., *Science* 246:1275-1281, 1989) to allow rapid and easy identification of monoclonal Fab fragments with the desired specificity.

[0186] The protocols for detecting and measuring the expression of the gene
20 products, using the above mentioned antibodies, are well known in the art. Such methods include, but are not limited to, dot blotting, western blotting, competitive and noncompetitive protein binding assays, enzyme-linked immunosorbant assays (ELISA), immunohistochemistry, fluorescence-activated cell sorting (FACS), and others commonly used and widely described in scientific and patent literature, and many employed
25 commercially.

[0187] One method, for ease of detection, is the sandwich ELISA, of which a number of variations exist, all of which are intended to be encompassed by the present invention. For example, in a typical forward assay, unlabeled antibody is immobilized on a solid substrate and the sample to be tested is brought into contact with the bound molecule
30 and incubated for a period of time sufficient to allow formation of an antibody-antigen binary complex. At this point, a second antibody, labeled with a reporter molecule capable

of inducing a detectable signal, is then added and incubated, allowing time sufficient for the formation of a ternary complex of antibody-antigen-labeled antibody. Any unreacted material is washed away, and the presence of the antigen is determined by observation of a signal, or may be quantitated by comparing with a control sample containing known amounts of antigen. Variations on the forward assay include the simultaneous assay, in which both sample and antibody are added simultaneously to the bound antibody, or a reverse assay, in which the labeled antibody and sample to be tested are first combined, incubated and added to the unlabeled surface bound antibody. These techniques are well known to those skilled in the art, and the possibility of minor variations will be readily apparent. As used herein, "sandwich assay" is intended to encompass all variations on the basic two-site technique. For the immunoassays of the present invention, the only limiting factor is that the labeled antibody be an antibody which is specific for the protein expressed by the gene of interest.

[0188] The most commonly used reporter molecules in this type of assay are either enzymes, fluorophore- or radionuclide-containing molecules. In the case of an enzyme immunoassay, an enzyme is conjugated to the second antibody, usually by means of glutaraldehyde or periodate. As will be readily recognized, however, a wide variety of different ligation techniques exist which are well-known to the skilled artisan. Commonly used enzymes include horseradish peroxidase, glucose oxidase, beta-galactosidase and alkaline phosphatase, among others. The substrates to be used with the specific enzymes are generally chosen for the production, upon hydrolysis by the corresponding enzyme, of a detectable color change. For example, p-nitrophenyl phosphate is suitable for use with alkaline phosphatase conjugates; for peroxidase conjugates, 1,2-phenylenediamine or toluidine are commonly used. It is also possible to employ fluorogenic substrates, which yield a fluorescent product, rather than the chromogenic substrates noted above. A solution containing the appropriate substrate is then added to the tertiary complex. The substrate reacts with the enzyme linked to the second antibody, giving a qualitative visual signal, which may be further quantitated, usually spectrophotometrically, to give an evaluation of the amount of PLAB which is present in the serum sample.

[0189] Alternately, fluorescent compounds, such as fluorescein and rhodamine, can be chemically coupled to antibodies without altering their binding capacity. When activated by illumination with light of a particular wavelength, the fluorochrome-labeled antibody

absorbs the light energy, inducing a state of excitability in the molecule, followed by emission of the light at a characteristic longer wavelength. The emission appears as a characteristic color visually detectable with a light microscope. Immunofluorescence and EIA techniques are both very well established in the art and are particularly preferred for the present method. However, other reporter molecules, such as radioisotopes, chemiluminescent or bioluminescent molecules may also be employed. It will be readily apparent to the skilled artisan how to vary the procedure to suit the required use.

CELL RESCUE AND THERAPEUTIC ADMINISTRATION

[0190] In one aspect, the invention includes rescue of a cell that is defective in function of one or more endogenous genes or polypeptides of Figures 1 and/or 2 (thus conferring the relevant phenotype of interest, e.g., breast cancer susceptibility, resistance, etc.). This can be accomplished simply by introducing a new copy of the gene (or a heterologous nucleic acid that expresses the relevant protein), i.e., a gene having an allele that is desired, into the cell. Other approaches, such as homologous recombination to repair the defective gene (e.g., via chimeraplasty) can also be performed. In any event, rescue of function can be measured, e.g., in any of the assays noted herein. Indeed, this method can be used as a general method of screening cells in vitro for expression or activity of any gene or gene product of Figures 1 and/or 2. Accordingly, in vitro rescue of function is useful in this context for the myriad in vitro screening methods noted above. The cells that are rescued can include cells in culture, (including primary or secondary cell culture from patients, as well as cultures of well-established cells). Where the cells are isolated from a patient, this has additional diagnostic utility in establishing which gene or product is defective in a patient that presents with a relevant phenotype.

[0191] In another aspect, the cell rescue occurs in a patient, e.g., a human, e.g., to remedy a defect. Thus, one aspect of the invention is gene therapy to remedy defects. In these applications, the nucleic acids of the invention are optionally cloned into appropriate gene therapy vectors (and/or are simply delivered as naked or liposome-conjugated nucleic acids), which are then delivered, optionally in combination with appropriate carriers or delivery agents. Proteins can also be delivered directly, but delivery of the nucleic acid is typically preferred in applications where stable expression is desired. Similarly, modulators of any defect identified by the methods herein can be used therapeutically.

[0192] Compositions for administration, e.g., comprise a therapeutically effective amount of the modulator, gene therapy vector or other relevant nucleic acid, and a pharmaceutically acceptable carrier or excipient. Such a carrier or excipient includes, but is not limited to, saline, buffered saline, dextrose, water, glycerol, ethanol, and/or combinations thereof. The formulation is made to suit the mode of administration. In general, methods of administering gene therapy vectors for topical use are well known in the art and can be applied to administration of the nucleic acids of the invention.

[0193] Therapeutic compositions comprising one or more modulator or gene therapy nucleic acid of the invention are optionally tested in one or more appropriate in vitro and/or in vivo animal model of disease, to confirm efficacy, tissue metabolism, and to estimate dosages, according to methods well known in the art. In particular, dosages can initially be determined by activity, stability or other suitable measures of the formulation.

[0194] Administration is by any of the routes normally used for introducing a molecule into ultimate contact with cells. Modulators and/or nucleic acids that encode a relevant sequence can be administered in any suitable manner, optionally with one or more pharmaceutically acceptable carriers. Suitable methods of administering such nucleic acids in the context of the present invention to a patient are available, and, although more than one route can be used to administer a particular composition, a particular route can often provide a more immediate and more effective action or reaction than another route.

[0195] Pharmaceutically acceptable carriers are determined in part by the particular composition being administered, as well as by the particular method used to administer the composition. Accordingly, there is a wide variety of suitable formulations of pharmaceutical compositions of the present invention. Compositions can be administered by a number of routes including, but not limited to: oral, intravenous, intraperitoneal, intramuscular, transdermal, subcutaneous, topical, sublingual, or rectal administration. Compositions can be administered via liposomes (e.g., topically), or via topical delivery of naked DNA or viral vectors. Such administration routes and appropriate formulations are generally known to those of skill in the art.

