

GENETIC VARIANTS FOR BREAST CANCER RISK ASSESSMENT

INTRODUCTION

- 5 Breast cancer is by far the most common cancer in women worldwide. Current global incidence is in excess of 1,151,000 new cases diagnosed each year [Parkin, et al., (2005), *CA Cancer J Clin*, **55**, 74-108]. Breast cancer incidence is highest in developed countries, particularly amongst populations of Northern European ethnic origin, and is increasing. In the United States the annual age-standardized incidence rate is approximately 125 cases per 100,000 populations,
- 10 more than three times the world average. Rates in Northern European countries are similarly high. In the year 2008 it is estimated that 184,450 new cases of invasive breast cancer will be diagnosed in the U.S.A. and 40,930 people will die from the disease [Jemal, et al., (2008), *CA Cancer J Clin*, **58**, 71-96]. To this figure must be added a further 67,770 ductal and lobular carcinoma in-situ diagnoses expected in 2008. From an individual perspective, the lifetime
- 15 probability of developing breast cancer is 12.3% in U.S. women (i.e., 1 in 8 women will develop breast cancer during their lives). As with most cancers, early detection and appropriate treatment are important factors. Overall, the 5-year survival rate for breast cancer is 89%. However, in individuals presenting with regionally invasive or metastatic disease, the rate declines to 84% and 27%, respectively [Jemal, et al., (2008), *CA Cancer J Clin*, **58**, 71-96].
- 20 Increasingly, emphasis is falling on the identification individuals who are at high risk for primary or recurrent breast cancer. Such individuals can be managed by more intensive screening, preventative chemotherapies, hormonal therapies and, in cases of individuals at extremely high risk, prophylactic surgery. Mass screening programs constitute a huge economic burden on health services, while preventative therapies have associated risks and quality of life
- 25 consequences.

Genetic Predisposition to Breast Cancer

- The two primary classes of known risk factors for breast cancer are endocrine factors and genetics. Regarding the latter, approximately 12% of breast cancer patients have one or more
- 30 first degree relatives with breast cancer [(2001), *Lancet*, **358**, 1389-99]. The well known, dominant breast cancer predisposition genes BRCA1 and BRCA2 confer greatly increased breast cancer risk to carriers, with lifetime penetrance estimates ranging from 40-80%. The presence of BRCA1 and BRCA2 mutations can account for the majority of families with 6 or more cases of breast cancer and for a large proportion of families comprising breast and ovarian or male breast

cancer. However such families are very rare indeed. BRCA1 and BRCA2 mutations are found much less frequently in families with fewer cases or in families characterized by breast cancer cases only. Together, mutations in BRCA1 and BRCA2 can account for 15-20% of the risk for familial breast cancer. In non-founder populations, if all common BRCA mutations could be
 5 detected, between 2-3% of incident breast cancer patients would be expected to harbor a mutation [Gorski, et al., (2005), *Breast Cancer Res Treat*, **92**, 19-24; (2000), *Br J Cancer*, **83**, 1301-8]. This low "chance to find" statistic precludes the responsible use of BRCA mutation testing outside families with an obvious hereditary predisposition (Anon[(2003), *J Clin Oncol*, **21**, 2397-406]). Rare, high penetrance mutations are known to occur in the TP53 and PTEN
 10 genes, however, these together account for no more than 5% of the total genetic risk for breast cancer [Easton, (1999), *Breast Cancer Res*, **1**, 14-7]. Linkage studies have been largely unsuccessful in identifying any more, widespread mutations conferring high risk for breast cancer [Smith, et al., (2006), *Genes Chromosomes Cancer*, **45**, 646-55].

Recent epidemiological studies have indicated that the majority of breast cancer cases arise in a
 15 predisposed, susceptible minority of the population [Antoniou, et al., (2002), *Br J Cancer*, **86**, 76-83; Pharoah, et al., (2002), *Nat Genet*, **31**, 33-6]. Data from twin studies and observations of the constant, high incidence of cancer in the contralateral breast of patients surviving primary breast cancer indicate that a substantial portion of the uncharacterized risk for breast cancer is related to endogenous factors, most probably genetic [Lichtenstein, et al., (2000), *N Engl J Med*,
 20 **343**, 78-85; Peto and Mack, (2000), *Nat Genet*, **26**, 411-4]. Knowledge of the genetic factors that underpin this widespread risk is very limited. Segregation analyses predict that the uncharacterized genetic risk for breast cancer is most likely to be polygenic in nature, with risk alleles that confer low to moderate risk and which may interact with each other and with
 25 hormonal risk factors. Nevertheless, these studies predict as much as 40-fold differences in relative risk between the highest and lowest quintiles of a distribution that could be defined by genetic profiling that captures these low to moderate risk alleles [Antoniou, et al., (2002), *Br J Cancer*, **86**, 76-83; Pharoah, et al., (2002), *Nat Genet*, **31**, 33-6]. 88% of all breast cancer cases are expected to arise amongst a predisposed 50% of the population and the 12% of the population at highest risk accounts for 50% of all breast cancer cases [Pharoah, et al., (2002),
 30 *Nat Genet*, **31**, 33-6; Pharoah, (2003), *Recent Results Cancer Res*, **163**, 7-18; discussion 264-6]. Much focus is therefore directed towards the identification of such genetically predisposed individuals and developing personalized medical management strategies for them.

We and others have shown that there is a significant familial risk of breast cancer in Iceland which extends to at least 5th degree relatives [Amundadottir, et al., (2004), *PLoS Med*, **1**, e65;
 35 Tullinius, et al., (2002), *J Med Genet*, **39**, 457-62]. The contribution of BRCA1 mutations to familial risk in Iceland is thought to be minimal [Arason, et al., (1998), *J Med Genet*, **35**, 446-9; Bergthorsson, et al., (1998), *Hum Mutat*, **Suppl 1**, S195-7]. A single founder mutation in the BRCA2 gene (999del5) is present at a carrier frequency of 0.6-0.8% in the general Icelandic population and 7.7-8.6% in female breast cancer patients [Thorlacius, et al., (1997), *Am J Hum*

Genet, **60**, 1079-84; Gudmundsson, et al., (1996), Am J Hum Genet, **58**, 749-56]. This single mutation is estimated to account for approximately 40% of the inherited breast cancer risk to first through third degree relatives [Tulinius, et al., (2002), J Med Genet, **39**, 457-62]. Although this estimate is higher than the 15-25% of familial risk attributed to all BRCA 1 and 2 mutations combined in non-founder populations, there is still some 60% of Icelandic familial breast cancer risk to be explained. First degree relatives of patients who test negative for BRCA2 999del5 remain at a 1.72 fold the population risk for breast cancer (95% CI 1.49-1.96) [Tulinius, et al., (2002), J Med Genet, **39**, 457-62].

Genetic risk is conferred by subtle differences in the genome among individuals in a population. Variations in the human genome are most frequently due to single nucleotide polymorphisms (SNPs), although other variations are also important. SNPs are located on average every 1000 base pairs in the human genome. Accordingly, a typical human gene containing 250,000 base pairs may contain 250 different SNPs. Only a minor number of SNPs are located in exons and alter the amino acid sequence of the protein encoded by the gene. Most SNPs may have little or no effect on gene function, while others may alter transcription, splicing, translation, or stability of the mRNA encoded by the gene. Additional genetic polymorphisms in the human genome are caused by insertions, deletions, translocations or inversion of either short or long stretches of DNA. Genetic polymorphisms conferring disease risk may directly alter the amino acid sequence of proteins, may increase the amount of protein produced from the gene, or may decrease the amount of protein produced by the gene.

As genetic polymorphisms conferring risk of common diseases are uncovered, genetic testing for such risk factors is becoming increasingly important for clinical medicine. Examples are apolipoprotein E testing to identify genetic carriers of the apoE4 polymorphism in dementia patients for the differential diagnosis of Alzheimer's disease, and of Factor V Leiden testing for predisposition to deep venous thrombosis. More importantly, in the treatment of cancer, diagnosis of genetic variants in tumor cells is used for the selection of the most appropriate treatment regime for the individual patient. In breast cancer, genetic variation in estrogen receptor expression or heregulin type 2 (Her2) receptor tyrosine kinase expression determine if anti-estrogenic drugs (tamoxifen) or anti-Her2 antibody (Herceptin) will be incorporated into the treatment plan. In chronic myeloid leukemia (CML) diagnosis of the Philadelphia chromosome genetic translocation fusing the genes encoding the Bcr and Abl receptor tyrosine kinases indicates that Gleevec (ST1571), a specific inhibitor of the Bcr-Abl kinase should be used for treatment of the cancer. For CML patients with such a genetic alteration, inhibition of the Bcr-Abl kinase leads to rapid elimination of the tumor cells and remission from leukemia. Furthermore, genetic testing services are now available, providing individuals with information about their disease risk based on the discovery that certain SNPs have been associated with risk of many of the common diseases.

Understanding of the genetic factors contributing to the residual genetic risk for breast cancer is limited. Variants in two genes have been rigorously confirmed as low penetrance breast cancer risk genes; CHEK2 and ATM [Renwick, et al., (2006), *Nat Genet*, **38**, 873-5; (2004), *Am J Hum Genet*, **74**, 1175-82]. Furthermore, a recent report establishes a link between variants on

5 chromosomes 2q35 and 16q12 and increased risk of estrogen receptor positive breast cancer (Simon, SN. et al. *Nat Genet* 39:865-9 (2007)). Further, variants in or near the FGFR2, TNRC9, MAP3K1 and LSP1 genes (Easton, D.F., et al. *Nature* 447:1087-93 (2007)), as well as the FGFR2 gene (Hunter, D.J., et al *Nat Genet* 39:870-4 (2007)) have been reported. Many other genes have been implicated however their contribution to breast cancer risk has not been confirmed in

10 analyses employing very large sample sets [Breast Cancer Association, (2006), *J Natl Cancer Inst*, **98**, 1382-96].

No universally successful method for the prevention or treatment of breast cancer is currently available. Management of breast cancer currently relies on a combination of primary prevention, early diagnosis, appropriate treatments and secondary prevention. There are clear clinical

15 imperatives for integrating genetic testing into all aspects of these management areas. Identification of cancer susceptibility genes may also reveal key molecular pathways that may be manipulated (e.g., using small or large molecular weight drugs) and may lead to more effective treatments. The present invention provides additional genetic variants for breast cancer than can be integrated in prevention programmes for breast cancer.

20

SUMMARY OF THE INVENTION

The present invention relates to methods of assessing a susceptibility to breast cancer. The invention includes methods of diagnosing an increased susceptibility to breast cancer, as well as methods of diagnosing a decreased susceptibility to breast cancer or diagnosing a protection

25 against cancer, by evaluating certain markers or haplotypes that have been found to be associated with increased or decreased susceptibility of breast cancer. The invention also relates to methods of assessing prognosis of individuals diagnosed with breast cancer, methods of assessing the probability of response to a breast cancer therapeutic agent or breast cancer therapy, as well as methods of monitoring progress of treatment of an individual diagnosed with

30 breast cancer.

In one aspect, the present invention relates to a method of diagnosing a susceptibility to breast cancer in a human individual, the method comprising determining the presence or absence of at least one allele of at least one polymorphic marker selected from the group consisting of

35 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs9956546 (SEQ ID

NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, in a nucleic acid sample obtained from the individual, wherein the presence of the at least one allele is indicative of a susceptibility to breast cancer. The invention also relates to a method of determining a susceptibility to breast cancer, by determining the presence or absence of at least one allele of at least one polymorphic selected from the group consisting of rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, wherein the determination of the presence of the at least one allele is indicative of a susceptibility to breast cancer.

In another aspect, the invention relates to a method of determining a susceptibility to breast cancer in a human individual, comprising determining whether at least one at-risk allele in at least one polymorphic marker is present in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from markers selected from the group consisting of rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, and wherein determination of the presence of the at least one at-risk allele is indicative of increased susceptibility to breast cancer in the individual.

The invention furthermore relates to a method for determining a susceptibility to breast cancer in a human individual, comprising determining whether at least one allele of at least one polymorphic marker is present in a nucleic acid sample obtained from the individual or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, and wherein the presence of the at least one allele is indicative of a susceptibility to breast cancer for the individual.

The genotype dataset comprises in one embodiment information about marker identity, and the allelic status of the individual, i.e. information about the identity of the two alleles carried by the individual for the marker. The genotype dataset may comprise allelic information about one or more marker, including two or more markers, three or more markers, five or more markers, one hundred or more markers, etc. In some embodiments, the genotype dataset comprises genotype information from a whole-genome assessment of the individual that may include hundreds of thousands of markers, or even one million or more markers.

In certain embodiments, the at least one polymorphic marker is associated with the PAX5 (PAIRED BOX GENE 5, also known as BSAP) gene, the TUB (Tubby homolog (mouse, rd5))

gene, the SERPINH1 (serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)) gene, the RAD51L1 (REC2; R51H2; hREC2; RAD51B; MGC34245) gene, the FHOD3 (formin homology 2 domain containing 3, also known as FHOS2; Formactin2) gene or the TNRC6B (trinucleotide repeat containing 6B or KIAA1093) gene.

5

In certain such embodiments, the at least one polymorphic marker is in linkage disequilibrium with any one of the above-mentioned genes, i.e. any one of the PAX5 (PAIRED BOX GENE 5, also known as BSAP) gene, the TUB (Tubby homolog (mouse, rd5)) gene, the SERPINH1 (serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)) gene, the RAD51L1 (REC2; R51H2; hREC2; RAD51B; MGC34245) gene, the FHOD3 (formin homology 2 domain containing 3, also known as FHOS2; Formactin2) gene and the TNRC6B (trinucleotide repeat containing 6B or KIAA1093) gene.

Another aspect of the invention relates to a method of determining a susceptibility to breast cancer in a human individual, the method comprising:

obtaining nucleic acid sequence data about a human individual identifying at least one allele of at least one polymorphic marker selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9), and markers in linkage disequilibrium therewith, wherein different alleles of the at least one polymorphic marker are associated with different susceptibilities to breast cancer in humans, and

determining a susceptibility to breast cancer from the nucleic acid sequence data.

In a general sense, genetic markers lead to alternate sequences at the nucleic acid level. If the nucleic acid marker changes the codon of a polypeptide encoded by the nucleic acid, then the marker will also result in alternate sequence at the amino acid level of the encoded polypeptide (polypeptide markers). Determination of the identity of particular alleles at polymorphic markers in a nucleic acid or particular alleles at polypeptide markers comprises whether particular alleles are present at a certain position in the sequence. Sequence data identifying a particular allele at a marker comprises sufficient sequence to detect the particular allele. For single nucleotide polymorphisms (SNPs) or amino acid polymorphisms described herein, sequence data can comprise sequence at a single position, i.e. the identity of a nucleotide or amino acid at a single position within a sequence. Nucleic acid sequence data identifying particular alleles of polymorphic markers is sometimes also referred to as genotype data. Nucleic acid sequence data can be obtained for example by analyzing sequence of the at least one polymorphic marker in a biological sample from the individual. Alternatively, nucleic acid sequence data can be obtained in a genotype dataset from the human individual and analyzing sequence of the at least one polymorphic marker in the dataset. Such analysis in certain embodiments comprises

determining the presence or absence of a particular allele of specific polymorphic markers. Identification of particular alleles in general terms should be taken to mean that determination of the presence or absence of the allele(s) is made. Usually, determination of both allelic copies in the genome of an individual is performed, by determining the occurrence of all possible alleles of the particular polymorphism in a particular individual (for SNPs, each of the two possible nucleotides possible for the allelic site). It is also possible to determine whether only particular alleles are present or not. For example, in certain embodiments, determination of the presence or absence of certain alleles that have been shown to associate with risk of glaucoma is made, but not necessarily other alleles of the particular marker, and a determination of susceptibility is made based on such determination.

In certain embodiments, it may be useful to determine the nucleic acid sequence for at least two polymorphic markers. In other embodiments, the nucleic acid sequence for at least three, at least four or at least five or more polymorphic markers is determined. Haplotype information can be derived from an analysis of two or more polymorphic markers. Thus, in certain embodiments, a further step is performed, whereby haplotype information is derived based on sequence data for at least two polymorphic markers.

The invention also provides a method of determining a susceptibility to breast cancer in a human individual, the method comprising obtaining nucleic acid sequence data about a human individual identifying both alleles of at least two polymorphic markers selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, determine the identity of at least one haplotype based on the sequence data, and determine a susceptibility to breast cancer from the haplotype data.

In certain embodiments, determination of a susceptibility comprises comparing the nucleic acid sequence data to a database containing correlation data between the at least one polymorphic marker and susceptibility to breast cancer. In some embodiments, the database comprises at least one risk measure of susceptibility to breast cancer for the at least one marker. The sequence database can for example be provided as a look-up table that contains data that indicates the susceptibility of breast cancer for any one, or a plurality of, particular polymorphisms. The database may also contain data that indicates the susceptibility for a particular haplotype that comprises at least two polymorphic markers.

Obtaining nucleic acid sequence data can in certain embodiments comprise obtaining a biological sample from the human individual and analyzing sequence of the at least one polymorphic marker in nucleic acid in the sample. Analyzing sequence can comprise determining the presence or absence of at least one allele of the at least one polymorphic marker. Determination of the presence of a particular susceptibility allele (e.g., an at-risk allele) is indicative of

susceptibility to breast cancer in the human individual. Determination of the absence of a particular susceptibility allele is indicative that the particular susceptibility is not present in the individual.

In some embodiments, obtaining nucleic acid sequence data comprises obtaining nucleic acid sequence information from a preexisting record. The preexisting record can for example be a computer file or database containing sequence data, such as genotype data, for the human individual, for at least one polymorphic marker.

Susceptibility determined by the diagnostic methods of the invention can be reported to a particular entity. In some embodiments, the at least one entity is selected from the group consisting of the individual, a guardian of the individual, a genetic service provider, a physician, a medical organization, and a medical insurer.

In certain embodiments of the invention, determination of a susceptibility comprises comparing the nucleic acid sequence data to a database containing correlation data between the at least one polymorphic marker and susceptibility to breast cancer. In one such embodiment, the database comprises at least one risk measure of susceptibility to breast cancer for the at least one polymorphic marker. In another embodiment, the database comprises a look-up table containing at least one risk measure of the at least one condition for the at least one polymorphic marker.

In certain embodiments, obtaining nucleic acid sequence data comprises obtaining a biological sample from the human individual and analyzing sequence of the at least one polymorphic marker in nucleic acid in the sample. Analyzing sequence of the at least one polymorphic marker can comprise determining the presence or absence of at least one allele of the at least one polymorphic marker. Obtaining nucleic acid sequence data can also comprise obtaining nucleic acid sequence information from a preexisting record.

Certain embodiments of the invention relate to obtaining nucleic acid sequence data about at least two polymorphic markers selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith.

In certain embodiments of the invention, the at least one polymorphic marker is selected from the markers set forth in Table 4. In one embodiment, the at least one polymorphic marker is selected from the markers as set forth in SEQ ID NO:1-562. In one embodiment, the at least one marker is in linkage disequilibrium with at least one of rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954. In one embodiment, markers in linkage disequilibrium with rs999737 are selected from the group consisting of

rs999737, rs10134446, rs10138140, rs10146772, rs10467820, rs10483812, rs10483813, rs11158749, rs11158751, rs11621276, rs11624097, rs11624164, rs11624333, rs11628293, rs11846360, rs11847185, rs11849916, rs12878761, rs12879200, rs12886864, rs12889251, rs12894230, rs1468279, rs1468280, rs1547012, rs17105675, rs17755657, rs17755734, 5 rs17755752, rs17755925, rs17756000, rs17828691, rs17828721, rs17828763, rs17828907, rs17828955, rs1956534, rs2074563, rs2074565, rs2097800, rs2107340, rs2145157, rs2158357, rs2189517, rs2253317, rs2257111, rs2257116, rs2257127, rs2331701, rs2331705, rs2331775, rs2525503, rs2525530, rs2842327, rs3784121, rs4531674, rs4899246, rs4902604, rs4902606, rs4902608, rs5004090, rs6573837, rs7140266, rs7146456, rs7153476, rs739874, rs746663, 10 rs8007194, rs8010439, rs8012610, rs9323512, rs9323513, and rs9323514.

In one embodiment, markers in linkage disequilibrium with rs2005154 are selected from the group consisting of rs2005154, rs4878662, and rs4880019.

In one embodiment, markers in linkage disequilibrium with rs2184380 are selected from the group consisting of rs2184380, rs10466295, rs10508363, rs10508364, rs10508365, 15 rs10795670, rs10905411, rs10905414, rs10905415, rs10905430, rs10905437, rs10905439, rs10905440, rs10905443, rs10905444, rs10905445, rs10905446, rs10905447, rs10905454, rs11255764, rs11255776, rs11255777, rs11255778, rs11255779, rs11255790, rs11255795, rs11255797, rs11255800, rs11255804, rs11255805, rs11255820, rs11255821, rs11255822, rs11255832, rs11255836, rs11255840, rs11255858, rs11255862, rs11255869, rs11255870, 20 rs11255871, rs11255882, rs11255884, rs12049705, rs12218610, rs12250379, rs12259226, rs1325874, rs1334549, rs1334550, rs1334559, rs1360749, rs1413678, rs1413683, rs1537601, rs1537602, rs1537603, rs17407711, rs17407781, rs17407830, rs17408204, rs17408337, rs17408580, rs17484150, rs17485426, rs17485998, rs17486082, rs17486795, rs17486816, rs1970170, rs1999638, rs2031561, rs2182292, rs2388821, rs2388825, rs2388826, rs2892613, 25 rs4112287, rs4112288, rs4345867, rs4454616, rs4747806, rs4749805, rs4749807, rs4749812, rs6602328, rs6602329, rs7069110, rs7080765, rs7083359, rs7477023, rs7904921, rs7912413, rs7912704, rs7912831, rs827389, and rs966562.

In one embodiment, markers in linkage disequilibrium with rs2224696 are selected from the group consisting of rs2224696, rs10905509, rs11256045, rs12761213, rs12761461, 30 rs12766048, rs12772042, rs12776383, rs12778120, rs12780218, rs12781427, rs1475189, rs1573109, rs1573110, rs17145088, rs17145095, rs17145118, rs17145120, rs17145151, rs17145164, rs17145169, rs17145188, rs17145193, rs17145221, rs17363338, rs1775559, rs1857230, rs1891532, rs1935813, rs2013364, rs2025289, rs2057442, rs2093625, rs2093626, rs2146598, rs2185817, rs2397336, rs2760204, rs2797266, rs391733, rs4550140, rs7081544, 35 rs852273, rs860418, rs861172, rs962993, and rs965307.

In one embodiment, markers in linkage disequilibrium with rs2242503 are selected from the group consisting of rs2242503, rs10431029, rs1055233, rs10734629, rs10743052, rs10743053,

rs10743054, rs10743055, rs10769872, rs10769873, rs10769878, rs10769882, rs10839976, rs10839984, rs11041740, rs11041742, rs11041788, rs11041791, rs11041794, rs1108277, rs12146654, rs12808387, rs1528125, rs1569128, rs1970880, rs1997262, rs2049684, rs2141321, rs2242501, rs2272383, rs3750955, rs3752898, rs3849986, rs3849990, rs3911309, 5 rs3911310, rs4340037, rs4343012, rs4385931, rs4575312, rs4578424, rs4636658, rs4758040, rs4758042, rs4758287, rs4758309, rs4758310, rs7103334, rs7112519, rs7115706, andrs7122690, rs7127738, rs7358396, rs7479156, rs7479738, rs7480804, rs7481667, rs7481683, rs7482611, rs7927368, and rs7940668.

10 In one embodiment, markers in linkage disequilibrium with rs12291026 are selected from the group consisting of rs12291026, rs1004856, rs10899091, rs11236449, rs11236452, rs11236454, rs12362081, rs1540210, rs1540211, rs1557471, rs1631470, rs1783551, rs1783556, rs1783559, rs1790144, rs1790152, rs1790307, rs1793396, rs1793397, rs1793398, rs1793399, rs1793414, rs1938800, rs2853066, rs499613, rs504793, rs514477, rs549034, rs550881, rs581007, rs589724, rs600387, rs606460, rs617617, rs618202, rs628972, rs640649, 15 rs662279, rs667410, rs667531, rs670100, rs670491, rs682292, rs7128888, rs7129014, rs7129150, andrs947844.

In one embodiment, markers in linkage disequilibrium with rs11912922 are selected from the group consisting of rs11912922, rs11089967, rs11704971, rs11705454, rs17406386, rs17406434, rs2071771, rs2958650, rs2958651, rs2958659, rs7284488, rs7285507, rs7291782, 20 rs739145, rs9611246, and rs9611265.

In one embodiment, markers in linkage disequilibrium with rs6001954 are selected from the group consisting of rs6001954, rs10483203, rs10483204, rs10483205, rs10483206, rs1106673, rs11913132, rs12158399, rs12158872, rs12159200, rs12159970, rs12484697, rs12627881, rs133036, rs133038, rs16985899, rs17001846, rs17001868, rs17001943, rs17001974, 25 rs17001977, rs17001993, rs17001994, rs17001997, rs17002019, rs17002020, rs17002026, rs17002027, rs17002034, rs17002036, rs17002038, rs17002069, rs2075764, rs2187832, rs2235318, rs2280790, rs2294348, rs2294350, rs2294352, rs2413624, rs3788577, rs3788578, rs3788579, rs3827381, rs3827382, rs4140512, rs470113, rs5750957, rs5750960, rs5750966, rs5757976, rs5757998, rs5758001, rs5995849, rs5995856, rs5995870, rs5995871, rs5995886, 30 rs6001900, rs6001910, rs6001911, rs6001912, rs6001913, rs6001930, rs6001931, rs6001932, rs6001935, rs6001950, rs6001974, rs6001980, rs6001990, rs6002000, rs718193, rs7292804, rs7293100, rs742140, rs760700, rs760701, rs9306345, rs932379, rs9607721, rs9611310, rs9611311, rs9611312, rs9611316, rs9611318, rs9611324, rs9611325, rs9611328, and rs9611329.

35 In certain embodiments of the invention, a further step of assessing the frequency of at least one haplotype in the individual is performed. In such embodiments, two or more markers, including three, four, five, six, seven, eight, nine or ten or more markers can be included in the

haplotype. In certain embodiments, the at least one haplotype comprises markers that are all in LD with at least one of rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9). In such
 5 embodiments, the at least one haplotype is representative of the genomic structure of a particular genomic region (LD block) to which any one of the above-mentioned markers reside. In one embodiment, the haplotype comprises markers in linkage disequilibrium with rs999737.

The markers conferring risk of breast cancer, as described herein, can be combined with other genetic markers for breast cancer. Such markers are typically not in linkage disequilibrium with
 10 any one of the markers described herein, in particular markers rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith. Any of the methods described herein can be practiced by combining the genetic risk factors described
 15 herein with additional genetic risk factors for breast cancer.

Thus, in certain embodiments, a further step is included, comprising determining whether at least one at-risk allele of at least one at-risk variant for breast cancer not in linkage disequilibrium with any one of the markers set forth in Table 1 and/or Table 4 is present in an individual. In other words, genetic markers in other locations in the genome can be useful in
 20 combination with the markers of the present invention, so as to determine overall risk of breast cancer based on multiple genetic factors. Selection of markers that are not in linkage disequilibrium (not in LD) can be based on a suitable measure for linkage disequilibrium, as described further herein. In certain embodiments, markers that are not in linkage disequilibrium have values for the LD measure r^2 between the markers of less than 0.2. In certain other
 25 embodiments, markers that are not in LD have values for r^2 between the markers of less than 0.15, including less than 0.10, less than 0.05, less than 0.02 and less than 0.01. Other suitable cutoff values for establishing that markers are not in LD are contemplated, including values bridging any of these values. Examples of such genetic risk factors include markers on chromosome 5p12 and chromosome 10q26, e.g., marker rs10941679 and marker rs1219648.
 30 Alternatively, markers in LD with any one of these markers could be assessed. Other markers known to confer risk of breast cancer can also be assessed together with the markers described herein, include markers on chromosome 2q14 (e.g., marker rs4848543 or markers in linkage disequilibrium therewith), 2q35 (e.g., marker rs13387042, or markers in linkage disequilibrium therewith), and chromosome 16 (e.g., marker rs3803662, or markers in linkage disequilibrium
 35 therewith).

In certain embodiments, multiple markers as described herein are determined to determine overall risk of breast cancer. Thus, in certain embodiments, an additional step is included, the step comprising determining whether at least one allele in each of at least two polymorphic

markers is present in a sample comprising genomic DNA from a human individual or a genotype dataset derived from a human individual, wherein the presence of the at least one allele in the at least two polymorphic markers is indicative of an increased susceptibility to breast cancer. In one embodiment, the markers are selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5),
 5 rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith.

Risk assessment based on the markers of the present invention can also be combined with assessment for the presence or absence of at least one high penetrant genetic factor for breast cancer in a nucleic acid sample obtained from the individual or in a genotype dataset derived from the individual. The high penetrant genetic factor for breast cancer can for example be a BRCA1 mutation, a BRCA2 mutation, a TP53 mutation or a PTEN mutation. Together, mutations in BRCA1 and BRCA2 can account for 15-20% of the risk for familial breast cancer, and these can account for between 2-3% of incident breast cancer patients [Gorski, et al., (2005), Breast
 10 Cancer Res Treat, **92**, 19-24; (2000), Br J Cancer, **83**, 1301-8]. Known mutations in the TP53 and PTEN genes account for about 5% of the total genetic risk for breast cancer [Easton, (1999), Breast Cancer Res, **1**, 14-7]. In one embodiment, the high penetrant genetic factor is BRCA2 999del5.

The genetic markers of the invention can also be combined with non-genetic information to establish overall risk for an individual. Thus, in certain embodiments, a further step is included, comprising analyzing non-genetic information to make risk assessment, diagnosis, or prognosis of the individual. The non-genetic information can be any information pertaining to the disease status of the individual or other information that can influence the estimate of overall risk of breast cancer for the individual. In one embodiment, the non-genetic information is selected
 20 from age, gender, ethnicity, socioeconomic status, previous disease diagnosis, medical history of subject, family history of breast cancer, biochemical measurements, and clinical measurements.

In another aspect, the invention relates to a method of assessing risk of developing at least a second primary tumor in an individual previously diagnosed with breast cancer, the method comprising determining the presence or absence of at least one allele of at least one polymorphic marker in a nucleic acid sample obtained from the individual, wherein the at least one
 30 polymorphic marker is selected from the group consisting of the polymorphic markers listed in Tables 1 and 4, and markers in linkage disequilibrium therewith, wherein the presence of the at least one allele is indicative of risk of developing at least a second primary tumor. Alternatively, the invention relates to a method of determining risk of developing at least a second primary tumor in an individual previously diagnosed with breast cancer, the method comprising
 35 determining whether at least one allele of at least one polymorphic marker is present in a nucleic acid sample obtained from the individual, or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs2005154 (SEQ ID NO:1),

rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, and wherein the presence of the at least one allele is indicative of risk of developing at least a second primary tumor. In one such embodiment, the at least one polymorphic marker is selected from the markers set forth in Table 4.

The invention also provides computer-implemented aspects. In one such aspect, the invention provides a computer-readable medium having computer executable instructions for determining susceptibility to breast cancer in an individual, the computer readable medium comprising:
 10 data representing at least one polymorphic marker; and
 a routine stored on the computer readable medium and adapted to be executed by a processor to determine susceptibility to breast cancer in an individual based on the allelic status of at least one allele of said at least one polymorphic marker in the individual.

In one embodiment, said data representing at least one polymorphic marker comprises at least one parameter indicative of the susceptibility to breast cancer linked to said at least one polymorphic marker. In another embodiment, said data represents at least one polymorphic marker comprises data indicative of the allelic status of at least one allele of said at least one allelic marker in said individual. In another embodiment, said routine is adapted to receive input data indicative of the allelic status for at least one allele of said at least one allelic marker in said individual. In a preferred embodiment, the at least one marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith. In another preferred embodiment, the at least one polymorphic marker is selected from the markers set forth in Table 4.

The invention further provides an apparatus for determining a genetic indicator for breast cancer in a human individual, comprising:

a processor,
 a computer readable memory having computer executable instructions adapted to be executed on the processor to analyze marker and/or haplotype information for at least one human individual with respect to breast cancer, and
 generate an output based on the marker or haplotype information, wherein the output comprises a risk measure of the at least one marker or haplotype as a genetic indicator of breast cancer for the human individual. In one embodiment, the computer readable memory comprises data
 35 indicative of the frequency of at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of individuals diagnosed with breast cancer, and data indicative of

the frequency of at the least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference individuals, and wherein a risk measure is based on a comparison of the at least one marker and/or haplotype status for the human individual to the data indicative of the frequency of the at least one marker and/or haplotype information for the plurality of individuals diagnosed with breast cancer. In an alternative embodiment, the computer readable memory further comprises data indicative of the risk of developing breast cancer associated with at least one allele of at least one polymorphic marker or at least one haplotype, and wherein a risk measure for the human individual is based on a comparison of the genotype status for the human individual to the risk associated with the at least one allele of the at least one polymorphic marker or the at least one haplotype. In another embodiment, the computer readable memory further comprises data indicative of the frequency of at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of individuals diagnosed with breast cancer, and data indicative of the frequency of at the least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference individuals, and wherein risk of developing breast cancer is based on a comparison of the frequency of the at least one allele or haplotype in individuals diagnosed with breast cancer, and reference individuals. In a preferred embodiment, the at least one marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith. In another preferred embodiment, the at least one polymorphic marker is selected from the markers set forth in Tables 4.

In another aspect, the invention relates to a method of identification of a marker for use in assessing susceptibility to breast cancer, the method comprising: identifying at least one polymorphic marker in linkage disequilibrium with at least one of rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9); determining the genotype status of a sample of individuals diagnosed with, or having a susceptibility to, breast cancer; and determining the genotype status of a sample of control individuals; wherein a significant difference in frequency of at least one allele in at least one polymorphism in individuals diagnosed with, or having a susceptibility to, breast cancer, as compared with the frequency of the at least one allele in the control sample is indicative of the at least one polymorphism being useful for assessing susceptibility to breast cancer. Significant difference can be estimated on statistical analysis of allelic counts at certain polymorphic markers in breast cancer patients and controls. In one embodiment, a significant difference is based on a calculated *P*-value between breast cancer patients and controls of less than 0.05. In other embodiments, a significant difference is based on a lower value of the calculated *P*-value, such as less than 0.005, 0.0005, or 0.00005. In one embodiment, an increase in frequency of the at least one allele in the at least one polymorphism in individuals diagnosed with, or having a susceptibility to, breast cancer, as compared with the

frequency of the at least one allele in the control sample is indicative of the at least one polymorphism being useful for assessing increased susceptibility to breast cancer. In another embodiment, a decrease in frequency of the at least one allele in the at least one polymorphism in individuals diagnosed with, or having a susceptibility to, breast cancer, as compared with the frequency of the at least one allele in the control sample is indicative of the at least one polymorphism being useful for assessing decreased susceptibility to, or protection against, breast cancer.

The invention also relates to a method of genotyping a nucleic acid sample obtained from a human individual comprising determining whether at least one allele of at least one polymorphic marker is present in a nucleic acid sample from the individual sample, wherein the at least one marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, and wherein determination of the presence of the at least one allele in the sample is indicative of a susceptibility to breast cancer in the individual. In one embodiment, determination of the presence of rs2005154 allele T, rs2184380 allele G, rs2224696 allele T, rs2242503 allele C, rs12291026 allele G, rs999737 allele C, rs9956546 allele A, rs11912922 allele T and rs6001954 allele G is indicative of increased susceptibility of breast cancer in the individual. In one embodiment, genotyping comprises amplifying a segment of a nucleic acid that comprises the at least one polymorphic marker by Polymerase Chain Reaction (PCR), using a nucleotide primer pair flanking the at least one polymorphic marker. In another embodiment, genotyping is performed using a process selected from allele-specific probe hybridization, allele-specific primer extension, allele-specific amplification, nucleic acid sequencing, 5'-exonuclease digestion, molecular beacon assay, oligonucleotide ligation assay, size analysis, single-stranded conformation analysis and microarray technology. In one embodiment, the microarray technology is Molecular Inversion Probe array technology or BeadArray Technologies. In one embodiment, the process comprises allele-specific probe hybridization. In another embodiment, the process comprises microarray technology. One preferred embodiment comprises the steps of (1) contacting copies of the nucleic acid with a detection oligonucleotide probe and an enhancer oligonucleotide probe under conditions for specific hybridization of the oligonucleotide probe with the nucleic acid; wherein (a) the detection oligonucleotide probe is from 5-100 nucleotides in length and specifically hybridizes to a first segment of a nucleic acid whose nucleotide sequence is given by any one of SEQ ID NO:1-562; (b) the detection oligonucleotide probe comprises a detectable label at its 3' terminus and a quenching moiety at its 5' terminus; (c) the enhancer oligonucleotide is from 5-100 nucleotides in length and is complementary to a second segment of the nucleotide sequence that is 5' relative to the oligonucleotide probe, such that the enhancer oligonucleotide is located 3' relative to the detection oligonucleotide probe when both oligonucleotides are hybridized to the nucleic acid; and (d) a single base gap exists between the first segment and the second segment, such that when the oligonucleotide probe and the enhancer oligonucleotide probe are both hybridized

to the nucleic acid, a single base gap exists between the oligonucleotides; (2) treating the nucleic acid with an endonuclease that will cleave the detectable label from the 3' terminus of the detection probe to release free detectable label when the detection probe is hybridized to the nucleic acid; and (3) measuring free detectable label, wherein the presence of the free
 5 detectable label indicates that the detection probe specifically hybridizes to the first segment of the nucleic acid, and indicates the sequence of the polymorphic site as the complement of the detection probe.

A further aspect of the invention pertains to a method of assessing an individual for probability of response to a breast cancer therapeutic agent, comprising: determining whether at least one
 10 allele of at least one polymorphic marker is present in a nucleic acid sample obtained from the individual, or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID
 15 NO:9) and markers in linkage disequilibrium therewith, wherein the presence of the at least one allele of the at least one marker is indicative of a probability of a positive response to the therapeutic agent

The invention in another aspect relates to a method of predicting prognosis of an individual diagnosed with breast cancer, the method comprising determining whether at least one allele of
 20 at least one polymorphic marker is present in a nucleic acid sample obtained from the individual, or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9), and
 25 markers in linkage disequilibrium therewith, wherein the presence of the at least one allele is indicative of a worse prognosis of the breast cancer in the individual.

