GENETIC VARIANTS ON CHR 5p12 AND 10q26 AS MARKERS FOR USE IN BREAST CANCER RISK ASSESSMENT, DIAGNOSIS, PROGNOSIS AND TREATMENT

BACKGROUND OF THE INVENTION

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 population, 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 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].

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 chemotheraples, hormonal theraples and, in cases of individuals at extremely high risk, prophylactic surgery. Mass screening programs constitute a huge economic burden on health services, while preventative theraples have associated risks and quality of life 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 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 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 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 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, **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 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), 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; Tulinius, 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 within a population. Genes differ between individuals most frequently due to single nucleotide polymorphisms (SNP), although other variations are also important. SNP 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 SNP. 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 polymorphism in the human genome is caused by insertion, deletion, translocation, or inversion of either short or long stretches of DNA. Genetic polymorphisms conferring disease risk may therefore 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 disease are uncovered, genetic testing for such risk factors becomes important for clinical medicine. Recent 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 (STI571), 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.

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 chromosomes 2q35 and 16q12 and increased risk of estrogen receptor positive breast cancer (Simon, SN. *et al. Nat Genet* 39:865-9 (2007)). Many other genes have been implicated however their contribution to breast cancer risk has not been confirmed in 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 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.

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 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 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 on chromosome 5p12 or on chromosome 10q26 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 marker on chromosome 5p12 or on chromosome 10q26 in a nucleic acid sample from the individual, 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 within chromosome 5p12, 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237), 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 FGF10 gene, the HCN1 gene, the MRPS30 gene, and/or the FGFR2 gene. In certain such embodiments, the at least one polymorphic marker is in linkage disequilibrium with the FGF10 gene, the HCN1 gene, the MRPS30 gene, and/or the FGFR2 gene. In certain other embodiments, the at least one polymorphic marker is selected from the group of markers located within the chromosomal segment spanning position 44,666,047 and 44,976,797, in NCBI Build 34, and markers in linkage disequilibrium therewith. In another embodiment, the at least one polymorphic marker is selected from the group consisting of the polymorphic markers listed in Table 1 and Table 3, and markers in linkage disequilibrium therewith.

In certain embodiments, the at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14. In one embodiment, the at least one polymorphic marker is selected from the markers as set forth in SEQ ID NO:1-237. In one embodiment, the markers in linkage disequilibrium with marker rs4415084 are selected from the markers set forth in Table 12. In another embodiment, the markers in linkage disequilibrium with marker rs10941679 are selected from the markers set forth in Table 13. In another embodiment, the

markers in linkage disequilibrium with marker rs1219648 are selected from the markers set forth in Table 14.

In certain embodiments, 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 one embodiment, the haplotype comprises markers in the chromosome 5p12 region. In another embodiment, the haplotype comprises markers in the chromosome 10q26 region. In certain embodiments, the haplotype comprises markers in linkage disequilibrium with rs4415084. In certain other embodiments, the haplotype comprises markers in linkage disequilibrium with rs10941679. In certain other embodiments, the haplotype comprises markers in linkage disequilibrium with rs1219648.

The markers conferring risk of breast cancer, as described herein, can be combined with other genetic markers 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 12, Table 13 and Table 14 is present in a sample comprising genomic DNA from a human individual or a genotype dataset derived from a human individual. In other words, genetic markers in other locations in the genome can be useful in 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 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.

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 rs4415084 (SEQ ID NO:235), rs10941679 (SEQ ID NO:236) and rs1219648 (SEQ ID NO:237), 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 of 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 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 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 polymorphic marker is selected from the group consisting of the polymorphic markers listed in Tables 12, 13 and 14, 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 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, 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 12, Table 13 and Table 14.

The invention also relates to an apparatus for determining a genetic indicator for breast cancer in a human individual, comprising: a computer readable memory; and a routine stored on the computer readable memory; wherein the routine is adapted to be executed on a processor to analyze marker and/or haplotype information for at least one human individual with respect to at least one polymorphic marker selected from rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and markers in linkage disequilibrium therewith, and