[0196] The compositions, alone or in combination with other suitable components, can also be made into aerosol formulations (i.e., they can be "nebulized") to be administered via inhalation. Aerosol formulations can be placed into pressurized acceptable propellants,

such as dichlorodifluoromethane, propane, nitrogen, and the like. Formulations suitable for parenteral administration, such as, for example, by intraarticular (in the joints), intravenous, intramuscular, intradermal, intraperitoneal, and subcutaneous routes, include aqueous and non-aqueous, isotonic sterile injection solutions, which can contain antioxidants, buffers, bacteriostats, and solutes that render the formulation isotonic with the blood of the intended recipient, and aqueous and non-aqueous sterile suspensions that can include suspending agents, solubilizers, thickening agents, stabilizers, and preservatives. The formulations of packaged nucleic acid can be presented in unit-dose or multi-dose sealed containers, such as ampules and vials.

10 [0197] The dose administered to a patient, in the context of the present invention, is sufficient to effect a beneficial therapeutic response in the patient over time. The dose is determined by the efficacy of the particular vector, or other formulation, and the activity, stability or serum half-life of the polypeptide which is expressed, and the condition of the patient, as well as the body weight or surface area of the patient to be treated. The size of
15 the dose is also determined by the existence, nature, and extent of any adverse side-effects that accompany the administration of a particular vector, formulation, or the like in a particular patient. In determining the effective amount of the vector or formulation to be administered in the treatment of disease, the physician evaluates local expression, or circulating plasma levels, formulation toxicities, progression of the relevant disease, and/or
20 where relevant, the production of antibodies to proteins encoded by the polynucleotides. The dose administered, e.g., to a 70 kilogram patient are typically in the range equivalent to dosages of currently-used therapeutic proteins, adjusted for the altered activity or serum half-life of the relevant composition. The vectors of this invention can supplement treatment conditions by any known conventional therapy.

25 [0198] For administration, formulations of the present invention are administered at a rate determined by the LD-50 of the relevant formulation, and/or observation of any side-effects of the vectors of the invention at various concentrations, e.g., as applied to the mass or topical delivery area and overall health of the patient. Administration can be accomplished via single or divided doses.

30 [0199] If a patient undergoing treatment develops fevers, chills, or muscle aches, he/she receives the appropriate dose of aspirin, ibuprofen, acetaminophen or other pain/fever controlling drug. Patients who experience reactions to the compositions, such as

fever, muscle aches, and chills are premedicated 30 minutes prior to the future infusions with either aspirin, acetaminophen, or, e.g., diphenhydramine. Meperidine is used for more severe chills and muscle aches that do not quickly respond to antipyretics and antihistamines. Treatment is slowed or discontinued depending upon the severity of the reaction.

DIAGNOSTIC AND PROGNOSTIC ASSAYS

[0200] The nucleic acids, polypeptides, antibodies and other compositions herein may be utilized as reagents (e.g., in pre-packaged kits) for prognosis and diagnosis of susceptibility or resistance to breast cancer phenotypes. The methods can be practiced on subjects known to have one or more symptoms of a breast cancer phenotype as part of a differential diagnosis or prognosis of other diseases. The methods can also be practiced on subjects having a known susceptibility to a breast cancer phenotype. The polymorphic profile of such an individual can increase or decrease the assessment of susceptibility. For example, an individual having two siblings with breast cancer is known to be at increased susceptibility to the disease compared with the general population. A finding of additional factors favoring susceptibility increases the risk whereas finding factors favoring resistance decreases the risk.

[0201] The invention provides methods of determining the polymorphic profile of an individual at one or more of SNPs of the invention. The SNPs includes those shown in Figures 1 and 2. and those in linkage disequilibrium with them. Those in linkage disequilibrium with them usually occur in the same genes or within 100 or 50 or 20 kb of the same genes. SNPs in linkage disequilibrium with the SNPs in the Figures herein can be determined by haplotype mapping. Haplotypes can be determined by fusing diploid cells from different species. The resulting cells are partially haploid, allowing determination of haplotypes on haploid chromosomes (*see*, e.g., US 20030099964). Alternatively, SNPs in linkage disequilibrium with exemplified SNPs can be determined by similar association studies to those described in the examples below.

[0202] The polymorphic profile constitutes the polymorphic forms occupying the various polymorphic sites in an individual. In a diploid genome, two polymorphic forms, the same or different from each other, usually occupy each polymorphic site. Thus, the polymorphic profile at sites X and Y can be represented in the form X(x1, x1), and Y (y1,

y2), wherein x1, x1 represents two copies of allele x1 occupying site X and y1, y2 represent heterozygous alleles occupying site Y.

[0203] The polymorphic profile of an individual can be scored by comparison with the polymorphic forms associated with resistance or susceptibility to breast cancer

phenotypes occurring at each site as shown in the Figures herein. The comparison can be performed on at least, e.g., 1, 2, 5, 10, 25, 50, or all of the polymorphic sites, and optionally, others in linkage disequilibrium with them. The polymorphic sites can be analyzed in combination with other polymorphic sites. However, the total number of polymorphic sites analyzed is usually fewer than 10,000, 1000, 100, 50 or 25 and can be about 10 or less, about 5 or less, or about 2 or less.

[0204] The number of resistance or susceptibility alleles present in a particular individual can be combined additively or as ratio to provide an overall score for the individual's genetic propensity to breast cancer phenotypes (see USSN 60,566,302, filed

April 28, 2004, USSN 60/590,534, filed July 22, 2004, USSN 10/956,224 filed September 30, 2004, and PCT US05/07375 filed March 3, 2005). Resistance alleles can be arbitrarily

each scored as +1 and susceptibility alleles as -1 (or vice versa). For example, if an individual is typed at 100 polymorphic sites of the invention and is homozygous for resistance at all of them, he could be assigned a score of 100% genetic propensity to resistance to breast cancer phenotypes or 0% propensity to susceptibility to breast cancer

phenotypes. The reverse applies if the individual is homozygous for all susceptibility alleles. More typically, an individual is homozygous for resistance alleles at some loci, homozygous for susceptibility alleles at some loci, and heterozygous for resistance/susceptibility alleles at other loci. Such an individual's genetic propensity for

breast cancer phenotypes can be scored by assigning all resistance alleles a score of +1, and all susceptibility alleles a score of -1 (or vice versa) and combining the scores. For example, if an individual has 102 resistance alleles and 204 susceptibility alleles, the

individual can be scored as having a 33% genetic propensity to resistance and 67% genetic propensity to susceptibility. Alternatively, homozygous resistance alleles can be assigned a score of +1, heterozygous alleles a score of zero and homozygous susceptibility alleles a score of -1. The relative numbers of resistance alleles and susceptibility alleles can also be expressed as a percentage. Thus, an individual who is homozygous for resistance alleles at 30 polymorphic sites, homozygous for susceptibility alleles at 60 polymorphic sites, and

heterozygous at the remaining 63 sites is assigned a genetic propensity of 33% for resistance. As a further alternative, homozygosity for susceptibility can be scored as +2, heterozygosity, as +1 and homozygosity for resistance as 0.

[0205] The individual's score, and the nature of the polymorphic profile are useful in

5 prognosis or diagnosis of an individual's susceptibility to breast cancer phenotypes.