Yet another aspect of the invention relates to a method of monitoring progress of treatment of an individual undergoing treatment for breast cancer, the method comprising determining whether at least one allele of at least one polymorphic marker is present in a nucleic acid sample
 30 obtained from the individual, or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, wherein the presence of the at least one
 35 allele is indicative of the treatment outcome of the individual. In one embodiment, the treatment is treatment by surgery, treatment by radiation therapy, or treatment by drug administration.

The invention also relates to the use of an oligonucleotide probe in the manufacture of a reagent for diagnosing and/or assessing susceptibility to breast cancer in a human individual, wherein the probe hybridizes to a segment of a nucleic acid with nucleotide sequence as set forth in any one of SEQ ID NO:1-562, wherein the probe is 15-500 nucleotides in length. In certain
 5 embodiments, the probe is about 16 to about 100 nucleotides in length. In certain other embodiments, the probe is about 20 to about 50 nucleotides in length. In certain other embodiments, the probe is about 20 to about 30 nucleotides in length.

Various diagnoses and categories of the breast cancer phenotype are within scope of the present invention. In its broadest sense, the invention relates to any breast cancer phenotype. Breast
 10 cancer, in certain embodiments, includes any clinical diagnosis of breast cancer, including, but not limited to: invasive ductal, invasive lobular, tubular, or as otherwise invasive or mixed invasive, medullary, DCIS (Ductal Carcinoma In-Situ), LCIS (Lobular Carcinoma In-Situ), or otherwise non-invasive; Invasive breast cancer, including stage 0, stage 1, stage 2 (including stage 2a and stage 2b), stage 3 (including stage 3a, stage 3b and stage 3c) and stage 4 breast
 15 cancer. In certain embodiments, the breast cancer phenotype is selected from All Breast Cancer, Multiple Primary Breast Cancer, and early onset Breast Cancer. In some embodiments, the markers of the invention are associated with risk of breast cancer in individuals with a family history of breast cancer. In one such embodiment, the summed family history (FHS) is the phenotype associated with breast cancer. In another embodiment, the breast cancer associated
 20 with the variants of the invention is estrogen receptor (ER) positive and/or progesterone receptor (PR) positive breast cancer. In one embodiment, the breast cancer associated with the variants of the invention is estrogen receptor (ER) positive. In another embodiment, the breast cancer associated with the variants of the invention is progesterone receptor (ER) positive. In
 25 one such embodiment, the markers described herein to be associated with increased risk or susceptibility of breast cancer confers increased risk or susceptibility of ER-positive and/or PR-positive breast cancer. Thus, in certain embodiments, presence of at least one of the at-risk variants of the invention is predictive of ER positive or PR positive breast cancer in the individual.

In some embodiments of the methods of the invention, the susceptibility determined in the method is increased susceptibility. In one such embodiment, the increased susceptibility is
 30 characterized by a relative risk (RR) of at least 1.10. In another embodiment, the increased susceptibility is characterized by a relative risk of at least 1.20. In another embodiment, the increased susceptibility is characterized by a relative risk of at least 1.30. In another embodiment, the increased susceptibility is characterized by a relative risk of at least 1.40. In yet another embodiment, the increased susceptibility is characterized by a relative risk of at least
 35 1.50. In a further embodiment, the increased susceptibility is characterized by a relative risk of at least 1.70. In yet another embodiment, the increased susceptibility is characterized by a relative risk of at least 2.0. Other embodiments are characterized by relative risk of at least 1.10, 1.11, 1.12, 1.13, 1.14, 1.15, 1.16, 1.17, 1.18, 1.19, 1.20, 1.21, 1.22, 1.23, 1.24, 1.25, 1.26, 1.27, 1.28, 1.29, 1.30, 1.31, 1.32, 1.33, 1.34, 1.35. Other numeric values for risk

bridging any of these above-mentioned values are also possible, and these are also within scope of the invention.

In some embodiments of the methods of the invention, the susceptibility determined in the method is decreased susceptibility. In one such embodiment, the decreased susceptibility is characterized by a relative risk (RR) of less than 0.9. In another embodiment, the decreased susceptibility is characterized by a relative risk (RR) of less than 0.8. In another embodiment, the decreased susceptibility is characterized by a relative risk (RR) of less than 0.7. In yet another embodiment, the decreased susceptibility is characterized by a relative risk (RR) of less than 0.5. Other cutoffs, such as relative risk of less than 0.89, 0.88, 0.87, 0.86, 0.85, 0.84, 0.83, 0.82, 0.81, 0.80, 0.79, 0.78, 0.77, 0.76, 0.75, 0.74, 0.73, 0.72, 0.71, 0.70, and so on, are within scope of the invention.

The invention also relates to kits. In one such aspect, the invention relates to a kit for assessing susceptibility to breast cancer in a human individual, the kit comprising reagents necessary for selectively detecting at least one allele of at least one polymorphic marker selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8), rs6001954 (SEQ ID NO:9) and markers in linkage disequilibrium therewith, in the genome of the individual, wherein the presence of the at least one allele is indicative of increased susceptibility to breast cancer. In one embodiment, the kit further comprises a collection of data comprising correlation data between the at least one polymorphism and susceptibility to breast cancer. The correlation data may be in any suitable formation, for example as a Relative Risk measure (RR), odds ratio (OR), or other convenient measure known to the skilled person. In one embodiment, the collection of data is on a computer-readable medium. In another aspect, the invention relates to a kit for assessing susceptibility to breast cancer in a human individual, the kit comprising reagents for selectively detecting at least one allele of at least one polymorphic marker in the genome of the individual, wherein the polymorphic marker is selected from rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5), rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9), and wherein the presence of the at least one allele is indicative of a susceptibility to breast cancer. In one embodiment, the at least one polymorphic marker is selected from the markers set forth in Table 4.

Kit reagents may in one embodiment comprise at least one contiguous oligonucleotide that hybridizes to a fragment of the genome of the individual comprising the at least one polymorphic marker. In another embodiment, the kit comprises at least one pair of oligonucleotides that hybridize to opposite strands of a genomic segment obtained from the subject, wherein each oligonucleotide primer pair is designed to selectively amplify a fragment of the genome of the individual that includes one polymorphism, wherein the polymorphism is selected from the group

consisting of the polymorphisms as defined in Table 4, and wherein the fragment is at least 20 base pairs in size. In one embodiment, the oligonucleotide is completely complementary to the genome of the individual. In another embodiment, the kit further contains buffer and enzyme for amplifying said segment. In another embodiment, the reagents further comprise a label for
 5 detecting said fragment.

In one preferred embodiment, the kit comprises: a detection oligonucleotide probe that is from 5-100 nucleotides in length; an enhancer oligonucleotide probe that is from 5-100 nucleotides in length; and an endonuclease enzyme; wherein the detection oligonucleotide probe specifically hybridizes to a first segment of the nucleic acid whose nucleotide sequence is set forth in any
 10 one of SEQ ID NO:1-562, and wherein the detection oligonucleotide probe comprises a detectable label at its 3' terminus and a quenching moiety at its 5' terminus; wherein the enhancer oligonucleotide is from 5-100 nucleotides in length and is complementary to a second segment of the nucleotide sequence that is 5' relative to the oligonucleotide probe, such that the enhancer oligonucleotide is located 3' relative to the detection oligonucleotide probe when both
 15 oligonucleotides are hybridized to the nucleic acid; wherein a single base gap exists between the first segment and the second segment, such that when the oligonucleotide probe and the enhancer oligonucleotide probe are both hybridized to the nucleic acid, a single base gap exists between the oligonucleotides; and wherein treating the nucleic acid with the endonuclease will cleave the detectable label from the 3' terminus of the detection probe to release free detectable
 20 label when the detection probe is hybridized to the nucleic acid.

Kits according to the present invention may also be used in the other methods of the invention, including methods of assessing risk of developing at least a second primary tumor in an individual previously diagnosed with breast cancer, methods of assessing an individual for probability of response to a breast cancer therapeutic agent, and methods of monitoring
 25 progress of a treatment of an individual diagnosed with breast cancer and given a treatment for the disease.

The markers that are described herein to be associated with breast cancer can all be used in the various aspects of the invention, including the methods, kits, uses, apparatus, procedures described herein. In a general sense, the invention relates to markers associated with any one
 30 of LD block C09, LD block C10A, LD block 10B, LD block C11A, LD block C11B, LD block C14, LD block C18, LD block C22A, and LD block C22B as defined herein. In certain embodiments, the invention relates to the markers set forth in Table 1 or Table 4, and markers in linkage disequilibrium therewith. In certain other embodiments, the invention relates to the markers set forth in Table 4. In certain other embodiments, the invention relates to markers rs2005154,
 35 rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954, and markers in linkage disequilibrium therewith. In some other preferred embodiments, the invention relates to any one of rs2005154 (SEQ ID NO:1), rs2184380 (SEQ ID NO:2), rs2224696 (SEQ ID NO:3), rs2242503 (SEQ ID NO:4), rs12291026 (SEQ ID NO:5),

rs999737 (SEQ ID NO:6), rs9956546 (SEQ ID NO:7), rs11912922 (SEQ ID NO:8) and rs6001954 (SEQ ID NO:9).

In certain embodiments, the at least one marker allele conferring increased risk of breast cancer is selected from rs2005154 allele T, rs2184380 allele G, rs2224696 allele T, rs2242503 allele C, rs12291026 allele G, rs999737 allele C, rs9956546 allele A, rs11912922 allele T and rs6001954 allele G. In these embodiments, the presence of the allele (the at-risk allele) is indicative of increased risk of breast cancer.

In certain embodiments of the invention, linkage disequilibrium is determined using the linkage disequilibrium measures r^2 and $|D'|$, which give a quantitative measure of the extent of linkage disequilibrium (LD) between two genetic element (e.g., polymorphic markers). Certain numerical values of these measures for particular markers are indicative of the markers being in linkage disequilibrium, as described further herein. In one embodiment of the invention, linkage disequilibrium between markers (i.e., LD values indicative of the markers being in linkage disequilibrium) is defined as $r^2 > 0.1$. In another embodiment, linkage disequilibrium is defined as $r^2 > 0.2$. Other embodiments can include other definitions of linkage disequilibrium, such as $r^2 > 0.25$, $r^2 > 0.3$, $r^2 > 0.35$, $r^2 > 0.4$, $r^2 > 0.45$, $r^2 > 0.5$, $r^2 > 0.55$, $r^2 > 0.6$, $r^2 > 0.65$, $r^2 > 0.7$, $r^2 > 0.75$, $r^2 > 0.8$, $r^2 > 0.85$, $r^2 > 0.9$, $r^2 > 0.95$, $r^2 > 0.96$, $r^2 > 0.97$, $r^2 > 0.98$, or $r^2 > 0.99$. Linkage disequilibrium can in certain embodiments also be defined as $|D'| > 0.2$, or as $|D'| > 0.3$, $|D'| > 0.4$, $|D'| > 0.5$, $|D'| > 0.6$, $|D'| > 0.7$, $|D'| > 0.8$, $|D'| > 0.9$, $|D'| > 0.95$, $|D'| > 0.98$ or $|D'| > 0.99$. In certain embodiments, linkage disequilibrium is defined as fulfilling two criteria of r^2 and $|D'|$, such as $r^2 > 0.2$ and $|D'| > 0.8$. Other combinations of values for r^2 and $|D'|$ are also possible and within scope of the present invention, including but not limited to the values for these parameters set forth in the above.

It should be understood that all combinations of features described herein are contemplated, even if the combination of feature is not specifically found in the same sentence or paragraph herein. This includes in particular the use of all markers disclosed herein, alone or in combination, for analysis individually or in haplotypes, in all aspects of the invention as described herein.

This description may contain subject matter that falls outside the scope of the claims. This subject matter is included to aid understanding of the invention.

BRIEF DESCRIPTION OF THE DRAWING

The foregoing and other objects, features and advantages of the invention will be apparent from the following more particular description of preferred embodiments of the invention.

FIG 1 provides a diagram illustrating a computer-implemented system utilizing risk variants as described herein.

DETAILED DESCRIPTION

- 5 The present invention discloses polymorphic variants and haplotypes that have been found to be associated with breast cancer. Particular alleles at polymorphic markers on chromosomes 9, 10, 11, 14, 18 and 22 have been found to be associated with risk of developing breast cancer. Such markers and haplotypes are useful for diagnostic purposes, for methods of predicting drug response, and methods for predicting treatment progress, as described in further detail herein.
- 10 Further applications of the present invention includes methods for assessing response to breast cancer therapy by surgery or radiation utilizing the polymorphic markers of the invention, as well as kits for use in the methods of the invention.

Definitions

- 15 Unless otherwise indicated, nucleic acid sequences 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 the ordinary person skilled in the art to which the invention pertains.
- 20 The following terms shall, in the present context, have the meaning as indicated:
- A "polymorphic marker", sometimes referred to as a "marker", as described herein, refers to a genomic polymorphic site. Each polymorphic marker has at least two sequence variations characteristic of particular alleles at the polymorphic site. Thus, genetic association to a polymorphic marker implies that there is association to at least one specific allele of that
- 25 particular polymorphic marker. The marker can comprise any allele of any variant type found in the genome, including single nucleotide polymorphisms (SNPs), mini- or microsatellites, translocations and copy number variations (insertions, deletions, duplications). Polymorphic markers can be of any measurable frequency in the population. For mapping of disease genes, polymorphic markers with population frequency higher than 5-10% are in general most useful.
- 30 However, polymorphic markers may also have lower population frequencies, such as 1-5% frequency, or even lower frequency, in particular copy number variations (CNVs). The term shall, in the present context, be taken to include polymorphic markers with any population frequency.

An "allele" refers to the nucleotide sequence of a given locus (position) on a chromosome. A polymorphic marker allele thus refers to the composition (i.e., sequence) of the marker on a chromosome. Genomic DNA from an individual contains two alleles for any given polymorphic marker, representative of each copy of the marker on each chromosome. Sequence codes for nucleotides used herein are: A = 1, C = 2, G = 3, T = 4. For microsatellite alleles, the CEPH sample (Centre d'Etudes du Polymorphisme Humain, genomics repository, CEPH sample 1347-02) is used as a reference, the shorter allele of each microsatellite in this sample is set as 0 and all other alleles in other samples are numbered in relation to this reference. Thus, e.g., allele 1 is 1 bp longer than the shorter allele in the CEPH sample, allele 2 is 2 bp longer than the shorter allele in the CEPH sample, allele 3 is 3 bp longer than the lower allele in the CEPH sample, etc., and allele -1 is 1 bp shorter than the shorter allele in the CEPH sample, allele -2 is 2 bp shorter than the shorter allele in the CEPH sample, etc.

A "Single Nucleotide Polymorphism" or "SNP" is a DNA sequence variation occurring when a single nucleotide at a specific location in the genome differs between members of a species or between paired chromosomes in an individual. Most SNP polymorphisms have two alleles. Each individual is in this instance either homozygous for one allele of the polymorphism (i.e. both chromosomal copies of the individual have the same nucleotide at the SNP location), or the individual is heterozygous (i.e. the two sister chromosomes of the individual contain different nucleotides). The SNP nomenclature as reported herein refers to the official Reference SNP (rs) ID identification tag as assigned to each unique SNP by the National Center for Biotechnological Information (NCBI).

Sequence conucleotide ambiguity as described herein is as proposed by IUPAC-IUB. These codes are compatible with the codes used by the EMBL, GenBank, and PIR databases.

IUB code	Meaning
A	Adenosine
C	Cytidine
G	Guanine
T	Thymidine
R	G or A
Y	T or C
K	G or T
M	A or C
S	G or C
W	A or T
B	C G or T
D	A G or T
H	A C or T
V	A C or G
N	A C G or T (Any base)

The sequence listing presented herein provides flanking sequence for the polymorphic markers shown herein in Tables 1-4, with the polymorphic site indicated in the sequence using the sequence conucleotide ambiguity code as shown above.

- 5 A nucleotide position at which more than one sequence is possible in a population (either a natural population or a synthetic population, *e.g.*, a library of synthetic molecules) is referred to herein as a "polymorphic site".

A "variant", as described herein, refers to a segment of DNA that differs from the reference DNA. A "marker" or a "polymorphic marker", as defined herein, is a variant. Alleles that differ from the reference are referred to as "variant" alleles.

- 10 A "microsatellite" is a polymorphic marker that has multiple small repeats of bases that are 2-8 nucleotides in length (such as CA repeats) at a particular site, in which the number of repeat lengths varies in the general population.

An "indel" is a common form of polymorphism comprising a small insertion or deletion that is typically only a few nucleotides long.

- 15 A "haplotype," as described herein, refers to a segment of genomic DNA within one strand of DNA that is characterized by a specific combination of alleles arranged along the segment. For diploid organisms such as humans, a haplotype comprises one member of the pair of alleles for each polymorphic marker or locus. In a certain embodiment, the haplotype can comprise two or more alleles, three or more alleles, four or more alleles, or five or more alleles.

- 20 The term "susceptibility", as described herein, encompasses both increased susceptibility and decreased susceptibility. Thus, particular polymorphic markers and/or haplotypes of the invention may be characteristic of increased susceptibility (*i.e.*, increased risk) of breast cancer, as characterized by a relative risk (RR) of greater than one, or as an odds ratio (OR) of greater than one. Alternatively, the markers and/or haplotypes of the invention are characteristic of
 25 decreased susceptibility (*i.e.*, decreased risk) of breast cancer, as characterized by a relative risk of less than one, or an odds ratio of less than one. Haplotypes are described herein in the context of the marker name and the allele of the marker in that haplotype, *e.g.*, "A rs9956546" refers to the A allele of marker rs9956546 being in the haplotype, and this nomenclature is equivalent to "rs9956546 allele A" and "A- rs9956546". Furthermore, allelic codes in haplotypes
 30 are as for individual markers, *i.e.* 1 = A, 2 = C, 3 = G and 4 = T.

- The term "susceptibility", as described herein, refers to the proneness of an individual towards the development of a certain state (*e.g.*, a certain trait, phenotype or disease, *e.g.*, breast cancer), or towards being less able to resist a particular state than the average individual. The term encompasses both increased susceptibility and decreased susceptibility. Thus, particular
 35 alleles at polymorphic markers and/or haplotypes of the invention as described herein may be

characteristic of increased susceptibility (i.e., increased risk) of breast cancer, as characterized by a relative risk (RR) or odds ratio (OR) of greater than one for the particular allele or haplotype. Alternatively, the markers and/or haplotypes of the invention are characteristic of decreased susceptibility (i.e., decreased risk) of breast cancer, as characterized by a relative risk of less than one.

The term "and/or" shall in the present context be understood to indicate that either or both of the items connected by it are involved. In other words, the term herein shall be taken to mean "one or the other or both".

The term "look-up table", as described herein, is a table that correlates one form of data to another form, or one or more forms of data to a predicted outcome to which the data is relevant, such as phenotype or trait. For example, a look-up table can comprise a correlation between allelic data for at least one polymorphic marker and a particular trait or phenotype, such as a particular disease diagnosis, that an individual who comprises the particular allelic data is likely to display, or is more likely to display than individuals who do not comprise the particular allelic data. Look-up tables can be multidimensional, i.e. they can contain information about multiple alleles for single markers simultaneously, or they can contain information about multiple markers, and they may also comprise other factors, such as particulars about diseases diagnoses, racial information, biomarkers, biochemical measurements, therapeutic methods or drugs, etc.

A "computer-readable medium", is an information storage medium that can be accessed by a computer using a commercially available or custom-made interface. Exemplary computer-readable media include memory (e.g., RAM, ROM, flash memory, etc.), optical storage media (e.g., CD-ROM), magnetic storage media (e.g., computer hard drives, floppy disks, etc.), punch cards, or other commercially available media. Information may be transferred between a system of interest and a medium, between computers, or between computers and the computer-readable medium for storage or access of stored information. Such transmission can be electrical, or by other available methods, such as IR links, wireless connections, etc.

A "nucleic acid sample" is a sample obtained from an individual that contains nucleic acid (DNA or RNA). In certain embodiments, i.e. the detection of specific polymorphic markers and/or haplotypes, the nucleic acid sample comprises genomic DNA. Such a nucleic acid sample can be obtained from any source that contains genomic DNA, including as a blood sample, sample of amniotic fluid, sample of cerebrospinal fluid, or tissue sample from skin, muscle, buccal or conjunctival mucosa, placenta, gastrointestinal tract or other organs.

The term "breast cancer therapeutic agent" refers to an agent that can be used to ameliorate or prevent symptoms associated with breast cancer.

The term "breast cancer-associated nucleic acid", as described herein, refers to a nucleic acid that has been found to be associated to breast cancer. This includes, but is not limited to, the markers and haplotypes described herein and markers and haplotypes in strong linkage disequilibrium (LD) therewith.

- 5 The term "Breast Cancer", as described herein, refers to any clinical diagnosis of breast cancer, and includes any and all particular subphenotypes of breast cancer. For example, breast cancer is sometimes categorized as estrogen receptor (ER) positive breast or estrogen receptor negative breast cancer; breast cancer is sometimes also categorized as progesterone receptor (PR) positive or negative. Breast cancer is furthermore sometimes diagnosed as invasive ductal, as
 10 invasive lobular, as tubular, or as otherwise invasive or mixed invasive. Breast cancer can also be categorized as medullary DCIS (Ductal Carcinoma In-Situ) or LCIS (Lobular Carcinoma In-Situ, or otherwise non-invasive. Invasive breast cancer can also be defined as stage 0, stage 1, stage 2 (including stage 2a and stage 2b), stage 3 (including stage 3a, stage 3b and stage 3c) or stage 4 breast cancer. In the present context, "breast cancer" can include any of these
 15 subphenotypes of breast cancer, and also includes any other clinically applicable subphenotypes of breast cancer.

The term "All Breast Cancer", or "All BC", refers to all individuals diagnosed with breast cancer.

- The term "Medium Predisposition" breast cancer or "MedPre" breast cancer, refers to a sub-phenotype of breast cancer. The definition of this phenotype requires that the proband fulfills at
 20 least one of the following criteria:

The proband is a member of a cluster of breast cancer cases containing 3 or more affected relatives within a genetic distance of 3 meiotic events (3M).

The proband is a member of an affected pair related within 3M, one of whom was diagnosed when aged 50 or younger.

- 25 The proband is a member of an affected pair related within 3M, one of whom was diagnosed with a second primary tumor of any type.

The proband has been diagnosed with a second primary tumor of any type.

- The term "Multiple Primary Breast Tumor", or "MPBC", as described herein, refers to cases where at least one Primary tumor is diagnosed in addition to the first breast cancer diagnosis, and the
 30 two tumors confirmed both clinically and by histology to be independent primary tumors, arising simultaneously or subsequently to the first breast cancer and occurring in the contralateral or ipsilateral breast.

The term "family history score" or "FHS", as described herein, is defined based on the number of relatives affected with breast cancer for a proband with the disease. For each proband, a score of 1 is assigned for each affected first-degree relative, 0.5 for each affected second degree relative, and 0.25 for each third-degree relative. The total sum thus obtained over all affected relatives represents the summed family history score or FHS.

The term "estrogen receptor positive breast cancer", or "ER-positive breast cancer", as described herein, refers to tumors determined to be positive for estrogen receptor. In the present context, ER levels of greater than or equal to 10 fmol/mg and/or an immunohistochemical observation of greater than or equal to 10% positive nuclei is considered to be ER positive. Breast cancer that does not fulfill the criteria of being ER positive is defined herein as "ER negative" or "estrogen receptor negative".

The term "progesterone receptor positive breast cancer", or "PR-positive breast cancer", as described herein, refers to tumors determined to be positive for progesterone receptor. In the present context, PR levels of greater than or equal to 10 fmol/mg and/or an immunohistochemical observation of greater than or equal to 10% positive nuclei is considered to be PR positive. Breast cancer that does not fulfill the criteria of being PR positive is defined herein as "PR negative" or "progesterone receptor negative".

The term "PAX5" or "PAX5 gene", as described herein, refers to the PAIRED BOX GENE 5 gene, also known as BSAP on human chromosome 9p13.

The term "TUB" or "rd5 gene", as described herein, refers to the Tubby homolog (mouse) gene on human chromosome 11p15.5.

The term "SERPINH1" as described herein, refers to serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1) gene on human chromosome 11.

Also, the term "RAD51L1 X-gene" also termed; "REC2; R51H2; hREC2; RAD51B or MGC34245" as described herein, refers to a protein coding gene located on human chromosome 14q23-24.

Furthermore, the term "FHOD3gene", also known as "FHOS2gene" or "Formactin2" as described herein, stands for formin homology 2 domain containing 3 gene, a protein coding gene located on human chromosome 18q12.

Finally, the term "TNRC6B" or "TNRC6B gene", also known as "KIAA1093" as described herein, refers to trinucleotide repeat containing 6B, a protein coding gene located on human chromosome 22q13.

Through association analysis of a population of individuals diagnosed with breast cancer according to the present invention, it has been discovered that certain alleles at certain polymorphic markers on chromosomes 9, 10, 11, 14, 18 and 22 are associated with breast cancer. A genome-wide analysis for variants associated with cancer revealed association of
 5 breast cancer to nine chromosomal regions i.e.

- chromosome 9, between positions 36,806,001 and 36,859,001 (LD block C09);
- chromosome 10, between positions 8,643,001 and 8,817,001 (LD block C10A);
- chromosome 10, between positions 9,077,001 and 9,264,001 (LD block C10B);
- chromosome 11, between positions 8,053,268 and 8,191,268 (LD block C11A);
- 10 chromosome 11, between positions 74,886,341 and 74,971,341 (LD block C11B);
- chromosome 14, between positions 68,035,712 and 68,130,712 (LD block C14);
- chromosome 18, between positions 32,110,012 and 32,145,012 (LD block C18);
- chromosome 22, between positions 38,704,907 and 38,859,907 (LD block C22A);
- chromosome 22, between positions 38,859,907 and 39,411,907 (LD block C22B);
- 15 wherein all positions correspond to NCBI Build 36 coordinates).

Particular markers within these regions were found to be associated with an increased risk of breast cancer in these locations.

Through genotyping of approximately 1840 Icelandic breast cancer patients and an average of 30,350 controls using the Illumina HumanHap300 microarray technology, a large number of
 20 markers at several chromosomal locations were found to show association to breast cancer (Table1). In particular, nine SNPs; rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954 were found to be associated with an increased risk of breast cancer.

Follow-up analysis in additional cohorts from Iceland, Holland, Spain and Sweden showed that
 25 the association signals of the nine markers are indeed significant (Table 3). These markers, and surrogate markers in linkage disequilibrium with any one of these markers, are therefore useful for predicting risk of breast cancer in individuals. Exemplary surrogate markers are presented in Table 4 herein.

These markers thus identify nine chromosomal regions that in particular are expected to contain markers predictive of breast cancer, by virtue of their LD with any one of the above-mentioned markers. These regions are also called herein LD block C09, LD block C10A, LD block 10B, LD block C11A, LD block C11B, LD block C14, LD block C18, LD block C22A, and LD block C22B.

- 5 The skilled person will appreciate that markers in LD with any of the nine anchor markers rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954 can be located outside the LD blocks as defined herein. This is a consequence of the fact that LD can extend beyond the apparent physical boundaries of LD blocks as usually defined (boundaries usually defined by regions of high recombination rates). Such surrogate
10 markers in LD with these nine markers are specifically also contemplated to be useful for the present invention, and are therefore also within scope of the present invention.

Assessment for markers and haplotypes

- The genomic sequence within populations is not identical when individuals are compared.
- 15 Rather, the genome exhibits sequence variability between individuals at many locations in the genome. Such variations in sequence are commonly referred to as polymorphisms, and there are many such sites within each genome. For example, the human genome exhibits sequence variations which occur on average every 500 base pairs. The most common sequence variant consists of base variations at a single base position in the genome, and such sequence variants,
20 or polymorphisms, are commonly called Single Nucleotide Polymorphisms ("SNPs"). These SNPs are believed to have arisen by a single mutational event, and therefore there are usually two possible alleles possible at each SNPs site; the original allele and the mutated (alternate) allele. Due to natural genetic drift and possibly also selective pressure, the original mutation has resulted in a polymorphism characterized by a particular frequency of its alleles in any given
25 population. Many other types of sequence variants are found in the human genome, including mini- and microsatellites, and insertions, deletions, inversions (also called copy number variations (CNVs)). A polymorphic microsatellite has multiple small repeats of bases (such as CA repeats, TG on the complementary strand) at a particular site in which the number of repeat lengths varies in the general population. In general terms, each version of the sequence with
30 respect to the polymorphic site represents a specific allele of the polymorphic site. All sequence variants can be referred to as polymorphisms, occurring at specific polymorphic sites characteristic of the sequence variant in question. In general, polymorphisms can comprise any number of specific alleles within the population, although each human individual has two alleles at each polymorphic site – one maternal and one paternal allele. Thus in one embodiment of
35 the invention, the polymorphism is characterized by the presence of two or more alleles in a population. In another embodiment, the polymorphism is characterized by the presence of three or more alleles. In other embodiments, the polymorphism is characterized by four or more

alleles, five or more alleles, six or more alleles, seven or more alleles, nine or more alleles, or ten or more alleles. All such polymorphisms can be utilized in the methods and kits of the present invention, and are thus within the scope of the invention.

Due to their abundance, SNPs account for a majority of sequence variation in the human genome. Over 6 million human SNPs have been validated to date (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi). However, CNVs are receiving increased attention. These large-scale polymorphisms (typically 1kb or larger) account for polymorphic variation affecting a substantial proportion of the assembled human genome; known CNVs cover over 15% of the human genome sequence (Estivill, X Armengol; L., PloS Genetics 3:1787-99 (2007). http://projects.tcag.ca/variation/). Most of these polymorphisms are however very rare, and on average affect only a fraction of the genomic sequence of each individual. CNVs are known to affect gene expression, phenotypic variation and adaptation by disrupting gene dosage, and are also known to cause disease (microdeletion and microduplication disorders) and confer risk of common complex diseases, including HIV-1 infection and glomerulonephritis (Redon, R., et al. Nature 23:444-454 (2006)). It is thus possible that either previously described or unknown CNVs represent causative variants in linkage disequilibrium with the disease-associated markers described herein. Methods for detecting CNVs include comparative genomic hybridization (CGH) and genotyping, including use of genotyping arrays, as described by Carter (Nature Genetics 39:S16-S21 (2007)). The Database of Genomic Variants (http://projects.tcag.ca/variation/) contains updated information about the location, type and size of described CNVs. The database currently contains data for over 21,000 CNVs.

In some instances, reference is made to different alleles at a polymorphic site without choosing a reference allele. Alternatively, a reference sequence can be referred to for a particular polymorphic site. The reference allele is sometimes referred to as the "wild-type" allele and it usually is chosen as either the first sequenced allele or as the allele from a "non-affected" individual (e.g., an individual that does not display a trait or disease phenotype).

Alleles for SNP markers as referred to herein refer to the bases A, C, G or T as they occur at the polymorphic site. The allele codes for SNPs used herein are as follows: 1= A, 2=C, 3=G, 4=T. Since human DNA is double-stranded, the person skilled in the art will realize that by assaying or reading the opposite DNA strand, the complementary allele can in each case be measured. Thus, for a polymorphic site (polymorphic marker) characterized by an A/G polymorphism, the methodology employed to detect the marker may be designed to specifically detect the presence of one or both of the two bases possible, i.e. A and G. Alternatively, by designing an assay that is designed to detect the opposite strand on the DNA template, the presence of the complementary bases T and C can be measured. Quantitatively (for example, in terms of relative risk), identical results would be obtained from measurement of either DNA strand (+ strand or - strand).

- Typically, a reference sequence is referred to for a particular sequence. Alleles that differ from the reference are sometimes referred to as "variant" alleles. A variant sequence, as used herein, refers to a sequence that differs from the reference sequence but is otherwise substantially similar. Alleles at the polymorphic genetic markers described herein are variants. Variants can
- 5 include changes that affect a polypeptide. Sequence differences, when compared to a reference nucleotide sequence, can include the insertion or deletion of a single nucleotide, or of more than one nucleotide, resulting in a frame shift; the change of at least one nucleotide, resulting in a change in the encoded amino acid; the change of at least one nucleotide, resulting in the generation of a premature stop codon; the deletion of several nucleotides, resulting in a deletion
- 10 of one or more amino acids encoded by the nucleotides; the insertion of one or several nucleotides, such as by unequal recombination or gene conversion, resulting in an interruption of the coding sequence of a reading frame; duplication of all or a part of a sequence; transposition; or a rearrangement of a nucleotide sequence. Such sequence changes can alter the polypeptide encoded by the nucleic acid. For example, if the change in the nucleic acid sequence causes a
- 15 frame shift, the frame shift can result in a change in the encoded amino acids, and/or can result in the generation of a premature stop codon, causing generation of a truncated polypeptide. Alternatively, a polymorphism can be a synonymous change in one or more nucleotides (*i.e.*, a change that does not result in a change in the amino acid sequence). Such a polymorphism can, for example, alter splice sites, affect the stability or transport of mRNA, or otherwise affect the
- 20 transcription or translation of an encoded polypeptide. It can also alter DNA to increase the possibility that structural changes, such as amplifications or deletions, occur at the somatic level. The polypeptide encoded by the reference nucleotide sequence is the "reference" polypeptide with a particular reference amino acid sequence, and polypeptides encoded by variant alleles are referred to as "variant" polypeptides with variant amino acid sequences.
- 25 A haplotype refers to a single-stranded segment of DNA that is characterized by a specific combination of alleles arranged along the segment. For diploid organisms such as humans, a haplotype comprises one member of the pair of alleles for each polymorphic marker or locus. In a certain embodiment, the haplotype can comprise two or more alleles, three or more alleles, four or more alleles, or five or more alleles, each allele corresponding to a specific polymorphic
- 30 marker along the segment. Haplotypes can comprise a combination of various polymorphic markers, *e.g.*, SNPs and microsatellites, having particular alleles at the polymorphic sites. The haplotypes thus comprise a combination of alleles at various genetic markers.

- Detecting specific polymorphic markers and/or haplotypes can be accomplished by methods known in the art for detecting sequences at polymorphic sites. For example, standard
- 35 techniques for genotyping for the presence of SNPs and/or microsatellite markers can be used, such as fluorescence-based techniques (Chen, X. *et al.*, *Genome Res.* 9(5): 492-98 (1999)), utilizing PCR, LCR, Nested PCR and other techniques for nucleic acid amplification. Specific commercial methodologies available for SNP genotyping include, but are not limited to, TaqMan genotyping assays and SNiPlex platforms (Applied Biosystems), gel electrophoresis (Applied

Biosystems), mass spectrometry (e.g., MassARRAY system from Sequenom), minisequencing methods, real-time PCR, Bio-Plex system (BioRad), CEQ and SNPstream systems (Beckman), array hybridization technology (e.g., Affymetrix GeneChip; Perlegen), BeadArray Technologies (e.g., Illumina GoldenGate and Infinium assays), array tag technology (e.g., Parallele), and endonuclease-based fluorescence hybridization technology (Invader; Third Wave). Some of the available array platforms, including Affymetrix SNP Array 6.0 and Illumina CNV370-Duo and 1M BeadChips, include SNPs that tag certain CNVs. This allows detection of CNVs via surrogate SNPs included in these platforms. Thus, by use of these or other methods available to the person skilled in the art, one or more alleles at polymorphic markers, including microsatellites, SNPs or other types of polymorphic markers, can be identified.

In certain embodiments, polymorphic markers are detected by sequencing technologies. Obtaining sequence information about an individual identifies particular nucleotides in the context of a sequence. For SNPs, sequence information about a single unique sequence site is sufficient to identify alleles at that particular SNP. For markers comprising more than one nucleotide, sequence information about the nucleotides of the individual that contain the polymorphic site identifies the alleles of the individual for the particular site. The sequence information can be obtained from a sample from the individual. In certain embodiments, the sample is a nucleic acid sample. In certain other embodiments, the sample is a protein sample.

Various methods for obtaining nucleic acid sequence are known to the skilled person, and all such methods are useful for practicing the invention. Sanger sequencing is a well-known method for generating nucleic acid sequence information. Recent methods for obtaining large amounts of sequence data have been developed, and such methods are also contemplated to be useful for obtaining sequence information. These include pyrosequencing technology (Ronaghi, M. *et al. Anal Biochem* 267:65-71 (1999); Ronaghi, *et al. Biotechniques* 25:876-878 (1998)), e.g. 454 pyrosequencing (Nyren, P., *et al. Anal Biochem* 208:171-175 (1993)), Illumina/Solexa sequencing technology (<http://www.illumina.com>; see also Strausberg, RL, *et al Drug Disc Today* 13:569-577 (2008)), and Supported Oligonucleotide Ligation and Detection Platform (SOLiD) technology (Applied Biosystems, <http://www.appliedbiosystems.com>); Strausberg, RL, *et al Drug Disc Today* 13:569-577 (2008).

It is possible to impute or predict genotypes for un-genotyped relatives of genotyped individuals. For every un-genotyped case, it is possible to calculate the probability of the genotypes of its relatives given its four possible phased genotypes. In practice it may be preferable to include only the genotypes of the case's parents, children, siblings, half-siblings (and the half-sibling's parents), grand-parents, grand-children (and the grand-children's parents) and spouses. It will be assumed that the individuals in the small sub-pedigrees created around each case are not related through any path not included in the pedigree. It is also assumed that alleles that are not transmitted to the case have the same frequency – the population allele frequency. Let us

consider a SNP marker with the alleles A and G. The probability of the genotypes of the case's relatives can then be computed by:

$$\Pr(\text{genotypes of relatives}; \theta) = \sum_{h \in \{AA, AG, GA, GG\}} \Pr(h; \theta) \Pr(\text{genotypes of relatives} | h),$$

where θ denotes the A allele's frequency in the cases. Assuming the genotypes of each set of
5 relatives are independent, this allows us to write down a likelihood function for θ :

$$L(\theta) = \prod_i \Pr(\text{genotypes of relatives of case } i; \theta). \quad (*)$$

This assumption of independence is usually not correct. Accounting for the dependence between individuals is a difficult and potentially prohibitively expensive computational task. The likelihood
10 function in (*) may be thought of as a pseudolikelihood approximation of the full likelihood function for θ which properly accounts for all dependencies. In general, the genotyped cases and controls in a case-control association study are not independent and applying the case-control method to related cases and controls is an analogous approximation. The method of genomic control (Devlin, B. et al., *Nat Genet* 36, 1129-30; author reply 1131 (2004)) has proven to be
15 successful at adjusting case-control test statistics for relatedness. We therefore apply the method of genomic control to account for the dependence between the terms in our pseudolikelihood and produce a valid test statistic.

Fisher's information can be used to estimate the effective sample size of the part of the pseudolikelihood due to un-genotyped cases. Breaking the total Fisher information, I , into the
20 part due to genotyped cases, I_g , and the part due to ungenotyped cases, I_u , $I = I_g + I_u$, and denoting the number of genotyped cases with N , the effective sample size due to the un-genotyped cases is estimated by $\frac{I_u}{I_g} N$.