generate an output based on the marker or haplotype information, wherein the output comprises an individual 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 at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14. In one embodiment, the routine further comprises a risk measure for breast cancer associated with the at least one marker allele and/or haplotype, wherein the risk measure is based on a comparison of the frequency of at least one allele of at least one polymorphic marker and/or haplotype in a plurality of individuals diagnosed with breast cancer and an indicator of the frequency of the at least one allele of at least one polymorphic marker and/or haplotype in a plurality of reference individuals, and wherein the individual risk for the human individual is based on a comparison of the carrier status of the individual for the at least one marker allele and/or haplotype and the risk measure for the at least one marker allele and/or haplotype. For example, the risk measure may in certain embodiments be a measure of risk conferred by each copy of an at-risk variant for breast cancer in a population of individuals with breast cancer, compared with controls. Based on such reference data, risk for a particular individual can be estimated, by determining his/her genotype status at the particular marker and calculate a risk for the individual based thereupon. If the individual carries one copy of the genetic risk variant in his/her genome, the calculated risk can be based on the risk conferred by a single copy of the risk variant. If the individual carries two copies of the genetic risk variants, i.e. the individual is homozygous for the at-risk variant, then the risk estimate for the individual can be based on the risk based on a group of individuals, compared with controls. Normally, risk for homozygous carriers will be the risk for a single copy of the variant squared. Other methods for reporting or estimating risk for the individual based on genotype status at particular markers are also possible, and within the scope of the present invention.

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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237); 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 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, 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 allele T in rs4415084 (SEQ ID NO:235), allele G in rs10941679 (SEQ ID NO:236) and/or allele G in rs1219648 (SEQ ID NO:237) 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 microrray 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:1-237; (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

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 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, 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 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and 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 obtained from the individual, or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237), and markers in linkage disequilibrium therewith, wherein the presence of the at least one 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:1-237, wherein the probe is 15-500 nucleotides in length. In certain 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.

The invention also relates to computer-readable media. In one aspect, the invention relates to a medium on which is stored: an identifier for at least one polymorphic marker; an indicator of the frequency of at least one allele of said at least one polymorphic marker in a plurality of individuals diagnosed with breast cancer; and an indicator of the frequency of the least one allele of said at least one polymorphic markers in a plurality of reference individuals; wherein the at least one polymorphic marker is selected rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and polymorphic markers in linkage disequilibrium therewith. In one embodiment, the polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14. In another embodiment, the medium further comprises information about the ancestry of the plurality of individuals.

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 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 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 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 one such embodiment, the markers described herein to be associated with increased risk or susceptibility of breast cancer confer 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 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 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 numberic 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 on chromosome 5p12 or 10q26 in the genome of the individual, wherein the presence of the at least one allele is indicative of increased susceptibility to breast cancer. 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and markers in linkage disequilibrium therewith, 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 12, Table 13 and Table 14.

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 Tables 12, 13 and 14, 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 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 one of SEQ ID NO:1-237, 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 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 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 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 certain embodiments, the invention relates to use of markers within chromosome 5p12. In certain other embodiments, the invention relates to markers within chromosome 10q26. In certain embodiments, the invention relates to the markers set forth in Table 1 or Table 3, and markers in linkage disequilibrium therewith. In certain other embodiments, the invention relates to the markers set forth in Table 3. In certain other embodimens, the invention relates marker rs10941679, rs7703618, rs4415084, rs2067980, rs10035564, rs11743392, rs7716600, and rs1219648, and markers in linkage disequilibrium therewith. In some preferred embodiments, the invention relates to markers rs4415084, rs10941679 and rs1219648, and markers in linkage disequilibrium therewith. In some other preferred embodiments, the invention relates to markers as set forth in Table 12, Table 13 and Table 14 herein. In other preferred embodiments, the invention relates to rs4415084 and markers in linkage disequilibrium therewith (*e.g.*, markers as set forth in Table 12). In other preferred embodiments, the invention relates to rs10941679 and markers in linkage disequilibrium therewith (*e.g.*, markers as set forth in Table 13). In other preferred embodiments, the invention relates to rs1219648 and markers in linkage disequilibrium therewith (*e.g.*, markers as set forth in Table 14). In one embodiment, the invention relates to

marker rs4415084. In another embodiment, the invention relates to rs10941679. In another embodiment, the invention relates to rs1219648.

In certain embodiments, the at least one marker allele conferring increased risk of breast cancer is selected fromof rs10941679 allele G, rs7703618 allele T, rs4415084 allele G, rs2067980 allele G, rs10035564 allele G, rs11743392 allele T, rs7716600 allele A, and rs1219648 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 marker (*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.

BRIEF DESCRIPTION OF THE DRAWINGS

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 shows a map of Association Data on 5p12 from the Iceland 1 Cohort. The upper panel shows the P-values for the association signals derived from the Illumina Hap300 data frm the Iceland 1 cohort of 1660 breast cancer patients and 11,563 controls, plotted according to their physical location (NCBI Build 34). The signals from the key SNPs defining the 6 equivalence classes in the region are labelled A-F. In the lower panel are shown the locations of recombination hotspots, chromosome bands, exons of known genes and recombination rates. At the bottom are plotted pairwise r^2 values derived from HapMap Phase II data (release 19). The intensity of the dots is proportional to the magnitude of the pairwise r^2 value. Recombination

hotspots and recombination rates are derived using methods described by McVean et al. 2004 (see text).