Optionally, a patient can be informed of susceptibility to a breast cancer phenotype

indicated by the genetic profile. Presence of a high genetic propensity to breast cancer

phenotypes can be treated as a warning to commence prophylactic or therapeutic treatment.

For example, individuals with elevated risk of developing a breast cancer phenotype may be

10 monitored differently (e.g., more frequent mammography) or may be treated

prophylactically (e.g., with one or more drugs). Presence of a high propensity to a breast

cancer phenotype also indicates the utility of performing secondary testing, such as a

biopsy.

[0206] Polymorphic profiling is useful, for example, in selecting agents to effect

15 treatment or prophylaxis of breast cancer phenotypes in a given individual. Individuals

having similar polymorphic profiles are likely to respond to agents in a similar way.

[0207] Polymorphic profiling is also useful for stratifying individuals in clinical

trials of agents being tested for capacity to treat breast cancer phenotypes or related

conditions. Such trials are performed on treated or control populations having similar or

20 identical polymorphic profiles (see EP99965095.5), for example, a polymorphic profile

indicating an individual has an increased risk of developing a breast cancer phenotype. Use

of genetically matched populations eliminates or reduces variation in treatment outcome due to genetic factors, leading to a more accurate assessment of the efficacy of a potential drug.

Computer-implemented algorithms can be used to identify more genetically homogenous

25 subpopulations in which treatment or prophylaxis has a significant effect notwithstanding

that the treatment or prophylaxis is ineffective in more heterogeneous larger populations. In

such methods, data are provided for a first population with a breast cancer phenotype treated

with an agent, and a second population also with the breast cancer phenotype but treated

with a placebo. The polymorphic profile of individuals in the two populations is determined

30 in at least one polymorphic site in or within 100 kb or 50 kb or 20 kb of a gene selected

from those shown in Figures 1 and/or 2. Data are also provided as to whether each patient

in the populations reaches a desired endpoint indicative of successful treatment or

prophylaxis. Subpopulations of each of the first and second populations are then selected such that the individuals in the subpopulations have greater similarity of polymorphic profiles with each other than do the individuals in the original first and second populations. There are many criteria by which similarity can be assessed. For example, one criterion is to require that individuals in the subpopulations have at least one susceptibility allele at each of at least ten of the above genes. Another criterion is that individuals in the subpopulations have at least 75% susceptibility alleles for each of the polymorphic sites at which the polymorphic profile is determined. Regardless of the criteria used to assess similarity, the endpoint data of the subpopulations are compared to determine whether treatment or prophylaxis has achieved a statistically significant result in the subpopulations. As a result of computer implementation, billions of criteria for similarity can be analyzed to identify one or a few subpopulations showing statistical significance.

[0208] Polymorphic profiling is also useful for excluding individuals with no predisposition to breast cancer phenotypes from clinical trials. Including such individuals in the trial increases the size of the population needed to achieve a statistically significant result. Individuals with no predisposition to breast cancer phenotypes can be identified by determining the numbers of resistances and susceptibility alleles in a polymorphic profile as described above. For example, if a subject is genotyped at ten sites in ten genes of the invention associated with breast cancer phenotypes, twenty alleles are determined in total. If over 50% and preferably over 60% or 75% percent of these are resistance genes, the individual is unlikely to develop a breast cancer phenotype and can be excluded from the trial.

[0209] In other embodiments, stratifying individuals in clinical trials may be accomplished using polymorphic profiling in combination with other stratification methods, including, but not limited to, family history, risk models (e.g., Gail Score, Claus model), clinical phenotypes (e.g., atypical lesions and breast density), and specific candidate biomarkers (e.g., IGF1, IFG2, IGFBP3, Ki-67, and estradiol). For example, stratification of higher risk in chemoprevention trials that includes stratification based on polymorphic profiles can improve outcomes. In particular, markers linked to FGFR2 can be used to stratify anti-VEGF or anti-angiogenesis therapy response, and markers linked to PKHD1 can be used to stratify anti-EGF therapy efficacy (anti-EGF therapies are active in patients with polycystic kidney and hepatic disease).

[0210] Polymorphic profiles can also be used after the completion of a clinical trial to elucidated differences in response to a given treatment. For example, the set of polymorphisms can be used to stratify the enrolled patients into disease sub-types or classes. It is also possible to use the polymorphisms to identify subsets of patients with similar polymorphic profiles who have unusual (high or low) response to treatment or who do not respond at all (non-responders). In this way, information about the underlying genetic factors influencing response to treatment can be used in many aspects of the development of treatment (these range from the identification of new targets, through the design of new trials to product labeling and patient targeting). Additionally, the polymorphisms can be used to identify the genetic factors involved in adverse response to treatment (adverse events). For example, patients who show adverse response may have more similar polymorphic profiles than would be expected by chance. This allows the early identification and exclusion of such individuals from treatment. It also provides information that can be used to understand the biological causes of adverse events and to modify the treatment to avoid such outcomes.

[0211] Polymorphic profiles can also be used for other purposes, including paternity testing and forensic analysis as described by US 6,525,185. In forensic analysis, the polymorphic profile from a sample at the scene of a crime is compared with that of a suspect. A match between the two is evidence that the suspect in fact committed the crime, whereas lack of a match excludes the suspect. The present polymorphic sites can be used in such methods, as can other polymorphic sites in the human genome.

EXAMPLES

[0212] The following examples are offered to illustrate, but not to limit the claimed invention. One of skill will recognize a variety of non-critical parameters that can be altered within the scope of the invention.

EXAMPLE 1: STRATEGIES FOR IDENTIFICATION OF BREAST CANCER MARKERS

Introduction: Identifying common genetic variants

[0213] There are important applications to public health in the identification of breast cancer marker alleles. Where genetic variation is due to many loci, risks to individuals vary widely, depending upon the number of high-risk alleles inherited at

susceptibility loci. Our analyses based on the model of Antoniou et al¹³ suggest that there may be as much as 40-fold difference in risk between the top and bottom 20% of the population. Under the same model, half of all breast cancers occur in the 12% of women at greatest risk, and these women have risks of at least 1 in 8 by age 70. By contrast, the 50% of women at least risk have only 12% of the cancers, and individual risks of less than 1 in 30¹⁴. Genes that are identified as being correlated to breast cancer risk can be used for estimation of associated and individual risks. The practical consequences of this risk estimation are substantial.

[0214] Common genetic variants that confer modest degrees of risk have individually important effects at the population level. For example, a common variant, with frequency 20%, which increases individual risk by only 1.5-fold would account for 15% of the population burden of a cancer. The analogy is with moderate elevations of blood pressure, or of cholesterol, in cardiovascular disease. If the variant indicates a feasible mechanism for intervention, this also provides novel possibilities for targeted prevention.

[0215] In addition to these practical outcomes, the identification of cancer susceptibility genes helps to clarify mechanisms of carcinogenesis (as has already happened, for example, with *BRCA1* and *BRCA2*). Extending beyond known candidates to a whole genome search has the great advantage that totally novel mechanisms emerge. These mechanisms also provide new therapeutic targets.

[0216] Finally, knowledge of susceptibility genes allows us to clarify the effects of lifestyle risk factors by studying the effects of genes and these risk factors in combination, using for example the EPIC cohort.