In the present context, an individual who is at an increased susceptibility (i.e., increased risk) for a disease, is an individual in whom at least one specific allele at one or more polymorphic marker or haplotype conferring increased susceptibility (increased risk) for the disease is identified (i.e.,
25 at-risk marker alleles or haplotypes). The at-risk marker or haplotype is one that confers an increased risk (increased susceptibility) of the disease. In one embodiment, significance associated with a marker or haplotype is measured by a relative risk (RR). In another embodiment, significance associated with a marker or haplotype is measured by an odds ratio (OR). In a further embodiment, the significance is measured by a percentage. In one
30 embodiment, a significant increased risk is measured as a risk (relative risk and/or odds ratio) of at least 1.2, including but not limited to: at least 1.2, at least 1.3, at least 1.4, at least 1.5, at least 1.6, at least 1.7, 1.8, at least 1.9, at least 2.0, at least 2.5, at least 3.0, at least 4.0, and

at least 5.0. In a particular embodiment, a risk (relative risk and/or odds ratio) of at least 1.2 is significant. In another particular embodiment, a risk of at least 1.3 is significant. In yet another embodiment, a risk of at least 1.4 is significant. In a further embodiment, a relative risk of at least 1.5 is significant. In another further embodiment, a significant increase in risk is at least 1.7 is significant. However, other cutoffs are also contemplated, *e.g.*, at least 1.15, 1.25, 1.35, and so on, and such cutoffs are also within scope of the present invention. In other embodiments, a significant increase in risk is at least about 20%, including but not limited to about 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 100%, 150%, 200%, 300%, and 500%. In one particular embodiment, a significant increase in risk is at least 20%. In other embodiments, a significant increase in risk is at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90% and at least 100%. Other cutoffs or ranges as deemed suitable by the person skilled in the art to characterize the invention are however also contemplated, and those are also within scope of the present invention. In certain embodiments, a significant increase in risk is characterized by a p-value, such as a p-value of less than 0.05, less than 0.01, less than 0.001, less than 0.0001, less than 0.00001, less than 0.000001, less than 0.0000001, or less than 0.00000001.

An at-risk polymorphic marker or haplotype of the present invention is one where at least one allele of at least one marker or haplotype is more frequently present in an individual at risk for the disease or trait (affected), or diagnosed with the disease or trait, compared to the frequency of its presence in a comparison group (control), such that the presence of the marker or haplotype is indicative of susceptibility to the disease or trait (*e.g.*, breast cancer). The control group may in one embodiment be a population sample, *i.e.* a random sample from the general population. In another embodiment, the control group is represented by a group of individuals who are disease-free, *i.e.* individuals who have not been diagnosed with breast cancer. Such disease-free control may in one embodiment be characterized by the absence of one or more specific disease-associated symptoms. In another embodiment, the disease-free control group is characterized by the absence of one or more disease-specific risk factors. Such risk factors are in one embodiment at least one environmental risk factor. Representative environmental factors are natural products, minerals or other chemicals which are known to affect, or contemplated to affect, the risk of developing the specific disease or trait. Other environmental risk factors are risk factors related to lifestyle, including but not limited to food and drink habits, geographical location of main habitat, and occupational risk factors. In another embodiment, the risk factors are at least one genetic risk factor.

As an example of a simple test for correlation would be a Fisher-exact test on a two by two table. Given a cohort of chromosomes, the two by two table is constructed out of the number of chromosomes that include both of the markers or haplotypes, one of the markers or haplotypes but not the other and neither of the markers or haplotypes. Other statistical tests of association known to the skilled person are also contemplated and are also within scope of the invention.

The person skilled in the art will appreciate that for markers with two alleles present in the population being studied (such as SNPs), and wherein one allele is found in increased frequency in a group of individuals with a trait or disease in the population, compared with controls, the other allele of the marker will be found in decreased frequency in the group of individuals with the trait or disease, compared with controls. In such a case, one allele of the marker (the one found in increased frequency in individuals with the trait or disease) will be the at-risk allele, while the other allele will be a protective allele.

Thus is other embodiments of the invention, an individual who is at a decreased susceptibility (i.e., at a decreased risk) for a disease or trait is an individual in whom at least one specific allele at one or more polymorphic marker or haplotype conferring decreased susceptibility for the disease or trait is identified. The marker alleles and/or haplotypes conferring decreased risk are also said to be protective. In one aspect, the protective marker or haplotype is one that confers a significant decreased risk (or susceptibility) of the disease or trait. In one embodiment, significant decreased risk is measured as a relative risk of less than 0.90, including but not limited to less than 0.85, less than 0.80, less than 0.75, less than 0.7, less than 0.6, less than 0.5, less than 0.4, less than 0.3, less than 0.2 and less than 0.1. In one particular embodiment, significant decreased risk is less than 0.90. In another embodiment, significant decreased risk is less than 0.85. In yet another embodiment, significant decreased risk is less than 0.80. In another embodiment, the decrease in risk (or susceptibility) is at least 10%, including but not limited to at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 80%, at least 85%, at least 90%, at least 95% and at least 98%. In one particular embodiment, a significant decrease in risk is at least about 15%. In another embodiment, a significant decrease in risk at least about 20%. In another embodiment, the decrease in risk is at least about 25%. Other cutoffs or ranges as deemed suitable by the person skilled in the art to characterize the invention are however also contemplated, and those are also within scope of the present invention.

A genetic variant associated with a disease or a trait (e.g. breast cancer) can be used alone to predict the risk of the disease for a given genotype. For a biallelic marker, such as a SNP, there are 3 possible genotypes: homozygote for the at risk variant, heterozygote, and non carrier of the at risk variant. Risk associated with variants at multiple loci can be used to estimate overall risk. For multiple SNP variants, there are k possible genotypes $k = 3^n \times 2^p$; where n is the number autosomal loci and p the number of gonosomal (sex chromosomal) loci. Overall risk assessment calculations usually assume that the relative risks of different genetic variants multiply, i.e. the overall risk (e.g., RR or OR) associated with a particular genotype combination is the product of the risk values for the genotype at each locus. If the risk presented is the relative risk for a person, or a specific genotype for a person, compared to a reference population with matched gender and ethnicity, then the combined risk is the product of the locus specific risk values and also corresponds to an overall risk estimate compared with the population. If the risk for a person is based on a comparison to non-carriers of the at risk allele,

then the combined risk corresponds to an estimate that compares the person with a given combination of genotypes at all loci to a group of individuals who do not carry risk variants at any of those loci. The group of non-carriers of any at risk variant has the lowest estimated risk, and has a combined risk, compared with itself (*i.e.*, non-carriers) of 1.0, but has an overall risk, compare with the population, of less than 1.0. It should be noted that the group of non-carriers can potentially be very small, especially for large number of loci, and in that case, its relevance is correspondingly small.

The multiplicative model is a parsimonious model that usually fits the data of complex traits reasonably well. Deviations from multiplicity have been rarely described in the context of common variants for common diseases, and if reported are usually only suggestive since very large sample sizes are usually required to be able to demonstrate statistical interactions between loci.

By way of an example, let us consider the case were a total of seven variants that have been associated with breast cancer. One such example is provided by the markers rs13387042, rs4415084, rs1219648, rs3803662, rs13281615, rs3817198 and rs889312, all of which are used in the marketed deCODE BreastCancer test for breast cancer susceptibility (<http://www.decodediagnostics.com>). The total number of theoretical genotypic combinations is then $3^7 = 2187$. Some of those genotypic classes are very rare, but are still possible, and should be considered for overall risk assessment. It is likely that the multiplicative model applied in the case of multiple genetic variant will also be valid in conjunction with non-genetic risk variants assuming that the genetic variant does not clearly correlate with the "environmental" factor. In other words, genetic and non-genetic at-risk variants can be assessed under the multiplicative model to estimate combined risk, assuming that the non-genetic and genetic risk factors do not interact.

Using the same quantitative approach, the combined or overall risk associated with any plurality of these and other variants associated with breast cancer may be assessed. This includes the variants that are shown and claimed herein to be predictive of breast cancer risk.

Linkage Disequilibrium

The natural phenomenon of recombination, which occurs on average once for each chromosomal pair during each meiotic event, represents one way in which nature provides variations in sequence (and biological function by consequence). It has been discovered that recombination does not occur randomly in the genome; rather, there are large variations in the frequency of recombination rates, resulting in small regions of high recombination frequency (also called recombination hotspots) and larger regions of low recombination frequency, which are commonly referred to as Linkage Disequilibrium (LD) blocks (Myers, S, *et al.*, *Biochem Soc Trans* 34:526-

530 (2006); Jeffreys, A.J., et al., *Nature Genet* 29:217-222 (2001); May, C.A., et al., *Nature Genet* 31:272-275(2002)).

Linkage Disequilibrium (LD) refers to a non-random assortment of two genetic elements. For example, if a particular genetic element (e.g., an allele of a polymorphic marker, or a haplotype) occurs in a population at a frequency of 0.50 (50%) and another element occurs at a frequency of 0.50 (50%), then the predicted occurrence of a person's having both elements is 0.25 (25%), assuming a random distribution of the elements. However, if it is discovered that the two elements occur together at a frequency higher than 0.125, then the elements are said to be in linkage disequilibrium, since they tend to be inherited together at a higher rate than what their independent frequencies of occurrence (e.g., allele or haplotype frequencies) would predict. Roughly speaking, LD is generally correlated with the frequency of recombination events between the two elements. Allele or haplotype frequencies can be determined in a population by genotyping individuals in a population and determining the frequency of the occurrence of each allele or haplotype in the population. For populations of diploids, e.g., human populations, individuals will typically have two alleles for each genetic element (e.g., a marker, haplotype or gene).

Many different measures have been proposed for assessing the strength of linkage disequilibrium (LD; reviewed in Devlin, B. & Risch, N., *Genomics* 29:311-22 (1995)). Most capture the strength of association between pairs of biallelic sites. Two important pairwise measures of LD are r^2 (sometimes denoted Δ^2) and $|D'|$ (Lewontin, R., *Genetics* 49:49-67 (1964); Hill, W.G. & Robertson, A. *Theor. Appl. Genet.* 22:226-231 (1968)). Both measures range from 0 (no disequilibrium) to 1 ('complete' disequilibrium), but their interpretation is slightly different. $|D'|$ is defined in such a way that it is equal to 1 if just two or three of the possible haplotypes for two markers are present, and it is <1 if all four possible haplotypes are present. Therefore, a value of $|D'|$ that is <1 indicates that historical recombination may have occurred between two sites (recurrent mutation can also cause $|D'|$ to be <1 , but for single nucleotide polymorphisms (SNPs) this is usually regarded as being less likely than recombination). The measure r^2 represents the statistical correlation between two sites, and takes the value of 1 if only two haplotypes are present.

The r^2 measure is arguably the most relevant measure for association mapping, because there is a simple inverse relationship between r^2 and the sample size required to detect association between susceptibility loci and particular SNPs. These measures are defined for pairs of sites, but for some applications a determination of how strong LD is across an entire region that contains many polymorphic sites might be desirable (e.g., testing whether the strength of LD differs significantly among loci or across populations, or whether there is more or less LD in a region than predicted under a particular model). Roughly speaking, r measures how much recombination would be required under a particular population model to generate the LD that is seen in the data. This type of method can potentially also provide a statistically rigorous

approach to the problem of determining whether LD data provide evidence for the presence of recombination hotspots. For the methods described herein, a significant r^2 value between markers indicative of the markers being in linkage disequilibrium can be at least 0.1, such as at least 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, or at least 0.99. In one preferred embodiment, the significant r^2 value can be at least 0.2. Alternatively, markers in linkage disequilibrium are characterized by values of $|D'|$ of at least 0.2, such as 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 0.96, 0.97, 0.98, or at least 0.99. Thus, linkage disequilibrium represents a correlation between alleles of distinct markers. In certain embodiments, linkage disequilibrium is defined in terms of values for both the r^2 and $|D'|$ measures. In one such embodiment, a significant linkage disequilibrium is defined as $r^2 > 0.1$ and $|D'| > 0.8$, and markers fulfilling these criteria are said to be in linkage disequilibrium. In another embodiment, a significant linkage disequilibrium is defined as $r^2 > 0.2$ and $|D'| > 0.9$. Other combinations and permutations of values of r^2 and $|D'|$ for determining linkage disequilibrium are also contemplated, and are also within the scope of the invention. Linkage disequilibrium can be determined in a single human population, as defined herein, or it can be determined in a collection of samples comprising individuals from more than one human population. In one embodiment of the invention, LD is determined in a sample from one or more of the HapMap populations (Caucasian, African (Yuroban), Japanese, Chinese), as defined (<http://www.hapmap.org>). In one such embodiment, LD is determined in the CEU population of the HapMap samples (Utah residents with ancestry from northern and western Europe). In another embodiment, LD is determined in the YRI population of the HapMap samples (Yuroba in Ibadan, Nigeria). In another embodiment, LD is determined in the CHB population of the HapMap samples (Han Chinese from Beijing, China). In another embodiment, LD is determined in the JPT population of the HapMap samples (Japanese from Tokyo, Japan). In yet another embodiment, LD is determined in samples from the Icelandic population.

If all polymorphisms in the genome were identical at the population level, then every single one of them would need to be investigated in association studies. However, due to linkage disequilibrium between polymorphisms, tightly linked polymorphisms are strongly correlated, which reduces the number of polymorphisms that need to be investigated in an association study to observe a significant association. Another consequence of LD is that many polymorphisms may give an association signal due to the fact that these polymorphisms are strongly correlated.

Genomic LD maps have been generated across the genome, and such LD maps have been proposed to serve as framework for mapping disease-genes (Risch, N. & Merikangas, K, *Science* 273:1516-1517 (1996); Maniatis, N., et al., *Proc Natl Acad Sci USA* 99:2228-2233 (2002); Reich, DE et al, *Nature* 411:199-204 (2001)).

It is now established that many portions of the human genome can be broken into series of discrete haplotype blocks containing a few common haplotypes; for these blocks, linkage

disequilibrium data provides little evidence indicating recombination (see, e.g., Wall, J.D. and Pritchard, J.K., *Nature Reviews Genetics* 4:587-597 (2003); Daly, M. et al., *Nature Genet.* 29:229-232 (2001); Gabriel, S.B. et al., *Science* 296:2225-2229 (2002); Patil, N. et al., *Science* 294:1719-1723 (2001); Dawson, E. et al., *Nature* 418:544-548 (2002); Phillips, M.S. et al.,
 5 *Nature Genet.* 33:382-387 (2003)).

There are two main methods for defining these haplotype blocks: blocks can be defined as regions of DNA that have limited haplotype diversity (see, e.g., Daly, M. et al., *Nature Genet.* 29:229-232 (2001); Patil, N. et al., *Science* 294:1719-1723 (2001); Dawson, E. et al., *Nature* 418:544-548 (2002); Zhang, K. et al., *Proc. Natl. Acad. Sci. USA* 99:7335-7339 (2002)), or as
 10 regions between transition zones having extensive historical recombination, identified using linkage disequilibrium (see, e.g., Gabriel, S.B. et al., *Science* 296:2225-2229 (2002); Phillips, M.S. et al., *Nature Genet.* 33:382-387 (2003); Wang, N. et al., *Am. J. Hum. Genet.* 71:1227-1234 (2002); Stumpf, M.P., and Goldstein, D.B., *Curr. Biol.* 13:1-8 (2003)). More recently, a
 15 fine-scale map of recombination rates and corresponding hotspots across the human genome has been generated (Myers, S., et al., *Science* 310:321-32324 (2005); Myers, S. et al., *Biochem Soc Trans* 34:526530 (2006)). The map reveals the enormous variation in recombination across the genome, with recombination rates as high as 10-60 cM/Mb in hotspots, while closer to 0 in intervening regions, which thus represent regions of limited haplotype diversity and high LD. The map can therefore be used to define haplotype blocks/LD blocks as regions flanked by
 20 recombination hotspots. As used herein, the terms "haplotype block" or "LD block" includes blocks defined by any of the above described characteristics, or other alternative methods used by the person skilled in the art to define such regions.

For example The term "LD block C09", as described herein, refers to the Linkage Disequilibrium (LD) block on Chromosome 9 between positions 36,806,001 and 36,859,001 of NCBI (National
 25 Center for Biotechnology Information) Build 36. The following table defines chorosomal positions of the nine LD blocks in NCBI Build 36, referred to in this application:

Table of LD-blocks (positions from NCBI Build 36)

Chromosome	LD Block	Block Start B36	Block End B36
C09	LD block C09	36,806,001	36,859,001
C10	LD block C10A	8,643,001	8,817,001
C10	LD block C10B	9,077,001	9,264,001
C11	LD block C11A	8,053,268	8,191,268
C11	LD block C11B	74,886,341	74,971,341
C14	LD block C14	68,035,712	68,130,712
C18	LD block C18	32,110,012	32,145,012
C22	LD block C22A	38,704,907	38,859,907
C22	LD block C22B	38,859,907	39,411,907

Haplotype blocks (LD blocks) can be used to map associations between phenotype and haplotype status, using single markers or haplotypes comprising a plurality of markers. The main haplotypes can be identified in each haplotype block, and then a set of "tagging" SNPs or markers (the smallest set of SNPs or markers needed to distinguish among the haplotypes) can then be identified. These tagging SNPs or markers can then be used in assessment of samples from groups of individuals, in order to identify association between phenotype and haplotype. Markers shown herein to be associated with breast cancer are such tagging markers. If desired, neighboring haplotype blocks can be assessed concurrently, as there may also exist linkage disequilibrium among the haplotype blocks.

- 10 It has thus become apparent that for any given observed association to a polymorphic marker in the genome, that additional markers in the genome also show association. This is a natural consequence of the uneven distribution of LD across the genome, as observed by the large variation in recombination rates. The markers used to detect association thus in a sense represent "tags" for a genomic region (i.e., a haplotype block or LD block) that is associating with a given disease or trait. One or more causative (functional) variants or mutations may reside within the region found to be associating to the disease or trait. The functional variant may be another SNP, a tandem repeat polymorphism (such as a minisatellite or a microsatellite), a transposable element, or a copy number variation, such as an inversion, deletion or insertion. Such variants in LD with the variants described herein may confer a higher relative risk (RR) or odds ratio (OR) than observed for the tagging markers used to detect the association. The present invention thus refers to the markers used for detecting association to the disease, as described herein, as well as markers in linkage disequilibrium with the markers. Thus, in certain embodiments of the invention, markers that are in LD with the markers originally used to detect an association may be used as surrogate markers. The surrogate markers have in one embodiment relative risk (RR) and/or odds ratio (OR) values smaller than originally detected. In other embodiments, the surrogate markers have RR or OR values greater than those initially determined for the markers initially found to be associating with the disease. An example of such an embodiment would be a rare, or relatively rare (< 10% allelic population frequency) variant in LD with a more common variant (> 10% population frequency) initially found to be associating with the disease. Identifying and using such surrogate markers for detecting the association can be performed by routine methods well known to the person skilled in the art, and are therefore within the scope of the present invention.

Determination of haplotype frequency

- 35 The frequencies of haplotypes in patient and control groups can be estimated using an expectation-maximization algorithm (Dempster A. *et al.*, *J. R. Stat. Soc. B*, 39:1-38 (1977)). An implementation of this algorithm that can handle missing genotypes and uncertainty with the

phase can be used. Under the null hypothesis, the patients and the controls are assumed to have identical frequencies. Using a likelihood approach, an alternative hypothesis is tested, where a candidate at-risk-haplotype, which can include the markers described herein, is allowed to have a higher frequency in patients than controls, while the ratios of the frequencies of other
 5 haplotypes are assumed to be the same in both groups. Likelihoods are maximized separately under both hypotheses and a corresponding 1-df likelihood ratio statistic is used to evaluate the statistical significance.

To look for at-risk and protective markers and haplotypes within a susceptibility region, for example, within an LD block region, association of all possible combinations of genotyped
 10 markers within the region is studied. The combined patient and control groups can be randomly divided into two sets, equal in size to the original group of patients and controls. The marker and haplotype analysis is then repeated and the most significant p-value registered is determined. This randomization scheme can be repeated, for example, over 100 times to construct an empirical distribution of p-values. In a preferred embodiment, a p-value of <0.05 is
 15 indicative of an significant marker and/or haplotype association.

Haplotype Analysis

One general approach to haplotype analysis involves using likelihood-based inference applied to NEstED MOdeLS (Gretarsdottir S., et al., Nat. Genet. 35:131-38 (2003)). The method is
 20 implemented in the program NEMO, which allows for many polymorphic markers, SNPs and microsatellites. The method and software are specifically designed for case-control studies where the purpose is to identify haplotype groups that confer different risks. It is also a tool for studying LD structures. In NEMO, maximum likelihood estimates, likelihood ratios and p-values are calculated directly, with the aid of the EM algorithm, for the observed data treating it as a
 25 missing-data problem.

Even though likelihood ratio tests based on likelihoods computed directly for the observed data, which have captured the information loss due to uncertainty in phase and missing genotypes, can be relied on to give valid p-values, it would still be of interest to know how much information had been lost due to the information being incomplete. The information measure for haplotype
 30 analysis is described in Nicolae and Kong (Technical Report 537, Department of Statistics, University of Statistics, University of Chicago; *Biometrics*, 60(2):368-75 (2004)) as a natural extension of information measures defined for linkage analysis, and is implemented in NEM

Association analysis

For single marker association to a disease, the Fisher exact test can be used to calculate two-sided p-values for each individual allele. Correcting for relatedness among patients can be done by extending a variance adjustment procedure previously described (Risch, N. & Teng, J.

- 5 *Genome Res.*, 8:1273-1288 (1998)) for sibships so that it can be applied to general familial relationships. The method of genomic controls (Devlin, B. & Roeder, K. *Biometrics* 55:997 (1999)) can also be used to adjust for the relatedness of the individuals and possible stratification.

- 10 For both single-marker and haplotype analyses, relative risk (RR) and the population attributable risk (PAR) can be calculated assuming a multiplicative model (haplotype relative risk model) (Terwilliger, J.D. & Ott, J., *Hum. Hered.* 42:337-46 (1992) and Falk, C.T. & Rubinstein, P., *Ann. Hum. Genet.* 51 (Pt 3):227-33 (1987)), i.e., that the risks of the two alleles/haplotypes a person carries multiply. For example, if RR is the risk of A relative to a, then the risk of a person homozygote AA will be RR times that of a heterozygote Aa and RR^2 times that of a homozygote aa. The multiplicative model has a nice property that simplifies analysis and computations — haplotypes are independent, i.e., in Hardy-Weinberg equilibrium, within the affected population as well as within the control population. As a consequence, haplotype counts of the affecteds and controls each have multinomial distributions, but with different haplotype frequencies under the alternative hypothesis. Specifically, for two haplotypes, h_i and h_j , $\text{risk}(h_i)/\text{risk}(h_j) =$
15 $(f_i/p_i)/(f_j/p_j)$, where f and p denote, respectively, frequencies in the affected population and in the control population. While there is some power loss if the true model is not multiplicative, the loss tends to be mild except for extreme cases. Most importantly, p-values are always valid since they are computed with respect to null hypothesis.

- 25 An association signal detected in one association study may be replicated in a second cohort, ideally from a different population (e.g., different region of same country, or a different country) of the same or different ethnicity. The advantage of replication studies is that the number of tests performed in the replication study is usually quite small, and hence the less stringent the statistical measure that needs to be applied. For example, for a genome-wide search for susceptibility variants for a particular disease or trait using 300,000 SNPs, a correction for the 300,000 tests performed (one for each SNP) can be performed. Since many SNPs on the arrays typically used are correlated (i.e., in LD), they are not independent. Thus, the correction is conservative. Nevertheless, applying this correction factor requires an observed P-value of less than $0.05/300,000 = 1.7 \times 10^{-7}$ for the signal to be considered significant applying this conservative test on results from a single study cohort. Obviously, signals found in a genome-wide association study with P-values less than this conservative threshold (i.e., more significant)
35 are a measure of a true genetic effect, and replication in additional cohorts is not necessarily from a statistical point of view. Importantly, however, signals with P-values that are greater than this threshold may also be due to a true genetic effect. The sample size in the first study

may not have been sufficiently large to provide an observed P-value that meets the conservative threshold for genome-wide significance, or the first study may not have reached genome-wide significance due to inherent fluctuations due to sampling. Since the correction factor depends on the number of statistical tests performed, if one signal (one SNP) from an initial study is replicated in a second case-control cohort, the appropriate statistical test for significance is that for a single statistical test, i.e., P-value less than 0.05. Replication studies in one or even several additional case-control cohorts have the added advantage of providing assessment of the association signal in additional populations, thus simultaneously confirming the initial finding and providing an assessment of the overall significance of the genetic variant(s) being tested in human populations in general.

The results from several case-control cohorts can also be combined to provide an overall assessment of the underlying effect. The methodology commonly used to combine results from multiple genetic association studies is the Mantel-Haenszel model (Mantel and Haenszel, *J Natl Cancer Inst* 22:719-48 (1959)). The model is designed to deal with the situation where association results from different populations, with each possibly having a different population frequency of the genetic variant, are combined. The model combines the results assuming that the effect of the variant on the risk of the disease, as measured by the OR or RR, is the same in all populations, while the frequency of the variant may differ between the populations. Combining the results from several populations has the added advantage that the overall power to detect a real underlying association signal is increased, due to the increased statistical power provided by the combined cohorts. Furthermore, any deficiencies in individual studies, for example due to unequal matching of cases and controls or population stratification will tend to balance out when results from multiple cohorts are combined, again providing a better estimate of the true underlying genetic effect.

25

Risk assessment and Diagnostics

Within any given population, there is an absolute risk of developing a disease or trait, defined as the chance of a person developing the specific disease or trait over a specified time-period. For example, a woman's lifetime absolute risk of breast cancer is one in nine. That is to say, one woman in every nine will develop breast cancer at some point in their lives. Risk is typically measured by looking at very large numbers of people, rather than at a particular individual. Risk is often presented in terms of Absolute Risk (AR) and Relative Risk (RR). Relative Risk is used to compare risks associating with two variants or the risks of two different groups of people. For example, it can be used to compare a group of people with a certain genotype with another group having a different genotype. For a disease, a relative risk of 2 means that one group has twice the chance of developing a disease as the other group. The Risk presented is usually the relative risk for a person, or a specific genotype of a person, compared to the population with

matched gender and ethnicity. Risks of two individuals of the same gender and ethnicity could be compared in a simple manner. For example, if, compared to the population, the first individual has relative risk 1.5 and the second has relative risk 0.5, then the risk of the first individual compared to the second individual is $1.5/0.5 = 3$.

5 Risk Calculations

The creation of a model to calculate the overall genetic risk involves two steps: i) conversion of odds-ratios for a single genetic variant into relative risk and ii) combination of risk from multiple variants in different genetic loci into a single relative risk value.

Deriving risk from odds-ratios

- 10 Most gene discovery studies for complex diseases that have been published to date in authoritative journals have employed a case-control design because of their retrospective setup. These studies sample and genotype a selected set of cases (people who have the specified disease condition) and control individuals. The interest is in genetic variants (alleles) which frequency in cases and controls differ significantly.
- 15 The results are typically reported in odds ratios, that is the ratio between the fraction (probability) with the risk variant (carriers) versus the non-risk variant (non-carriers) in the groups of affected versus the controls, i.e. expressed in terms of probabilities conditional on the affection status:

$$OR = (Pr(c|A)/Pr(nc|A)) / (Pr(c|C)/Pr(nc|C))$$

- 20 Sometimes it is however the absolute risk for the disease that we are interested in, i.e. the fraction of those individuals carrying the risk variant who get the disease or in other words the probability of getting the disease. This number cannot be directly measured in case-control studies, in part, because the ratio of cases versus controls is typically not the same as that in the general population. However, under certain assumption, we can estimate the risk from the odds ratio.
- 25

- It is well known that under the rare disease assumption, the relative risk of a disease can be approximated by the odds ratio. This assumption may however not hold for many common diseases. Still, it turns out that the risk of one genotype variant relative to another can be estimated from the odds ratio expressed above. The calculation is particularly simple under the assumption of random population controls where the controls are random samples from the same population as the cases, including affected people rather than being strictly unaffected individuals. To increase sample size and power, many of the large genome-wide association and replication studies use controls that were neither age-matched with the cases, nor were they carefully scrutinized to ensure that they did not have the disease at the time of the study.
- 30

Hence, while not exactly, they often approximate a random sample from the general population. It is noted that this assumption is rarely expected to be satisfied exactly, but the risk estimates are usually robust to moderate deviations from this assumption.

- 5 Calculations show that for the dominant and the recessive models, where we have a risk variant carrier, "c", and a non-carrier, "nc", the odds ratio of individuals is the same as the risk ratio between these variants:

$$OR = Pr(A|c)/Pr(A|nc) = r$$

And likewise for the multiplicative model, where the risk is the product of the risk associated with the two allele copies, the allelic odds ratio equals the risk factor:

10 $OR = Pr(A|aa)/Pr(A|ab) = Pr(A|ab)/Pr(A|bb) = r$

Here "a" denotes the risk allele and "b" the non-risk allele. The factor "r" is therefore the relative risk between the allele types.

- 15 For many of the studies published in the last few years, reporting common variants associated with complex diseases, the multiplicative model has been found to summarize the effect adequately and most often provide a fit to the data superior to alternative models such as the dominant and recessive models.

The risk relative to the average population risk

- 20 It is most convenient to represent the risk of a genetic variant relative to the average population since it makes it easier to communicate the lifetime risk for developing the disease compared with the baseline population risk. For example, in the multiplicative model we can calculate the relative population risk for variant "aa" as:

$$RR(aa) = Pr(A|aa)/Pr(A) = (Pr(A|aa)/Pr(A|bb))/(Pr(A)/Pr(A|bb)) =$$

$$r^2/(Pr(aa) r^2 + Pr(ab) r + Pr(bb)) = r^2/(p^2 r^2 + 2pq r + q^2) = r^2/R$$

- 25 Here "p" and "q" are the allele frequencies of "a" and "b" respectively. Likewise, we get the $RR(ab) = r/R$ and $RR(bb) = 1/R$. The allele frequency estimates may be obtained from the publications that report the odds-ratios and from the HapMap database. Note that in the case where we do not know the genotypes of an individual, the relative genetic risk for that test or marker is simply equal to one.

As an example, for marker rs999737 on chromosome 14, allele C has an allelic OR for breast cancer of 1.15 and a frequency (p) around 0.76 in Caucasian populations. The genotype relative risk compared to genotype TT are estimated based on the multiplicative model.

For CC it is $1.15 \times 1.15 = 1.32$; for CT it is simply the OR 1.15, and for TT it is 1.0 by definition.

- 5 The frequency of allele T is $q = 1 - p = 1 - 0.76 = 0.24$. Population frequency of each of the three possible genotypes at this marker is:

$$\Pr(CC) = p^2 = 0.58, \Pr(CT) = 2pq = 0.36, \text{ and } \Pr(TT) = q^2 = 0.06$$

The average population risk relative to genotype TT (which is defined to have a risk of one) is:

$$R = 0.50 \times 1.32 + 0.36 \times 1.15 + 0.06 \times 1 = 1.13$$

- 10 Therefore, the risk relative to the general population (RR) for individuals who have one of the following genotypes at this marker is:

$$RR(CC) = 1.32/1.13 = 1.17, RR(CT) = 1.15/1.13 = 1.02, RR(TT) = 1/1.13 = 0.88.$$

Combining the risk from multiple markers:

- 15 When genotypes of many SNP variants are used to estimate the risk for an individual a multiplicative model for risk can generally be assumed. This means that the combined genetic risk relative to the population is calculated as the product of the corresponding estimates for individual markers, e.g. for two markers g1 and g2:

$$RR(g1, g2) = RR(g1)RR(g2)$$

- 20 The underlying assumption is that the risk factors occur and behave independently, i.e. that the joint conditional probabilities can be represented as products:

$$\Pr(A|g1, g2) = \Pr(A|g1)\Pr(A|g2)/\Pr(A) \text{ and } \Pr(g1, g2) = \Pr(g1)\Pr(g2)$$

- 25 Obvious violations to this assumption are markers that are closely spaced on the genome, i.e. in linkage disequilibrium, such that the concurrence of two or more risk alleles is correlated. In such cases, we can use so called haplotype modeling where the odds-ratios are defined for all allele combinations of the correlated SNPs.

As is in most situations where a statistical model is utilized, the model applied is not expected to be exactly true since it is not based on an underlying bio-physical model. However, the multiplicative model has so far been found to fit the data adequately, i.e. no significant

deviations are detected for many common diseases for which many risk variants have been discovered.

As an example, an individual who has the following genotypes at 4 hypothetical markers associated with a particular disease along with the risk relative to the population at each marker:

Marker	Genotype	Calculated risk
M1	CC	1.03
M2	GG	1.30
M3	AG	0.88
M4	TT	1.54

- 5 Combined, the overall risk relative to the population for this individual is: $1.03 \times 1.30 \times 0.88 \times 1.54 = 1.81$. In an analogous fashion, overall risk for any plurality of markers (or haplotypes) may be assessed.

Adjusted life-time risk

- 10 The lifetime risk of an individual is derived by multiplying the overall genetic risk relative to the population with the average life-time risk of the disease in the general population of the same ethnicity and gender and in the region of the individual's geographical origin. As there are usually several epidemiologic studies to choose from when defining the general population risk, we will pick studies that are well-powered for the disease definition that has been used for the
- 15 genetic variants.

For example, for a particular disease, if the overall genetic risk relative to the population is 1.8 for an individual, and if the average life-time risk of the disease for individuals of his demographic is 20%, then the adjusted lifetime risk for him is $20\% \times 1.8 = 36\%$.

- Note that since the average RR for a population is one, this multiplication model provides
- 20 same average adjusted life-time risk of the disease. Furthermore, since the actual life-time risk cannot exceed 100%, there must be an upper limit to the genetic RR.

Risk assessment for breast cancer

- As described herein, certain polymorphic markers and haplotypes comprising such markers are
- 25 found to be useful for risk assessment of breast cancer. Risk assessment can involve the use of the markers for diagnosing a susceptibility to breast cancer. Particular alleles of certain polymorphic markers are found more frequently in individuals with breast cancer, than in individuals without diagnosis of breast cancer. Therefore, these marker alleles have predictive

value for detecting breast cancer, or a susceptibility to breast cancer, in an individual. Tagging markers in linkage disequilibrium with at-risk variants (or protective variants) described herein, can be used as surrogates for these markers (and/or haplotypes). Such surrogate markers can be located within a particular haplotype block or LD block. Such surrogate markers can also

5 sometimes be located outside the physical boundaries of such a haplotype block or LD block, either in close vicinity of the LD block/haplotype block, but possibly also located in a more distant genomic location.

Long-distance LD can for example arise if particular genomic regions (e.g., genes) are in a functional relationship. For example, if two genes encode proteins that play a role in a shared

10 metabolic pathway, then particular variants in one gene may have a direct impact on observed variants for the other gene. Let us consider the case where a variant in one gene leads to increased expression of the gene product. To counteract this effect and preserve overall flux of the particular pathway, this variant may have led to selection of one (or more) variants at a

15 second gene that confers decreased expression levels of that gene. These two genes may be located in different genomic locations, possibly on different chromosomes, but variants within the genes are in apparent LD, not because of their shared physical location within a region of high LD, but rather due to evolutionary forces. Such LD is also contemplated and within scope of the present invention. The skilled person will appreciate that many other scenarios of functional gene-gene interaction are possible, and the particular example discussed here represents only

20 one such possible scenario.

Markers with values of r^2 equal to 1 are perfect surrogates for the at-risk variants (anchor variants), i.e. genotypes for one marker perfectly predicts genotypes for the other. Markers with smaller values of r^2 than 1 can also be surrogates for the at-risk variant, or alternatively

25 represent variants with relative risk values as high as or possibly even higher than the at-risk variant. In certain preferred embodiments, markers with values of r^2 to the at-risk anchor variant are useful surrogate markers. The at-risk variant identified may not be the functional variant itself, but is in this instance in linkage disequilibrium with the true functional variant. The functional variant may be a SNP, but may also for example be a tandem repeat, such as a minisatellite or a microsatellite, a transposable element (e.g., an *Alu* element), or a structural

30 alteration, such as a deletion, insertion or inversion (sometimes also called copy number variations, or CNVs). The present invention encompasses the assessment of such surrogate markers for the markers as disclosed herein. Such markers are annotated, mapped and listed in public databases, as well known to the skilled person, or can alternatively be readily identified by sequencing the region or a part of the region identified by the markers of the present

35 invention in a group of individuals, and identify polymorphisms in the resulting group of sequences. As a consequence, the person skilled in the art can readily and without undue experimentation identify and genotype surrogate markers in linkage disequilibrium with the markers and/or haplotypes as described herein. The tagging or surrogate markers in LD with the at-risk variants detected also have predictive value.

The present invention can in certain embodiments be practiced by assessing a sample comprising genomic DNA from an individual for the presence of certain variants described herein to be associated with breast cancer. Such assessment includes steps of detecting the presence or absence of at least one allele of at least one polymorphic marker, using methods well known to the skilled person and further described herein, and based on the outcome of such assessment, determine whether the individual from whom the sample is derived is at increased or decreased risk (increased or decreased susceptibility) of breast cancer. Alternatively, the invention can be practiced utilizing a dataset comprising information about the genotype status of at least one polymorphic marker described herein to be associated with breast cancer (or markers in linkage disequilibrium with at least one marker shown herein to be associated with breast cancer). In other words, a dataset containing information about such genetic status, for example in the form of genotype counts at a certain polymorphic marker, or a plurality of markers (e.g., an indication of the presence or absence of certain at-risk alleles), or actual genotypes for one or more markers, can be queried for the presence or absence of certain at-risk alleles at certain polymorphic markers shown by the present inventors to be associated with breast cancer. A positive result for a variant (e.g., marker allele) associated with increased risk of breast cancer, as shown herein, is indicative of the individual from which the dataset is derived is at increased susceptibility (increased risk) of breast cancer.

In certain embodiments of the invention, a polymorphic marker is correlated to breast cancer, by referencing genotype data for the polymorphic marker to a database, such as a look-up table that comprises correlations data between at least one allele of the polymorphism and breast cancer. In some embodiments, the table comprises a correlation for one polymorphism. In other embodiments, the table comprises a correlation for a plurality of polymorphisms. In both scenarios, by referencing to a look-up table that gives an indication of a correlation between a marker and breast cancer, a risk for breast cancer, or a susceptibility to breast cancer, can be identified in the individual from whom the sample is derived. In some embodiments, the correlation is reported as a statistical measure. The statistical measure may be reported as risk measure, such as a relative risk (RR), an absolute risk (AR) or an odds ratio (OR).