DETAILED DESCRIPTION OF THE INVENTION

The present invention discloses polymorphic variants and haplotypes that have been found to be associated with breast cancer. Particular alleles at polymorphic markers on chromosome 5p12 have been found to be associated with 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. 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

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.

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 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. 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 nonucleotide 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)

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.

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.

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.

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 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.*, "T *rs4415084*" refers to the T allele of marker rs4415084 being in the haplotype, and this nomenclature is equivalent to "rs4415084 allele T" and "T-rs4415084". Furthermore, allelic codes in haplotypes 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 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 the 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 compute-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 acess 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.

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 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 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 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.

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 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 "chromosome 5p12", as described herein, refers to the region on Chromosome 5 between positions 44,094,392 and 46,393,984 of NCBI (National Center for Biotechnology Information) Build 34.

The term "FGF10" or "FGF10 gene", as described herein, refers to the Fibroblast Growth Factor 10 gene on human chromosome 5p.

The term "MRPS30" or "MRPS30 gene", as described herein, refers to the Mitochondrial Ribosomal Protein S30 gene on human chromosome 5p. This gene is also called programmed cell death protein 9 (PDCD9), and encodes a mitochondrial S28 subunit.

The term "FGFR2" or "FGFR2 gene", as described herein, refers to the Fibroblast Growth Factor Receptor 2 gene on human chromosome 10q26. This gene is also called Protein Tyrosine Kinase Receptor Like 14 (TK14), Keratinocyte Growth Factor Receptor (KGFR), and Fibroblast Growth Factor Receptor BEK.

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 chromosome 5p12 are associated with breast cancer. A genome-wide analysis for variants associated with cancer revealed association of breast cancer to a region of chromosome 5, between positions 44,094,392 and 46,393,984 (NCBI Build 34 coordinates), referred to herein as chromosome 5p12 region. Particular markers were found to be associated with an increased risk of breast cancer in this region.

Through genotyping of approximately 1,600 Icelandic breast cancer patients and 11,563 controls using the Illumina HumanHap300 microarray technology, a large number of markers on

chromosome 5p were found to show association to breast cancer (Table 1). In particular, the T allele of marker rs4415084 and the G allele of marker rs7703618 were found to be associated with an increased risk of breast cancer. The association of marker rs7703618 was replicated in a second Icelandic cohort, showing that the association signal is indeed significant.

A comparison of the Iceland discovery cohort with the public CGEMS data set revealed that association to rs4415084 is also replicated in this cohort. In fact, the association signal to this marker (p-value 9.02E-06 in the Icelandic discovery cohort) is significant at the genome-wide level (after Bonferroni correction), with a nominal p-value of 1.38E-07 when the two data sets are merged. This SNP had an unremarkable P-value of 2.21E-03 in the CGEMS data set alone, but does replicate the original finding in the Icelandic population.

Marker rs10941679, which is correlated with marker rs4415084 (D' = 0.99, r2 = 0.51), has an even stronger correlation with breast cancer (OR = 1.19, p-value 2.2E-06). Follow-up analysis has shown that the signal due to rs4415084 and rs10941679 in cohorts from Sweden, Holland, Spain and the US (see Table 6).

The present invention also shows evidence of allelic heterogeneity in the Chr5p12 region, and six equivalence classes, represented by the key markers rs7703618, rs4415084, rs2067980, rs10035564, rs11743392 and rs7716600, have been identified. Further analysis has established that the observed association signal is mostly accounted for by markers rs4415084 and rs10941679.

There are three known genes of note in the region identified by the present invention as harboring markers and haplotypes associating with breast cancer. These genes are FGF10, MRPS30, and HCN1, along with the poorly characterized gene LOC441070. Two of these genes, FGF10 and MRPS30, are compelling candidates for an involvement in breast cancer predisposition.

Thus, FGF10 is required for normal embryonic development of the breast [Howard and Ashworth, (2006), PLoS Genet, 2, e112], and FGF10 has been implicated as an oncogene in mouse models of breast cancer by MMTV insertional mutagenesis and FGF10 is over expressed in around 10% of human breast cancers [Theodorou, et al., (2004), Oncogene, 23, 6047-55]. The FGF10 gene is separated from the main clusters of association signals by a recombination hotspot. However key elements controlling regulation of FGF10 may be present in the region where the strong association signals occur. Alternatively, the association signals may be in linkage disequilibrium with pathogenic mutations within the FGF10 gene itself.