Breast cancer

[0217] Although arguments may be made for association studies in many cancers, there are several reasons why it is particularly appropriate to carry out a study in breast cancer. It is the commonest cancer in women, and its aetiology is still poorly understood. The genetic basis for the disease has been more thoroughly investigated than for any other common cancer. As a result, the evidence in favor of a polygenic basis is clearer than for other cancers. Long-term studies have assembled sufficiently large series of cases to identify susceptibility loci reliably. In addition, cases with a strong family history are available through cancer genetics clinics and can provide substantial gains in efficiency in

the association study (see 'Research Proposal'). Finally, there are interventions that can be offered to women found to be at increased risk. For example, prophylactic oophorectomy can reduce substantially the subsequent risk of breast cancer¹⁵. Recent studies suggest that screening by MRI may provide much greater sensitivity than mammography but at significantly greater cost¹⁶.

Study design

[0218] Genotypes for 200,000 single nucleotide polymorphisms (SNPs) are determined in a set of 400 familial breast cancer cases and 400 controls from the EPIC cohort; 5% of these SNPs showing the strongest association are analyzed in further population-based series of 4,600 cases and 4,600 controls. Positive associations at this stage are confirmed in further large case-control series.

[0219] In addition to the breast cancer endpoint, many quantitative phenotypes are available in the control set and provide additional data for genetic analysis. These include phenotypes (e.g. mammographic patterns, hormone levels) that are related to cancer risk.

Yield

[0220] The scan evaluates single nucleotide polymorphisms of frequency 10% or greater (and some in the 5-10% range) across the entire genome outside repetitive sequences. It has approximately 80% power to detect any common variant within these regions that accounts for 2% or more of the overall inherited component of breast cancer.

The design of studies to search for common susceptibility variants

[0221] An efficient design to identify common low risk alleles is a case/control study. Variants that are associated with susceptibility are identified by their occurrence at a significantly higher frequency in cancer cases than in controls matched for genetic background. In this study, the variants are single nucleotide polymorphisms (SNPs). Most often, the active or functional variant that might be relevant to disease susceptibility is not known, and so the search relies on a set of 'tagging' SNPs that can report on the presumed (but unknown) active variants.

[0222] The case-control association study approach has already been used extensively in breast cancer on a "candidate gene" basis. Polymorphisms in the coding region and introns of close to 100 genes have previously been studied in this way. These include, for example, genes involved in sex-steroid hormone metabolism, cell-cycle control

and DNA repair. Although some associations have been suggested for common variants, none has been definitively established. The results to date suggest that the majority of the variation in breast cancer risk is not due to variants in the intragenic DNA of the genes that would be a first choice as breast cancer candidates. There are, moreover, serious limitations to a candidate gene approach. It is slow and relatively expensive, being dependent on developing assays on a SNP by SNP basis for each gene to be tested; it is incomplete in its coverage even of the candidate genes, in particular ignoring, in most cases, potential regulatory variation; and it is restricted by current knowledge of the biology of the disease. A genome-wide search, by contrast, has the potential to identify active common variants without any prior knowledge of function or location.

Genome scanning SNPs

[0223] A requirement for a genome scan is to define a set of SNPs that provides the best compromise between completeness and cost in reporting on the set of all other SNPs across the genome. Perlegen Sciences ([www\(dot\)perlegen\(dot\)com](http://www.perlegen.com)) have identified 1.1 million common SNP markers (a density of 1 SNP per 2 kb) by resequencing the non-repetitive sequences of the human genome in 20 to 50 haploid genomes segregated in human/rodent somatic cell hybrids. This SNP search is based on a similar strategy to that reported in a study of chromosome 21 reported by Patil et al¹⁷. From this they have defined, using a dynamic programming algorithm¹⁸, a set of 200,000 tagging SNPs that report unambiguously more than 80% of common haplotypes defined by the complete set of 1.1 million SNPs. It is this set of 200,000 tagging SNPs that were used in this example.

[0224] The SNPs are typed on high-density oligonucleotide arrays developed by Affymetrix, which have been extensively validated in routine use. Briefly, the array design uses 80 features (25-mer oligonucleotides) to query each SNP. The 80 features comprise 10 overlapping feature sets where each feature set includes 4 features specific for the reference allele (one perfect match and 3 mismatch features) and 4 similar features for the alternative allele. By comparing the fluorescence intensities of perfect match features for the reference allele with those that are perfect matches for the alternative allele, the three possible SNP genotypes (common homozygote, heterozygote and rare homozygote) can be distinguished. To carry out the genotyping assay, regions of the sample genome containing SNPs of interest are specifically amplified using multiplex (78-plex) short range PCR. The PCR products from each individual are pooled and labeled with biotin to create target DNA. The

target DNA is hybridized to the SNP-typing high density oligonucleotide arrays. After overnight hybridization, the arrays are washed, stained and scanned for fluorescence intensities.

[0225] In this example, the features to genotype the 200,000 SNPs are arrayed onto a series of 6 high-density arrays, requiring that target DNA with a complexity of approximately 30,000 SNPs be used for hybridization. A target of this complexity gives a call rate of 97.3%. Comparison of the high density array genotyping technology with other technologies (real-time PCR and fluorescence polarization) shows a consistent concordance of greater than 99% in approximately 20,000 genotypings with up to 20 different SNPs. The technology has proven to be robust when used with DNA from a variety of collaborating clinical and research laboratories, and to work well with genome amplified DNA.

[0226] A property of the 200,000 tagging SNPs to be used in this study, is their ability to 'report' on all other SNP variants within the genome. For given power, the required sample size is proportional to $1/r^2$, where r is the coefficient of linkage disequilibrium between the functional variant which is being sought and the most closely linked tagging SNP¹⁹.

[0227] The distribution of r^2 between non-tagging SNPs and their corresponding tagging SNP was determined empirically by genotyping 1608 SNPs, evenly spaced over a 4Mb region of genomic DNA, in 28 unrelated individuals (56 chromosomes): see Table 1. The mean r^2 for all 988 tester SNPs, each with a minor allele frequency of >10%, was 67%, with 69% of SNPs having an r^2 greater than 0.5.

Table 1

[0228] **Table 1** Distribution of r^2 values between 417 selected 'tagging' SNPs and 988 'tester' SNPs, determined in a 4Mb segment of chromosome 21 in 28 individuals distinct from those used for SNP discovery. 15% of the total set of SNPs had minor allele frequencies 1-10%; 85% had frequencies greater than 10% with an even distribution between 10 and 50%. All tester SNPs had a minor allele frequency >10%. The average spacing of the 417 tagging SNP set was 1 per 10 kb, similar to that in the 200,000 SNP set to be used in the genome scan.

Table 1

	r^2	% of tester SNPs
5	0.9-1.0	34%
	0.8-0.89	10%
	0.7-0.79	9%
	0.6-0.69	9%
	0.5-0.59	8%
10	0.4-0.49	6%
	0.3-0.39	7%
	0.2-0.29	7%
	0.1-0.19	8%
	0.0-0.09	2%
15		

[0229] Increasing the set of 200,000 tagging SNPs increases the proportion of SNPs in the total genome that are reported, but at increased cost. Even a set of 1.1 million SNPs would not provide complete coverage, since some bases could not be assayed and some common SNPs are missed due to the limited number of chromosomes surveyed. We conclude that the 200,000 tagging SNP set provides a good compromise between coverage and cost.