Risk markers may be useful for risk assessment and diagnostic purposes, either alone or in combination. Results of disease risk assessment based on the markers described herein can also be combined with data for other genetic markers or risk factors for the disease, to establish overall risk. Thus, even in cases where the increase in risk by individual markers is relatively modest, e.g. on the order of 10-30%, the association may have significant implications when combined with other risk markers. Thus, relatively common variants may have significant contribution to the overall risk (Population Attributable Risk is high), or combination of markers can be used to define groups of individual who, based on the combined risk of the markers, is at significant combined risk of developing the disease. For example, combined risk can be assessed based on genotype results for any one of, or combinations of rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922, and rs6001954. Such

combinations can also include other susceptibility markers for breast cancer, such as markers on chromosome 5p12 and chromosome 10q26, e.g., marker rs10941679 and marker rs1219648 (Stacey, S.N. *et al Nat Genet* 40:703-6 (2008)). Alternatively, markers in LD with any one of these markers could be assessed. Other markers known to confer risk of breast cancer can also
 5 be assessed together with the markers described herein, such as markers on chromosome 2q14 (e.g., marker rs4848543 or markers in linkage disequilibrium therewith), 2q35 (e.g., marker rs13387042, or markers in linkage disequilibrium therewith), and chromosome 16 (e.g., marker rs3803662, or markers in linkage disequilibrium therewith) (Stacey, S.N. *et al. Nat Genet* 39:865-9 (2007)).

10 Thus, in certain embodiment of the invention, a plurality of variants (markers and/or haplotypes) is used for overall risk assessment. These variants are in one embodiment selected from the variants as disclosed herein. Other embodiments include the use of the variants of the present invention in combination with other variants known to be useful for diagnosing a susceptibility to breast cancer. In such embodiments, the genotype status of a plurality of markers and/or
 15 haplotypes is determined in an individual, and the status of the individual compared with the population frequency of the associated variants, or the frequency of the variants in clinically healthy subjects, such as age-matched and sex-matched subjects. Methods known in the art, such as multivariate analyses or joint risk analyses, such as those described herein, or other methods known to the skilled person, may subsequently be used to determine the overall risk
 20 conferred based on the genotype status at the multiple loci. Assessment of risk based on such analysis may subsequently be used in the methods, uses and kits of the invention, as described herein.

In a general sense, the methods and kits described herein can be utilized from samples containing nucleic acid material (DNA or RNA) from any source and from any individual, or from
 25 genotype or sequence data derived from such samples. In preferred embodiments, the individual is a human individual. The individual can be an adult, child, or fetus. The nucleic acid source may be any sample comprising nucleic acid material, including biological samples, or a sample comprising nucleic acid material derived there from. The present invention also provides for assessing markers and/or haplotypes in individuals who are members of a target population.
 30 Such a target population is in one embodiment a population or group of individuals at risk of developing the disease, based on other genetic factors, biomarkers, biophysical parameters (e.g., weight, BMD, blood pressure), or general health and/or lifestyle parameters (e.g., history of cancer, history of breast cancer, previous diagnosis of disease, family history of cancer, family history of breast cancer).

35 The invention provides for embodiments that include individuals from specific age subgroups, such as those over the age of 40, over age of 45, or over age of 50, 55, 60, 65, 70, 75, 80, or

85. Other embodiments of the invention pertain to other age groups, such as individuals aged less than 85, such as less than age 80, less than age 75, or less than age 70, 65, 60, 55, 50, 45, 40, 35, or age 30. Other embodiments relate to individuals with age at onset of breast cancer in any of the age ranges described in the above. It is also contemplated that a range of ages may be relevant in certain embodiments, such as age at onset at more than age 45 but less than age 60. Other age ranges are however also contemplated, including all age ranges bracketed by the age values listed in the above. The invention furthermore relates to Individuals of either gender, males or females.

The Icelandic population is a Caucasian population of Northern European ancestry. A large number of studies reporting results of genetic linkage and association in the Icelandic population have been published in the last few years. Many of those studies show replication of variants, originally identified in the Icelandic population as being associating with a particular disease, in other populations (Sulem, P., *et al. Nat Genet* May 17 2009 (Epub ahead of print); Rafnar, T., *et al. Nat Genet* 41:221-7 (2009); Gretarsdottir, S., *et al. Ann Neurol* 64:402-9 (2008); Stacey, S.N., *et al. Nat Genet* 40:1313-18 (2008); Gudbjartsson, D.F., *et al. Nat Genet* 40:886-91 (2008); Styrkarsdottir, U., *et al. N Engl J Med* 358:2355-65 (2008); Thorgeirsson, T., *et al. Nature* 452:638-42 (2008); Gudmundsson, J., *et al. Nat Genet.* 40:281-3 (2008); Stacey, S.N., *et al., Nat Genet.* 39:865-69 (2007); Helgadóttir, A., *et al., Science* 316:1491-93 (2007); Steinthorsdóttir, V., *et al., Nat Genet.* 39:770-75 (2007); Gudmundsson, J., *et al., Nat Genet.* 39:631-37 (2007); Frayling, TM, *Nature Reviews Genet* 8:657-662 (2007); Amundadóttir, L.T., *et al., Nat Genet.* 38:652-58 (2006); Grant, S.F., *et al., Nat Genet.* 38:320-23 (2006)). Thus, genetic findings in the Icelandic population have in general been replicated in other populations, including populations from Africa and Asia.

It is thus believed that the markers described herein to be associated with risk of breast cancer will show similar association in other human populations. Particular embodiments comprising individual human populations are thus also contemplated and within the scope of the invention. Such embodiments relate to human subjects that are from one or more human population including, but not limited to, Caucasian populations, European populations, American populations, Eurasian populations, Asian populations, Central/South Asian populations, East Asian populations, Middle Eastern populations, African populations, Hispanic populations, or Oceanian populations. European populations include, but are not limited to, Swedish, Norwegian, Finnish, Russian, Danish, Icelandic, Irish, Kelt, English, Scottish, Dutch, Belgian, French, German, Spanish, Portuguese, Italian, Polish, Bulgarian, Slavic, Serbian, Bosnian, C Greek and Turkish populations. In one embodiment, the invention relates to individuals of Caucasian origin.

The racial contribution in individual subjects may also be determined by genetic analysis. Genetic analysis of ancestry may be carried out using unlinked microsatellite markers such those set out in Smith *et al. (Am J Hum Genet* 74, 1001-13 (2004)).

In certain embodiments, the invention relates to markers and/or haplotypes identified in specific populations, as described in the above. The person skilled in the art will appreciate that measures of linkage disequilibrium (LD) may give different results when applied to different populations. This is due to different population history of different human populations as well as differential selective pressures that may have led to differences in LD in specific genomic regions. It is also well known to the person skilled in the art that certain markers, e.g. SNP markers, are polymorphic in one population but not in another. The person skilled in the art will however apply the methods available and as thought herein to practice the present invention in any given human population. This may include assessment of polymorphic markers in the LD region of the present invention, so as to identify those markers that give strongest association within the specific population. Thus, the at-risk variants of the present invention may reside on different haplotype background and in different frequencies in various human populations. However, utilizing methods known in the art and the markers of the present invention, the invention can be practiced in any given human population.

15

Models to predict inherited risk for breast cancer

The goal of breast cancer risk assessment is to provide a rational framework for the development of personalized medical management strategies for all women with the aim of increasing survival and quality of life in high-risk women while minimizing costs, unnecessary interventions and anxiety in women at lower risk. Risk prediction models attempt to estimate the risk for breast cancer in an individual who has a given set of congenital risk characteristics (e.g., family history, prior benign breast lesion, previous breast tumor). The breast cancer risk assessment models most commonly employed in clinical practice estimate inherited risk factors by considering family history. The risk estimates are based on the observations of increased risk to individuals with one or more close relatives previously diagnosed with breast cancer. They do not take into account complex pedigree structures. These models have the further disadvantage of not being able to differentiate between carriers and non-carriers of genes with breast cancer predisposing mutations.

More sophisticated risk models have better mechanisms to deal with specific family histories and have an ability to take into account carrier status for BRCA1 and BRCA2 mutations. For example, the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) (Antoniou et al., 2004) takes into account family history based on individual pedigree structures through the pedigree analysis program MENDEL. Information on known BRCA1 and BRCA2 status is also taken into account. The main limitations of the BOADICEA and all other breast cancer risk models currently in use are that they do not incorporate genotypic information from other predisposition genes. Current models depend strongly on family history to act as a surrogate to compensate for the lack of knowledge of non-BRCA genetic

determinants of risk. Therefore the available models are limited to situations where there is a known family history of disease. Lower penetrance breast cancer predisposition genes may be relatively common in the population and may not show such strong tendencies to drive familial clustering as do the BRCA1 and BRCA2 genes. Patients with a relatively high genetic load of predisposition alleles may show little or no family history of disease. There is a need therefore to construct models which incorporate Inherited susceptibility data obtained directly through gene-based testing. In addition to making the models more precise, this will reduce the dependency on family history parameters and assist in the extension of the risk profiling into the wider at-risk population where family history is not such a key factor.

10

Integration of Improved Genetic Risk Models into Clinical Management of Breast Cancer Primary prevention

Clinical primary prevention options currently can be classified as chemopreventative (or hormonal) treatments and prophylactic surgery. Patients identified as high risk can be prescribed long-term courses of chemopreventative therapies. This concept is well accepted in the field of cardiovascular medicine, but is only now beginning to make an impact in clinical oncology. The most widely used oncology chemopreventative is Tamoxifen, a Selective Estrogen Receptor Modulator (SERM). Initially used as an adjuvant therapy directed against breast cancer recurrence, Tamoxifen now has proven efficacy as a breast cancer preventative agent [Cuzick, et al., (2003), Lancet, 361, 296-300][Martino, et al., (2004), Oncologist, 9, 116-25]. The FDA has approved the use of Tamoxifen as a chemopreventative agent in certain high risk women.

Unfortunately, long term Tamoxifen use increases risks for endometrial cancer approximately 2.5-fold, the risk of venous thrombosis approximately 2.0-fold. Risks for pulmonary embolism, stroke, and cataracts are also increased [Cuzick, et al., (2003), Lancet, 361, 296-300]. Accordingly, the benefits in Tamoxifen use for reducing breast cancer incidence may not be easily translated into corresponding decreases in overall mortality. Another SERM called Raloxifene may be more efficacious in a preventative mode, and does not carry the same risk for endometrial cancer. However risk for thrombosis is still elevated in patients treated long-term with Raloxifene [Cuzick, et al., (2003), Lancet, 361, 296-300; Martino, et al., (2004), Oncologist, 9, 116-25]. Moreover, both Tamoxifen and Raloxifene have quality of life issues associated with them. To make a rational risk:benefit analysis of SERM therapy in a chemopreventative mode, there is a clinical need to identify individuals who are most at risk for breast cancer. Given that a substantial proportion of risk for breast cancer is genetic, there is a clear clinical need for genetic tests to quantify individuals' risks in this context. One can anticipate similar issues arising from any future cancer chemo-preventative therapies that may become available, such as the aromatase inhibitors. Moreover, as chemopreventative therapies

become safer, there is an increased need to identify patients who are genetically predisposed, but do not have massively elevated risks associated with BRCA1 & 2 mutation carriers.

Patients who are identified as being at high risk for breast cancer are considered for prophylactic surgery; either bilateral mastectomy or oophorectomy or both. Clearly such drastic treatments are recommended only for patients who are perceived to be at extremely high risk. In practice, such risks can currently be identified only in individuals who carry mutations in BRCA1, BRCA2 or genes known to be involved in rare breast cancer predisposition syndromes like p53 in Li-Fraumeni Syndrome, PTEN in Cowden's Syndrome.

Estimates of the penetrance of BRCA1 and BRCA2 mutations tend to be higher when they are derived from multiple-case families than when they are derived from population-based estimates. This is because different mutation-carrying families exhibit different penetrances for breast cancer (see [Thorlacius, et al., (1997), Am J Hum Genet, 60, 1079-84] for example). One of the major factors contributing to this variation is the action of as yet unknown predisposition genes whose effects modify the penetrance of BRCA1 and BRCA2 mutations. Therefore the absolute risk to an individual who carries a mutation in the BRCA1 or BRCA2 genes cannot be accurately quantified in the absence of knowledge of the existence and action of modifying genes. Since the treatment options for BRCA1 and BRCA2 carriers can be severe, it is important in this context to quantify the risks to individual BRCA carriers with the greatest accuracy possible. There is a need, therefore, to identify predisposition genes whose effects modify the penetrance of breast cancer in BRCA1 and BRCA2 carriers and to develop improved risk assessment models based on these genes.

Furthermore, there are individuals who are perceived to be at very high risk for breast cancer, perhaps because of a strong family history of breast cancer, but in whom no mutations in known predisposition genes can be identified. Consideration of prophylactic surgery is difficult in such cases because one cannot test the individual to discover whether or not she has inherited a high penetrance predisposition gene. Accordingly, the individual's risk cannot be assessed accurately. There is a clear clinical need, therefore, to identify any high penetrance predisposition genes that remain undiscovered and to develop associated genetic tests for use in primary prevention strategies. Such genes may for example be the genes disclosed herein to be associated with risk of breast cancer. Although the variants shown herein to be associated with risk of breast cancer are fairly common, and conferring a relatively low risk of breast cancer, it is quite possible that higher risk variants exist within one or more of these genes. It is thus contemplated that high-risk genetic variants within, or associated with, one or more of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and/or *TNRC6B* genes could be useful for determining whether an individual is a carrier of a high risk (and high penetrance) genetic factor for breast cancer.

Early Diagnosis

Clinical screening for breast cancer in most western countries consists of periodic clinical breast examination (CBE) and X-ray mammography. There is good evidence to indicate that CBE has little added benefit when used in the context of a good mammographic screening program. In the United Kingdom, women between the ages of 50 and 70 are invited to undergo screening mammography every three years. The situation in the United States varies depending on healthcare provider, however the American Cancer Society recommends annual mammographic screening from age 40. Mammographic screening has proven effectiveness in reducing mortality amongst screened women over the age of 50.

- 10 It is unlikely that genetic testing would ever be employed as a means of reducing access to existing mammographic screening programs. However, mammographic screening is not without shortcomings and it is conceivable that genetic testing should be used to select people for augmented screening programs. One of the drawbacks of mammographic screening is that it has thus far not been possible to demonstrate a significant effect on improved survival for women screened under 50 years of age.

- One reason that mammography is less effective in women under 50 may be that the density of breast tissue is higher in younger women, making mammographic detection of tumors more difficult. However, breast cancers in predisposed individuals tend to occur at early ages groups and there is a clear association between high breast density and breast cancer risk. Therefore there is a problem with simple increases in mammographic screening for individuals with high predisposition because they would be managed by a technique that performs sub-optimally in the group at highest risk. Recent studies have shown that contrast-enhanced magnetic resonance imaging (CE-MRI) is more sensitive and detects tumors at an earlier stage in this high-risk group than mammographic screening does [Warner, et al., (2004), *Jama*, 292, 1317-25; Leach, et al., (2005), *Lancet*, 365, 1769-78]. CE-MRI strategies work particularly well when used in combination with routine X-ray mammography [Leach, et al., (2005), *Lancet*, 365, 1769-78]. Because CE-MRI requires specialist centers that incur high costs, screening of under-50's must be restricted to those individuals at the highest risk. Present CE-MRI trials restrict entry to those individuals with BRCA1, BRCA2 or p53 mutations or very strong family histories of disease.
- 30 The extension of this screening modality to a wider range of high-risk patients would be greatly assisted by the provision of gene-based risk profiling tools.

- There is good evidence to support the notion that early-onset breast cancers and cancers occurring in genetically predisposed women grow faster than cancers in older, less strongly predisposed women. This comes from observations of higher rates of interval cancers in younger women, that is, cancers that arise in the intervals between screening visits in a well-screened population are higher amongst younger women. Therefore there are suggestions that screening intervals, by whatever method, should be reduced for younger women. There is a paradox here

In that more frequent screening using more expensive methodologies seems to be required for an age group in which the overall rates of breast cancer are comparatively low. There is a clear clinical need here to identify those young individuals who are most strongly predisposed to develop the disease early, and channel them into more expensive and extensive screening regimes. The variants disclosed herein to confer risk of breast cancer can be useful for identification of individuals who are at particularly high risk of developing breast cancer. Such individuals are likely to most benefit from early and aggressive screening programs, so as to maximizing the likelihood of early identification of the cancer.

10 *Treatment*

Currently, primary breast cancer is treated by surgery, adjuvant chemotherapy, radiotherapy, followed by long term hormonal therapy. Often combinations of three or four therapies are used.

Breast cancer patients with the same stage of disease can have very different responses to adjuvant chemotherapy resulting in a broad variation in overall treatment outcomes. Consensus guidelines (the St Gallen and NIH criteria) have been developed for determining the eligibility of breast cancer patients for adjuvant chemotherapy treatment. However, even the strongest clinical and histological predictors of metastasis fail to predict accurately the clinical responses of breast tumors [Goldhirsch, et al., (1998), *J Natl Cancer Inst*, 90, 1601-8; Eifel, et al., (2001), *J Natl Cancer Inst*, 93, 979-89]. Chemotherapy or hormonal therapy reduces the risk of metastasis only by approximately 1/3, however 70-80% of patients receiving this treatment would have survived without it. Therefore the majority of breast cancer patients are currently offered treatment that is either ineffective or unnecessary. There is a clear clinical need for improvements in the development of prognostic measures which will allow clinicians to tailor treatments more appropriately to those who will best benefit. It is reasonable to expect that profiling individuals for genetic predisposition may reveal information relevant to their treatment outcome and thereby aid in rational treatment planning. The markers of the present invention, conferring risk of breast cancer, are contemplated to be useful in this context.

Several previous studies exemplify this concept: Breast cancer patients who are BRCA mutation carriers appear to show better clinical response rates and survival when treated with adjuvant chemotherapies [Chappuis, et al., (2002), *J Med Genet*, 39, 608-10; Goffin, et al., (2003), *Cancer*, 97, 527-36]. BRCA mutation carriers demonstrate improved responses to platinum chemotherapy for ovarian cancer than non-carriers [Cass, et al., (2003), *Cancer*, 97, 2187-95]. Similar considerations may apply to predisposed patients in whom the genes involved are not known. For example, infiltrating lobular breast carcinoma (ILBC) is known to have a strong familial component but the genetic variants involved have not yet been identified. Patients with

ILBC demonstrate poorer responses to common chemotherapy regimes [Mathieu, et al., (2004), Eur J Cancer, 40, 342-51].

Genetic predisposition models may not only aid in the individualization of treatment strategies, but may play an integral role in the design of these strategies. For example, BRCA1 and BRCA2 mutant tumor cells have been found to be profoundly sensitive to poly (ADP-ribose) polymerase (PARP) inhibitors as a result of their defective DNA repair pathway [Farmer, et al., (2005), Nature, 434, 917-21]. This has stimulated development of small molecule drugs targeted on PARP with a view to their use specifically in BRCA carrier patients. From this example it is clear that knowledge of genetic predisposition may identify drug targets that lead to the development of personalized chemotherapy regimes to be used in combination with genetic risk profiling. Similarly, the markers of the present invention may aid in the identification of novel drugs that target, for example, one or more of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and/or *TNRC6B* genes.

Cancer chemotherapy has well known, dose-limiting side effects on normal tissues particularly the highly proliferative hemopoietic and gut epithelial cell compartments. It can be anticipated that genetically-based individual differences exist in sensitivities of normal tissues to cytotoxic drugs. An understanding of these factors might aid in rational treatment planning and in the development of drugs designed to protect normal tissues from the adverse effects of chemotherapy.

Genetic profiling may also contribute to improved radiotherapy approaches: Within groups of breast cancer patients undergoing standard radiotherapy regimes, a proportion of patients will experience adverse reactions to doses of radiation that are normally tolerated. Acute reactions include erythema, moist desquamation, edema and radiation pneumonitis. Long term reactions including telangiectasia, edema, pulmonary fibrosis and breast fibrosis may arise many years after radiotherapy. Both acute and long-term reactions are considerable sources of morbidity and can be fatal. In one study, 87% of patients were found to have some adverse side effects to radiotherapy while 11% had serious adverse reactions (LENT/SOMA Grade 3-4); [Hoeller, et al., (2003), Int J Radiat Oncol Biol Phys, 55, 1013-8]. The probability of experiencing an adverse reaction to radiotherapy is due primarily to constitutive individual differences in normal tissue reactions and there is a suspicion that these have a strong genetic component. Several of the known breast cancer predisposition genes (e.g. BRCA1, BRCA2, ATM) affect pathways of DNA double strand break repair. DNA double strand breaks are the primary cytotoxic lesion induced by radiotherapy. This has led to concern that individuals who are genetically predisposed to breast cancer through carriage of variants in genes belonging to these pathways might also be at higher risk of suffering excessive normal tissue damage from radiotherapy. It is contemplated that the genetic variants described herein to confer risk of breast cancer, for example through one or more of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and/or *TNRC6B* genes, may be useful for identifying individuals at particular risk of adverse reaction to radiotherapy.

The existence of constitutively radiosensitive individuals in the population means that radiotherapy dose rates for the majority of the patient population must be restricted, in order to keep the frequency of adverse reactions to an acceptable level. There is a clinical need, therefore, for reliable tests that can identify individuals who are at elevated risk for adverse reactions to radiotherapy. Such tests would indicate conservative or alternative treatments for individuals who are radiosensitive, while permitting escalation of radiotherapeutic doses for the majority of patients who are relatively radioresistant. It has been estimated that the dose escalations made possible by a test to triage breast cancer patients simply into radiosensitive, intermediate and radioresistant categories would result in an approximately 35% increase in local tumor control and consequent improvements in survival rates [Burnet, et al., (1996), Clin Oncol (R Coll Radiol), 8, 25-34].

Exposure to ionizing radiation is a proven factor contributing to oncogenesis in the breast [Dumitrescu and Cotarla, (2005), J Cell Mol Med, 9, 208-21]. Known breast cancer predisposition genes encode pathway components of the cellular response to radiation-induced DNA damage [Narod and Foulkes, (2004), Nat Rev Cancer, 4, 665-76]. Accordingly, there is concern that the risk for second primary breast tumors may be increased by irradiation of normal tissues within the radiotherapy field. There does not appear to be any measurable increased risk for BRCA carriers from radiotherapy, however their risk for second primary tumors is already exceptionally high. There is evidence to suggest that risk for second primary tumors is increased in carriers in breast cancer predisposing alleles of the ATM and CHEK2 genes who are treated with radiotherapy [Bernstein, et al., (2004), Breast Cancer Res, 6, R199-214; Broeks, et al., (2004), Breast Cancer Res Treat, 83, 91-3]. It is expected that the risk of second primary tumors from radiotherapy (and, possibly, from intensive mammographic screening) will be better defined by obtaining accurate genetic risk profiles from patients during the treatment planning stage.

Secondary Prevention

Approximately 30% of patients who are diagnosed with a stage 1 or 2 breast cancer will experience either a loco-regional or distant metastatic recurrence of their original tumor. Patients who have had a primary breast cancer are also at greatly increased risk for being diagnosed with a second primary tumor, either in the contralateral breast or in the ipsilateral breast when breast-conserving surgery has been carried out. Secondary prevention refers to methods used to prevent recurrences or second primary tumors from developing. Methods currently in use comprise; long-term treatment with Tamoxifen or another SERM either alone or alternated with an aromatase inhibitor, risk-reducing mastectomy of the contralateral breast, and risk-reducing oophorectomy (in patients who are at risk for familial breast-ovarian cancer). Considerations regarding the use of Tamoxifen have been discussed above. With risk-reducing

surgical options, it is clear that the risk needs to be quantified as well as possible in order to make an informed cost versus benefit analysis.

There are some indications that patients with known genetic predispositions to breast cancer fare worse than the majority of patients. Patients carrying the CHEK2 gene 1100delC variant have an estimated 2.8-fold increased risk of distant metastasis and a 3.9-fold increased risk of disease recurrence compared to non-carriers [de Bock, et al., (2004), *J Med Genet*, 41, 731-5]. Patients with BRCA1 node-negative tumors have a greater risk of metastasis than similar patients who do not carry a BRCA1 mutation [Goffin, et al., (2003), *Cancer*, 97, 527-36; Moller, et al., (2002), *Int J Cancer*, 101, 555-9; Eerola, et al., (2001), *Int J Cancer*, 93, 368-72]. Genetic profiling can therefore be used to help assess the risk of local recurrence and metastasis, thereby guiding the choice of secondary preventative treatment. Genetic profiling based on the variants described herein may be useful in this context. In certain embodiments, such profiling may be based on one or more of the variants described herein. In other embodiments, such profiling may include one or several other known genetic risk factors for breast cancer. Such risk factors may be well established high-penetrant risk factors, or they may be one or more of the common, lower penetrance risk factors that have been previously described (e.g., markers rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954 or markers in linkage disequilibrium therewith, e.g. markers in Table 4).

In general, patients with a primary tumor diagnosis are at risk for second primary tumors at a constant annual incidence of 0.7% [Peto and Mack, (2000), *Nat Genet*, 26, 411-4]. Patients with BRCA mutations are at significantly greater risks for second primary tumors than most breast cancer patients, with absolute risks in the range 40-60% [Easton, (1999), *Breast Cancer Res*, 1, 14-7]. Carriers of BRCA mutations have a greatly increased risk for second primary tumors [Stacey, et al., (2006), *PLoS Med*, 3, e217; Metcalfe, et al., (2004), *J Clin Oncol*, 22, 2328-35]. Patients with mutations in the CHEK2 gene have an estimated 5.7-fold increased risk of contralateral breast cancer [de Bock, et al., (2004), *J Med Genet*, 41, 731-5]. Carriers of the BARD1 Cys557Ser variant are 2.7 fold more likely to be diagnosed with a second primary tumor [Stacey, et al., (2006), *PLoS Med*, 3, e217]. Genetic risk profiling can be used to assess the risk of second primary tumors in patients and will inform decisions on how aggressive the preventative measures should be.

METHODS

Methods for disease risk assessment and risk management are described herein and are encompassed by the invention. The invention also encompasses methods of assessing an individual for probability of response to a therapeutic agents, methods for predicting the effectiveness of a therapeutic agents, nucleic acids, polypeptides and antibodies and computer-

implemented functions. Kits for use in the various methods presented herein are also encompassed by the invention.

Diagnostic and screening methods

5 In certain embodiments, the present invention pertains to methods of diagnosing, or aiding in the diagnosis of, breast cancer or a susceptibility to breast cancer, by detecting particular alleles at genetic markers that appear more frequently in breast cancer subjects or subjects who are susceptible to breast cancer. In particular embodiments, the invention is a method of determining a susceptibility to breast cancer by detecting at least one allele of at least one
10 polymorphic marker (e.g., the markers described herein). In other embodiments, the invention relates to a method of diagnosing a susceptibility to breast cancer by detecting at least one allele of at least one polymorphic marker. The present invention describes methods whereby detection of particular alleles of particular markers or haplotypes is indicative of a susceptibility to breast cancer. Such prognostic or predictive assays can also be used to determine prophylactic
15 treatment of a subject prior to the onset of symptoms associated with breast cancer.

The present invention pertains in some embodiments to methods of clinical applications of diagnosis, e.g., diagnosis performed by a medical professional. In other embodiments, the invention pertains to methods of diagnosis or determination of a susceptibility performed by a layman. The layman can be the customer of a genotyping service. The layman may also be a
20 genotype service provider, who performs genotype analysis on a DNA sample from an individual, in order to provide service related to genetic risk factors for particular traits or diseases, based on the genotype status of the individual (i.e., the customer). Recent technological advances in genotyping technologies, including high-throughput genotyping of SNP markers, such as Molecular Inversion Probe array technology (e.g., Affymetrix GeneChip), and BeadArray
25 Technologies (e.g., Illumina GoldenGate and Infinium assays) have made it possible for individuals to have their own genome assessed for up to one million SNPs simultaneously, at relatively little cost. The resulting genotype information, which can be made available to the individual, can be compared to information about disease or trait risk associated with various SNPs, including information from public literature and scientific publications. The diagnostic
30 application of disease-associated alleles as described herein, can thus for example be performed by the individual, through analysis of his/her genotype data, by a health professional based on results of a clinical test, or by a third party, including the genotype service provider. The third party may also be service provider who interprets genotype information from the customer to provide service related to specific genetic risk factors, including the genetic markers described
35 herein. In other words, the diagnosis or determination of a susceptibility of genetic risk can be made by health professionals, genetic counselors, third parties providing genotyping service, third parties providing risk assessment service or by the layman (e.g., the individual), based on

information about the genotype status of an individual and knowledge about the risk conferred by particular genetic risk factors (e.g., particular SNPs). In the present context, the term "diagnosing", "diagnose a susceptibility" and "determine a susceptibility" is meant to refer to any available diagnostic method, including those mentioned above.

- 5 In certain embodiments, a sample containing genomic DNA from an individual is collected. Such sample can for example be a buccal swab, a saliva sample, a blood sample, or other suitable samples containing genomic DNA, as described further herein. The genomic DNA is then analyzed using any common technique available to the skilled person, such as high-throughput array technologies. Results from such genotyping are stored in a convenient data storage unit, such as a data carrier, including computer databases, data storage disks, or by other convenient data storage means. In certain embodiments, the computer database is an object database, a relational database or a post-relational database. The genotype data is subsequently analyzed for the presence of certain variants known to be susceptibility variants for a particular human conditions, such as the genetic variants described herein. Genotype data can be retrieved from the data storage unit using any convenient data query method. Calculating risk conferred by a particular genotype for the individual can be based on comparing the genotype of the individual to previously determined risk (expressed as a relative risk (RR) or and odds ratio (OR), for example) for the genotype, for example for an heterozygous carrier of an at-risk variant for a particular disease or trait. The calculated risk for the individual can be the relative risk for a person, or for a specific genotype of a person, compared to the average population with matched gender and ethnicity. The average population risk can be expressed as a weighted average of the risks of different genotypes, using results from a reference population, and the appropriate calculations to calculate the risk of a genotype group relative to the population can then be performed. Alternatively, the risk for an individual is based on a comparison of particular genotypes, for example heterozygous carriers of an at-risk allele of a marker compared with non-carriers of the at-risk allele. Using the population average may in certain embodiments be more convenient, since it provides a measure which is easy to interpret for the user, i.e. a measure that gives the risk for the individual, based on his/her genotype, compared with the average in the population. The calculated risk estimated can be made available to the customer via a website, preferably a secure website.

- In certain embodiments, a service provider will include in the provided service all of the steps of isolating genomic DNA from a sample provided by the customer, performing genotyping of the isolated DNA, calculating genetic risk based on the genotype data, and report the risk to the customer. In some other embodiments, the service provider will include in the service the interpretation of genotype data for the individual, i.e., risk estimates for particular genetic variants based on the genotype data for the individual. In some other embodiments, the service provider may include service that includes genotyping service and interpretation of the genotype data, starting from a sample of isolated DNA from the individual (the customer).

Overall risk for multiple risk variants can be performed using standard methodology. For example, assuming a multiplicative model, i.e. assuming that the risk of individual risk variants multiply to establish the overall effect, allows for a straight-forward calculation of the overall risk for multiple markers.

- 5 In addition, in certain other embodiments, the present invention pertains to methods of diagnosing, or aiding in the diagnosis of, a decreased susceptibility to breast cancer, by detecting particular genetic marker alleles or haplotypes that appear less frequently in breast cancer patients than in individual not diagnosed with breast cancer or in the general population.

- As described and exemplified herein, particular marker alleles or haplotypes are associated with risk of breast cancer. In one embodiment, the marker allele or haplotype is one that confers a significant risk or susceptibility to breast cancer. In another embodiment, the invention relates to a method of determining a susceptibility to breast cancer in a human individual, the method comprising determining the presence or absence of at least one allele of at least one polymorphic marker in a nucleic acid sample obtained from the individual. In another embodiment, the invention pertains to methods of determining a susceptibility to breast cancer in a human individual, by screening for at least one marker allele or haplotype as described herein. In another embodiment, the marker allele or haplotype is more frequently present in a subject having, or who is susceptible to, breast cancer (affected), as compared to the frequency of its presence in a healthy subject (control, such as population controls). In certain embodiments, the significance of association of the at least one marker allele or haplotype is characterized by a p value < 0.05 . In other embodiments, the significance of association is characterized by smaller p-values, such as < 0.01 , < 0.001 , < 0.0001 , < 0.00001 , < 0.000001 , < 0.0000001 , < 0.00000001 or < 0.000000001 .

- In these embodiments, determination of the presence of the at least one marker allele or haplotype is indicative of a susceptibility to breast cancer. These diagnostic methods involve detecting the presence or absence of at least one marker allele or haplotype that is associated with breast cancer. The detection of the particular genetic marker alleles that make up the particular haplotypes can be performed by a variety of methods described herein and/or known in the art. For example, genetic markers can be detected at the nucleic acid level (e.g., by direct nucleotide sequencing or by other means known to the skilled in the art) or at the amino acid level if the genetic marker affects the coding sequence of a protein encoded by a breast cancer-associated nucleic acid (e.g., by protein sequencing or by immunoassays using antibodies that recognize such a protein). The marker alleles or haplotypes correspond to fragments of a genomic DNA sequence associated with breast cancer. Such fragments encompass the DNA sequence of the polymorphic marker or haplotype in question, but may also include DNA segments in strong LD (linkage disequilibrium) with the marker or haplotype. In one embodiment, such segments comprises segments in LD with the marker or haplotype (as determined by a value of r^2 greater than 0.1 and/or $|D'| > 0.8$).

In one embodiment, diagnosis of a susceptibility to breast cancer can be accomplished using hybridization methods (see Current Protocols in Molecular Biology, Ausubel, F. *et al.*, eds., John Wiley & Sons, including all supplements). A biological sample from a test subject or individual (a "test sample") of genomic DNA, RNA, or cDNA is obtained from a subject suspected of having, being susceptible to, or predisposed for breast cancer (the "test subject"). The subject can be an adult, child, or fetus. The test sample can be from any source that contains genomic DNA, such as a blood sample, sample of amniotic fluid, sample of cerebrospinal fluid, or tissue sample from skin, muscle, buccal or conjunctival mucosa, placenta, gastrointestinal tract or other organs. A test sample of DNA from fetal cells or tissue can be obtained by appropriate methods, such as by amniocentesis or chorionic villus sampling. The DNA, RNA, or cDNA sample is then examined. The presence of a specific marker allele can be indicated by sequence-specific hybridization of a nucleic acid probe specific for the particular allele. The presence of more than one specific marker allele or a specific haplotype can be indicated by using several sequence-specific nucleic acid probes, each being specific for a particular allele. In one embodiment, a haplotype can be indicated by a single nucleic acid probe that is specific for the specific haplotype (i.e., hybridizes specifically to a DNA strand comprising the specific marker alleles characteristic of the haplotype). A sequence-specific probe can be directed to hybridize to genomic DNA, RNA, or cDNA. A "nucleic acid probe", as used herein, can be a DNA probe or an RNA probe that hybridizes to a complementary sequence. One of skill in the art would know how to design such a probe so that sequence specific hybridization will occur only if a particular allele is present in a genomic sequence from a test sample. The invention can also be reduced to practice using any convenient genotyping method, including commercially available technologies and methods for genotyping particular polymorphic markers.

To determine a susceptibility to breast cancer, a hybridization sample can be formed by contacting the test sample, such as a genomic DNA sample, with at least one nucleic acid probe. A non-limiting example of a probe for detecting mRNA or genomic DNA is a labeled nucleic acid probe that is capable of hybridizing to mRNA or genomic DNA sequences described herein. The nucleic acid probe can be, for example, a full-length nucleic acid molecule, or a portion thereof, such as an oligonucleotide of at least 15, 30, 50, 100, 250 or 500 nucleotides in length that is sufficient to specifically hybridize under stringent conditions to appropriate mRNA or genomic DNA. In certain embodiments, the oligonucleotide is from about 15 to about 100 nucleotides in length. In certain other embodiments, the oligonucleotide is from about 20 to about 50 nucleotides in length. The nucleic acid probe can comprise all or a portion of a nucleotide sequence of any one of LD block C09, LD block C10A, LD block 10B, LD block C11A, LD block C11B, LD block C14, LD block C18, LD block C22A, and LD block C22B as defined herein; alternatively, the nucleic acid probe can comprise all or a portion of a nucleotide sequence comprising the markers set forth in Tables 1 and 4 herein (SEQ ID NO:1-562), or a nucleotide sequence comprising any one of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes, as described herein, optionally comprising at least one allele of a marker described herein, or at least one haplotype described herein, or the probe can be the complementary

sequence of such a sequence. In a particular embodiment, the nucleic acid probe is a portion of the nucleotide sequence of comprising the markers listed in any one of Tables 1, 2, 3 and 4 (SEQ ID NO:1-562), or a nucleotide sequence comprising any one of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes or fragments thereof, as described herein, optionally comprising at least one allele of a marker described herein, or at least one allele of one polymorphic marker or haplotype comprising at least one polymorphic marker described herein; or the probe can be the complementary sequence of such a sequence. Other suitable probes for use in the diagnostic assays of the invention are described herein. Hybridization can be performed by methods well known to the person skilled in the art (see, e.g., Current Protocols in Molecular Biology, Ausubel, F. *et al.*, eds., John Wiley & Sons, including all supplements). In one embodiment, hybridization refers to specific hybridization, i.e., hybridization with no mismatches (exact hybridization). In one embodiment, the hybridization conditions for specific hybridization are high stringency.

Specific hybridization, if present, is detected using standard methods. If specific hybridization occurs between the nucleic acid probe and the nucleic acid in the test sample, then the sample contains the allele that is complementary to the nucleotide that is present in the nucleic acid probe. The process can be repeated for any markers of the present invention, or markers that make up a haplotype of the present invention, or multiple probes can be used concurrently to detect more than one marker alleles at a time. It is also possible to design a single probe containing more than one marker alleles of a particular haplotype (e.g., a probe containing alleles complementary to 2, 3, 4, 5 or all of the markers that make up a particular haplotype). Detection of the particular markers of the haplotype in the sample is indicative that the source of the sample has the particular haplotype (e.g., a haplotype) and therefore is susceptible to breast cancer.

In one preferred embodiment, a method utilizing a detection oligonucleotide probe comprising a fluorescent moiety or group at its 3' terminus and a quencher at its 5' terminus, and an enhancer oligonucleotide, is employed, as described by Kutyavin *et al.* (*Nucleic Acid Res.* 34:e128 (2006)). The fluorescent moiety can be Gig Harbor Green or Yakima Yellow, or other suitable fluorescent moieties. The detection probe is designed to hybridize to a short nucleotide sequence that includes the SNP polymorphism to be detected. Preferably, the SNP is anywhere from the terminal residue to -6 residues from the 3' end of the detection probe. The enhancer is a short oligonucleotide probe which hybridizes to the DNA template 3' relative to the detection probe. The probes are designed such that a single nucleotide gap exists between the detection probe and the enhancer nucleotide probe when both are bound to the template. The gap creates a synthetic abasic site that is recognized by an endonuclease, such as Endonuclease IV. The enzyme cleaves the dye off the fully complementary detection probe, but cannot cleave a detection probe containing a mismatch. Thus, by measuring the fluorescence of the released fluorescent moiety, assessment of the presence of a particular allele defined by nucleotide sequence of the detection probe can be performed.

The detection probe can be of any suitable size, although preferably the probe is relatively short. In one embodiment, the probe is from 5-100 nucleotides in length. In another embodiment, the probe is from 10-50 nucleotides in length, and in another embodiment, the probe is from 12-30 nucleotides in length. Other lengths of the probe are possible and within scope of the skill of the average person skilled in the art.

In a preferred embodiment, the DNA template containing the SNP polymorphism is amplified by Polymerase Chain Reaction (PCR) prior to detection. In such an embodiment, the amplified DNA serves as the template for the detection probe and the enhancer probe.

Certain embodiments of the detection probe, the enhancer probe, and/or the primers used for amplification of the template by PCR include the use of modified bases, including modified A and modified G. The use of modified bases can be useful for adjusting the melting temperature of the nucleotide molecule (probe and/or primer) to the template DNA, for example for increasing the melting temperature in regions containing a low percentage of G or C bases, in which modified A with the capability of forming three hydrogen bonds to its complementary T can be used, or for decreasing the melting temperature in regions containing a high percentage of G or C bases, for example by using modified G bases that form only two hydrogen bonds to their complementary C base in a double stranded DNA molecule. In a preferred embodiment, modified bases are used in the design of the detection nucleotide probe. Any modified base known to the skilled person can be selected in these methods, and the selection of suitable bases is well within the scope of the skilled person based on the teachings herein and known bases available from commercial sources as known to the skilled person.

Additionally, or alternatively, a peptide nucleic acid (PNA) probe can be used in addition to, instead of, a nucleic acid probe in the hybridization methods described herein. A PNA is a DNA mimic having a peptide-like, inorganic backbone, such as N-(2-aminoethyl)glycine units, with an organic base (A, G, C, T or U) attached to the glycine nitrogen via a methylene carbonyl linker (see, for example, Nielsen, P., *et al.*, *Bioconjug. Chem.* 5:3-7 (1994)). The PNA probe can be designed to specifically hybridize to a molecule in a sample suspected of containing one or more of the marker alleles or haplotypes that are associated with breast cancer. In one embodiment of the invention, a test sample containing genomic DNA obtained from the subject is collected and the polymerase chain reaction (PCR) is used to amplify a fragment comprising one or more markers or haplotypes of the present invention. As described herein, identification of a particular marker allele or haplotype associated with breast cancer can be accomplished using a variety of methods (e.g., sequence analysis, analysis by restriction digestion, specific hybridization, single stranded conformation polymorphism assays (SSCP), electrophoretic analysis, etc.). In another embodiment, diagnosis is accomplished by expression analysis, for example by using quantitative PCR (kinetic thermal cycling). This technique can, for example, utilize commercially available technologies, such as TaqMan® (Applied Biosystems, Foster City, CA). The technique can assess the presence of an alteration in the expression or composition of

a polypeptide or splicing variant(s) that is encoded by a nucleic acid associated with breast cancer. Further, the expression of the variant(s) can be quantified as physically or functionally different.

In another method of the invention, analysis by restriction digestion can be used to detect a particular allele if the allele results in the creation or elimination of a restriction site relative to a reference sequence. Restriction fragment length polymorphism (RFLP) analysis can be conducted, e.g., as described in Current Protocols in Molecular Biology, *supra*. The digestion pattern of the relevant DNA fragment indicates the presence or absence of the particular allele in the sample.

Sequence analysis can also be used to detect specific alleles or haplotypes. Therefore, in one embodiment, determination of the presence or absence of a particular marker alleles or haplotypes comprises sequence analysis of a test sample of DNA or RNA obtained from a subject or individual. PCR or other appropriate methods can be used to amplify a portion of a nucleic acid associated with breast cancer, and the presence of a specific allele can then be detected directly by sequencing the polymorphic site (or multiple polymorphic sites in a haplotype) of the genomic DNA in the sample.

In another embodiment, arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from a subject, can be used to identify particular alleles at polymorphic sites. For example, an oligonucleotide array can be used. Oligonucleotide arrays typically comprise a plurality of different oligonucleotide probes that are coupled to a surface of a substrate in different known locations. These arrays can generally be produced using mechanical synthesis methods or light directed synthesis methods that incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis methods, or by other methods known to the person skilled in the art (see, e.g., Bier, F.F., *et al. Adv Biochem Eng Biotechnol* 109:433-53 (2008); Hoheisel, J.D., *Nat Rev Genet* 7:200-10 (2006); Fan, J.B., *et al. Methods Enzymol* 410:57-73 (2006); Raquousis, J. & Elvidge, G., *Expert Rev Mol Diagn* 6:145-52 (2006); Mockler, T.C., *et al Genomics* 85:1-15 (2005), and references cited therein, the entire teachings of each of which are incorporated by reference herein). Many additional descriptions of the preparation and use of oligonucleotide arrays for detection of polymorphisms can be found, for example, in US 6,858,394, US 6,429,027, US 5,445,934, US 5,700,637, US 5,744,305, US 5,945,334, US 6,054,270, US 6,300,063, US 6,733,977, US 7,364,858, EP 619 321, and EP 373 203, the entire teachings of which are incorporated by reference herein.

Other methods of nucleic acid analysis that are available to those skilled in the art can be used to detect a particular allele at a polymorphic site associated with breast cancer. Representative methods include, for example, direct manual sequencing (Church and Gilbert, *Proc. Natl. Acad. Sci. USA*, 81: 1991-1995 (1988); Sanger, F., *et al., Proc. Natl. Acad. Sci. USA*, 74:5463-5467 (1977); Beavis, *et al.*, U.S. Patent No. 5,288,644); automated fluorescent sequencing; single-

stranded conformation polymorphism assays (SSCP); clamped denaturing gel electrophoresis (CDGE); denaturing gradient gel electrophoresis (DGGE) (Sheffield, V., *et al.*, *Proc. Natl. Acad. Sci. USA*, 86:232-236 (1989)), mobility shift analysis (Orita, M., *et al.*, *Proc. Natl. Acad. Sci. USA*, 86:2766-2770 (1989)), restriction enzyme analysis (Flavell, R., *et al.*, *Cell*, 15:25-41 (1978); Geever, R., *et al.*, *Proc. Natl. Acad. Sci. USA*, 78:5081-5085 (1981)); heteroduplex analysis; chemical mismatch cleavage (CMC) (Cotton, R., *et al.*, *Proc. Natl. Acad. Sci. USA*, 85:4397-4401 (1988)); RNase protection assays (Myers, R., *et al.*, *Science*, 230:1242-1246 (1985)); use of polypeptides that recognize nucleotide mismatches, such as *E. coli* mutS protein; and allele-specific PCR.

10 In another embodiment of the invention, determination of a susceptibility to breast cancer can be made by examining expression and/or composition of a polypeptide encoded by a nucleic acid associated with breast cancer in those instances where the genetic marker(s) or haplotype(s) of the present invention result in a change in the composition or expression of the polypeptide. Thus, diagnosis of a susceptibility to breast cancer can be made by examining expression and/or
15 composition of one of these polypeptides, or another polypeptide encoded by a nucleic acid associated with breast cancer, in those instances where the genetic marker or haplotype of the present invention results in a change in the composition or expression of the polypeptide (e.g., one or more of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes). The markers described herein that show association to breast cancer may also affect expression of nearby
20 genes. It is well known that regulatory element affecting gene expression may be located far away, even as far as tenths or hundreds of kilobases away, from the promoter region of a gene. By assaying for the presence or absence of at least one allele of at least one polymorphic marker of the present invention, it is thus possible to assess the expression level of such nearby genes. Possible mechanisms affecting these genes include, e.g., effects on transcription, effects on RNA
25 splicing, alterations in relative amounts of alternative splice forms of mRNA, effects on RNA stability, effects on transport from the nucleus to cytoplasm, and effects on the efficiency and accuracy of translation.

A variety of methods can be used for detecting protein expression levels, including enzyme linked immunosorbent assays (ELISA), Western blots, immunoprecipitations and
30 immunofluorescence. A test sample from a subject is assessed for the presence of an alteration in the expression and/or an alteration in composition of the polypeptide encoded by a nucleic acid associated with breast cancer. An alteration in expression of a polypeptide encoded by a nucleic acid associated with breast cancer can be, for example, an alteration in the quantitative polypeptide expression (i.e., the amount of polypeptide produced). An alteration in the
35 composition of a polypeptide encoded by a nucleic acid associated with breast cancer is an alteration in the qualitative polypeptide expression (e.g., expression of a mutant polypeptide or of a different splicing variant). In one embodiment, diagnosis of a susceptibility to breast cancer is made by detecting a particular splicing variant encoded by a nucleic acid associated with

breast cancer, or a particular pattern of splicing variants (e.g., the nucleic acids encoding the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes).

- Both such alterations (quantitative and qualitative) can also be present. An "alteration" in the polypeptide expression or composition, as used herein, refers to an alteration in expression or composition in a test sample, as compared to the expression or composition of the polypeptide in a control sample. A control sample is a sample that corresponds to the test sample (e.g., is from the same type of cells), and is from a subject who is not affected by, and/or who does not have a susceptibility to, breast cancer. In one embodiment, the control sample is from a subject that does not possess a marker allele or haplotype as described herein. Similarly, the presence of one or more different splicing variants in the test sample, or the presence of significantly different amounts of different splicing variants in the test sample, as compared with the control sample, can be indicative of a susceptibility to breast cancer. An alteration in the expression or composition of the polypeptide in the test sample, as compared with the control sample, can be indicative of a specific allele in the instance where the allele alters a splice site relative to the reference in the control sample. Various means of examining expression or composition of a polypeptide encoded by a nucleic acid are known to the person skilled in the art and can be used, including spectroscopy, colorimetry, electrophoresis, isoelectric focusing, and immunoassays (e.g., David *et al.*, U.S. Pat. No. 4,376,110) such as immunoblotting (see, e.g., Current Protocols in Molecular Biology, particularly chapter 10, *supra*).
- For example, in one embodiment, an antibody (e.g., an antibody with a detectable label) that is capable of binding to a polypeptide encoded by a nucleic acid associated with breast cancer can be used. Antibodies can be polyclonal or monoclonal. An intact antibody, or a fragment thereof (e.g., Fv, Fab, Fab', F(ab')₂) can be used. The term "labeled", with regard to the probe or antibody, is intended to encompass direct labeling of the probe or antibody by coupling (i.e., physically linking) a detectable substance to the probe or antibody, as well as indirect labeling of the probe or antibody by reactivity with another reagent that is directly labeled. Examples of indirect labeling include detection of a primary antibody using a labeled secondary antibody (e.g., a fluorescently-labeled secondary antibody) and end-labeling of a DNA probe with biotin such that it can be detected with fluorescently-labeled streptavidin.
- In one embodiment of this method, the level or amount of polypeptide encoded by a nucleic acid associated with breast cancer (e.g., *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and/or *TNRC6B* genes) in a test sample is compared with the level or amount of the polypeptide in a control sample. A level or amount of the polypeptide in the test sample that is higher or lower than the level or amount of the polypeptide in the control sample, such that the difference is statistically significant, is indicative of an alteration in the expression of the polypeptide encoded by the nucleic acid, and is diagnostic for a particular allele or haplotype responsible for causing the difference in expression. Alternatively, the composition of the polypeptide in a test sample is compared with the composition of the polypeptide in a control sample. In another embodiment,

both the level or amount and the composition of the polypeptide can be assessed in the test sample and in the control sample.

In another embodiment, determination of a susceptibility to breast cancer is made by detecting at least one marker or haplotype of the present invention, in combination with an additional
 5 protein-based, RNA-based or DNA-based assay.

Kits

Kits useful in the methods of the invention comprise components useful in any of the methods described herein, including for example, primers for nucleic acid amplification, hybridization
 10 probes, restriction enzymes (e.g., for RFLP analysis), allele-specific oligonucleotides, antibodies that bind to an altered polypeptide encoded by a nucleic acid of the invention as described herein (e.g., a genomic segment comprising at least one polymorphic marker and/or haplotype of the present invention) or to a non-altered (native) polypeptide encoded by a nucleic acid of the invention as described herein, means for amplification of a nucleic acid associated with breast
 15 cancer, means for analyzing the nucleic acid sequence of a nucleic acid associated with breast cancer, means for analyzing the amino acid sequence of a polypeptide encoded by a nucleic acid associated with breast cancer, etc. The kits can for example include necessary buffers, nucleic acid primers for amplifying nucleic acids of the invention (e.g., one or more of the polymorphic markers as described herein), and reagents for allele-specific detection of the fragments
 20 amplified using such primers and necessary enzymes (e.g., DNA polymerase). Additionally, kits can provide reagents for assays to be used in combination with the methods of the present invention, e.g., reagents for use with breast cancer diagnostic assays.

In one embodiment, the invention is a kit for assaying a sample from a subject to detect the presence of a breast cancer or a susceptibility breast cancer in a subject, wherein the kit
 25 comprises reagents necessary for selectively detecting at least one allele of at least one polymorphism of the present invention in the genome of the individual. In a particular embodiment, the reagents comprise at least one contiguous oligonucleotide that hybridizes to a fragment of the genome of the individual comprising at least one polymorphism of the present invention. In another embodiment, the reagents comprise at least one pair of oligonucleotides
 30 that hybridize to opposite strands of a genomic segment obtained from a subject, wherein each oligonucleotide primer pair is designed to selectively amplify a fragment of the genome of the individual that includes at least one polymorphism, wherein the polymorphism is selected from the group consisting of the polymorphisms as listed in Tables 1 and 4 (SEQ ID NO:1-562), and polymorphic markers in linkage disequilibrium therewith. In yet another embodiment the
 35 fragment is at least 20 base pairs in size. Such oligonucleotides or nucleic acids (e.g., oligonucleotide primers) can be designed using portions of the nucleic acid sequence flanking

polymorphisms (e.g., SNPs or microsatellites) that are indicative of breast cancer. In another embodiment, the kit comprises one or more labeled nucleic acids capable of allele-specific detection of one or more specific polymorphic markers or haplotypes associated with breast cancer, and reagents for detection of the label. Suitable labels include, e.g., a radioisotope, a fluorescent label, an enzyme label, an enzyme co-factor label, a magnetic label, a spin label, an epitope label.

In particular embodiments, the polymorphic marker or haplotype to be detected by the reagents of the kit comprises one or more markers, two or more markers, three or more markers, four or more markers or five or more markers selected from the group consisting of the markers in Tables 1 and 4. In another embodiment, the marker or haplotype to be detected comprises at least one marker from the group of markers in strong linkage disequilibrium, as defined by values of r^2 greater than 0.2, to at least one of the group of markers consisting of the markers listed in Tables 1, 2, 3 and 4. In another embodiment, the marker or haplotype to be detected comprises at least one marker selected from the group of markers consisting of markers rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954.

In a preferred embodiment, the DNA template containing the SNP polymorphism is amplified by Polymerase Chain Reaction (PCR) prior to detection, and primers for such amplification are included in the reagent kit. In such an embodiment, the amplified DNA serves as the template for the detection probe and the enhancer probe.

In one embodiment, the DNA template is amplified by means of Whole Genome Amplification (WGA) methods, prior to assessment for the presence of specific polymorphic markers as described herein. Standard methods well known to the skilled person for performing WGA may be utilized, and are within scope of the invention. In one such embodiment, reagents for performing WGA are included in the reagent kit.

In certain embodiments, determination of the presence of a particular marker allele or haplotype is indicative of a susceptibility (increased susceptibility or decreased susceptibility) to breast cancer. In another embodiment, determination of the presence of the marker allele or haplotype is indicative of response to a therapeutic agent for breast cancer. In another embodiment, the presence of the marker allele or haplotype is indicative of breast cancer prognosis. In yet another embodiment, the presence of the marker allele or haplotype is indicative of progress of breast cancer treatment. Such treatment may include intervention by surgery, medication or by other means (e.g., lifestyle changes).

In a further aspect of the present invention, a pharmaceutical pack (kit) is provided, the pack comprising a therapeutic agent and a set of instructions for administration of the therapeutic agent to humans diagnostically tested for one or more variants of the present invention, as disclosed herein. The therapeutic agent can be a small molecule drug, an antibody, a peptide,

an antisense or rnai molecule, or other therapeutic molecules. In one embodiment, an individual identified as a carrier of at least one variant of the present invention is instructed to take a prescribed dose of the therapeutic agent. In one such embodiment, an individual identified as a homozygous carrier of at least one variant of the present invention is instructed to take a prescribed dose of the therapeutic agent. In another embodiment, an individual identified as a non-carrier of at least one variant of the present invention is instructed to take a prescribed dose of the therapeutic agent.

In certain embodiments, the kit further comprises a set of instructions for using the reagents comprising the kit. In certain embodiments, the kit further comprises a collection of data comprising correlation data between the polymorphic markers assessed by the kit and susceptibility to breast cancer.

Therapeutic agents

The risk variants for breast cancer presented herein can be useful in the identification of novel therapeutic targets for breast cancer. For example, genes containing, or in linkage disequilibrium with, variants (markers and/or haplotypes) associated with breast cancer (e.g., one or more of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes, or their products, as well as genes or their products that are directly or indirectly regulated by or interact with these variant genes or their products, can be targeted for the development of therapeutic agents to treat breast cancer, or prevent or delay onset of symptoms associated with breast cancer. Therapeutic agents may comprise one or more of, for example, small non-protein and non-nucleic acid molecules, proteins, peptides, protein fragments, nucleic acids (DNA, RNA), PNA (peptide nucleic acids), or their derivatives or mimetics which can modulate the function and/or levels of the target genes or their gene products.

The nucleic acids and/or variants described herein, or nucleic acids comprising their complementary sequence, may be used as antisense constructs to control gene expression in cells, tissues or organs. The methodology associated with antisense techniques is well known to the skilled artisan, and is for example described and reviewed in *Antisense Drug Technology: Principles, Strategies, and Applications*, Crooke, ed., Marcel Dekker Inc., New York (2001). In general, antisense agents (antisense oligonucleotides) are comprised of single stranded oligonucleotides (RNA or DNA) that are capable of binding to a complimentary nucleotide segment. By binding the appropriate target sequence, an RNA-RNA, DNA-DNA or RNA-DNA duplex is formed. The antisense oligonucleotides are complementary to the sense or coding strand of a gene. It is also possible to form a triple helix, where the antisense oligonucleotide binds to duplex DNA.

Several classes of antisense oligonucleotide are known to those skilled in the art, including cleavers and blockers. The former bind to target RNA sites, activate intracellular nucleases (e.g., RnaseH or Rnase L), that cleave the target RNA. Blockers bind to target RNA, inhibit protein translation by steric hindrance of the ribosomes. Examples of blockers include nucleic acids, morpholino compounds, locked nucleic acids and methylphosphonates (Thompson, *Drug Discovery Today*, 7:912-917 (2002)). Antisense oligonucleotides are useful directly as therapeutic agents, and are also useful for determining and validating gene function, for example by gene knock-out or gene knock-down experiments. Antisense technology is further described in Lavery *et al.*, *Curr. Opin. Drug Discov. Devel.* 6:561-569 (2003), Stephens *et al.*, *Curr. Opin. Mol. Ther.* 5:118-122 (2003), Kurreck, *Eur. J. Biochem.* 270:1628-44 (2003), Dias *et al.*, *Mol. Cancer Ther.* 1:347-55 (2002), Chen, *Methods Mol. Med.* 75:621-636 (2003), Wang *et al.*, *Curr. Cancer Drug Targets* 1:177-96 (2001), and Bennett, *Antisense Nucleic Acid Drug. Dev.* 12:215-24 (2002).

In certain embodiments, the antisense agent is an oligonucleotide that is capable of binding to a particular nucleotide segment. In certain embodiments, the nucleotide segment comprises any of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes. In certain other embodiments, the antisense nucleotide is capable of binding to a nucleotide segment of as set forth in SEQ ID NO:1-562. Antisense nucleotides can be from 5-500 nucleotides in length, including 5-200 nucleotides, 5-100 nucleotides, 10-50 nucleotides, and 10-30 nucleotides. In certain preferred embodiments, the antisense nucleotides are from 14-50 nucleotides in length, including 14-40 nucleotides and 14-30 nucleotides. The variants described herein can also be used for the selection and design of antisense reagents that are specific for particular variants. Using information about the variants described herein, antisense oligonucleotides or other antisense molecules that specifically target mRNA molecules that contain one or more variants of the invention can be designed. In this manner, expression of mRNA molecules that contain one or more variant of the present invention (i.e. certain marker alleles and/or haplotypes) can be inhibited or blocked. In one embodiment, the antisense molecules are designed to specifically bind a particular allelic form (i.e., one or several variants (alleles and/or haplotypes)) of the target nucleic acid, thereby inhibiting translation of a product originating from this specific allele or haplotype, but which do not bind other or alternate variants at the specific polymorphic sites of the target nucleic acid molecule. As antisense molecules can be used to inactivate mRNA so as to inhibit gene expression, and thus protein expression, the molecules can be used for disease treatment. The methodology can involve cleavage by means of ribozymes containing nucleotide sequences complementary to one or more regions in the mRNA that attenuate the ability of the mRNA to be translated. Such mRNA regions include, for example, protein-coding regions, in particular protein-coding regions corresponding to catalytic activity, substrate and/or ligand binding sites, or other functional domains of a protein.

The phenomenon of RNA Interference (RNAi) has been actively studied for the last decade, since its original discovery in *C. elegans* (Fire *et al.*, *Nature* 391:806-11 (1998)), and in recent years its

potential use in treatment of human disease has been actively pursued (reviewed in Kim & Rossi, *Nature Rev. Genet.* 8:173-204 (2007)). RNA interference (RNAi), also called gene silencing, is based on using double-stranded RNA molecules (dsRNA) to turn off specific genes. In the cell, cytoplasmic double-stranded RNA molecules (dsRNA) are processed by cellular complexes into small interfering RNA (siRNA). The siRNA guide the targeting of a protein-RNA complex to specific sites on a target mRNA, leading to cleavage of the mRNA (Thompson, *Drug Discovery Today*, 7:912-917 (2002)). The siRNA molecules are typically about 20, 21, 22 or 23 nucleotides in length. Thus, one aspect of the invention relates to isolated nucleic acid molecules, and the use of those molecules for RNA interference, i.e. as small interfering RNA molecules (siRNA). In one embodiment, the isolated nucleic acid molecules are 18-26 nucleotides in length, preferably 19-25 nucleotides in length, more preferably 20-24 nucleotides in length, and more preferably 21, 22 or 23 nucleotides in length.

Another pathway for RNAi-mediated gene silencing originates in endogenously encoded primary microRNA (pri-miRNA) transcripts, which are processed in the cell to generate precursor miRNA (pre-miRNA). These miRNA molecules are exported from the nucleus to the cytoplasm, where they undergo processing to generate mature miRNA molecules (miRNA), which direct translational inhibition by recognizing target sites in the 3' untranslated regions of mRNAs, and subsequent mRNA degradation by processing P-bodies (reviewed in Kim & Rossi, *Nature Rev. Genet.* 8:173-204 (2007)).

Clinical applications of RNAi include the incorporation of synthetic siRNA duplexes, which preferably are approximately 20-23 nucleotides in size, and preferably have 3' overlaps of 2 nucleotides. Knockdown of gene expression is established by sequence-specific design for the target mRNA. Several commercial sites for optimal design and synthesis of such molecules are known to those skilled in the art.

Other applications provide longer siRNA molecules (typically 25-30 nucleotides in length, preferably about 27 nucleotides), as well as small hairpin RNAs (shRNAs; typically about 29 nucleotides in length). The latter are naturally expressed, as described in Amarzguioui *et al.* (*FEBS Lett.* 579:5974-81 (2005)). Chemically synthetic siRNAs and shRNAs are substrates for *in vivo* processing, and in some cases provide more potent gene-silencing than shorter designs (Kim *et al.*, *Nature Biotechnol.* 23:222-226 (2005); Siolas *et al.*, *Nature Biotechnol.* 23:227-231 (2005)). In general siRNAs provide for transient silencing of gene expression, because their intracellular concentration is diluted by subsequent cell divisions. By contrast, expressed shRNAs mediate long-term, stable knockdown of target transcripts, for as long as transcription of the shRNA takes place (Marques *et al.*, *Nature Biotechnol.* 23:559-565 (2006); Brummelkamp *et al.*, *Science* 296: 550-553 (2002)).

Since RNAi molecules, including siRNA, miRNA and shRNA, act in a sequence-dependent manner, the variants presented herein can be used to design RNAi reagents that recognize specific nucleic

acid molecules comprising specific alleles and/or haplotypes (e.g., the alleles and/or haplotypes of the present invention), while not recognizing nucleic acid molecules comprising other alleles or haplotypes. These RNAi reagents can thus recognize and destroy the target nucleic acid molecules. As with antisense reagents, RNAi reagents can be useful as therapeutic agents (i.e., for turning off disease-associated genes or disease-associated gene variants), but may also be useful for characterizing and validating gene function (e.g., by gene knock-out or gene knock-down experiments).

Delivery of RNAi may be performed by a range of methodologies known to those skilled in the art. Methods utilizing non-viral delivery include cholesterol, stable nucleic acid-lipid particle (SNALP), heavy-chain antibody fragment (Fab), aptamers and nanoparticles. Viral delivery methods include use of lentivirus, adenovirus and adeno-associated virus. The siRNA molecules are in some embodiments chemically modified to increase their stability. This can include modifications at the 2' position of the ribose, including 2'-O-methylpurines and 2'-fluoropyrimidines, which provide resistance to Rnase activity. Other chemical modifications are possible and known to those skilled in the art.

The following references provide a further summary of RNAi, and possibilities for targeting specific genes using RNAi: Kim & Rossi, *Nat. Rev. Genet.* 8:173-184 (2007), Chen & Rajewsky, *Nat. Rev. Genet.* 8: 93-103 (2007), Reynolds, et al., *Nat. Biotechnol.* 22:326-330 (2004), Chi et al., *Proc. Natl. Acad. Sci. USA* 100:6343-6346 (2003), Vickers et al., *J. Biol. Chem.* 278:7108-7118 (2003), Agami, *Curr. Opin. Chem. Biol.* 6:829-834 (2002), Lavery, et al., *Curr. Opin. Drug Discov. Devel.* 6:561-569 (2003), Shi, *Trends Genet.* 19:9-12 (2003), Shuey et al., *Drug Discov. Today* 7:1040-46 (2002), McManus et al., *Nat. Rev. Genet.* 3:737-747 (2002), Xia et al., *Nat. Biotechnol.* 20:1006-10 (2002), Plasterk et al., *curr. Opin. Genet. Dev.* 10:562-7 (2000), Bosher et al., *Nat. Cell Biol.* 2:E31-6 (2000), and Hunter, *Curr. Biol.* 9:R440-442 (1999).

A genetic defect leading to increased predisposition or risk for development of breast cancer, or a defect causing breast cancer, may be corrected permanently by administering to a subject carrying the defect a nucleic acid fragment that incorporates a repair sequence that supplies the normal/wild-type nucleotide(s) at the site of the genetic defect. Such site-specific repair sequence may encompass an RNA/DNA oligonucleotide that operates to promote endogenous repair of a subject's genomic DNA. The administration of the repair sequence may be performed by an appropriate vehicle, such as a complex with polyethelenimine, encapsulated in anionic liposomes, a viral vector such as an adenovirus vector, or other pharmaceutical compositions suitable for promoting intracellular uptake of the administered nucleic acid. The genetic defect may then be overcome, since the chimeric oligonucleotides induce the incorporation of the normal sequence into the genome of the subject, leading to expression of the normal/wild-type gene product. The replacement is propagated, thus rendering a permanent repair and alleviation of the symptoms associated with the disease or condition.

The present invention provides methods for identifying compounds or agents that can be used to treat breast cancer. Thus, the variants of the invention are useful as targets for the identification and/or development of therapeutic agents. In certain embodiments, such methods include assaying the ability of an agent or compound to modulate the activity and/or expression of a nucleic acid that includes at least one of the variants (markers and/or haplotypes) of the present invention, or the encoded product of the nucleic acid. This includes, for example, one or more of the FGF10, MRPS30, HCN1 and FGFR2 genes, and their gene products. This in turn can be used to identify agents or compounds that inhibit or alter the undesired activity or expression of the encoded nucleic acid product. Assays for performing such experiments can be performed in cell-based systems or in cell-free systems, as known to the skilled person. Cell-based systems include cells naturally expressing the nucleic acid molecules of interest, or recombinant cells that have been genetically modified so as to express a certain desired nucleic acid molecule.

Variant gene expression in a patient can be assessed by expression of a variant-containing nucleic acid sequence (for example, a gene containing at least one variant of the present invention, which can be transcribed into RNA containing the at least one variant, and in turn translated into protein), or by altered expression of a normal/wild-type nucleic acid sequence due to variants affecting the level or pattern of expression of the normal transcripts, for example variants in the regulatory or control region of the gene. Assays for gene expression include direct nucleic acid assays (mRNA), assays for expressed protein levels, or assays of collateral compounds involved in a pathway, for example a signal pathway. Furthermore, the expression of genes that are up- or down-regulated in response to the signal pathway can also be assayed. One embodiment includes operably linking a reporter gene, such as luciferase, to the regulatory region of the gene(s) of interest.

Modulators of gene expression can in one embodiment be identified when a cell is contacted with a candidate compound or agent, and the expression of mRNA is determined. The expression level of mRNA in the presence of the candidate compound or agent is compared to the expression level in the absence of the compound or agent. Based on this comparison, candidate compounds or agents for treating breast cancer can be identified as those modulating the gene expression of the variant gene. When expression of mRNA or the encoded protein is statistically significantly greater in the presence of the candidate compound or agent than in its absence, then the candidate compound or agent is identified as a stimulator or up-regulator of expression of the nucleic acid. When nucleic acid expression or protein level is statistically significantly less in the presence of the candidate compound or agent than in its absence, then the candidate compound is identified as an inhibitor or down-regulator of the nucleic acid expression.

The invention further provides methods of treatment using a compound identified through drug (compound and/or agent) screening as a gene modulator (i.e. stimulator and/or inhibitor of gene expression).

Methods of Assessing Probability of Response to Therapeutic Agents, Methods of Monitoring Progress of Treatment and Methods for Treating Breast Cancer

As is known in the art, individuals can have differential responses to a particular therapy (e.g., a therapeutic agent or therapeutic method). The basis of the differential response may be genetically determined in part. Pharmacogenomics addresses the issue of how genetic variations (e.g., the variants (markers and/or haplotypes) of the present invention) affect drug response, due to altered drug disposition and/or abnormal or altered action of the drug. Thus, the basis of the differential response may be genetically determined in part. Clinical outcomes due to genetic variations affecting drug response may result in toxicity of the drug in certain individuals (e.g., carriers or non-carriers of the genetic variants of the present invention), or therapeutic failure of the drug. Therefore, the variants of the present invention may determine the manner in which a therapeutic agent and/or method acts on the body, or the way in which the body metabolizes the therapeutic agent.

Accordingly, in one embodiment, the presence of a particular allele at a polymorphic site or haplotype is indicative of a different response rate to a particular treatment modality. This means that a patient diagnosed with breast cancer, and carrying a certain allele at a polymorphic site or haplotype of the present invention (e.g., the at-risk and protective alleles and/or haplotypes of the invention) would respond better to, or worse to, a specific therapeutic, drug and/or other therapy used to treat the disease. Therefore, the presence or absence of the marker allele or haplotype could aid in deciding what treatment should be used for the patient. For example, for a newly diagnosed patient, the presence of a marker or haplotype of the present invention may be assessed (e.g., through testing DNA derived from a blood sample, as described herein). If the patient is positive for a marker allele or haplotype (that is, at least one specific allele of the marker, or haplotype, is present), then the physician recommends one particular therapy, while if the patient is negative for the at least one allele of a marker, or a haplotype, then a different course of therapy may be recommended (which may include recommending that no immediate therapy, other than serial monitoring for progression of the disease, be performed). Thus, the patient's carrier status could be used to help determine whether a particular treatment modality should be administered. The value lies within the possibilities of being able to diagnose the disease at an early stage, to select the most appropriate treatment, and provide information to the clinician about prognosis/aggressiveness of the disease in order to be able to apply the most appropriate treatment.

As described further herein, current clinical preventive options for breast cancer are mainly chemopreventive (chemotherapy, or hormonal therapy) and prophylactic surgery. The most common chemopreventive is Tamoxifen and Raloxifene; other options include other Selective Estrogen Receptor Modulator (SERM) and aromatase inhibitors. Treatment options also include

radiation therapy, for which a proportion of patients experience adverse symptoms. The markers of the invention, as described herein, may be used to assess response to these therapeutic options, or to predict the progress of therapy using any one of these treatment options. Thus, genetic profiling can be used to select the appropriate treatment strategy based on the genetic status of the individual, or it may be used to predict the outcome of the particular treatment option, and thus be useful in the strategic selection of treatment options or a combination of available treatment options.

The present invention also relates to methods of monitoring progress or effectiveness of a treatment for a breast cancer. This can be done based on the genotype and/or haplotype status of the markers and haplotypes of the present invention, i.e., by assessing the absence or presence of at least one allele of at least one polymorphic marker as disclosed herein, or by monitoring expression of genes that are associated with the variants (markers and haplotypes) of the present invention. The risk gene mRNA or the encoded polypeptide can be measured in a tissue sample (e.g., a peripheral blood sample, or a biopsy sample). Expression levels and/or mRNA levels can thus be determined before and during treatment to monitor its effectiveness. Alternatively, or concomitantly, the genotype and/or haplotype status of at least one risk variant for breast cancer as presented herein is determined before and during treatment to monitor its effectiveness.

Alternatively, biological networks or metabolic pathways related to the markers and haplotypes of the present invention can be monitored by determining mRNA and/or polypeptide levels. This can be done for example, by monitoring expression levels or polypeptides for several genes belonging to the network and/or pathway, in samples taken before and during treatment. Alternatively, metabolites belonging to the biological network or metabolic pathway can be determined before and during treatment. Effectiveness of the treatment is determined by comparing observed changes in expression levels/metabolite levels during treatment to corresponding data from healthy subjects.

In a further aspect, the markers of the present invention can be used to increase power and effectiveness of clinical trials. Thus, individuals who are carriers of the at-risk variants of the present invention, i.e. individuals who are carriers of at least one allele of at least one polymorphic marker conferring increased risk of developing breast cancer may be more likely to respond to a particular treatment modality. In one embodiment, individuals who carry at-risk variants for gene(s) in a pathway and/or metabolic network for which a particular treatment (e.g., small molecule drug) is targeting are more likely to be responders to the treatment. In another embodiment, individuals who carry at-risk variants for a gene, which expression and/or function is altered by the at-risk variant, are more likely to be responders to a treatment modality targeting that gene, its expression or its gene product.

In a further aspect, the markers and haplotypes of the present invention can be used for targeting the selection of pharmaceutical agents for specific individuals. Personalized selection of treatment modalities, lifestyle changes or combination of the two, can be realized by the utilization of the at-risk variants of the present invention. Thus, the knowledge of an individual's status for particular markers of the present invention, can be useful for selection of treatment options that target genes or gene products affected by the at-risk variants of the invention. Certain combinations of variants may be suitable for one selection of treatment options, while other gene variant combinations may target other treatment options. Such combination of variant may include one variant, two variants, three variants, or four or more variants, as needed to determine with clinically reliable accuracy the selection of treatment module.

Computer-implemented aspects

As understood by those of ordinary skill in the art, the methods and information described herein may be implemented, in all or in part, as computer executable instructions on known computer readable media. For example, the methods described herein may be implemented in hardware. Alternatively, the method may be implemented in software stored in, for example, one or more memories or other computer readable medium and implemented on one or more processors. As is known, the processors may be associated with one or more controllers, calculation units and/or other units of a computer system, or implanted in firmware as desired. If implemented in software, the routines may be stored in any computer readable memory such as in RAM, ROM, flash memory, a magnetic disk, a laser disk, or other storage medium, as is also known. Likewise, this software may be delivered to a computing device via any known delivery method including, for example, over a communication channel such as a telephone line, the Internet, a wireless connection, etc., or via a transportable medium, such as a computer readable disk, flash drive, etc.

More generally, and as understood by those of ordinary skill in the art, the various steps described above may be implemented as various blocks, operations, tools, modules and techniques which, in turn, may be implemented in hardware, firmware, software, or any combination of hardware, firmware, and/or software. When implemented in hardware, some or all of the blocks, operations, techniques, etc. may be implemented in, for example, a custom integrated circuit (IC), an application specific integrated circuit (ASIC), a field programmable logic array (FPGA), a programmable logic array (PLA), etc.

When implemented in software, the software may be stored in any known computer readable medium such as on a magnetic disk, an optical disk, or other storage medium, in a RAM or ROM or flash memory of a computer, processor, hard disk drive, optical disk drive, tape drive, etc. Likewise, the software may be delivered to a user or a computing system via any known delivery

method including, for example, on a computer readable disk or other transportable computer storage mechanism.

Fig. 1 illustrates an example of a suitable computing system environment 100 on which a system for the steps of the claimed method and apparatus may be implemented. The computing system environment 100 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the method or apparatus of the claims. Neither should the computing environment 100 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the exemplary operating environment 100.

The steps of the claimed method and system are operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well known computing systems, environments, and/or configurations that may be suitable for use with the methods or system of the claims include, but are not limited to, personal computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above systems or devices, and the like.

The steps of the claimed method and system may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that performs particular tasks or implement particular abstract data types. The methods and apparatus may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In both integrated and distributed computing environments, program modules may be located in both local and remote computer storage media including memory storage devices.

With reference to Fig. 1, an exemplary system for implementing the steps of the claimed method and system includes a general purpose computing device in the form of a computer 110. Components of computer 110 may include, but are not limited to, a processing unit 120, a system memory 130, and a system bus 121 that couples various system components including the system memory to the processing unit 120. The system bus 121 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus also known as Mezzanine bus.

Computer 110 typically includes a variety of computer readable media. Computer readable media can be any available media that can be accessed by computer 110 and includes both volatile and nonvolatile media, removable and non-removable media. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer 110. Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of the any of the above should also be included within the scope of computer readable media.