[0230] Previous studies focus on the genetics of cancer predisposition, including an increasing focus on low penetrance susceptibility^{13,14,20-23}. Relevant topics for include: (1) assembly of case and control sets; (2) development of genetic models for breast cancer susceptibility; (3) establishment of laboratory facilities for association studies.

(1) Sample sets

[0231] *Breast cancer cases – population based set.* We have assembled a population based set of (currently 4900) cases of invasive breast cancer diagnosed before the age of 70, ascertained through the local Anglian Cancer Registry. The median time from diagnosis to completion of recruitment is 6 months (interquartile range 3 to 9 months). 65% of all eligible cases have provided a blood sample. This set provides some of the familial cases to be analyzed in stage 1 of our study, and the population based series of cases for stage 2. In addition to a blood sample, subjects complete a questionnaire which includes family history to second degree relatives, reproductive history, breast feeding, oral contraceptive and HRT use, benign breast disease, medical history including other cancers, smoking, alcohol, education and ethnic group. Registry data include clinical stage, pathological grade and stage, simple treatment data and follow-up for survival. Paraffin

blocks of tumour are available from 800 cases currently, and can in principle be collected for the majority of the entire set if funding is available.

[0232] *Breast cancer cases – familial set.* Through the familial breast cancer clinic in Cambridge and the population-based set described above, a set of over 200 cases with
5 either a strong family history of breast cancer or bilateral primary breast cancers has been assembled, that have tested negative for *BRCA1* and *BRCA2* mutations. These cases, together with similar cases obtained by collaboration with other CR UK groups, forms the ‘genetically enriched’ case set for the first stage of the association study.

[0233] *Controls.* Control DNAs for both stage 1 and stage 2 of the study are
10 obtained from the EPIC-Norfolk cohort²⁴. This is part of the Multicentre European Prospective Investigation of Cancer, a 450,000 strong population based cohort of men and women aged 45 to 70 at recruitment from whom blood, extensive epidemiological information and follow up are available. The 25,000 participants in EPIC-Norfolk are
15 volunteers ascertained from family medical practices in Norfolk, which is within the same Anglian region from which the breast cancer cases are obtained. In addition to providing the controls for the gene discovery phase of this project, the larger EPIC cohort provides samples and data for confirmation of positive associations and for the investigation of gene/lifestyle interactions at the follow-up stage.

[0234] The study population is relatively homogeneous ethnically, with more than
20 95% of the population recorded as white and born in the U.K. Evidence for population stratification in this population has been evaluated by genotyping 1655 controls for SNPs in 23 unlinked genes. No significant association between unlinked loci was found, indicating no evidence of stratification²⁵.

[0235] *Additional phenotypic information in the EPIC control set.* Of relevance to
25 the study, extensive phenotypic information is either available (in subsets of individuals) or can readily be obtained, from the 400 EPIC controls who will be genotyped for 200,000 SNPs in stage 1. These quantitative or semi-quantitative phenotypes are evaluated for genotype associations. Phenotypes relevant to breast cancer include: mammographic density, heel bone density, body mass index, and a range of measurements in serum, of
30 which so far estrogen metabolites, SHBG, IGF-1, and some cytokines are already available in different sets of individuals. Other phenotypes include blood pressure, lipid profiles, C-

reactive protein, fibrinogen, full blood count, glycated haemoglobin and thyroid function. In 2004 and 2005 limited recalls are planned for re-interview and further blood sampling, with the possibility of additional phenotyping and the collection of fresh serum and viable cells.

(2) Genetic models for breast cancer susceptibility.

- 5 [0236] We analysed the first 1500 breast cancers ascertained in the Anglian population-based study for mutations in *BRCA1* and *BRCA2*, and found, consistent with Peto et al, that only 15% of the familial clustering of breast cancer is attributable to mutations in these genes. Our segregation analysis of the patterns of breast cancer in the families of cases in this study (subsequently tested on other series) led to the polygenic
10 model summarised earlier¹³. This model in turn underlies the calculations of the increased power for association studies provided by using familial rather than unselected cases, which is the basis of the proposed two-stage design used in this example²⁶.

(3) Laboratory set up for sample processing and SNP genotyping.

- [0237] A moderate throughput genotyping laboratory based on the 384-well Taqman
15 platform, is used for candidate gene association studies. Genotyping capacity is ~100,000 SNPs per week.

- [0238] In brief, the laboratory set up is as follows. Study participants are given a code- number at recruitment which remains attached to all their data and their biological sample tubes as a bar-code. Within the laboratory, samples are tracked with a Laboratory
20 Information Management System (Thermo, Altringham UK). DNA is extracted, in batches of 96 subjects, from whole blood in coded tubes by Whatman Ltd (Ely, UK) and returned in coded arrays with DNA normalised to 40ng/μl. Pre-amplification of the whole genome is performed on the normalised arrays and the products stored in aliquots. 384-well working-stocks for genotyping are created from equal numbers of cases and controls interleaved with
25 blank wells as negative controls. 3% of samples from a study are duplicated. Thus, the cases and controls (described above) are held in 13 plates – 12 of unique samples and a 13th of duplicates. Genotyping is carried out on all study-plates simultaneously – reagents are added by robot (Matrix, UK), thermal cycling on MJ Tetrads (GRI, UK) and end-point fluorescence detection by 7900 Sequence Detector (ABI, Warrington, UK). Genotypes are
30 exported to a database and linked to the phenotypic data on each subject. Control genotypes are tested for departure from Hardy-Weinberg Equilibrium as a final quality control step.

RESEARCH DESIGN

[0239] The study is organised in stages:

[0240] Stage 1. The full set of 200,000 tagging SNPs is analysed in 400 unrelated breast cancer cases enriched for family history, and 400 female controls drawn from the EPIC study. The breast cancer cases will have been screened negative for *BRCA1/2* mutations.

[0241] Stage 2. SNPs that show a significant difference in frequency between the cancer series and the control series, at the $p < 0.05$ level, are re-evaluated in a further 4600 breast cancer cases and 4600 matched controls.

10 Rationale for the research design

[0242] The staged design is chosen to minimise the amount of genotyping required, while retaining a high power to detect SNPs with a modest effect on risk. With the proposed thresholds, approximately 10,000 SNPs will go forward to stage 2, while substantially fewer (depending on the number of “true” associations) are expected to be put forward at the end of stage 2 for additional verification in other studies. Calculations have shown that such a staged design is very efficient compared with genotyping all samples for all SNPs²⁷.

15 Cases – stage 1

[0243] Cases are women with invasive breast cancer with at least two first degree relatives with breast cancer, or equivalently strong family history (for example, one first and two second degree relatives affected). These women are selected from cancer genetics centres in the U.K. or from the Anglian Breast Cancer Study. Women whose ethnic group is not recorded as white will be excluded.

[0244] We have previously demonstrated that the power to detect associations is strongly related to the degree of family history of the cases²⁶. The use of cases with two affected first degree relatives reduces the required sample size by at least fourfold, as compared with using unselected cases. From amongst all available cases, we will select the four hundred cases with the strongest family history. If more than one case is available from the same family we will choose the case with the strongest immediate family history, so that all cases in the set will be unrelated.

[0245] All cases are screened (and are negative) for mutations in *BRCA1* and *BRCA2*. This screening includes, screening of all exons and splice junctions by a sensitive screening technique (e.g. CSGE). This is done because it is unknown whether low penetrance alleles that influence breast cancer risk in non-carriers of *BRCA1* or *BRCA2* mutations will also influence the risk in carriers. The analysis of Antoniou et al¹³ suggests a similar “polygenic” component in carriers. However, it is possible that genes modifying the risk in *BRCA1* and *BRCA2* carriers may be different, particularly given the distinctive pathology of *BRCA1* tumours. *BRCA1* and *BRCA2* mutations would be present in more than 20% of cases selected by the criteria used in stage 1. If a polymorphism of interest did not influence the disease risk in carriers, inclusion of carriers could reduce the power of the study. The study, thus, conservatively screens for and excludes known *BRCA1* and *BRCA2* mutation carriers. This approach is estimated to exclude approximately 70% of *BRCA1* mutations and 90% of *BRCA2* mutations, so that less than 5% of cases in the final set are likely to harbour unidentified mutations.

Cases – stage 2

[0246] Cases at stage 2 consist of 4,600 cases drawn from the population-based Anglian Breast Cancer (ABC) Study.

[0247] While there is an argument for the use of “enriched” cases to maximise power and minimise costs at stage 1, the case is more finely balanced at stage 2. Use of familial cases increases power, but the gain is less marked because the main determinant of power is the efficiency of stage 1. At the same time, the population-based case-control sets are already in use for candidate gene association studies, and DNA samples are already arrayed from stage 1. Developing a new set of familial cases on this scale would entail considerable delay and expense. Secondly, the cases are closely matched geographically to the controls, providing more protection against false positive associations due to regional variation in allele frequencies. Thirdly, the population-based series provides a direct estimate of the relative risk associated with each SNP or haplotype. Fourthly, the ABC study has collected systematically information on lifestyle risk factors and clinical outcome of the cancers. This provides the potential for further analyses to study associations with survival and interactions with lifestyle risk factors. The same quality of information could not be obtained on familial cases.

[0248] In summary, by utilizing series of both enriched and population-based case series, we are optimizing the power to detect true associations, while at the same time gaining the added value of genotyping in a well-characterized population-based case-control study.

5 Controls

[0249] Controls for both stages 1 and 2 will be women from EPIC study as described above. The age distribution of the controls is similar to that of the cases. Women who are known to have developed cancer, or who are non-white, will be excluded. Controls for stage 1 will be sampled from a subcohort of 2,000 postmenopausal women for whom
10 detailed analyses of sex steroid hormones and mammographic density have been conducted²⁸.

[0250] Ethical approvals that cover the use of both case and control samples for genetic association studies have been obtained. Both cases and controls have given informed consent for their DNA to be used for such genetic studies.

15 STATISTICAL CONSIDERATIONS

Statistical Analysis

[0251] The primary analysis, at both stages 1 and 2, is to evaluate the association of each SNP individually with breast cancer. Epidemiological studies suggest little or no difference in the relative risk of breast cancer between mothers of and sisters of cases,
20 except possibly at very young ages^{1,12}, indicating that most susceptibility alleles have little recessive component (as in the polygenic model of Antoniou et al¹³). The primary analyses are, therefore, based on a trend test for the difference in allele frequencies between cases and controls²⁹. Cases are weighted by family history to improve the efficiency of the test.

[0252] In principle, haplotype analysis or joint genotype analysis can provide some
25 improvement in power³⁰. In the current design, however, haplotype analysis is largely redundant since only a small minority of SNPs will be taken through to stage 2, and the power calculations have therefore assumed single SNP analyses. The cost of taking all tagging SNPs in a block through to stage 2 to allow full haplotype analyses would outweigh any gain in power. Haplotype analysis is utilized in those cases where more than one linked
30 SNP in the same LD block is typed at stage 2, and will be utilized extensively in follow-up studies.

Power calculations

[0253] The power of the study has been derived on the basis of a significance level of $p=10^{-4}$ over both stages combined. Approximately 12 loci would be expected to be significant at this level by chance (given the staged design), leaving a manageable number of loci to retest in larger series and a favourable ratio of “true”: “false” positives.

[0254] The power calculations assume that the cases in stage 1 have a family history of two affected first degree relatives. In practice, the power is somewhat greater since this is the minimum criterion and many of the cases will have a stronger family history.

Examples of the power to detect a disease susceptibility allele are given in Table 2, for different values of the disease allele frequency and relative risk, assuming the estimated distribution of r^2 from the tagging set. (For polymorphisms with alleles of frequency .05, the power has been calculated by assuming that the polymorphism is in LD, at $D'=1$, with a randomly chosen common polymorphism from the set). For common alleles, the power is principally dependent on the contribution of the locus to the overall genetic variance, and is at least 50% for loci explaining 1% of the variance and approximately 80% for loci explaining 2% of the variance. In contrast, for alleles with frequency less than 5%, power is poor unless the effect size is very large.

Table 2

[0255] **Table 2.** Estimated overall power to detect a dominant susceptibility locus with a given allele frequency conferring a given relative risk ($P<.0001$ after two stages) assuming the distribution of r^2 between the susceptibility locus and a tagging SNP based on previously reported data (Table 1). Percentage of overall genetic variance explained in brackets.

Relative risk	Allele frequency					
	.2		.1		.05	
1.2	53%	(0.9%)	27%	(0.6%)	<1%	(0.3%)
1.3	79%	(2.0%)	61%	(1.2%)	<1%	(0.7%)
1.5	91%	(5.2%)	86%	(3.2%)	6%	(1.8%)
2.0	99%	(17.8%)	97%	(11.5%)	29%	(6.7%)

[0256] For quantitative traits measured in the controls, the power is approximately 50% to detect loci explaining at least 5% of the phenotypic variance, at the 5% significance level. These loci would be available for further evaluation in future large studies within the EPIC cohort.

Detailed evaluation of susceptibility loci

[0257] Once a significant association is identified, further polymorphisms are evaluated in the region in an attempt to establish the most strongly associated variant or haplotype. The general procedure will be similar to that used for investigating candidate genes. The available databases are searched for known SNPs. Where no systematic search for all available SNPs has been conducted, in-house resequencing of a limited number of individuals (e.g., n=48) is used. After excluding SNPs that are in complete LD, informative SNPs are genotyped in the cases and controls used in stages 1 and 2. Multiple logistic regression is used to investigate the joint effect of multiple SNPs. Further investigations can be performed to identify functional variant(s).

Further Evaluation of Quantitative Trait Loci

[0258] SNPs that exhibit significant associations at stage 1 with quantitative traits are available for replication in further series. Since the number of quantitative traits is large, prioritization is performed on the basis of the strength of the association, the plausibility of the locus and the importance of the phenotype. For example, associations with serum sex steroid hormone levels and mammographic density are prioritized favorably, because these are related to breast cancer risk. SNPs associated with these phenotypes are typed in a further 1,600 samples from postmenopausal women in EPIC. If the associations are replicated, they are pursued as above.