The system memory 130 includes computer storage media in the form of volatile and/or nonvolatile memory such as read only memory (ROM) 131 and random access memory (RAM) 132. A basic input/output system 133 (BIOS), containing the basic routines that help to transfer information between elements within computer 110, such as during start-up, is typically stored in ROM 131. RAM 132 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 120. By way of example and not limitation, Fig. 1 illustrates operating system 134, application programs 135, other program modules 136, and program data 137.

The computer 110 may also include other removable/non-removable, volatile/nonvolatile computer storage media. By way of example only, Fig. 1 illustrates a hard disk drive 140 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 151 that reads from or writes to a removable, nonvolatile magnetic disk 152, and an optical disk drive 155 that reads from or writes to a removable, nonvolatile optical disk 156 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 141 is typically connected to the system bus 121 through a non-removable memory interface such as interface

140, and magnetic disk drive 151 and optical disk drive 155 are typically connected to the system bus 121 by a removable memory interface, such as Interface 150.

The drives and their associated computer storage media discussed above and illustrated in Fig. 1, provide storage of computer readable instructions, data structures, program modules and other data for the computer 110. In Fig. 1, for example, hard disk drive 141 is illustrated as storing operating system 144, application programs 145, other program modules 146, and program data 147. Note that these components can either be the same as or different from operating system 134, application programs 135, other program modules 136, and program data 137. Operating system 144, application programs 145, other program modules 146, and program data 147 are given different numbers here to illustrate that, at a minimum, they are different copies. A user may enter commands and information into the computer 20 through input devices such as a keyboard 162 and pointing device 161, commonly referred to as a mouse, trackball or touch pad. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 120 through a user input interface 160 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor 191 or other type of display device is also connected to the system bus 121 via an interface, such as a video interface 190. In addition to the monitor, computers may also include other peripheral output devices such as speakers 197 and printer 196, which may be connected through an output peripheral interface 190.

The computer 110 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be a personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 110, although only a memory storage device 181 has been illustrated in Fig. 1. The logical connections depicted in Fig. 1 include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the computer 110 is connected to the LAN 171 through a network interface or adapter 170. When used in a WAN networking environment, the computer 110 typically includes a modem 172 or other means for establishing communications over the WAN 173, such as the Internet. The modem 172, which may be internal or external, may be connected to the system bus 121 via the user input interface 160, or other appropriate mechanism. In a networked environment, program modules depicted relative to the computer 110, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, Fig. 1 illustrates remote application programs 185 as residing on memory device 181. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

Although the forgoing text sets forth a detailed description of numerous different embodiments of the invention, it should be understood that the scope of the invention is defined by the words of the claims set forth at the end of this patent. The detailed description is to be construed as exemplary only and does not describe every possible embodiment of the invention because
5 describing every possible embodiment would be impractical, if not impossible. Numerous alternative embodiments could be implemented, using either current technology or technology developed after the filing date of this patent, which would still fall within the scope of the claims defining the invention.

While the risk evaluation system and method, and other elements, have been described as
10 preferably being implemented in software, they may be implemented in hardware, firmware, etc., and may be implemented by any other processor. Thus, the elements described herein may be implemented in a standard multi-purpose CPU or on specifically designed hardware or firmware such as an application-specific integrated circuit (ASIC) or other hard-wired device as desired, including, but not limited to, the computer 110 of Fig. 1. When implemented in
15 software, the software routine may be stored in any computer readable memory such as on a magnetic disk, a laser disk, or other storage medium, in a RAM or ROM of a computer or processor, in any database, etc. Likewise, this software may be delivered to a user or a diagnostic system via any known or desired delivery method including, for example, on a computer readable disk or other transportable computer storage mechanism or over a
20 communication channel such as a telephone line, the internet, wireless communication, etc. (which are viewed as being the same as or interchangeable with providing such software via a transportable storage medium).

Thus, many modifications and variations may be made in the techniques and structures described and illustrated herein without departing from the spirit and scope of the present
25 invention. Thus, it should be understood that the methods and apparatus described herein are illustrative only and are not limiting upon the scope of the invention.

Accordingly, the invention relates to computer-implemented applications of the polymorphic markers and haplotypes described herein to be associated with breast cancer. Such applications can be useful for storing, manipulating or otherwise analyzing genotype data that is useful in the
30 methods of the invention. One example pertains to storing genotype information derived from an individual on readable media, so as to be able to provide the genotype information to a third party (e.g., the individual, a guardian of the individual, a health care provider or genetic analysis service provider), or for deriving information from the genotype data, e.g., by comparing the genotype data to information about genetic risk factors contributing to increased susceptibility to
35 the cancer, and reporting results based on such comparison.

In general terms, computer-readable media has capabilities of storing (i) identifier information for at least one polymorphic marker or a haplotype, as described herein; (ii) an indicator of the

frequency of at least one allele of said at least one marker, or the frequency of a haplotype, in individuals with the disease; and an indicator of the frequency of at least one allele of said at least one marker, or the frequency of a haplotype, in a reference population. The reference population can be a disease-free population of individuals. Alternatively, the reference population is a random sample from the general population, and is thus representative of the population at large. The frequency indicator may be a calculated frequency, a count of alleles and/or haplotype copies, or normalized or otherwise manipulated values of the actual frequencies that are suitable for the particular medium.

The markers and haplotypes described herein to be associated with increased susceptibility (increased risk) of breast cancer, are in certain embodiments useful for interpretation and/or analysis of genotype data. Thus in certain embodiments, determination of the presence of an at-risk allele for breast cancer, as shown herein, or determination of the presence of an allele at a polymorphic marker in LD with any such risk allele, is indicative of the individual from whom the genotype data originates is at increased risk of breast cancer. In one such embodiment, genotype data is generated for at least one polymorphic marker shown herein to be associated with breast cancer, or a marker in linkage disequilibrium therewith. The genotype data is subsequently made available to a third party, such as the individual from whom the data originates, his/her guardian or representative, a physician or health care worker, genetic counsellor, or insurance agent, for example via a user interface accessible over the internet, together with an interpretation of the genotype data, e.g., in the form of a risk measure (such as an absolute risk (AR), risk ratio (RR) or odds ratio (OR)) for the disease. In another embodiment, at-risk markers identified in a genotype dataset derived from an individual are assessed and results from the assessment of the risk conferred by the presence of such at-risk variants in the dataset are made available to the third party, for example via a secure web interface, or by other communication means. The results of such risk assessment can be reported in numeric form (e.g., by risk values, such as absolute risk, relative risk, and/or an odds ratio, or by a percentage increase in risk compared with a reference), by graphical means, or by other means suitable to illustrate the risk to the individual from whom the genotype data is derived.

30

Nucleic Acids and Polypeptides

The nucleic acids and polypeptides described herein can be used in methods and kits of present invention, as described in the above. An "isolated" nucleic acid molecule, as used herein, is one that is separated from nucleic acids that normally flank the gene or nucleotide sequence (as in genomic sequences) and/or has been completely or partially purified from other transcribed sequences (e.g., as in an RNA library). For example, an isolated nucleic acid of the invention can be substantially isolated with respect to the complex cellular milieu in which it

35

naturally occurs, or culture medium when produced by recombinant techniques, or chemical precursors or other chemicals when chemically synthesized. In some instances, the isolated material will form part of a composition (for example, a crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material can be purified to essential homogeneity, for example as determined by polyacrylamide gel electrophoresis (PAGE) or column chromatography (e.g., HPLC). An isolated nucleic acid molecule of the invention can comprise at least about 50%, at least about 80% or at least about 90% (on a molar basis) of all macromolecular species present. With regard to genomic DNA, the term "isolated" also can refer to nucleic acid molecules that are separated from the chromosome with which the genomic DNA is naturally associated. For example, the isolated nucleic acid molecule can contain less than about 250 kb, 200 kb, 150 kb, 100 kb, 75 kb, 50 kb, 25 kb, 10 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of the nucleotides that flank the nucleic acid molecule in the genomic DNA of the cell from which the nucleic acid molecule is derived.

The nucleic acid molecule can be fused to other coding or regulatory sequences and still be considered isolated. Thus, recombinant DNA contained in a vector is included in the definition of "isolated" as used herein. Also, isolated nucleic acid molecules include recombinant DNA molecules in heterologous host cells or heterologous organisms, as well as partially or substantially purified DNA molecules in solution. "Isolated" nucleic acid molecules also encompass *in vivo* and *in vitro* RNA transcripts of the DNA molecules of the present invention. An isolated nucleic acid molecule or nucleotide sequence can include a nucleic acid molecule or nucleotide sequence that is synthesized chemically or by recombinant means. Such isolated nucleotide sequences are useful, for example, in the manufacture of the encoded polypeptide, as probes for isolating homologous sequences (e.g., from other mammalian species), for gene mapping (e.g., by *in situ* hybridization with chromosomes), or for detecting expression of the gene in tissue (e.g., human tissue), such as by Northern blot analysis or other hybridization techniques.

The invention also pertains to nucleic acid molecules that hybridize under high stringency hybridization conditions, such as for selective hybridization, to a nucleotide sequence described herein (e.g., nucleic acid molecules that specifically hybridize to a nucleotide sequence containing a polymorphic site associated with a marker or haplotype described herein). Such nucleic acid molecules can be detected and/or isolated by allele- or sequence-specific hybridization (e.g., under high stringency conditions). Stringency conditions and methods for nucleic acid hybridizations are well known to the skilled person (see, e.g., *Current Protocols in Molecular Biology*, Ausubel, F. et al, John Wiley & Sons, (1998), and Kraus, M. and Aaronson, S., *Methods Enzymol.*, 200:546-556 (1991), the entire teachings of which are incorporated by reference herein.

The percent identity of two nucleotide or amino acid sequences can be determined by aligning the sequences for optimal comparison purposes (e.g., gaps can be introduced in the sequence of

a first sequence). The nucleotides or amino acids at corresponding positions are then compared, and the percent identity between the two sequences is a function of the number of identical positions shared by the sequences (i.e., % identity = # of identical positions/total # of positions x 100). In certain embodiments, the length of a sequence aligned for comparison purposes is at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, or at least 95%, of the length of the reference sequence. The actual comparison of the two sequences can be accomplished by well-known methods, for example, using a mathematical algorithm. A non-limiting example of such a mathematical algorithm is described in Karlin, S. and Altschul, S., *Proc. Natl. Acad. Sci. USA*, 90:5873-5877 (1993). Such an algorithm is incorporated into the NBLAST and XBLAST programs (version 2.0), as described in Altschul, S. et al., *Nucleic Acids Res.*, 25:3389-3402 (1997). When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (e.g., NBLAST) can be used. See the website on the World Wide Web at ncbi.nlm.nih.gov. In one embodiment, parameters for sequence comparison can be set at score=100, wordlength=12, or can be varied (e.g., W=5 or W=20). Another example of an algorithm is BLAT (Kent, W.J. *Genome Res.* 12:656-64 (2002)).

Other examples include the algorithm of Myers and Miller, CABIOS (1989), ADVANCE and ADAM as described in Torellis, A. and Robotti, C., *Comput. Appl. Biosci.* 10:3-5 (1994); and FASTA described in Pearson, W. and Lipman, D., *Proc. Natl. Acad. Sci. USA*, 85:2444-48 (1988).

In another embodiment, the percent identity between two amino acid sequences can be accomplished using the GAP program in the GCG software package (Accelrys, Cambridge, UK).

The present invention also provides isolated nucleic acid molecules that contain a fragment or portion that hybridizes under highly stringent conditions to a nucleic acid that comprises, or consists of, a nucleotide sequence comprising all or a portion of LD block C09, LD block C10A, LD block 10B, LD block C11A, LD block C11B, LD block C14, LD block C18, LD block C22A, and LD block C22B, as defined herein; or a nucleic acid that comprises at least one of the polymorphic markers listed in and Table 4 herein (SEQ ID NO:1-562); or a nucleotide sequence of any one of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes; or a nucleotide sequence comprising, or consisting of, the complement of a nucleotide sequence comprising all or a portion of LD block C09, LD block C10A, LD block 10B, LD block C11A, LD block C11B, LD block C14, LD block C18, LD block C22A, and LD block C22B, as defined herein; or a nucleic acid that comprises at least one of the polymorphic markers listed in and Table 4 herein (SEQ ID NO:1-562); or a nucleotide sequence of any one of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes.

The nucleic acid fragments of the invention are at least about 15, at least about 18, 20, 23 or 25 nucleotides, and can be 30, 40, 50, 100, 200, 500, 1000, 10,000 or more nucleotides in length.

The nucleic acid fragments of the invention are used as probes or primers in assays such as those described herein. "Probes" or "primers" are oligonucleotides that hybridize in a base-specific manner to a complementary strand of a nucleic acid molecule. In addition to DNA and RNA, such probes and primers include polypeptide nucleic acids (PNA), as described in Nielsen, P. *et al.*, *Science* 254:1497-1500 (1991). A probe or primer comprises a region of nucleotide sequence that hybridizes to at least about 15, typically about 20-25, and in certain embodiments about 40, 50 or 75, consecutive nucleotides of a nucleic acid molecule. In one embodiment, the probe or primer comprises at least one allele of at least one polymorphic marker or at least one haplotype described herein, or the complement thereof. In particular embodiments, a probe or primer can comprise 100 or fewer nucleotides; for example, in certain embodiments from 6 to 50 nucleotides, or, for example, from 12 to 30 nucleotides. In other embodiments, the probe or primer is at least 70% identical, at least 80% identical, at least 85% identical, at least 90% identical, or at least 95% identical, to the contiguous nucleotide sequence or to the complement of the contiguous nucleotide sequence. In another embodiment, the probe or primer is capable of selectively hybridizing to the contiguous nucleotide sequence or to the complement of the contiguous nucleotide sequence. Often, the probe or primer further comprises a label, e.g., a radioisotope, a fluorescent label, an enzyme label, an enzyme co-factor label, a magnetic label, a spin label, an epitope label.

The nucleic acid molecules of the invention, such as those described above, can be identified and isolated using standard molecular biology techniques well known to the skilled person. The amplified DNA can be labeled (e.g., radiolabeled) and used as a probe for screening a cDNA library derived from human cells. The cDNA can be derived from mRNA and contained in a suitable vector. Corresponding clones can be isolated, DNA can be obtained following *in vivo* excision, and the cloned insert can be sequenced in either or both orientations by art-recognized methods to identify the correct reading frame encoding a polypeptide of the appropriate molecular weight. Using these or similar methods, the polypeptide and the DNA encoding the polypeptide can be isolated, sequenced and further characterized.

Antibodies

The invention also provides antibodies which bind to an epitope comprising either a variant amino acid sequence (e.g., comprising an amino acid substitution) encoded by a variant allele or the reference amino acid sequence encoded by the corresponding non-variant or wild-type allele. The term "antibody" as used herein refers to immunoglobulin molecules and immunologically active portions of immunoglobulin molecules, *i.e.*, molecules that contain antigen-binding sites that specifically bind an antigen. A molecule that specifically binds to a polypeptide of the invention is a molecule that binds to that polypeptide or a fragment thereof, but does not substantially bind other molecules in a sample, e.g., a biological sample, which naturally contains

the polypeptide. Examples of immunologically active portions of immunoglobulin molecules include F(ab) and F(ab')₂ fragments which can be generated by treating the antibody with an enzyme such as pepsin. The invention provides polyclonal and monoclonal antibodies that bind to a polypeptide of the invention. The term "monoclonal antibody" or "monoclonal antibody composition", as used herein, refers to a population of antibody molecules that contain only one species of an antigen binding site capable of immunoreacting with a particular epitope of a polypeptide of the invention. A monoclonal antibody composition thus typically displays a single binding affinity for a particular polypeptide of the invention with which it immunoreacts.

Polyclonal antibodies can be prepared as described above by immunizing a suitable subject with a desired immunogen, e.g., polypeptide of the invention or a fragment thereof. The antibody titer in the immunized subject can be monitored over time by standard techniques, such as with an enzyme linked immunosorbent assay (ELISA) using immobilized polypeptide. If desired, the antibody molecules directed against the polypeptide can be isolated from the mammal (e.g., from the blood) and further purified by well-known techniques, such as protein A chromatography to obtain the IgG fraction. At an appropriate time after immunization, e.g., when the antibody titers are highest, antibody-producing cells can be obtained from the subject and used to prepare monoclonal antibodies by standard techniques, such as the hybridoma technique originally described by Kohler and Milstein, *Nature* 256:495-497 (1975), the human B cell hybridoma technique (Kozbor *et al.*, *Immunol. Today* 4: 72 (1983)), the EBV-hybridoma technique (Cole *et al.*, *Monoclonal Antibodies and Cancer Therapy*, Alan R. Liss, 1985, Inc., pp. 77-96) or trioma techniques. The technology for producing hybridomas is well known (see generally *Current Protocols in Immunology* (1994) Coligan *et al.*, (eds.) John Wiley & Sons, Inc., New York, NY). Briefly, an immortal cell line (typically a myeloma) is fused to lymphocytes (typically splenocytes) from a mammal immunized with an immunogen as described above, and the culture supernatants of the resulting hybridoma cells are screened to identify a hybridoma producing a monoclonal antibody that binds a polypeptide of the invention.

Any of the many well known protocols used for fusing lymphocytes and immortalized cell lines can be applied for the purpose of generating a monoclonal antibody to a polypeptide of the invention (see, e.g., *Current Protocols in Immunology*, *supra*; Galfre *et al.*, *Nature* 266:55052 (1977); R.H. Kenneth, in *Monoclonal Antibodies: A New Dimension In Biological Analyses*, Plenum Publishing Corp., New York, New York (1980); and Lerner, *Yale J. Biol. Med.* 54:387-402 (1981)). Moreover, the ordinarily skilled worker will appreciate that there are many variations of such methods that also would be useful.

Alternative to preparing monoclonal antibody-secreting hybridomas, a monoclonal antibody to a polypeptide of the invention can be identified and isolated by screening a recombinant combinatorial immunoglobulin library (e.g., an antibody phage display library) with the polypeptide to thereby isolate immunoglobulin library members that bind the polypeptide. Kits for generating and screening phage display libraries are commercially available (e.g., the

- Pharmacia Recombinant Phage Antibody System, Catalog No. 27-9400-01; and the Stratagene SurfZAP™ Phage Display Kit, Catalog No. 240612). Additionally, examples of methods and reagents particularly amenable for use in generating and screening antibody display library can be found in, for example, U.S. Patent No. 5,223,409; PCT Publication No. WO 92/18619; PCT Publication No. WO 91/17271; PCT Publication No. WO 92/20791; PCT Publication No. WO 92/15679; PCT Publication No. WO 93/01288; PCT Publication No. WO 92/01047; PCT Publication No. WO 92/09690; PCT Publication No. WO 90/02809; Fuchs *et al.*, *Bio/Technology* 9: 1370-1372 (1991); Hay *et al.*, *Hum. Antibod. Hybridomas* 3:81-85 (1992); Huse *et al.*, *Science* 246: 1275-1281 (1989); and Griffiths *et al.*, *EMBO J.* 12:725-734 (1993).
- 10 Additionally, recombinant antibodies, such as chimeric and humanized monoclonal antibodies comprising both human and non-human portions, which can be made using standard recombinant DNA techniques, are within the scope of the invention. Such chimeric and humanized monoclonal antibodies can be produced by recombinant DNA techniques known in the art.
- 15 In general, antibodies of the invention (*e.g.*, a monoclonal antibody) can be used to isolate a polypeptide of the invention (*e.g.* a polypeptide encoded by any one of the *PAX5*, *TUB*, *SERPINH1*, *RAD51L1*, *FHOD3* and *TNRC6B* genes) by standard techniques, such as affinity chromatography or immunoprecipitation. A polypeptide-specific antibody can facilitate the purification of natural polypeptide from cells and of recombinantly produced polypeptide expressed in host cells.
- 20 Moreover, an antibody specific for a polypeptide of the invention can be used to detect the polypeptide (*e.g.*, in a cellular lysate, cell supernatant, or tissue sample) in order to evaluate the abundance and pattern of expression of the polypeptide. Antibodies can be used diagnostically to monitor protein levels in tissue as part of a clinical testing procedure, *e.g.*, to, for example, determine the efficacy of a given treatment regimen. The antibody can be coupled to a
- 25 detectable substance to facilitate its detection. Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase, beta-galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin;
- 30 examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include ^{125}I , ^{131}I , ^{35}S or ^3H .
- 35 Antibodies may also be useful in pharmacogenomic analysis. In such embodiments, antibodies against variant proteins encoded by nucleic acids according to the invention such as variant proteins that are encoded by nucleic acids that contain at least one polymorphic marker of the invention, can be used to identify individuals that require modified treatment modalities.

Antibodies can furthermore be useful for assessing expression of variant proteins in disease states, such as in active stages of a disease, or in an individual with a predisposition to a disease related to the function of the protein, in particular breast cancer. Antibodies specific for a variant protein of the present invention that is encoded by a nucleic acid that comprises at least one polymorphic marker or haplotype as described herein can be used to screen for the presence of the variant protein, for example to screen for a predisposition to breast cancer as indicated by the presence of the variant protein.

Antibodies can be used in other methods. Thus, antibodies are useful as diagnostic tools for evaluating proteins, such as variant proteins of the invention in conjunction with analysis by electrophoretic mobility, isoelectric point, tryptic or other protease digest, or for use in other physical assays known to those skilled in the art. Antibodies may also be used in tissue typing. In one such embodiment, a specific variant protein has been correlated with expression in a specific tissue type, and antibodies specific for the variant protein can then be used to identify the specific tissue type.

Subcellular localization of proteins, including variant proteins, can also be determined using antibodies, and can be applied to assess aberrant subcellular localization of the protein in cells in various tissues. Such use can be applied in genetic testing, but also in monitoring a particular treatment modality. In the case where treatment is aimed at correcting the expression level or presence of the variant protein or aberrant tissue distribution or developmental expression of the variant protein, antibodies specific for the variant protein or fragments thereof can be used to monitor therapeutic efficacy.

Antibodies are further useful for inhibiting variant protein function, for example by blocking the binding of a variant protein to a binding molecule or partner. Such uses can also be applied in a therapeutic context in which treatment involves inhibiting a variant protein's function. An antibody can be for example be used to block or competitively inhibit binding, thereby modulating (i.e., agonizing or antagonizing) the activity of the protein. Antibodies can be prepared against specific protein fragments containing sites required for specific function or against an intact protein that is associated with a cell or cell membrane. For administration *in vivo*, an antibody may be linked with an additional therapeutic payload, such as radionuclide, an enzyme, an immunogenic epitope, or a cytotoxic agent, including bacterial toxins (diphtheria or plant toxins, such as ricin). The *in vivo* half-life of an antibody or a fragment thereof may be increased by pegylation through conjugation to polyethylene glycol.

The present invention further relates to kits for using antibodies in the methods described herein. This includes, but is not limited to, kits for detecting the presence of a variant protein in a test sample. One preferred embodiment comprises antibodies such as a labeled or labelable antibody and a compound or agent for detecting variant proteins in a biological sample, means for determining the amount or the presence and/or absence of variant protein in the sample, and

means for comparing the amount of variant protein in the sample with a standard, as well as instructions for use of the kit.

The present invention will now be exemplified by the following non-limiting example.

EXAMPLIFICATION.

Identification of variants on nine chromosomal locations that associate with risk of breast cancer

In order to search widely for alleles of common SNPs associating to breast cancer susceptibility, we carried out a genome-wide SNP association study using Illumina HumanHap300 microarray technology. Genotyping was carried out on approximately 1835 Icelandic breast cancer patients and 30,320 controls. After removing SNPs that failed quality control checks, 311,524 SNPs remained and were tested for association with breast cancer. The results were adjusted for relatedness among individuals and potential population stratification using the method of genomic control [Devlin and Roeder, (1999), *Biometrics*, **55**, 997-1004](see Methods). Signals were ranked by P-value. A set of SNPs from various locations on the genome, most notably from Chromosomes 9,10 (2 regions A and B), 11 (2 regions A and B), 14,18 and 22 (2 regions A and B) showed the highest ranks. The regions of interest containing these markers (called LD Blocks) are defined in the LD Block Table; all co-ordinates herein being from NCBI Build 36. Results from genotyping of the Illumina SNPs in these regions are presented in Table 1.

In order to further investigate the signals related to the highly ranked nine markers; e.g. rs2005154 (C09), rs2184380 (C10A), rs2224696 (C10B), rs2242503 (C11A), rs12291026 (C11B), rs999737 (C14), rs9956546 (C18), rs11912922 (C22A) and rs6001954 (C22B), we generated and validated Centaurus assays for these SNPs. The SNP assays were used to genotype an additional sample of approximately 450 Icelandic Breast Cancer patients and over 5000 controls. The combined data from the Illumina and Centaurus assays for the Icelandic Breast Cancer patients (2280) and controls (35650) is shown in Table 2. All SNPs show a significant association with breast cancer in the combined Icelandic cohorts, confirming the original observations with the results in Table 1.

METHODS

Patient and Control Selection:

Collection of blood samples and medical information from study subjects was conducted with informed consent and ethical review board approval in accordance with the Declaration of Helsinki.

Iceland: Records of breast cancer diagnoses were obtained from the Icelandic Cancer Registry (ICR). The ICR contains all cases of invasive breast tumours and ductal or lobular carcinoma in-situ diagnosed in Iceland from January 1st 1955. All prevalent cases living in Iceland who had a

diagnosis entered into the ICR up to the end of December 2006 were eligible to participate in the study. The ICR contained records of 4785 individuals diagnosed during this period. Consent, samples and successful genotypes were obtained from approximately 2280 patients. Of these, genotypes were derived from Illumina Hap300 chips for 1835 patients and from Centaurus assays for 445 patients. The roughly 35500 Icelandic controls consisted of individuals selected from ongoing Illumina-based genome-wide association studies at deCODE genetics. Individuals with a diagnosis of breast cancer in the ICR were excluded. Both male and female genders were included. In the Icelandic controls (and the foreign replication control groups described below) there were no significant differences between genders in the frequencies of SNPs listed in Table 6. Therefore we considered that these control groups provided reasonable representations of the population frequencies of the SNPs under investigation.

Spain: The Spanish study patients were recruited from the Oncology Department of Zaragoza Hospital between March 2006 and August 2007. Genotyping was carried out satisfactorily on approximately 825 patients. The successfully genotyped controls (about 1730) had attended the University Hospital in Zaragoza for diseases other than cancer. Controls were questioned to rule out prior cancers before drawing the blood sample. All patients and controls were of European ethnicity.

Sweden: The Swedish sample sets consisted of Familial and Consecutive patient series. The Familial breast cancer recruitment group consisted of 347 breast cancer patients who had been referred to the oncogenetic counseling clinic of the Karolinska University Hospital, Stockholm for investigation of a family history of breast cancer. Each patient came from a distinct family. All cases who met the current criteria for BRCA mutation screening had tested negative. The Consecutive breast cancer recruitment group was comprised of 482 consecutively recruited patients who were treated surgically for primary invasive breast cancer at the Departments of Oncology at Huddinge and Söder Hospitals (covering the population of southern Stockholm) from October 1998 to May 2000. Family history was not taken into account in the selection of patients for recruitment. Controls were 1302 blood donors and 448 cancer-free individuals of both genders. All controls were collected at the Karolinska University Hospital, Stockholm. There was no evidence of significant heterogeneity between the Familial and Consecutive series for any of the SNPs tested.

Holland: Female patients diagnosed with breast cancer in the period 2005-2006 were selected from the regional cancer registry held by the Comprehensive Cancer Centre East in Nijmegen, the Netherlands. This cancer center keeps a population-based cancer registry and covers the eastern part of the Netherlands, a region with 1.3 million inhabitants. All patients diagnosed with breast cancer before the age of 70 were invited to participate in the study. The Comprehensive Cancer Centre East collected the clinical and pathology data for all patients in the cancer registry. These standard cancer registry data were supplemented with more detailed data by extraction from the medical files in the hospitals where the patients were treated.

Controls were collected in a survey in 2002-2003 by the Radboud University Nijmegen Medical Center. This survey, The Nijmegen Biomedical Study, was based on an age-stratified random sample of the population of Nijmegen. From this group 2034 control individuals, age-matched by frequency to the patient population, were selected and genotyped.

- 5 CGEMS (*The Cancer Genetics Markers of Susceptibility*): Is a project of the U.S. National Cancer Institute that has released data to the public domain on a genome-wide SNP association study for breast cancer susceptibility based on 1145 patients and 1142 controls genotyped with approximately 530,000 SNPs using the Illumina platform. These data are available at: <https://caintegrator.nci.nih.gov/cgems/>.

10

Genotyping

- Approximately 1840 Icelandic patients and 3020 controls were genotyped on Illumina Hap300 SNP arrays, as described previously [Stacey, et al., (2007), Nat Genet **39**:865-9] Results are shown in Table 1. All other genotyping was carried out using Nanongen Centaurus assays
- 15 [Kutyavin, et al., (2006), Nucleic Acids Res, **34**, e128] that were generated for SNPs shown in Tables 2 and 3. Centaurus SNP assays were validated by genotyping the HapMap CEU samples and comparing the genotypes with published data. Assays were rejected if they showed $\geq 1.5\%$ mismatches with the HapMap data. Approximately 10% of the Icelandic case samples were genotyped on both Illumina and Nanongen platforms and the observed mismatch rate was lower
 - 20 than 0.5%. All genotyping was carried out at the deCODE genetics facility. All physical coordinates are given according to NCBI Build 36.

Illumina Genotyping

- DNA samples were genotyped according to the manufacturer's instructions on Illumina Infinium HumanHap300 SNP bead microarrays (Illumina, San Diego, CA, USA), containing 317,503 SNPs derived from Phase I of the International HapMap project. This chip provides about 75% genomic coverage in the Utah CEPH (CEU) HapMap samples for common SNPs at $r^2 \geq 0.8$ [Barrett and Cardon, (2006), Nat Genet, **38**, 659-62]. Of the total number of SNPs on the chip, 5979 were deemed unsuitable either because they were monomorphic (i.e. the minor allele frequency in the combined patients and control set was less than 0.001), or had low (<95%) yield or showed a
- 30 very significant distortion from Hardy-Weinberg equilibrium in the controls ($P < 1 \times 10^{-10}$). All of these problematic SNPs were removed from the analysis. Thus 311,524 SNPs were used in the association analysis. Any chips with an overall call rate below 98% of the SNPs were also excluded from the genome-wide association analysis.

Centaurus SNP Genotyping

Centaurus assays [Kutyavin, et al., (2006), Nucleic Acids Res, 34, e128] were designed for all nine variants: e. g. rs2005154, rs2184380, rs2224696, rs2242503, rs12291026, rs999737, rs9956546, rs11912922 and rs6001954 and validated by genotyping the HapMap CEU sample and comparing the genotypes with published data. The assays gave <1.5% mis-matches with HapMap data. Table 2 shows the Seq ID Number reference for the sequence context for these SNPs, e.g. Seq ID No 1, Seq ID No 2, Seq ID No 3, Seq ID No 4, Seq ID No 5, Seq ID No 6, Seq ID No 7, Seq ID No 8 and Seq ID No 9.

10 *Statistical Methods*

We calculated the odds ratio (OR) of a SNP allele assuming the multiplicative model, i.e. assuming that the relative risk of the two alleles that a person carries multiplies. Allelic frequencies rather than carrier frequencies are presented for the markers. The associated P-values were calculated with a standard likelihood ratio Chi-squared statistic as implemented in the NEMO software package [Gretarsdottir, et al., (2003), Nat Genet, **35**, 131-8]. Confidence intervals were calculated assuming that the estimate of the OR has a log-normal distribution.

Some Icelandic patients and controls are related, both within and between groups, causing the Chi-squared test statistic to have a mean greater than one and a median larger than 0.675². We estimated the inflation factor for *Iceland 1* using a method of genomic control [Devlin and Roeder, (1999), Biometrics, **55**, 997-1004] by calculating the average of the observed Chi-squared statistics for the genome-wide SNP set, which accounts for relatedness and for potential population stratification. For *Iceland 2*, which was not typed with a genome-wide set of markers, the inflation factor was estimated by simulating genotypes through the Icelandic genealogy [Grant, et al., (2006), Nat Genet, **38**, 320-3]. The estimated inflation factors were 1.105 for *Iceland 1* and 1.11 for *Iceland 2*. The estimated inflation factor for the joint analyses of the *Iceland 1* and *Iceland 2* sample sets was 1.08, obtained by simulation.

All P-values are reported as two-sided.

Table 1: SNPs from the Illumina Infinium HumanHap300 chip found in association to breast cancer in Icelandic case- and control-samples. The table shows the rs-name of the SNP, P-value for the association with Breast Cancer, the associated risk, number and frequency of associated allele in affecteds and controls, identity of the at-risk allele, chromosome and position on the chromosome (NCBI Build36) of the associated marker, and reference to the sequence ID showing flanking sequence of the SNP. Allelic codes are: 1 = A; 2 = C; 3 = G; 4 = T.

SNP	P-value	OR	Aff	Aff.freq	Contrl.	Con.freq	Allele	Chr	Pos in Build 36	Seq ID No
rs2800256	0.1454055	1.0545	1835	0.323706	30311	0.312197	4	C09	36,806,714	12
rs2851696	0.1540843	1.0513	1836	0.630174	30363	0.618434	1	C09	36,809,706	15
rs2851695	0.4950720	1.024	1833	0.603382	30308	0.597697	4	C09	36,810,262	14
rs2800266	0.0940407	1.0605	1833	0.625477	30301	0.61163	1	C09	36,810,604	13
rs7848675	0.0424981	1.0789	1837	0.715024	30383	0.699306	1	C09	36,814,086	18
rs2800252	0.5399746	1.0477	1837	0.948013	30370	0.94567	3	C09	36,820,616	11
rs4880018	0.8570986	1.0154	1837	0.958628	30345	0.958016	2	C09	36,833,844	17
rs3780136	0.0670526	1.0972	1831	0.873839	30243	0.863258	2	C09	36,835,973	16
rs7852051	0.0515094	1.077	1837	0.279532	30380	0.264845	4	C09	36,841,358	19
rs1051193	0.2874298	1.056	1837	0.128198	30380	0.122235	4	C09	36,846,268	10
rs2005154	0.0009612	1.2026	1837	0.902831	30353	0.885398	4	C09	36,846,828	1
rs2031557	0.0127968	1.0884	1831	0.513381	30326	0.492201	1	C10	8,651,020	32
rs1325881	0.5529599	1.0381	1837	0.080022	30376	0.077314	1	C10	8,653,900	28
rs1079565	0.6949928	1.0221	1837	0.104246	30380	0.102222	2	C10	8,656,133	22
rs1125575	0.0423179	1.0728	1834	0.41385	30330	0.396917	3	C10	8,659,854	25
rs1090540	0.7733530	1.016	1837	0.106696	30378	0.105191	2	C10	8,663,595	23
rs7921734	0.3601882	1.0528	1836	0.103486	30381	0.098812	2	C10	8,668,422	40
rs7069464	0.4623737	1.0515	1836	0.067266	30375	0.064181	3	C10	8,677,428	38
rs4747806	0.0000274	1.1555	1837	0.595808	30379	0.560568	4	C10	8,678,241	35
rs1090541	0.0328102	1.0772	1837	0.39657	30382	0.378925	2	C10	8,685,729	24
rs1241208	0.0297294	1.1385	1835	0.092916	30225	0.082548	1	C10	8,691,311	27
rs6602329	0.0004012	1.1281	1836	0.541939	30380	0.511899	4	C10	8,700,101	36
rs4454616	0.0008570	1.1202	1837	0.47877	30369	0.450542	1	C10	8,705,153	33
rs4584486	0.6199739	1.0318	1837	0.079477	30388	0.077218	4	C10	8,708,704	34
rs1050836	0.1309007	1.0717	1837	0.167665	30388	0.15822	1	C10	8,714,527	20
rs7083359	1.76E-06	1.1923	1837	0.704954	30380	0.6671	2	C10	8,722,685	39
rs827402	0.7315722	1.0268	1829	0.051668	30287	0.050385	4	C10	8,726,332	41
rs827405	0.0002462	1.1624	1837	0.228634	30375	0.203177	4	C10	8,729,009	42
rs1334553	0.6242464	1.0218	1837	0.183723	30377	0.180515	3	C10	8,733,175	30
rs706771	0.0249670	1.0848	1837	0.33043	30365	0.312679	1	C10	8,736,452	37
rs1334549	4.86E-08	1.2428	1834	0.773991	30247	0.733726	1	C10	8,744,989	29
rs2184380	2.13E-08	1.2496	1836	0.773693	30351	0.732332	3	C10	8,754,080	2

SNP	P-value	OR	Aff	Aff.freq	Contrl.	Con.freq	Allele	Chr	Pos in Build 36	Seq ID No
rs1537603	0.0063633	1.0978	1833	0.455265	30293	0.432229	3	C10	8,774,301	31
rs1225037	5.32E-06	1.1696	1837	0.593087	30380	0.554789	4	C10	8,802,201	26
rs1075214	0.4819822	1.0344	1837	0.147795	30375	0.143588	1	C10	8,815,736	21
rs4749829	0.9501223	1.0042	1828	0.068107	30108	0.067839	3	C10	9,078,140	61
rs1912584	0.2426291	1.0945	1835	0.052589	30237	0.048269	4	C10	9,081,164	56
rs1775551	0.2937532	1.0477	1837	0.82417	30336	0.817313	1	C10	9,093,049	55
rs962993	0.2774384	1.0386	1836	0.393519	30288	0.384525	4	C10	9,093,138	67
rs7916897	0.2492789	1.0426	1832	0.674945	30120	0.66572	3	C10	9,101,110	64
rs2197415	0.2826170	1.0382	1834	0.39313	30349	0.38423	4	C10	9,102,862	57
rs1444788	0.2798396	1.0491	1837	0.82417	30368	0.81711	4	C10	9,104,126	54
rs2476175	0.8776101	1.0064	1837	0.217746	30376	0.216668	4	C10	9,114,133	58
rs2797266	0.3487364	1.0331	1836	0.396786	30364	0.389013	2	C10	9,129,166	59
rs860419	0.2755681	1.0382	1835	0.432425	30218	0.423258	4	C10	9,132,868	65
rs1276604	0.0466943	1.119	1835	0.104087	30349	0.094056	2	C10	9,133,679	52
rs1079569	0.1235447	1.064	1810	0.767956	30063	0.756711	2	C10	9,148,549	44
rs7067548	0.1945938	1.0599	1834	0.829062	30143	0.820655	1	C10	9,150,794	62
rs961945	0.1230478	1.0542	1834	0.445747	30375	0.432741	3	C10	9,166,649	66
rs1278142	0.0040593	1.1397	1837	0.173925	30379	0.15593	3	C10	9,167,711	53
rs2224696	0.0010448	1.1626	1837	0.170659	30377	0.150377	4	C10	9,168,781	3
rs1049094	0.4748388	1.0506	1836	0.936275	30382	0.933266	2	C10	9,183,517	43
rs1125606	0.0894290	1.0937	1837	0.120849	30372	0.111649	2	C10	9,187,738	45
rs372712	0.0755905	1.0628	1837	0.43767	30369	0.422734	1	C10	9,194,043	60
rs7898051	0.6463330	1.0192	1837	0.788242	30318	0.785045	3	C10	9,204,690	63
rs1149923	0.0228098	1.0815	1796	0.496381	28769	0.476815	4	C10	9,222,722	46
rs1149928	0.4126463	1.0305	1836	0.314815	30374	0.308372	3	C10	9,231,471	47
rs1149938	0.9114461	1.0038	1834	0.493184	30279	0.492239	4	C10	9,240,108	48
rs1149945	0.8799003	1.0052	1837	0.43767	30368	0.436397	3	C10	9,247,627	49
rs1149952	0.4948711	1.0248	1802	0.358213	29500	0.35261	2	C10	9,256,962	50
rs1242987	0.9668504	1.0014	1835	0.415259	30374	0.414911	3	C10	9,263,430	51
rs1540210	0.4912002	1.0262	1836	0.288126	30368	0.282847	3	C11	74,887,046	78
rs1790144	0.1544257	1.0705	1837	0.151062	30376	0.14253	4	C11	74,889,549	79
rs1793399	0.0609546	1.0906	1837	0.164671	30385	0.153085	2	C11	74,898,153	80
rs618202	0.0017623	1.1675	1837	0.142896	30362	0.124959	4	C11	74,916,133	86
rs633473	0.0053293	1.113	1836	0.740741	30362	0.71965	2	C11	74,930,524	87
rs1229102	0.0007803	1.1935	1837	0.12466	30384	0.106602	3	C11	74,932,878	5
rs1938800	0.1182986	1.0798	1836	0.141885	30372	0.132787	1	C11	74,933,770	81
rs662279	0.6695122	1.0159	1837	0.303756	30370	0.300428	2	C11	74,939,444	90

SNP	P-value	OR	Aff	Aff.freq	Contrl.	Con.freq	Allele	Chr	Pos in Build 36	Seq ID No
rs4945015	0.0274668	1.2223	1800	0.965556	29750	0.958218	1	C11	74,944,615	82
rs608585	0.9731233	1.0015	1836	0.826797	30383	0.826581	3	C11	74,946,352	84
rs640649	0.0066177	1.1668	1837	0.104246	30371	0.090695	4	C11	74,949,499	88
rs652054	0.8326146	1.008	1836	0.714052	30368	0.712428	2	C11	74,953,180	89
rs609309	0.2081718	1.0523	1837	0.774633	30209	0.765616	3	C11	74,964,989	85
rs585433	0.3707179	1.0312	1834	0.43675	30358	0.429211	1	C11	74,970,249	83
rs3794018	0.1981440	1.0455	1835	0.414169	30364	0.403422	2	C11	8,054,010	74
rs4758040	0.0129157	1.0895	1834	0.588604	30343	0.567709	4	C11	8,057,749	76
rs4578424	0.1068004	1.0722	1837	0.81301	30373	0.802176	2	C11	8,064,861	75
rs1881237	0.0243448	1.1965	1835	0.955041	30380	0.946675	4	C11	8,067,055	71
rs2242503	0.0001113	1.1522	1837	0.697877	30352	0.667205	2	C11	8,075,048	4
rs2242504	0.0277748	1.1826	1836	0.950436	30375	0.941909	2	C11	8,079,519	73
rs2049684	0.0030921	1.1133	1837	0.683179	30387	0.659509	3	C11	8,127,695	72
rs1083998	0.0239367	1.0939	1837	0.765378	30383	0.748873	1	C11	8,137,470	68
rs1104179	0.0101003	1.0987	1837	0.692433	30363	0.672035	4	C11	8,168,818	69
rs1881229	0.3637320	1.0428	1837	0.840501	30372	0.834798	3	C11	8,184,540	70
rs7951027	0.7460144	1.0117	1837	0.336962	30383	0.334365	1	C11	8,190,452	77
rs737387	0.7483800	1.011	1835	0.452316	30359	0.449603	1	C14	68,037,135	111
rs757369	0.2989297	1.0362	1836	0.559913	30378	0.551139	3	C14	68,043,299	112
rs1541390	0.6290307	1.021	1837	0.195427	30382	0.192186	3	C14	68,049,682	98
rs1023529	0.3683977	1.0323	1834	0.366957	30379	0.359607	2	C14	68,059,561	92
rs8012659	0.6278097	1.0364	1837	0.056886	30383	0.054998	4	C14	68,060,343	114
rs916962	0.3596030	1.0387	1836	0.215959	30363	0.209597	1	C14	68,063,331	115
rs1162127	0.0099335	1.1003	1837	0.707403	30367	0.687226	1	C14	68,066,219	96
rs1013444	0.6226353	1.0171	1837	0.418345	30370	0.414225	3	C14	68,067,388	91
rs727392	0.6708335	1.0309	1837	0.060425	30382	0.058719	2	C14	68,071,554	110
rs1115875	0.0561406	1.0802	1835	0.234877	30355	0.221298	2	C14	68,076,508	94
rs1710581	0.8079818	1.0104	1837	0.803484	30382	0.80184	4	C14	68,078,080	99
rs2842331	0.2212033	1.0527	1808	0.213772	29730	0.205264	2	C14	68,080,443	107
rs2158358	0.3127803	1.0405	1835	0.251226	30275	0.243832	2	C14	68,081,725	101
rs2842329	0.3199784	1.04	1827	0.250137	30229	0.242846	2	C14	68,087,348	106
rs2244814	0.3779656	1.0343	1822	0.276345	29822	0.26965	3	C14	68,095,623	102
rs2525503	0.0007823	1.1213	1837	0.550898	30365	0.522427	3	C14	68,098,356	105
rs1710583	0.0138473	1.1027	1837	0.247142	30373	0.229398	1	C14	68,098,494	100
rs1115875	0.8363170	1.0071	1837	0.468699	30376	0.466948	1	C14	68,101,178	95
rs999737	0.0017229	1.1326	1836	0.766885	30385	0.743887	2	C14	68,104,435	6
rs6573841	0.3860408	1.0386	1837	0.816277	30382	0.810529	2	C14	68,107,274	109

SNP	P-value	OR	Aff	Aff.freq	Contrl.	Con.freq	Allele	Chr	Pos in Build 36	Seq ID No
rs8009944	0.0319682	1.0851	1833	0.28096	30307	0.264757	2	C14	68,109,341	113
rs2285883	0.2432548	1.0508	1836	0.203431	30344	0.195525	1	C14	68,113,723	104
rs1290999	0.0505573	1.0811	1835	0.243869	30359	0.229784	1	C14	68,115,629	97
rs2256224	0.3955036	1.0321	1836	0.704793	30374	0.698179	1	C14	68,122,500	103
rs1048381	0.0804077	1.1317	1837	0.063963	30381	0.056943	1	C14	68,125,552	93
rs4899246	0.4842718	1.0287	1836	0.773148	30366	0.768145	1	C14	68,129,955	108
rs559325	0.5198271	1.0224	1837	0.581383	30369	0.575982	4	C18	32,116,417	121
rs1296146	0.7436940	1.0159	1833	0.14539	30306	0.143437	3	C18	32,119,988	117
rs4799863	0.5350597	1.0215	1837	0.434404	30382	0.429185	3	C18	32,127,413	119
rs9956546	0.0000702	1.2177	1836	0.144608	30375	0.121909	1	C18	32,136,446	7
rs680423	0.1990102	1.0817	1835	0.917711	30339	0.911582	4	C18	32,136,736	122
rs7243619	0.2091339	1.0454	1837	0.642352	30371	0.632083	4	C18	32,139,089	125
rs686784	0.5894600	1.0287	1837	0.120033	30374	0.117074	4	C18	32,142,076	123
rs492392	0.0498932	1.0841	1836	0.222767	30363	0.209103	1	C18	32,142,779	120
rs7240720	0.0271451	1.1376	1837	0.097441	30354	0.086677	3	C18	32,142,893	124
rs1272559	0.9364776	1.0031	1837	0.268645	30383	0.268045	3	C18	32,142,971	116
rs4531816	0.1177158	1.0929	1837	0.903375	30374	0.895338	4	C18	32,143,579	118
rs114607	0.4201611	1.0281	1835	0.426158	30375	0.419391	1	C22	38,706,329	126
rs2958658	0.7075628	1.0145	1837	0.731355	30360	0.728524	4	C22	38,718,953	127
rs3021222	0.0121019	1.0919	1837	0.62847	30368	0.607729	3	C22	38,722,550	128
rs3021274	0.0110828	1.0932	1837	0.631464	30372	0.610497	2	C22	38,725,030	129
rs1191292	0.0005138	1.1298	1837	0.383234	30385	0.35483	4	C22	38,733,117	8
rs6001722	0.6476742	1.0178	1836	0.739379	30361	0.735961	3	C22	38,758,379	130
rs9623117	0.0169431	1.1073	1837	0.808383	30374	0.792092	4	C22	38,782,065	134
rs6001749	0.0153071	1.1091	1837	0.808383	30382	0.791834	1	C22	38,805,272	131
rs6001762	0.0654330	1.085	1836	0.824891	30380	0.812788	2	C22	38,825,419	132
rs6001794	0.1618358	1.0563	1837	0.255035	30382	0.244767	4	C22	38,859,361	133
rs1108997	0.0832280	1.0697	1837	0.260751	30365	0.247966	4	C22	38,873,554	136
rs9611280	0.6602329	1.0241	1837	0.890038	30384	0.887688	3	C22	38,882,065	169
rs1108997	0.8500487	1.0072	1837	0.272727	30378	0.271298	4	C22	38,892,173	137
rs9611286	0.9363221	1.0031	1834	0.272083	30286	0.271479	2	C22	38,914,908	170
rs8137636	0.0133127	1.0931	1836	0.669662	30296	0.64969	1	C22	38,929,577	166
rs138021	0.8870477	1.0054	1837	0.272455	30365	0.271382	3	C22	38,942,920	147
rs739182	0.0242520	1.0817	1835	0.397275	30343	0.378638	4	C22	38,945,222	163
rs138027	0.8504863	1.0072	1837	0.272727	30372	0.271303	3	C22	38,946,058	148
rs138039	0.0145794	1.0918	1836	0.66939	30368	0.649681	3	C22	38,951,234	149
rs1305673	0.6304407	1.0295	1837	0.087099	30385	0.084812	4	C22	38,976,018	140

SNP	P-value	OR	Aff	Aff.freq	Contrl.	Con.freq	Allele	Chr	Pos in Build 36	Seq ID No
rs9611302	0.6750492	1.023	1837	0.890038	30378	0.887797	3	C22	38,980,230	171
rs733381	0.1003993	1.0658	1837	0.262657	30348	0.250494	3	C22	38,999,594	162
rs139921	0.0014652	1.1167	1834	0.607688	30282	0.581088	3	C22	39,056,708	150
rs470113	0.0017857	1.128	1835	0.27248	30377	0.249268	3	C22	39,059,560	155
rs1248469	0.0009138	1.1368	1836	0.270153	30368	0.24562	1	C22	39,066,418	139
rs8135371	0.4303623	1.037	1837	0.164126	30370	0.159203	2	C22	39,087,174	164
rs9611324	0.0008177	1.1381	1836	0.270697	30357	0.245924	2	C22	39,111,108	172
rs2072856	0.9278854	1.0031	1837	0.537289	30365	0.536522	3	C22	39,146,695	152
rs5757949	0.0045246	1.1225	1836	0.781863	30315	0.761521	4	C22	39,150,097	158
rs9306345	0.0014298	1.1507	1837	0.189439	30377	0.168812	3	C22	39,162,321	167
rs4507196	0.0163338	1.0897	1834	0.662214	30343	0.642735	1	C22	39,179,650	154
rs8135759	0.6212354	1.017	1836	0.535948	30354	0.531759	2	C22	39,188,950	165
rs9607721	0.0013221	1.1517	1837	0.189984	30380	0.16919	3	C22	39,192,066	168
rs6001932	0.0170002	1.1289	1837	0.133097	30380	0.119717	1	C22	39,207,581	161
rs3827382	0.0210574	1.1247	1835	0.132425	30372	0.119502	1	C22	39,211,349	153
rs5750948	0.5730917	1.0346	1837	0.914262	30353	0.911557	1	C22	39,216,690	156
rs9611367	0.0038990	1.1159	1837	0.731083	30360	0.708992	4	C22	39,244,895	173
rs6001954	0.0053978	1.1537	1829	0.131219	30092	0.115762	3	C22	39,251,626	9
rs5758001	0.0228856	1.0817	1833	0.423622	30282	0.40458	1	C22	39,276,253	159
rs5758008	0.0060923	1.1104	1837	0.735166	30371	0.714283	1	C22	39,289,206	160
rs5750960	0.4684336	1.0255	1836	0.603214	30342	0.597176	2	C22	39,289,866	157
rs1022533	0.0084935	1.1059	1832	0.735535	30222	0.715489	2	C22	39,302,493	135
rs1700203	0.0175391	1.1304	1834	0.12868	30350	0.115552	4	C22	39,326,313	151
rs133036	0.4223034	1.0282	1837	0.604518	30371	0.59784	2	C22	39,342,384	141
rs133047	0.3858756	1.0467	1836	0.883715	30385	0.878937	2	C22	39,357,765	142
rs1215920	0.0507958	1.0818	1837	0.236799	30368	0.222883	2	C22	39,372,037	138
rs133067	0.5254346	1.0263	1837	0.780348	30374	0.775861	4	C22	39,404,114	143
rs133072	0.7468165	1.0113	1826	0.59885	30062	0.596151	3	C22	39,405,489	144
rs133074	0.6315619	1.0164	1834	0.505725	30331	0.501648	2	C22	39,408,419	145
rs133076	0.0627455	1.0659	1836	0.572712	30373	0.557024	2	C22	39,411,507	146

Table 2: Anchor SNPs resulting from the Illumina Infinium HumanHap300 chip found in association to breast cancer in Icelandic case- and control-samples and found to replicate in several Caucasian cohorts. Icelandic data shown in the table are from analysis of data using Illumina and Centaurus assays combined.

SNP	Allele	Chromosome	Pos in Build 36	P-value	OR	# Aff	Aff. freq	# Con	Con. freq	Seq ID No
rs2005154	4	C09	36,846,828	5.21E-04	1.19	2282	0.90	35552	0.89	1
rs2184380	3	C10	8,754,080	1.33E-06	1.19	2282	0.77	35928	0.73	2
rs2224696	4	C10	9,168,781	3.48E-04	1.16	2277	0.17	35631	0.15	3
rs2242503	2	C11	8,075,048	9.39E-04	1.11	2280	0.69	35582	0.67	4
rs12291026	3	C11	74,932,878	3.98E-04	1.18	2280	0.12	35635	0.11	5
rs999737	2	C14	68,104,435	8.79E-05	1.15	2279	0.77	35638	0.74	6
rs9956546	1	C18	32,136,446	2.11E-06	1.23	2282	0.15	36015	0.12	7
rs11912922	4	C22	38,733,117	3.33E-05	1.14	2285	0.38	35631	0.35	8
rs6001954	3	C22	39,251,626	1.28E-03	1.16	2267	0.13	35210	0.12	9

Table 3: Association with breast cancer for nine variants on Chromosomes 9,10,11,14,18 and 22 shown for individual and combined Caucasian cohorts.

POPULATION	P	OR	# Affected	Aff. freq	# Con	Con freq	Allele	SNP	Chromosome	Cand. Gene
AllCauc	1.80E-04	1.14	NA	NA	NA	NA	4	rs2005154	C09	PAX5
AllCaucexclCGEMS	8.30E-04	1.13	NA	NA	NA	NA	4	rs2005154	C09	PAX5
CGEMS	7.05E-02	1.22	1144	0.88	1140	0.85	4	rs2005154	C09	PAX5
Holland	1.02E-01	1.16	735	0.88	2017	0.86	4	rs2005154	C09	PAX5
Iceland	5.21E-04	1.19	2282	0.90	35552	0.89	4	rs2005154	C09	PAX5
Spain	9.92E-01	1.00	633	0.89	1502	0.89	4	rs2005154	C09	PAX5
Sweden	4.52E-01	1.07	827	0.86	1738	0.86	4	rs2005154	C09	PAX5
AllCauc	3.00E-04	1.11	NA	NA	NA	NA	3	rs2184380	C10	
AllCaucexclCGEMS	2.50E-04	1.11	NA	NA	NA	NA	3	rs2184380	C10	
CGEMS	8.81E-01	0.96	1144	0.79	1141	0.80	3	rs2184380	C10	
Holland	8.80E-01	1.01	741	0.80	2019	0.80	3	rs2184380	C10	
Iceland	1.33E-06	1.19	2282	0.77	35928	0.73	3	rs2184380	C10	
Spain	1.89E-01	0.88	459	0.75	986	0.77	3	rs2184380	C10	
Sweden	2.47E-01	1.09	820	0.82	1717	0.80	3	rs2184380	C10	
AllCauc	3.70E-04	1.12	NA	NA	NA	NA	4	rs2224696	C10	
AllCaucexclCGEMS	1.10E-03	1.12	NA	NA	NA	NA	4	rs2224696	C10	
CGEMS	1.25E-01	1.19	1144	0.16	1141	0.14	4	rs2224696	C10	
Holland	2.56E-01	0.90	738	0.12	2025	0.13	4	rs2224696	C10	
Iceland	3.48E-04	1.16	2277	0.17	35631	0.15	4	rs2224696	C10	
Spain	3.32E-01	1.10	632	0.15	1507	0.14	4	rs2224696	C10	
Sweden	6.07E-02	1.18	831	0.15	1743	0.13	4	rs2224696	C10	
AllCauc	2.20E-05	1.08	NA	NA	NA	NA	2	rs2242503	C11	TUB
AllCaucexclCGEMS	1.90E-03	1.08	NA	NA	NA	NA	2	rs2242503	C11	TUB
CGEMS	3.93E-03	1.08	1144	0.67	1140	0.65	2	rs2242503	C11	TUB
Holland	9.86E-01	1.00	739	0.67	2005	0.67	2	rs2242503	C11	TUB
Iceland	9.39E-04	1.11	2280	0.69	35582	0.67	2	rs2242503	C11	TUB
Spain	2.79E-01	1.09	636	0.74	1485	0.72	2	rs2242503	C11	TUB
Sweden	3.04E-01	1.07	814	0.67	1729	0.66	2	rs2242503	C11	TUB
AllCauc	7.00E-04	1.13	NA	NA	NA	NA	3	rs12291026	C11	SERPINH1
AllCaucexclCGEMS	1.60E-03	1.12	NA	NA	NA	NA	3	rs12291026	C11	SERPINH1
CGEMS	1.91E-01	1.18	1139	0.13	1138	0.11	3	rs12291026	C11	SERPINH1
Holland	8.65E-01	0.98	737	0.12	2012	0.12	3	rs12291026	C11	SERPINH1
Iceland	3.98E-04	1.18	2280	0.12	35635	0.11	3	rs12291026	C11	SERPINH1

Spain	9.31E-02	1.18	641	0.14	1518	0.12			s12291026	C11	SERPINH1
Sweden	7.97E-01	1.03	829	0.10	1740	0.10			s12291026	C11	SERPINH1
AllCauc	1.70E-07	1.15	NA	NA	NA	NA			rs999737	C14	RAD51L1
AllCaucexcluCGEMS	3.50E-06	1.14	NA	NA	NA	NA			rs999737	C14	RAD51L1
CGEMS	1.31E-02	1.19	1143	0.79	1141	0.76			rs999737	C14	RAD51L1
Holland	3.18E-01	1.08	740	0.78	2006	0.77			rs999737	C14	RAD51L1
Iceland	8.79E-05	1.15	2279	0.77	35638	0.74			rs999737	C14	RAD51L1
Spain	2.49E-02	1.21	638	0.81	1514	0.78			rs999737	C14	RAD51L1
Sweden	8.90E-02	1.13	831	0.80	1733	0.78			rs999737	C14	RAD51L1
AllCauc	3.90E-06	1.17	NA	NA	NA	NA			s9956546	C18	FHOD3
AllCaucexcluCGEMS	1.50E-05	1.17	NA	NA	NA	NA			s9956546	C18	FHOD3
CGEMS	1.04E-01	1.21	1144	0.13	1140	0.11			s9956546	C18	FHOD3
Holland	4.71E-01	1.07	740	0.12	2022	0.11			s9956546	C18	FHOD3
Iceland	2.11E-06	1.23	2282	0.15	36015	0.12			s9956546	C18	FHOD3
Spain	7.07E-01	1.04	458	0.14	993	0.13			s9956546	C18	FHOD3
Sweden	2.56E-01	1.11	819	0.12	1732	0.11	1		rs9956546	C18	FHOD3
AllCauc	5.00E-06	1.11	NA	NA	NA	NA	4		rs11912922	C22	TNRC6B
AllCaucexcluCGEMS	3.40E-05	1.11	NA	NA	NA	NA	4		rs11912922	C22	TNRC6B
CGEMS	4.72E-02	1.16	1143	0.32	1141	0.29	4		rs11912922	C22	TNRC6B
Holland	4.80E-01	0.96	741	0.32	2011	0.33	4		rs11912922	C22	TNRC6B
Iceland	3.33E-05	1.14	2285	0.38	35631	0.35	4		rs11912922	C22	TNRC6B
Spain	3.77E-02	1.16	639	0.34	1529	0.30	4		rs11912922	C22	TNRC6B
Sweden	6.66E-02	1.12	826	0.36	1702	0.33	4		rs11912922	C22	TNRC6B
AllCauc	8.90E-06	1.17	NA	NA	NA	NA	3		rs6001954	C22	TNRC6B
AllCaucexcluCGEMS	3.70E-05	1.17	NA	NA	NA	NA	3		rs6001954	C22	TNRC6B
CGEMS	9.30E-02	1.22	1101	0.11	1105	0.09	3		rs6001954	C22	TNRC6B
Holland	1.35E-01	1.16	741	0.12	2017	0.10	3		rs6001954	C22	TNRC6B
Iceland	1.28E-03	1.16	2267	0.13	35210	0.12	3		rs6001954	C22	TNRC6B
Spain	6.68E-02	1.22	640	0.11	1533	0.10	3		rs6001954	C22	TNRC6B
Sweden	1.26E-01	1.16	825	0.12	1724	0.10	3		rs6001954	C22	TNRC6B

Allelic Odds Ratios calculated under the multiplicative model. All P values are two sided and have been adjusted for relatedness and other potential

Table 4: Surrogate markers for the nine Anchor SNPs. Markers with values of R^2 greater than 0.2 to an Anchor marker in the HapMap CEU dataset (<http://www.hapmap.org>) in a 1Mb interval flanking the marker were selected. Shown is the name of the correlated SNP, values for R^2 and D' to that marker and the corresponding P-value, as well as the position of the surrogate marker in NCBI Build 36 and a reference to the sequence ID containing flanking sequences for that marker.

Anchor	Surrogate	D'	R^2	P-value	Chromosome	Pos Build 36	Seq ID
rs2005154	rs2005154	1	1	-	C09	36,846,828	1
rs2005154	rs4878662	1	1	1.07E-16	C09	36,851,331	174
rs2005154	rs4880019	1	0.900826	3.20E-15	C09	36,853,450	175
rs2184380	rs2184380	1	1	-	C10	8,754,080	2
rs2184380	rs10466295	1	1	1.53E-28	C10	8,747,556	176
rs2184380	rs10508363	1	1	1.53E-28	C10	8,736,508	177
rs2184380	rs10508364	0.948936	0.856732	1.01E-22	C10	8,808,518	178
rs2184380	rs10508365	0.948936	0.856732	1.01E-22	C10	8,808,985	179
rs2184380	rs10795670	1	0.209877	4.71E-09	C10	8,765,045	180
rs2184380	rs10905411	0.874714	0.387964	3.45E-11	C10	8,678,685	181
rs2184380	rs10905414	0.870352	0.356105	3.50E-10	C10	8,680,134	182
rs2184380	rs10905415	0.91879	0.237472	4.68E-08	C10	8,680,223	183
rs2184380	rs10905430	0.844581	0.2373	1.96E-07	C10	8,704,489	184
rs2184380	rs10905437	1	0.304695	1.26E-11	C10	8,735,801	185
rs2184380	rs10905439	0.947367	0.897401	6.59E-23	C10	8,753,333	186
rs2184380	rs10905440	1	0.951417	1.10E-26	C10	8,755,150	187
rs2184380	rs10905443	1	1	1.53E-28	C10	8,757,369	188
rs2184380	rs10905444	1	1	1.53E-28	C10	8,757,533	189
rs2184380	rs10905445	1	1	1.53E-28	C10	8,758,377	190
rs2184380	rs10905446	1	1	1.53E-28	C10	8,758,390	191
rs2184380	rs10905447	1	1	1.53E-28	C10	8,758,955	192
rs2184380	rs10905454	0.895085	0.796777	4.82E-19	C10	8,807,933	193
rs2184380	rs11255764	0.885251	0.566844	9.13E-15	C10	8,676,745	194
rs2184380	rs11255776	0.844286	0.245302	2.21E-07	C10	8,701,838	195
rs2184380	rs11255777	0.847467	0.247154	1.12E-07	C10	8,702,017	196
rs2184380	rs11255778	0.848675	0.247226	9.24E-08	C10	8,703,356	197
rs2184380	rs11255779	0.840039	0.241626	3.17E-07	C10	8,703,789	198
rs2184380	rs11255790	0.937236	0.429299	1.11E-12	C10	8,722,186	199
rs2184380	rs11255795	1	1	1.93E-28	C10	8,731,572	200
rs2184380	rs11255797	1	1	1.53E-28	C10	8,737,159	201
rs2184380	rs11255800	1	1	3.91E-27	C10	8,742,581	202
rs2184380	rs11255804	0.94445	0.792459	3.17E-19	C10	8,747,207	203
rs2184380	rs11255805	1	1	1.53E-28	C10	8,747,459	204
rs2184380	rs11255820	1	1	1.53E-28	C10	8,759,525	205
rs2184380	rs11255821	1	1	9.41E-28	C10	8,759,540	206
rs2184380	rs11255822	1	1	2.45E-28	C10	8,759,613	207
rs2184380	rs11255832	1	0.949367	2.82E-26	C10	8,770,466	208
rs2184380	rs11255836	1	1	1.53E-28	C10	8,773,212	209
rs2184380	rs11255840	1	0.95	4.16E-26	C10	8,779,199	210
rs2184380	rs11255858	0.948936	0.856732	1.01E-22	C10	8,790,639	211
rs2184380	rs11255862	0.948936	0.856732	1.01E-22	C10	8,793,888	212
rs2184380	rs11255869	0.948925	0.855817	1.27E-22	C10	8,801,802	213
rs2184380	rs11255870	0.948936	0.856732	1.01E-22	C10	8,803,268	214
rs2184380	rs11255871	0.948936	0.856732	1.01E-22	C10	8,803,763	215
rs2184380	rs11255882	0.884394	0.542127	3.04E-14	C10	8,810,853	216
rs2184380	rs11255884	0.933202	0.665922	4.29E-14	C10	8,811,968	217
rs2184380	rs12049705	0.885669	0.565013	5.63E-15	C10	8,679,446	218
rs2184380	rs12218610	0.856757	0.38242	1.22E-09	C10	8,706,615	219
rs2184380	rs12250379	0.926804	0.305824	1.61E-09	C10	8,802,201	26
rs2184380	rs12259226	0.946719	0.85359	2.31E-21	C10	8,806,948	220
rs2184380	rs1325874	1	1	1.53E-28	C10	8,740,707	221
rs2184380	rs1334549	1	1	1.53E-28	C10	8,744,989	29
rs2184380	rs1334550	1	1	1.53E-28	C10	8,744,845	222
rs2184380	rs1334559	0.948936	0.856732	1.01E-22	C10	8,796,810	223

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs2184380	rs1360749	1	1	1.53E-28	C10	8,759,215	224
rs2184380	rs1413678	0.924198	0.284149	5.54E-09	C10	8,794,683	225
rs2184380	rs1413683	0.944299	0.850224	5.27E-20	C10	8,804,953	226
rs2184380	rs1537601	0.949468	0.901005	7.18E-24	C10	8,785,004	227
rs2184380	rs1537602	0.949474	0.9015	5.73E-24	C10	8,784,898	228
rs2184380	rs1537603	1	0.201238	7.61E-09	C10	8,774,301	31
rs2184380	rs17407711	1	1	1.53E-28	C10	8,769,001	229
rs2184380	rs17407781	1	1	1.93E-28	C10	8,773,088	230
rs2184380	rs17407830	0.943872	0.805805	5.02E-20	C10	8,775,487	231
rs2184380	rs17408204	0.948913	0.854883	1.59E-22	C10	8,790,006	232
rs2184380	rs17408337	0.948936	0.856732	1.01E-22	C10	8,791,931	233
rs2184380	rs17408580	0.946832	0.854684	4.55E-22	C10	8,794,975	234
rs2184380	rs17484150	0.874978	0.400207	3.72E-11	C10	8,707,041	235
rs2184380	rs17485426	1	1	1.53E-28	C10	8,743,013	236
rs2184380	rs17485998	1	1	1.53E-28	C10	8,759,878	237
rs2184380	rs17486082	1	1	1.53E-28	C10	8,767,474	238
rs2184380	rs17486795	0.948936	0.856732	1.01E-22	C10	8,792,167	239
rs2184380	rs17486816	0.948936	0.856732	1.01E-22	C10	8,792,191	240
rs2184380	rs1970170	0.949389	0.899443	2.68E-23	C10	8,785,426	241
rs2184380	rs1999638	1	0.95	4.16E-26	C10	8,783,743	242
rs2184380	rs2031561	0.874358	0.384902	4.27E-11	C10	8,683,505	243
rs2184380	rs2182292	0.872839	0.372331	7.65E-11	C10	8,683,484	244
rs2184380	rs2388821	1	0.318315	4.75E-12	C10	8,736,687	245
rs2184380	rs2388825	1	1	1.53E-28	C10	8,756,643	246
rs2184380	rs2388826	0.946678	0.850726	4.53E-21	C10	8,805,023	247
rs2184380	rs2892613	1	1	1.53E-28	C10	8,756,702	248
rs2184380	rs4112287	0.948936	0.856732	1.01E-22	C10	8,791,551	249
rs2184380	rs4112288	0.948936	0.856732	1.01E-22	C10	8,791,525	250
rs2184380	rs4345867	1	0.283927	4.34E-11	C10	8,736,698	251
rs2184380	rs4454616	0.911606	0.211521	4.36E-07	C10	8,705,153	33
rs2184380	rs4747806	0.870907	0.357423	1.65E-10	C10	8,678,241	35
rs2184380	rs4749805	1	1	7.43E-28	C10	8,751,692	252
rs2184380	rs4749807	1	1	1.53E-28	C10	8,751,763	253
rs2184380	rs4749812	1	0.201238	7.61E-09	C10	8,777,570	254
rs2184380	rs6602328	0.8255	0.21783	3.95E-06	C10	8,700,007	255
rs2184380	rs6602329	0.843632	0.234214	2.42E-07	C10	8,700,101	36
rs2184380	rs7069110	1	0.318315	4.75E-12	C10	8,735,184	256
rs2184380	rs7080765	0.922785	0.313611	2.97E-09	C10	8,797,824	257
rs2184380	rs7083359	0.937236	0.429299	1.11E-12	C10	8,722,685	39
rs2184380	rs7477023	1	1	1.53E-28	C10	8,753,932	258
rs2184380	rs7904921	0.948936	0.856732	1.01E-22	C10	8,792,331	259
rs2184380	rs7912413	1	1	2.03E-24	C10	8,762,126	260
rs2184380	rs7912704	1	0.201238	7.61E-09	C10	8,771,181	261
rs2184380	rs7912831	1	0.201238	7.61E-09	C10	8,771,261	262
rs2184380	rs827389	1	0.417761	1.41E-13	C10	8,740,261	263
rs2184380	rs9665623	0.948936	0.856732	1.01E-22	C10	8,797,122	264
rs2224696	rs2224696	1	1	-	C10	9,168,781	3
rs2224696	rs10905509	1	0.461682	4.95E-12	C10	9,160,368	265
rs2224696	rs11256045	0.807004	0.329629	6.26E-08	C10	9,157,239	266
rs2224696	rs12761213	0.88785	0.589831	1.21E-10	C10	9,106,345	267
rs2224696	rs12761461	0.906382	0.751312	1.87E-13	C10	9,158,901	268
rs2224696	rs12766048	1	0.66736	4.52E-12	C10	9,133,679	52
rs2224696	rs12772042	1	0.919893	4.43E-18	C10	9,164,748	269
rs2224696	rs12776383	1	0.727891	1.42E-12	C10	9,227,620	270
rs2224696	rs12778120	0.621966	0.221569	0.000014	C10	9,253,860	271
rs2224696	rs12780218	0.906382	0.751312	1.87E-13	C10	9,159,062	272
rs2224696	rs12781427	1	0.919192	4.98E-18	C10	9,167,711	53
rs2224696	rs1475189	1	0.830628	2.85E-15	C10	9,147,691	273
rs2224696	rs1573109	0.772073	0.519575	1.43E-08	C10	9,141,049	274
rs2224696	rs1573110	0.794588	0.591124	4.06E-10	C10	9,135,501	275
rs2224696	rs17145088	1	0.830628	2.85E-15	C10	9,116,440	276

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs2224696	rs17145095	1	0.830628	2.85E-15	C10	9,116,991	277
rs2224696	rs17145118	1	0.66736	5.30E-12	C10	9,121,203	278
rs2224696	rs17145120	1	0.830628	3.46E-15	C10	9,122,007	279
rs2224696	rs17145151	1	0.830628	2.85E-15	C10	9,130,711	280
rs2224696	rs17145164	1	0.830628	2.85E-15	C10	9,146,874	281
rs2224696	rs17145169	1	0.66736	4.90E-12	C10	9,147,141	282
rs2224696	rs17145188	0.621966	0.221569	0.000014	C10	9,148,524	283
rs2224696	rs17145193	0.883003	0.562854	1.15E-09	C10	9,155,959	284
rs2224696	rs17145221	1	0.89916	5.67E-15	C10	9,168,824	285
rs2224696	rs17363338	0.906382	0.751312	1.87E-13	C10	9,161,882	286
rs2224696	rs1775559	0.820762	0.495886	1.94E-10	C10	9,115,547	287
rs2224696	rs1857230	1	0.270147	1.73E-08	C10	9,127,246	288
rs2224696	rs1891532	1	0.209856	7.58E-08	C10	9,120,032	289
rs2224696	rs1935813	1	0.830628	2.85E-15	C10	9,118,559	290
rs2224696	rs2013364	1	1	1.16E-19	C10	9,168,159	291
rs2224696	rs2025289	0.906382	0.751312	1.87E-13	C10	9,161,060	292
rs2224696	rs2057442	1	0.242991	1.64E-08	C10	9,129,940	293
rs2224696	rs2093625	1	0.310553	1.63E-08	C10	9,126,798	294
rs2224696	rs2093626	1	0.252336	1.09E-08	C10	9,126,888	295
rs2224696	rs2146598	1	0.66736	4.52E-12	C10	9,118,227	296
rs2224696	rs2185817	0.796105	0.599318	1.59E-08	C10	9,140,970	297
rs2224696	rs2397336	1	0.267074	2.33E-08	C10	9,147,920	298
rs2224696	rs2760204	1	0.242991	1.64E-08	C10	9,139,981	299
rs2224696	rs2797266	1	0.242991	1.64E-08	C10	9,129,166	59
rs2224696	rs391733	0.780788	0.203685	0.00001	C10	9,191,976	300
rs2224696	rs4550140	1	0.25	1.23E-08	C10	9,145,853	301
rs2224696	rs7081544	1	0.830628	2.85E-15	C10	9,144,695	302
rs2224696	rs852273	1	0.276161	4.18E-09	C10	9,133,645	303
rs2224696	rs860418	1	0.245283	5.19E-08	C10	9,131,906	304
rs2224696	rs861172	1	0.240684	1.85E-08	C10	9,133,432	305
rs2224696	rs962993	0.885222	0.209643	1.52E-06	C10	9,093,138	67
rs2224696	rs965307	1	0.830628	2.85E-15	C10	9,119,005	306
rs2242503	rs2242503	1	1	-	C11	8,075,048	307
rs2242503	rs10431029	0.723593	0.468793	4.39E-13	C11	8,133,019	307
rs2242503	rs1055233	0.76309	0.517167	1.34E-14	C11	8,084,071	308
rs2242503	rs10734629	0.720552	0.46649	6.01E-13	C11	8,143,253	309
rs2242503	rs10743052	0.717443	0.464136	8.24E-13	C11	8,131,542	310
rs2242503	rs10743053	0.553357	0.295358	9.02E-09	C11	8,158,239	311
rs2242503	rs10743054	0.553357	0.295358	9.02E-09	C11	8,158,291	312
rs2242503	rs10743055	0.553357	0.295358	9.02E-09	C11	8,158,376	313
rs2242503	rs10769872	0.732431	0.432489	1.14E-11	C11	8,110,090	314
rs2242503	rs10769873	0.734884	0.322358	2.52E-09	C11	8,114,186	315
rs2242503	rs10769878	0.717939	0.348534	1.89E-08	C11	8,143,862	316
rs2242503	rs10769882	0.5497	0.229219	5.61E-06	C11	8,154,849	317
rs2242503	rs10839976	0.724865	0.308568	1.44E-08	C11	8,116,468	318
rs2242503	rs10839984	0.734884	0.322358	2.52E-09	C11	8,137,470	68
rs2242503	rs11041740	0.941774	0.461102	4.78E-14	C11	8,074,094	319
rs2242503	rs11041742	0.734884	0.322358	2.52E-09	C11	8,083,499	320
rs2242503	rs11041788	0.539249	0.212989	0.000025	C11	8,164,924	321
rs2242503	rs11041791	0.553934	0.235484	7.51E-07	C11	8,168,818	69
rs2242503	rs11041794	0.644371	0.240003	2.20E-06	C11	8,172,242	322
rs2242503	rs1108277	1	0.705954	1.08E-24	C11	8,077,794	323
rs2242503	rs12146654	0.64592	0.38833	1.53E-08	C11	8,152,036	324
rs2242503	rs12808387	0.553357	0.295358	9.02E-09	C11	8,159,414	325
rs2242503	rs1528125	0.553357	0.295358	9.02E-09	C11	8,156,792	326
rs2242503	rs1569128	0.723593	0.468793	4.39E-13	C11	8,089,461	327
rs2242503	rs1970880	0.544244	0.288701	1.80E-08	C11	8,160,092	328
rs2242503	rs1997262	0.522176	0.272668	3.97E-08	C11	8,158,988	329
rs2242503	rs2049684	0.723593	0.468793	4.39E-13	C11	8,127,695	72
rs2242503	rs2141321	0.789386	0.513662	2.21E-14	C11	8,086,752	330
rs2242503	rs2242501	0.737441	0.335833	2.13E-09	C11	8,074,679	331