[0259] References:

1. Collaborative Group in Hormonal Factors in Breast Cancer (2001) Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* 358:1389-1399.
2. Lichtenstein P et al (2000) Environmental and heritable factors in the causation of cancer - analyses of cohorts of twins from Sweden, Denmark, and Finland. *New Engl J Med* 243:78-85.
3. Peto J, Mack TM (2000) High constant incidence in twins and other relatives of women with breast cancer. *Nature Genet* 26:411-414.
4. Antoniou A et al (2003) Average risks of breast and ovarian cancer associated with mutations in *BRCA1* or *BRCA2* detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 72:1117-1130.

5. Peto J et al (1989) The prevalence of *BRCA1* and *BRCA2* mutations amongst early onset breast cancer cases in the U.K. *J Natl Cancer Inst* 91:943-949.
6. The Anglian Breast Cancer Study Group (2000) Prevalence of *BRCA1* and *BRCA2* mutations in a large population based series of breast cancer cases. *Br J Cancer* 83:1301-1308.
7. Easton DF (1999) How many more breast cancer predisposition genes are there? *Breast Cancer Res* 1:1-4.
8. Ford D et al (1998) Genetic heterogeneity and Penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. *Am J Hum Genet* 62:334-345.
9. Thompson D et al (2002) Evaluation of linkage of breast cancer to the putative *BRCA3* locus on chromosome 13q21 in 128 multiple case families from the Breast Cancer Linkage Consortium. *Proc Natl Acad Sci USA* 99:827-831.
10. Huusko P et al (2003) Genome-wide scanning for linkage in Finnish breast cancer families. *Eur J Hum Genet*, in press.
11. Antoniou AC et al (2001) Evidence for further breast cancer susceptibility genes in addition to *BRCA1* and *BRCA2* in a population based study. *Genet Epidemiol* 21:1-18.
12. Cui J et al (2000) After *BRCA1* and *BRCA2* – what next? Multifactorial segregation analysis of three-generational, population-based Australian female breast cancer families. *Am J Hum Genet* 68:420-431.
13. Antoniou AC et al (2002) A comprehensive model for familial breast cancer incorporating *BRCA1*, *BRCA2* and other genes. *Brit J Cancer* 86:76-83.
14. Pharoah PDP et al (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics* 31:33-36.
15. Titus-Ernstoff L et al (1998) Menstrual factors in relation to breast cancer risk. *Cancer Epidemiol Biomarkers Prev.* 7: 783-9.
16. Kriege et al (2003) MRI screening for breast cancer in women with high familial and genetic risk: First results of the Dutch MRI screening study (MRISC). *Proc Am Soc Clin Oncol* 22:A5.
17. Patil N et al (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719-1723.
18. Zhang K et al (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* 99: 7335-7339.

19. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 69: 1-14.
20. Dunning AM et al (1999). A systematic review of genetic polymorphisms and breast cancer risk. *Cancer Epidemiol Biomarkers Prevention* 8:843-854.
- 5 21. Healey CS et al (2000) A common variant in *BRCA2* is associated with both breast cancer risk and prenatal viability. *Nature Genet* 26:362-364.
22. Kuschel B et al (2002) Variants in DNA double strand break repair genes and breast cancer susceptibility. *Hum Mol Genet* 11:1399-1407.
23. Dunning AM et al (2003) A TGF β -1 signal peptide variant increases secretion *in*
 10 *vitro* and is associated with increased incidence of invasive breast cancer. *Cancer Res* 63:2610-15.
24. Day N et al (1999) EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer* 80 Suppl 1:95-103.
25. Goode EL et al (2001) Assessment of population stratification in a large population-
 15 based cohort. *Genet Epidemiol* 21:A126.
26. Antoniou A, Easton DF (2003) Polygenic inheritance of breast cancer: implications for design of association studies. *Genet Epidemiol* 25:190-202.
27. Satagopan JM et al (2002) Two-staged designs for gene-disease association studies. *Biometrics* 58:163-170.
- 20 28. Dunning AM et al (2004) Polymorphisms associated with circulating sex hormone levels in post-menopausal women. *J Natl Cancer Inst*, in press.
29. Sasieni PD (1997) From genotypes to genes: doubling the sample size. *Biometrics* 53:1253-1261.
30. Chapman JP et al (2003) Detecting disease associations due to linkage
 25 disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 56: 18-31.

EXAMPLE 2: MARKER POLYMORPHISMS ASSOCIATED WITH BREAST CANCER PREDISPOSITION

- [0260] SNPs identified as being associated with breast cancer risk (predisposition)
- 30 are set forth in Figures 1 and 2. Figure 1 provides the currently most preferred associations; Figure 2 provides additional associations.

[0261] Sequences for given dbSNP_rsID numbers (*see*, "REFSNP_ID," column 2 from the Figures) are found at: <http://www.ncbi.nlm.nih.gov/SNP/>.

In Figures 1 and 2, the second column is labeled "REFSNP_ID". The values in this column are SNP identification numbers according to the dbSNP database established and

5 maintained by NCBI of the US National Library of Medicine at the US National Institute of Health. The NCBI dbSNP database is publicly accessible and considerable additional information can be easily viewed by searching the database using the rsID numbers provided in the Figures by entering the number prefixed by "rs" in the database search window and clicking on "Search." The information provided can include, but is not limited to, alleles at the SNP locus, flanking nucleotide sequences, and submission information.

[0262] The SNP_ID column numbers (*See*, column 1, Figures 1 and 2) reference publicly available Perlegen SNP identification numbers, which can be viewed with associated information at Perlegen(dot)com, using the company's available genome browser at genome(dot)perlegen(dot)com/browser/index(dot)html, following the instructions
15 provided (Perlegen Sciences, Inc., Menlo Park, CA). As noted in the instructions provided, wild card characters (e.g., "*" symbols) can be added at the beginning of the SNP_ID to identify pertinent information for all alleles of the SNP. This database also links to the NCBI genomic database.

20 In the figures, the first row on each page is a header row with the column names. The columns are as follows:

Header	description of content
SNP_ID	Perlegen internal SNP identifier.
REFSNP_ID	The dbSNP RefSNP cluster ID (from NCBI) when available. Can be null.
SUBSNP_ID	The dbSNP submission ID for SNPs Perlegen submitted to dbSNP. Can be null.
ACCESSION_ID	The accession number from NCBI Build 35 of the contig to which the SNP aligns; may be null.
POSITION	Nucleotide position in NCBI build 35 contig of the reference base in the alignment; may be null.
ALLELE_1	The nucleotide code for allele 1 (Perlegen ref allele).
ALLELE_2	The nucleotide code for allele 2 (Perlegen alt allele).
Ptrend_weighted	The trend score p-value of the association, weighted by degree of family history for each sample
Gene_name	NCBI gene database symbol of a gene near SNP
geneDescription	NCBI gene database description of the gene

HIT_TYPE	The type of hit: exon, intron, up (within 10 kb upstream of transcription start site), down (within 10 kb downstream of the transcription stop). Numbers indicate distances greater than 10 kb, in addition to hit type.
----------	--

[0263] While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and apparatus described above
5 can be used in various combinations. All publications, patents, patent applications, and/or other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication, patent, patent application, and/or other document were individually indicated to be incorporated by reference for all
10 purposes.