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs2242503	rs2272383	0.63086	0.397985	3.43E-11	C11	8,080,075	332
rs2242503	rs3750955	0.784744	0.385656	1.72E-10	C11	8,082,818	333
rs2242503	rs3752898	0.900696	0.245488	5.00E-07	C11	8,073,496	334
rs2242503	rs3849986	0.522176	0.272668	3.97E-08	C11	8,159,303	335
rs2242503	rs3849990	0.665217	0.252525	3.28E-07	C11	8,175,019	336
rs2242503	rs3911309	0.554372	0.202829	2.81E-06	C11	8,173,902	337
rs2242503	rs3911310	0.554372	0.202829	2.81E-06	C11	8,174,324	338
rs2242503	rs4340037	0.782801	0.349689	2.40E-10	C11	8,084,734	339
rs2242503	rs4343012	0.734884	0.322358	2.52E-09	C11	8,089,751	340
rs2242503	rs4385931	0.770418	0.551609	6.82E-16	C11	8,066,224	341
rs2242503	rs4575312	0.669287	0.255625	1.73E-07	C11	8,065,299	342
rs2242503	rs4578424	0.642849	0.214843	2.18E-06	C11	8,064,861	75
rs2242503	rs4636658	0.590173	0.221903	1.11E-06	C11	8,178,862	343
rs2242503	rs4758040	0.834104	0.50807	1.52E-15	C11	8,057,749	76
rs2242503	rs4758042	0.717439	0.45831	2.74E-12	C11	8,085,872	344
rs2242503	rs4758287	0.63975	0.255278	1.71E-07	C11	8,056,489	345
rs2242503	rs4758309	0.544875	0.227845	2.57E-06	C11	8,165,308	346
rs2242503	rs4758310	0.553934	0.235484	7.51E-07	C11	8,165,969	347
rs2242503	rs7103334	0.546807	0.278288	2.54E-08	C11	8,157,501	348
rs2242503	rs7112519	0.535211	0.217411	3.18E-06	C11	8,169,345	349
rs2242503	rs7115706	0.544644	0.223928	1.40E-06	C11	8,169,421	350
rs2242503	rs7122690	0.522176	0.272668	3.97E-08	C11	8,159,397	351
rs2242503	rs7127738	0.527011	0.276798	1.88E-07	C11	8,154,963	352
rs2242503	rs7358396	0.71764	0.42377	2.13E-10	C11	8,141,858	353
rs2242503	rs7479156	0.551079	0.289269	1.57E-08	C11	8,157,993	354
rs2242503	rs7479738	0.553357	0.295358	9.02E-09	C11	8,157,931	355
rs2242503	rs7480804	1	0.236507	1.40E-08	C11	8,066,385	356
rs2242503	rs7481667	0.553357	0.295358	9.02E-09	C11	8,157,701	357
rs2242503	rs7481683	0.553357	0.295358	9.02E-09	C11	8,157,762	358
rs2242503	rs7482611	0.693202	0.457371	2.46E-10	C11	8,106,602	359
rs2242503	rs7927368	0.643226	0.370442	1.47E-10	C11	8,152,196	360
rs2242503	rs7940668	0.643226	0.370442	1.47E-10	C11	8,152,710	361
rs12291026	rs12291026	1	1	-	C11	74,932,878	5
rs12291026	rs1004856	0.795363	0.22224	3.36E-06	C11	74,858,615	362
rs12291026	rs10899091	1	1	1.43E-20	C11	74,933,440	363
rs12291026	rs11236449	1	0.919893	4.43E-18	C11	74,918,745	364
rs12291026	rs11236452	1	0.919893	4.98E-18	C11	74,927,534	365
rs12291026	rs11236454	1	0.906687	6.96E-16	C11	74,932,272	366
rs12291026	rs12362081	1	0.84127	3.43E-16	C11	74,899,292	367
rs12291026	rs1540210	0.800306	0.242864	1.38E-06	C11	74,887,046	78
rs12291026	rs1540211	0.750347	0.450541	1.04E-09	C11	74,887,332	368
rs12291026	rs1557471	0.79757	0.231043	2.21E-06	C11	74,857,322	369
rs12291026	rs1631470	0.799311	0.23844	1.70E-06	C11	74,864,790	370
rs12291026	rs1783551	0.837984	0.701135	2.54E-13	C11	74,908,860	371
rs12291026	rs1783556	0.876315	0.242246	0.000013	C11	74,897,710	372
rs12291026	rs1783559	0.820042	0.540279	1.73E-10	C11	74,894,676	373
rs12291026	rs1790144	0.750347	0.450541	1.04E-09	C11	74,889,549	79
rs12291026	rs1790152	0.79757	0.231043	2.21E-06	C11	74,862,499	374
rs12291026	rs1790307	0.830991	0.514897	7.92E-11	C11	74,898,758	375
rs12291026	rs1793396	0.895234	0.292441	6.02E-08	C11	74,890,048	376
rs12291026	rs1793397	0.750347	0.450541	1.04E-09	C11	74,890,155	377
rs12291026	rs1793398	0.906458	0.369032	1.79E-09	C11	74,893,069	378
rs12291026	rs1793399	0.831094	0.516953	7.12E-11	C11	74,898,153	80
rs12291026	rs1793414	0.792872	0.226475	4.83E-06	C11	74,860,335	379
rs12291026	rs1938800	1	0.748428	2.23E-16	C11	74,933,770	81
rs12291026	rs2853066	0.804722	0.264424	6.22E-07	C11	74,866,805	380
rs12291026	rs499613	0.792872	0.226475	4.83E-06	C11	74,872,133	381
rs12291026	rs504793	0.800306	0.242864	1.38E-06	C11	74,878,212	382
rs12291026	rs514477	0.800306	0.242864	1.38E-06	C11	74,875,650	383
rs12291026	rs549034	0.799814	0.24066	1.53E-06	C11	74,880,782	384
rs12291026	rs550881	0.800306	0.242864	1.38E-06	C11	74,880,612	385

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs12291026	rs581007	0.800306	0.242864	1.38E-06	C11	74,879,594	386
rs12291026	rs589724	0.805181	0.254846	6.34E-07	C11	74,876,829	387
rs12291026	rs600387	1	0.293756	1.11E-09	C11	74,935,675	388
rs12291026	rs606460	0.902797	0.333104	1.27E-08	C11	74,926,105	389
rs12291026	rs617617	1	0.425837	7.51E-12	C11	74,936,378	390
rs12291026	rs618202	0.838016	0.70227	2.28E-13	C11	74,916,133	86
rs12291026	rs628972	0.800306	0.242864	1.38E-06	C11	74,885,999	391
rs12291026	rs640649	0.793071	0.480582	5.31E-09	C11	74,949,499	88
rs12291026	rs662279	1	0.334073	2.08E-10	C11	74,939,444	90
rs12291026	rs667410	0.891296	0.289518	1.94E-07	C11	74,872,345	392
rs12291026	rs667531	0.721513	0.369086	1.84E-07	C11	74,950,364	393
rs12291026	rs670100	0.89343	0.292234	1.02E-07	C11	74,926,922	394
rs12291026	rs670491	0.902797	0.333104	1.27E-08	C11	74,926,833	395
rs12291026	rs682292	0.800306	0.242864	1.38E-06	C11	74,877,949	396
rs12291026	rs7128888	1	1	1.43E-20	C11	74,931,631	397
rs12291026	rs7129014	1	1	1.43E-20	C11	74,931,725	398
rs12291026	rs7129150	1	0.921773	7.97E-18	C11	74,931,825	399
rs12291026	rs947844	1	0.84127	3.43E-16	C11	74,906,532	400
rs999737	rs999737	1	1	-	C14	68,104,435	6
rs999737	rs10134446	1	0.22	1.39E-09	C14	68,067,388	91
rs999737	rs10138140	0.674626	0.269357	4.17E-08	C14	68,011,742	401
rs999737	rs10146772	0.659865	0.223224	3.27E-07	C14	67,948,122	402
rs999737	rs10467820	0.674626	0.269357	4.17E-08	C14	67,988,769	403
rs999737	rs10483812	1	0.914924	1.84E-27	C14	68,087,596	404
rs999737	rs10483813	1	1	1.04E-30	C14	68,101,037	405
rs999737	rs11158749	1	0.224097	1.07E-09	C14	68,067,049	406
rs999737	rs11158751	1	0.340659	4.73E-13	C14	68,101,178	95
rs999737	rs11621276	1	0.87513	3.15E-26	C14	68,066,219	96
rs999737	rs11624097	0.953506	0.831825	5.44E-24	C14	68,055,789	407
rs999737	rs11624164	0.953471	0.82933	9.30E-24	C14	68,055,910	408
rs999737	rs11624333	0.953506	0.831825	5.44E-24	C14	68,049,588	409
rs999737	rs11628293	1	0.954597	2.00E-28	C14	68,107,073	410
rs999737	rs11846360	0.528484	0.243633	1.31E-06	C14	67,997,698	411
rs999737	rs11847185	0.499138	0.226969	2.79E-06	C14	67,997,747	412
rs999737	rs11849916	0.499138	0.226969	2.79E-06	C14	67,957,791	413
rs999737	rs12878761	0.527934	0.255003	2.88E-07	C14	68,138,118	414
rs999737	rs12879200	0.739817	0.228092	4.95E-07	C14	68,137,863	415
rs999737	rs12886864	0.662607	0.269908	1.17E-07	C14	68,017,871	416
rs999737	rs12889251	1	0.876374	2.38E-26	C14	68,063,893	417
rs999737	rs12894230	0.544425	0.273137	2.77E-07	C14	68,137,669	418
rs999737	rs1468279	0.659722	0.26746	2.08E-07	C14	67,994,846	419
rs999737	rs1468280	1	0.310873	3.99E-12	C14	68,101,613	420
rs999737	rs1547012	0.575976	0.303524	1.93E-08	C14	68,133,183	421
rs999737	rs17105675	0.499138	0.226969	2.79E-06	C14	67,998,259	422
rs999737	rs17755657	0.511149	0.219511	2.63E-06	C14	67,950,390	423
rs999737	rs17755734	0.499138	0.226969	2.79E-06	C14	67,979,708	424
rs999737	rs17755752	0.499138	0.226969	2.79E-06	C14	67,979,904	425
rs999737	rs17755925	1	0.876374	2.38E-26	C14	68,081,699	426
rs999737	rs17756000	1	0.914924	1.84E-27	C14	68,089,291	427
rs999737	rs17828691	0.499138	0.226969	2.79E-06	C14	67,956,346	428
rs999737	rs17828721	0.499138	0.226969	2.79E-06	C14	67,979,791	429
rs999737	rs17828763	0.499138	0.226969	2.79E-06	C14	67,984,862	430
rs999737	rs17828907	1	0.914924	1.84E-27	C14	68,080,958	431
rs999737	rs17828955	1	0.914924	1.84E-27	C14	68,086,398	432
rs999737	rs1956534	0.587868	0.237655	4.20E-07	C14	67,963,825	433
rs999737	rs2074563	0.674626	0.269357	4.17E-08	C14	68,004,865	434
rs999737	rs2074565	0.674626	0.269357	4.17E-08	C14	68,013,900	435
rs999737	rs2097800	0.66475	0.25742	2.25E-07	C14	68,006,440	436
rs999737	rs2107340	0.674626	0.269357	4.17E-08	C14	68,000,418	437
rs999737	rs2145157	0.739817	0.228092	4.95E-07	C14	68,137,391	438
rs999737	rs2158357	1	0.956044	8.64E-29	C14	68,098,956	439

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs999737	rs2189517	1	0.219935	1.39E-09	C14	68,072,741	440
rs999737	rs2253317	0.923817	0.281194	1.03E-08	C14	68,122,156	441
rs999737	rs2257111	0.532356	0.267955	1.67E-07	C14	68,129,446	442
rs999737	rs2257116	0.533815	0.272433	1.21E-07	C14	68,129,544	443
rs999737	rs2257127	0.533815	0.272433	1.21E-07	C14	68,129,821	444
rs999737	rs2331701	0.669937	0.256122	8.50E-08	C14	67,952,384	445
rs999737	rs2331705	0.510976	0.244712	1.44E-06	C14	67,991,800	446
rs999737	rs2331775	0.739817	0.228092	4.95E-07	C14	68,136,385	447
rs999737	rs2525503	1	0.318681	1.88E-12	C14	68,098,356	105
rs999737	rs2525530	1	0.205178	4.00E-09	C14	68,084,205	448
rs999737	rs2842327	1	0.382208	4.73E-13	C14	68,098,285	449
rs999737	rs3784121	0.499138	0.226969	2.79E-06	C14	67,953,441	450
rs999737	rs4531674	0.527934	0.255003	2.88E-07	C14	68,136,496	451
rs999737	rs4899246	0.533815	0.272433	1.21E-07	C14	68,129,955	108
rs999737	rs4902604	0.561463	0.264852	2.16E-07	C14	68,124,238	452
rs999737	rs4902606	0.739817	0.228092	4.95E-07	C14	68,135,119	453
rs999737	rs4902608	0.739817	0.228092	4.95E-07	C14	68,138,576	454
rs999737	rs5004090	0.674626	0.269357	4.17E-08	C14	68,013,998	455
rs999737	rs6573837	0.77856	0.353726	5.97E-10	C14	67,999,862	456
rs999737	rs7140266	1	0.219935	1.39E-09	C14	68,074,558	457
rs999737	rs7146456	1	0.260433	1.04E-10	C14	68,108,734	458
rs999737	rs7153476	1	0.308233	3.65E-12	C14	68,102,983	459
rs999737	rs739874	0.952339	0.761982	7.28E-22	C14	68,044,163	460
rs999737	rs746663	0.61957	0.227187	5.55E-07	C14	67,954,305	461
rs999737	rs8007194	1	0.281975	2.77E-11	C14	68,105,178	462
rs999737	rs8010439	0.587868	0.237655	4.20E-07	C14	67,954,354	463
rs999737	rs8012610	0.659865	0.223224	3.27E-07	C14	67,972,552	464
rs999737	rs9323512	0.674626	0.269357	4.17E-08	C14	68,019,310	465
rs999737	rs9323513	1	0.220532	1.43E-09	C14	68,066,244	466
rs999737	rs9323514	1	0.315043	2.49E-12	C14	68,066,317	467
rs9956546	rs9956546	1	1	-	C18	32,136,446	47
rs9956546	rs16960059	1	0.66736	4.52E-12	C18	32,120,094	468
rs9956546	rs492392	1	0.399199	3.93E-11	C18	32,142,779	120
rs9956546	rs7239113	1	0.66736	4.52E-12	C18	32,133,395	469
rs9956546	rs7240720	0.524852	0.251926	0.000036	C18	32,142,893	124
rs9956546	rs9304157	1	0.211045	0.000182	C18	32,143,156	470
rs11912922	rs11912922	1	1	-	C22	38,733,117	48
rs11912922	rs11089967	1	1	9.56E-34	C22	38,739,085	471
rs11912922	rs11704971	1	1	1.57E-30	C22	38,734,239	472
rs11912922	rs11705454	1	1	6.56E-31	C22	38,731,841	473
rs11912922	rs17406386	1	1	1.37E-33	C22	38,735,844	474
rs11912922	rs17406434	1	1	3.28E-33	C22	38,737,520	475
rs11912922	rs2071771	1	0.214286	1.52E-09	C22	38,745,086	476
rs11912922	rs2958650	1	1	3.28E-33	C22	38,735,877	477
rs11912922	rs2958651	1	0.958241	2.48E-29	C22	38,735,709	478
rs11912922	rs2958659	1	1	9.56E-34	C22	38,730,858	479
rs11912922	rs7284488	1	0.239332	2.31E-10	C22	38,742,497	480
rs11912922	rs7285507	0.747862	0.497836	2.02E-12	C22	38,718,254	481
rs11912922	rs7291782	1	0.924812	2.56E-30	C22	38,745,563	482
rs11912922	rs739145	1	0.24812	1.20E-10	C22	38,728,791	483
rs11912922	rs9611246	1	0.214286	1.52E-09	C22	38,745,655	484
rs11912922	rs9611265	0.957937	0.846507	5.98E-24	C22	38,828,439	485
rs6001954	rs6001954	1	1	-	C22	39,251,626	49
rs6001954	rs10483203	0.922166	0.736052	7.16E-15	C22	39,159,899	486
rs6001954	rs10483204	1	0.865546	9.75E-19	C22	39,200,740	487
rs6001954	rs10483205	1	0.800222	4.35E-17	C22	39,213,545	488
rs6001954	rs10483206	1	0.865546	9.75E-19	C22	39,231,291	489
rs6001954	rs1106673	0.922166	0.736052	7.16E-15	C22	39,155,760	490
rs6001954	rs11913132	0.696801	0.304518	1.56E-07	C22	39,108,724	491
rs6001954	rs12158399	1	0.568501	2.92E-15	C22	39,378,887	492
rs6001954	rs12158872	0.922166	0.736052	7.16E-15	C22	39,162,447	493

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs6001954	rs12159200	1	0.542302	7.76E-15	C22	39,372,037	138
rs6001954	rs12159970	1	0.800222	4.35E-17	C22	39,149,100	494
rs6001954	rs12484697	0.624393	0.257388	1.79E-06	C22	39,066,418	139
rs6001954	rs12627881	0.925127	0.534379	1.11E-12	C22	39,374,539	495
rs6001954	rs133036	1	0.231068	6.07E-09	C22	39,342,384	141
rs6001954	rs133038	1	0.223301	9.26E-09	C22	39,345,215	496
rs6001954	rs16985899	1	0.932127	1.30E-20	C22	39,293,348	497
rs6001954	rs17001846	0.624393	0.257388	1.79E-06	C22	39,073,999	498
rs6001954	rs17001868	0.635691	0.328651	4.59E-07	C22	39,108,177	499
rs6001954	rs17001943	0.922166	0.736052	7.16E-15	C22	39,192,559	500
rs6001954	rs17001974	1	0.865546	9.75E-19	C22	39,207,330	501
rs6001954	rs17001977	1	0.779103	6.11E-16	C22	39,210,159	502
rs6001954	rs17001993	1	0.800222	4.35E-17	C22	39,230,023	503
rs6001954	rs17001994	1	0.788111	8.57E-16	C22	39,230,764	504
rs6001954	rs17001997	1	0.788111	4.07E-16	C22	39,235,018	505
rs6001954	rs17002019	1	0.932127	1.30E-20	C22	39,295,282	506
rs6001954	rs17002020	1	0.930025	1.15E-19	C22	39,295,389	507
rs6001954	rs17002026	1	0.930456	2.22E-19	C22	39,312,527	508
rs6001954	rs17002027	1	0.925154	2.26E-18	C22	39,314,517	509
rs6001954	rs17002034	1	0.865546	9.75E-19	C22	39,326,313	151
rs6001954	rs17002036	1	0.932127	1.30E-20	C22	39,327,057	510
rs6001954	rs17002038	1	0.925548	4.67E-19	C22	39,330,910	511
rs6001954	rs17002069	0.923881	0.509328	4.13E-12	C22	39,380,932	512
rs6001954	rs2075764	0.624393	0.257388	1.79E-06	C22	39,088,527	513
rs6001954	rs2187832	1	0.223301	9.26E-09	C22	39,328,395	514
rs6001954	rs2235318	0.628984	0.275382	9.11E-07	C22	39,130,490	515
rs6001954	rs2280790	0.624393	0.257388	1.79E-06	C22	39,075,127	516
rs6001954	rs2294348	0.694473	0.293289	2.75E-07	C22	39,113,726	517
rs6001954	rs2294350	0.624393	0.257388	1.79E-06	C22	39,114,224	518
rs6001954	rs2294352	0.922166	0.736052	7.16E-15	C22	39,157,265	519
rs6001954	rs2413624	1	0.260204	3.65E-09	C22	39,260,158	520
rs6001954	rs3788577	0.624393	0.257388	1.79E-06	C22	39,072,777	521
rs6001954	rs3788578	0.624393	0.257388	1.79E-06	C22	39,072,867	522
rs6001954	rs3788579	0.624453	0.267064	1.98E-06	C22	39,072,935	523
rs6001954	rs3827381	1	0.865546	9.75E-19	C22	39,211,348	524
rs6001954	rs3827382	1	0.800222	4.35E-17	C22	39,211,349	153
rs6001954	rs4140512	1	0.568501	2.92E-15	C22	39,362,809	525
rs6001954	rs4701113	0.624393	0.257388	1.79E-06	C22	39,059,560	155
rs6001954	rs5750957	1	0.265513	1.01E-09	C22	39,265,239	526
rs6001954	rs5750960	1	0.231068	6.07E-09	C22	39,289,866	157
rs6001954	rs5750966	1	0.213336	1.64E-08	C22	39,299,739	527
rs6001954	rs5757976	1	0.223301	9.26E-09	C22	39,245,664	528
rs6001954	rs5757998	1	0.265513	1.01E-09	C22	39,269,856	529
rs6001954	rs5758001	0.90345	0.224527	3.42E-07	C22	39,276,253	159
rs6001954	rs5995849	0.696801	0.304518	1.56E-07	C22	39,107,018	530
rs6001954	rs5995856	0.927089	0.801157	1.72E-16	C22	39,179,008	531
rs6001954	rs5995870	1	1	4.37E-23	C22	39,251,762	532
rs6001954	rs5995871	1	1	5.10E-23	C22	39,252,278	533
rs6001954	rs5995886	1	0.563253	6.45E-15	C22	39,363,747	534
rs6001954	rs6001900	0.693507	0.310492	4.91E-07	C22	39,120,948	535
rs6001954	rs6001910	1	0.800222	4.35E-17	C22	39,153,955	536
rs6001954	rs6001911	0.922166	0.736052	7.16E-15	C22	39,155,139	537
rs6001954	rs6001912	0.922166	0.736052	7.16E-15	C22	39,158,307	538
rs6001954	rs6001913	1	0.800222	4.35E-17	C22	39,166,699	539
rs6001954	rs6001930	1	0.865546	9.75E-19	C22	39,206,180	540
rs6001954	rs6001931	1	0.857313	9.26E-18	C22	39,207,460	541
rs6001954	rs6001932	1	0.865546	9.75E-19	C22	39,207,581	161
rs6001954	rs6001935	1	0.865546	9.75E-19	C22	39,216,939	542
rs6001954	rs6001950	1	0.865546	9.75E-19	C22	39,236,581	543
rs6001954	rs6001974	1	1	4.37E-23	C22	39,309,110	544
rs6001954	rs6001980	1	1	4.37E-23	C22	39,334,330	545

Anchor	Surrogate	D'	R ²	P-value	Chromosome	Pos Build 36	Seq ID
rs6001954	rs6001990	1	0.568501	2.92E-15	C22	39,364,131	546
rs6001954	rs6002000	1	0.568501	2.92E-15	C22	39,381,055	547
rs6001954	rs718193	1	0.932127	1.30E-20	C22	39,354,010	548
rs6001954	rs7292804	0.923881	0.509328	4.13E-12	C22	39,377,502	549
rs6001954	rs7293100	0.92505	0.531874	1.29E-12	C22	39,377,697	550
rs6001954	rs742140	0.624393	0.257388	1.79E-06	C22	39,111,532	551
rs6001954	rs760700	0.624393	0.257388	1.79E-06	C22	39,130,017	552
rs6001954	rs760701	0.628984	0.275382	9.11E-07	C22	39,130,464	553
rs6001954	rs9306345	0.780144	0.423648	8.54E-10	C22	39,162,321	167
rs6001954	rs932379	1	0.932127	1.30E-20	C22	39,282,480	554
rs6001954	rs9607721	0.851982	0.479217	3.33E-11	C22	39,192,066	168
rs6001954	rs9611310	0.624393	0.257388	1.79E-06	C22	39,063,723	555
rs6001954	rs9611311	0.694782	0.31869	2.91E-07	C22	39,063,897	556
rs6001954	rs9611312	0.61091	0.260998	7.16E-06	C22	39,064,736	557
rs6001954	rs9611316	0.673329	0.320213	8.89E-07	C22	39,066,973	558
rs6001954	rs9611318	0.624393	0.257388	1.79E-06	C22	39,069,360	559
rs6001954	rs9611324	0.696801	0.304518	1.56E-07	C22	39,111,108	172
rs6001954	rs9611325	0.624393	0.257388	1.79E-06	C22	39,112,436	560
rs6001954	rs9611328	0.624393	0.257388	1.79E-06	C22	39,121,918	561
rs6001954	rs9611329	0.624393	0.257388	1.79E-06	C22	39,121,984	562

Table 5: Association with breast cancer for rs 6001954 on Chromosome 22 is shown for individual cohorts including Taiwanese . Also presented are combined European results as well as results of all combined cohorts

Sample Set	SNP	Allele	P-value	OR	N cases	F cases	N controls	F controls	95% CI	Phet
ICELAND	rs6001954	3	7.20E-04	1.18				1.117		
DKFZ-Sweden	rs6001954	3	5.59E-01	1.06				1.130		
HOLLAND	rs6001954	3	2.90E-01	1.11				1.106		
MAYO-U.S.A.	rs6001954	3	7.29E-02	1.17				1.104		
SPAIN	rs6001954	3	2.99E-01	1.10				1.099		
SWEDEN-Karolinska	rs6001954	3	1.26E-01	1.16				1.105		
CGEMS	rs6001954	3	9.30E-02	1.22				1.091		
TAIWAN	rs6001954	3	1.02E-01	1.13				1.221		
ALL European	rs6001954	3	5.20E-06	1.15				N/A	(1.08, 1.22)	0.95
ALL Combined	rs6001954	3	1.30E-06	1.15				N/A	(1.08, 1.21)	0.98

WHAT IS CLAIMED IS

1. A method for determining a susceptibility to breast cancer in a human individual, comprising determining whether at least one allele of at least one polymorphic marker is present in a nucleic acid sample from the individual or in a genotype dataset from the individual, wherein the at least one polymorphic marker is selected from the group consisting of rs999737 (SEQ ID NO:6), and markers in linkage disequilibrium therewith, wherein linkage disequilibrium between markers is characterized by values of r^2 of at least 0.2, and wherein determination of the presence of the at least one allele is indicative of a susceptibility to breast cancer for the individual.
2. The method according to Claim 1, wherein the at least one polymorphic marker is selected from the markers rs999737, rs10134446, rs10138140, rs10146772, rs10467820, rs10483812, rs10483813, rs11158749, rs11158751, rs11621276, rs11624097, rs11624164, rs11624333, rs11628293, rs11846360, rs11847185, rs11849916, rs12878761, rs12879200, rs12886864, rs12889251, rs12894230, rs1468279, rs1468280, rs1547012, rs17105675, rs17755657, rs17755734, rs17755752, rs17755925, rs17756000, rs17828691, rs17828721, rs17828763, rs17828907, rs17828955, rs1956534, rs2074563, rs2074565, rs2097800, rs2107340, rs2145157, rs2158357, rs2189517, rs2253317, rs2257111, rs2257116, rs2257127, rs2331701, rs2331705, rs2331775, rs2525503, rs2525530, rs2842327, rs3784121, rs4531674, rs4899246, rs4902604, rs4902606, rs4902608, rs5004090, rs6573837, rs7140266, rs7146456, rs7153476, rs739874, rs746663, rs8007194, rs8010439, rs8012610, rs9323512, rs9323513, and rs9323514.
3. The method according to Claim 1 or Claim 2, wherein the at least one polymorphic marker is rs999737 (SEQ ID NO:6).
4. The method according to any one of the preceding Claims, further comprising assessing the frequency of at least one haplotype in the individual.
5. The method of any one of the preceding claims, wherein the susceptibility conferred by the presence of the at least one allele or haplotype is increased susceptibility.
6. The method according to Claim 6, wherein the presence of allele C in rs999737 is indicative of increased susceptibility to breast cancer in the individual.
7. The method according to Claim 5 or 6, wherein the presence of the at least one allele or haplotype is indicative of increased susceptibility to breast cancer with a relative risk (RR) or odds ratio (OR) of at least 1.08.

8. The method according to Claim 5 or 6, wherein the presence of the at least one allele or haplotype is indicative of increased susceptibility with a relative risk (RR) or odds ratio (OR) of at least 1.09, including at least 1.10, at least 1.11, at least 1.12, at least 1.13, at least 1.14, at least 1.15, at least 1.16, and at least 1.17.
9. The method according to any one of claims 1-4, wherein the susceptibility conferred by the presence of the at least one allele or haplotype is decreased susceptibility.
10. The method of any one of the preceding claims, further comprising determining whether at least one at-risk allele of at least one at-risk variant for breast cancer not in linkage disequilibrium with rs999737 is present in a sample comprising genomic DNA from a human individual or a genotype dataset derived from a human individual.
11. The method of any one of claims 1-9, comprising determining whether at least one allele in each of at least two polymorphic markers is present in a sample comprising genomic DNA from a human individual or a genotype dataset derived from a human individual, wherein the presence of the at least one allele in the at least two polymorphic markers is indicative of an increased susceptibility to breast cancer.
12. The method according to any one of the preceding claims, further comprising a step of determining the presence or absence of at least one high penetrant genetic factor for breast cancer in a nucleic acid sample obtained from the individual or in a genotype dataset derived from the individual.
13. The method according to claim 12, wherein the high penetrant genetic factor is a mutation in BRCA1, BRCA2, TP53 and/or PTEN.
14. The method of any one of the preceding Claims, further comprising analyzing non-genetic information selected from age, gender, ethnicity, socioeconomic status, previous disease diagnosis, medical history of subject, family history of breast cancer, biochemical measurements, and clinical measurements to make risk assessment, diagnosis, or prognosis of the individual.
15. The method of any one of the Claims 11 – 14, further comprising calculating combined risk.
16. A computer program for determining a genetic indicator for breast cancer in a human individual:

wherein the computer program is adapted to be executed on the processor to analyze marker and/or haplotype information for at least one human individual with respect to

at least one polymorphic marker selected from rs999737 (SEQ ID NO:6), and markers in linkage disequilibrium therewith, wherein linkage disequilibrium is characterized by values of $r^2 > 0.2$, and

generate an output based on the marker or haplotype information, wherein the output comprises a risk measure of the at least one marker or haplotype as a genetic indicator of breast cancer for the human individual.

17. The computer program according to Claim 16, wherein the computer readable memory further comprises data indicative of the risk of developing breast cancer associated with at least one allele of the at least one polymorphic marker or at least one haplotype, and wherein a risk measure for the human individual is based on a comparison of the genotype status for the human individual to the risk of developing breast cancer associated with the at least one allele or the at least one haplotype.
18. The computer program according to Claim 17, wherein the computer readable memory further comprises data indicative of the frequency of the at least one allele of at least one polymorphic marker or the at least one haplotype in a plurality of individuals diagnosed with breast cancer, and data indicative of the frequency of the at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference individuals, and wherein risk of developing breast cancer associated with the at least one allele or haplotype is based on a comparison of the frequency of the at least one allele or haplotype in individuals diagnosed with breast cancer and reference individuals.
19. The computer program according to any one of claims 16-18, wherein the at least one marker or haplotype comprises at least one marker selected from the markers rs999737, rs10134446, rs10138140, rs10146772, rs10467820, rs10483812, rs10483813, rs11158749, rs11158751, rs11621276, rs11624097, rs11624164, rs11624333, rs11628293, rs11846360, rs11847185, rs11849916, rs12878761, rs12879200, rs12886864, rs12889251, rs12894230, rs1468279, rs1468280, rs1547012, rs17105675, rs17755657, rs17755734, rs17755752, rs17755925, rs17756000, rs17828691, rs17828721, rs17828763, rs17828907, rs17828955, rs1956534, rs2074563, rs2074565, rs2097800, rs2107340, rs2145157, rs2158357, rs2189517, rs2253317, rs2257111, rs2257116, rs2257127, rs2331701, rs2331705, rs2331775, rs2525503, rs2525530, rs2842327, rs3784121, rs4531674, rs4899246, rs4902604, rs4902606, rs4902608, rs5004090, rs6573837, rs7140266, rs7146456, rs7153476, rs739874, rs746663, rs8007194, rs8010439, rs8012610, rs9323512, rs9323513, and rs9323514.

20. The computer program according to any one of Claims 16-19, wherein the at least one marker or haplotype is rs999737 (SEQ ID NO:6).
21. The computer program according to any one of the Claims 16-20, wherein the risk measure is characterized by an Odds Ratio (OR) or a Relative Risk (RR).
22. A method of identification of a marker for use in assessing susceptibility to breast cancer, the method comprising:
 - a. identifying at least one polymorphic marker in linkage disequilibrium with rs999737 (SEQ ID NO:6), wherein linkage disequilibrium is characterized by values of $r^2 > 0.2$;
 - b. determining the genotype status of samples from a sample of individuals diagnosed with, or having a susceptibility to, breast cancer, for the at least one polymorphic marker; and
 - c. determining the genotype status of samples from a sample of control individuals for the at least one polymorphic marker;

wherein a significant difference in frequency of at least one allele in at least one polymorphism in individuals diagnosed with, or having a susceptibility to, breast cancer, as compared with the frequency of the at least one allele in the control sample is indicative of the at least one polymorphism being useful for assessing susceptibility to breast cancer.
23. The method according to Claim 22, wherein an increase in frequency of the at least one allele in the at least one polymorphism in individuals diagnosed with, or having a susceptibility to, breast cancer, as compared with the frequency of the at least one allele in the control sample is indicative of the at least one polymorphism being useful for assessing increased susceptibility to breast cancer.
24. The method according to Claim 22 or Claim 23, wherein a decrease in frequency of the at least one allele in the at least one polymorphism in individuals diagnosed with, or having a susceptibility to, breast cancer, as compared with the frequency of the at least one allele in the control sample is indicative of the at least one polymorphism being useful for assessing decreased susceptibility to, or protection against, breast cancer.
25. A method of genotyping a nucleic acid sample obtained from a human individual comprising determining whether at least one allele of at least one polymorphic marker

is present in a nucleic acid sample from the individual sample, wherein the at least one marker is selected from rs999737 (SEQ ID NO:6) and markers in linkage disequilibrium therewith, wherein linkage disequilibrium is characterized by values of $r^2 > 0.2$, and wherein determination of the presence of the at least one allele in the sample is indicative of a susceptibility to breast cancer in the individual.

26. The method according to Claim 25, wherein determination of the presence of allele C in rs999737 is indicative of increased susceptibility of breast cancer in the individual.

27. A method of determining a susceptibility to breast cancer in a human individual, the method comprising:

using nucleic acid sequence data about a human individual to identify at least one allele of at least one polymorphic marker selected from rs999737 (SEQ ID NO:6), and markers in linkage disequilibrium therewith, wherein linkage disequilibrium is characterized by values of $r^2 > 0.2$, wherein different alleles of the at least one polymorphic marker are associated with different susceptibilities to breast cancer in humans, and

determining a susceptibility to breast cancer from the nucleic acid sequence data.

28. The method of claim 27, comprising obtaining nucleic acid sequence data about at least two polymorphic markers selected from rs999737, and markers in linkage disequilibrium therewith, wherein linkage disequilibrium is characterized by values of $r^2 > 0.2$.

29. The method of claim 27 or claim 28, wherein determination of a susceptibility comprises comparing the nucleic acid sequence data to a database containing correlation data between the at least one polymorphic marker and susceptibility to breast cancer.

30. The method of claim 29, wherein the database comprises at least one risk measure of susceptibility to breast cancer for the at least one polymorphic marker.

31. The method of claim 29, wherein the database comprises a look-up table containing at least one risk measure of the at least one condition for the at least one polymorphic marker.

32. The method of any of the claims 27-31, wherein using nucleic acid sequence data comprises analyzing sequence of the at least one polymorphic marker in nucleic acid in a biological sample from the individual.

33. The method of claim 32, wherein analyzing sequence of the at least one polymorphic marker comprises determining the presence or absence of at least one allele of the at least one polymorphic marker.
34. The method of any one of claims 27-31, wherein using nucleic acid sequence data comprises analyzing nucleic acid sequence information from a preexisting record.
35. The method of any one of the claims 27-34, further comprising reporting the susceptibility to at least one entity selected from the group consisting of the individual, a guardian of the individual, a genetic service provider, a physician, a medical organization, and a medical insurer.
36. The method of any one of the claims 27-35, wherein the at least one polymorphic marker is selected from the group consisting of the markers rs999737, rs10134446, rs10138140, rs10146772, rs10467820, rs10483812, rs10483813, rs11158749, rs11158751, rs11621276, rs11624097, rs11624164, rs11624333, rs11628293, rs11846360, rs11847185, rs11849916, rs12878761, rs12879200, rs12886864, rs12889251, rs12894230, rs1468279, rs1468280, rs1547012, rs17105675, rs17755657, rs17755734, rs17755752, rs17755925, rs17756000, rs17828691, rs17828721, rs17828763, rs17828907, rs17828955, rs1956534, rs2074563, rs2074565, rs2097800, rs2107340, rs2145157, rs2158357, rs2189517, rs2253317, rs2257111, rs2257116, rs2257127, rs2331701, rs2331705, rs2331775, rs2525503, rs2525530, rs2842327, rs3784121, rs4531674, rs4899246, rs4902604, rs4902606, rs4902608, rs5004090, rs6573837, rs7140266, rs7146456, rs7153476, rs739874, rs746663, rs8007194, rs8010439, rs8012610, rs9323512, rs9323513, and rs9323514.
37. The method of any one of the claims 27-36, wherein the at least one polymorphic marker is rs999737 (SEQ ID NO:6).
38. The method, use, medium or apparatus according to any of the preceding Claims, wherein determination of the presence of at least one risk allele for breast cancer allele is predictive of estrogen receptor positive or progesterone receptor positive breast cancer.
39. The method according to claim 1, substantially as herein described with reference to the Example and/or Figure 1.
40. The method according to any one of claims 1 to 15, substantially as herein described.

41. The computer program according to claim 16, substantially as herein described with reference to the Example and/or Figure 1.
42. The computer program according to any one of claims 16 to 21, substantially as herein described.
43. The method according to claim 22, substantially as herein described with reference to the Example and/or Figure 1.
44. The method according to any one of claims 22 to 24, substantially as herein described.
45. The method according to claim 25, substantially as herein described with reference to the Example and/or Figure 1.
46. The method according to claim 25 or 26, substantially as herein described.
47. The method according to claim 27, substantially as herein described with reference to the Example and/or Figure 1.
48. The method according to any one of claims 27 to 37, substantially as herein described.