WHAT IS CLAIMED IS:

1. A method of identifying a breast cancer phenotype for an organism or biological sample derived therefrom, the method comprising:

5 detecting, in the organism or biological sample, a polymorphism or a locus closely linked thereto, the polymorphism being selected from a polymorphism of Figures 1 or 2, wherein the polymorphism is associated with a breast cancer phenotype; and, correlating the polymorphism or locus to the phenotype.

2. The method of claim 1, wherein the organism is a mammal, or the biological sample is derived from a mammal.

10 3. The method of claim 1, wherein the organism is a human patient, or the biological sample is derived from a human patient.

4. The method of claim 1, wherein the detecting comprises amplifying the polymorphism, linked locus or a sequence associated therewith and detecting the resulting amplicon.

15 5. The method of claim 4, wherein the amplifying comprises:

a) admixing an amplification primer or amplification primer pair with a nucleic acid template isolated from the organism or biological sample, wherein the primer or primer pair is complementary or partially complementary to a region proximal to or including the polymorphism or linked locus, and is capable of initiating nucleic acid polymerization by a polymerase on the nucleic acid template; and,

20

b) extending the primer or primer pair in a DNA polymerization reaction comprising a polymerase and the template nucleic acid to generate the amplicon.

6. The method of claim 4, wherein the amplicon is detected by a process that includes one or more of: hybridizing the amplicon to an array, digesting the amplicon with a restriction enzyme, or real-time PCR analysis.

25

7. The method of claim 4, comprising partially or fully sequencing the amplicon.

8. The method of claim 4, wherein the amplifying comprises performing a polymerase chain reaction (PCR), reverse transcriptase PCR (RT-PCR), or ligase chain

reaction (LCR) using nucleic acid isolated from the organism or biological sample as a template in the PCR, RT-PCR, or LCR.

9. The method of claim 1, wherein the polymorphism or linked locus comprises a single nucleotide polymorphism (SNP).

5 10. The method of claim 1, wherein the polymorphism comprises an allele selected from the group consisting of: alleles in Figures 1 or 2.

11. The method of claim 1, wherein the polymorphism is a single nucleotide polymorphism selected from the group consisting of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, 10 SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418.

15 12. The method of claim 1, wherein the polymorphisim is in a gene or locus selected from the group consisting of: FGFR2, A2BP1, TNRC9, H19, FSTL5, LSP1, LOC388927, UNQ9391, HCN1, LOC441192, TNRC9, NR3C2, KIAA0826, FLJ31033, AACS, FRMD4A and SEC31L2.

20 13. The method of claim 12, wherein the method comprises detecting a plurality of polymorphisms or linked loci in a plurality of said genes.

14. The method of claim 12, wherein the method comprises detecting at least one polymorphism in each of: FGFR2, A2BP1, TNRC9, H19, and FSTL5.

15. The method of claim 12, wherein the method comprises detecting at least one polymorphism for each of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 25 1509710, and SNP ID 843029.

16. The method of claim 12, wherein the method comprises detecting at least one polymorphism for each of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 30 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID

4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418, or at least one polymorphism in each of: FGFR2, A2BP1, TNRC9, H19, FSTL5, LSP1, LOC388927, UNQ9391, HCN1, LOC441192, TNRC9,
5 NR3C2, KIAA0826, FLJ31033, AACCS, FRMD4A and SEC31L2.

17. The method of claim 1, wherein the linked locus is a closely linked locus that is about 5 cM or less from the polymorphism.

18. The method of claim 1, wherein correlating the polymorphism or linked locus to the breast cancer phenotype comprises referencing a look up table that comprises
10 correlation information for alleles of the polymorphism or linked locus and the breast cancer phenotype.

19. A composition comprising a plurality of marker probes or amplification primers that detect or amplify a plurality of polymorphisms selected from: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP
15 ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418.

20. The composition of claim 19, wherein the marker probes or amplification primers are free in solution.

21. A method of identifying a modulator of a breast cancer phenotype, the method comprising:

contacting a potential modulator to a gene or gene product, wherein the gene or gene
25 product comprises or is closely linked to a polymorphism of selected from SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID
30 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418; and,

detecting an effect of the potential modulator on the gene or gene product, thereby identifying whether the potential modulator modulates the phenotype.

22. The method of claim 21, wherein the gene or gene product comprises a polymorphism selected from SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150, SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418.

23. The method of claim 21, wherein the effect is selected from:

(a) increased or decreased expression of the gene or gene product in the presence of the modulator;

(b) increased or decreased activity of the gene product in the presence of the modulator; and,

(c) an altered expression pattern of the gene or gene product in the presence of the modulator.

24. A kit for treatment of a breast cancer phenotype, the kit comprising a modulator identified by the method of claim 21 and instructions for administering the modulator to a patient to treat the phenotype.

25. A system for identifying a breast cancer phenotype for an organism or biological sample derived therefrom, the system comprising:

a) a set of marker probes or primers configured to detect at least one allele of one or more polymorphism or linked locus, wherein the polymorphism is selected from SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, SNP ID 843029, SNP ID 1990126, SNP ID 604819, SNP ID 3025734, SNP ID 1152499, SNP ID 4415909, SNP ID 1732681, SNP ID 4281579, SNP ID 4454457, SNP ID 2616199, SNP ID 1720694, SNP ID 4077723, SNP ID 3711990, SNP ID 3337858, SNP ID 4093095, SNP ID 4213825, SNP ID 3488617, SNP ID 3610210, SNP ID 3451239, SNP ID 1582533, SNP ID 3488150,

SNP ID 2770052, SNP ID 4141351, SNP ID 1335030, SNP ID 2211665, and SNP ID 4538418;

b) a detector that is configured to detect one or more signal outputs from the set of marker probes or primers, or an amplicon produced from the set of marker probes or primers, thereby identifying the presence or absence of the allele; and,

c) system instructions that correlate the presence or absence of the allele with a predicted phenotype.

26. The system of claim 25, wherein the set of marker probes or primers detects a polymorphism in a gene selected from the group consisting of: FGFR2, A2BP1, TNRC9, H19, FSTL5, LSP1, LOC388927, UNQ9391, HCN1, LOC441192, TNRC9, NR3C2, KIAA0826, FLJ31033, AACS, FRMD4A and SEC31L2.

27. The system of claim 25, wherein the set of marker probes or primers detects a plurality of polymorphisms in a plurality of said genes or genetic loci.

28. The system of claim 25, wherein the set of marker probes or primers detects at least one polymorphism in each of: FGFR2, A2BP1, TNRC9, H19, and FSTL5.

29. The system of claim 25, wherein the set of marker probes or primers detects at least one polymorphism selected from a polymorphism of: SNP ID 2312116, SNP ID 1622530, SNP ID 3712013, SNP ID 1509710, and SNP ID 843029.

30. The system of claim 25, wherein the detector detects one or more light emission, wherein the light emission is indicative of the presence or absence of the allele.

31. The system of claim 25, wherein the instructions comprise at least one look-up table that includes a correlation between the presence or absence of the allele and the phenotype.

32. The system of claim 25, wherein the system comprises a sample.

33. The system of claim 32, wherein the sample comprises genomic DNA, amplified genomic DNA, cDNA, amplified cDNA, RNA, or amplified RNA.

34. The system of claim 32, wherein the sample is derived from a mammal.