The MRPS30 gene, also known as programmed cell death protein 9 (PDCD9), encodes a mitochondrial 28S ribosomal subunit. This gene is the mammalian counterpart of the Gallus gallus pro-apoptotic protein p52. It has been shown to induce apoptosis and activate the stress-responsive JNK1 pathway in mammalian cells. The protein appears to function in apoptosis at

least in part through the Bcl-2 pathway [Sun, et al., (1998), Gene, **208**, 157-66; Carim, et al., (1999), Cytogenet Cell Genet, **87**, 85-8; Cavdar Koc, et al., (2001), FEBS Lett, **492**, 166-70]. Although it has not been implicated previously in breast cancer, its involvement in the above pathways suggest that genetic variants in MRPS30 may be involved in modifying breast cancer risk.

It has also been discovered that marker rs1219648 at the *FGFR2* locus on chromosome 10 confers risk of breast cancer (Table 6), which is particularly associated with ER positive tumours (Table 10). It was also discovered that association to rs1219648 was more significant in node positive than node negative tumours, and that the association is stronger for individuals with a family history of breast cancer.

Assessment for markers and haplotypes

The genomic sequence within populations is not identical when individuals are compared. 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, or polymorphisms, are commonly called Single Nucelotide 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 SNPsite; 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 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 complimentary 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 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 terms, polymorphisms can comprise any number of specific alleles. Thus in one embodiment of the invention, the polymorphism is characterized by the presence of two or more alleles in any given 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 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 covery over 15% of the human genome sequence (Estivill, X Armengol; L., PloS Genetics 3:1787-99 (2007). A 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 markers described herein to be associated with breast cancer. 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 15,000 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 in the SNP assay employed. The allele codes for SNPs used herein are as follows: 1= A, 2=C, 3=G, 4=T. The person skilled in the art will however 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 assay employed 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 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 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 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 associated with a disease or trait 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 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.

A haplotype refers to a 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 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 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 SNPlex 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 methods described herein, an individual who is at an increased susceptibility (i.e., increased risk) for breast cancer, is an individual in whom at least one specific allele at one or more polymorphic marker or haplotype conferring increased susceptibility for breast cancer is identified (i.e., at-risk marker alleles or haplotypes). In one aspect, the at-risk marker or haplotype is one that confers a significant increased risk (or susceptibility) of breast cancer. 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 haplotye is measured by an odds ratio (OR). In a further embodiment, the significance is measured by a percentage. In one embodiment, a significant increased risk is measured as a risk (relative risk and/or odds ratio) of at least 1.10, including but not limited to: at least 1.11, at least 1.12, at least 1.13, at least 1.14, at least 1.15, at least 1.16, at least 1.17, at least 1.18, at least 1.19, at least 1.20, at least 1.21, at least 1.22, at least 1.23, at least 1.24, at least 1.25, at least 1.30, at least 1.35, at least 1.40, at least 1.50, at least 1.60, at least 1.70, 1.80, at least 1.90, 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 1.15 is significant. In another particular embodiment, a risk of at least 1.17 is significant. In yet another embodiment, a risk of at least 1.20 is significant. In a further embodiment, a relative risk of at least about 1.25 is significant. In another further embodiment, a significant increase in risk is at least about 1.30 is significant. However, other cutoffs are also contemplated, e.g. at least 1.16, 1.18, 1.19, 1.21, 1.22, 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 10%, including but not limted to about 15%, about 20%, about 25%, about 30%, about 35%, about 40%, about 45%, about 50%, about 55%, about 60%, about 65%, about 70%, about 75%, about 80%, about 85%, about about 90%, about 95%, and about 100%. In one particular embodiment, a significant increase in risk is at least 15%. In other embodiments, a significant increase in risk is at least 17%, at least 20%, at least 22%, at least 24%, at least 25%, at least 30%, at least 32% and at least 35%. 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 1×10^{-3} (0.001), less than 1×10^{-4} (0.0001), less than 1×10^{-5} (0.00001), less than 1×10^{-5} (0.000001), less than 1×10^{-6} (0.0000001), less than 1×10^{-7} (0.00000001), or less than 1×10^{-8} (0.000000001).

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.

In 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.

The person skilled in the art will appreciate that for polymorphic 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.

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 which 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 a total of eight variants that have been described to associate with prostate cancer (Gudmundsson, J., *et al., Nat Genet* **39**:631-7 (2007), Gudmundsson, J., *et al., Nat Genet* **39**:977-83 (2007); Yeager, M., *et al, Nat Genet* **39**:645-49 (2007), Amundadottir, L., *et al., Nat Genet* **38**:652-8 (2006); Haiman, C.A., *et al., Nat Genet* **39**:638-44 (2007)). Seven of these loci are on autosomes, and the remaining locus is on chromosome X. The total number of theoretical genotypic combinations is then $3^7 \times 2^1 = 4374$. 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 conjugation 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 a plurality of variants associated with breast cancer may be assessed.

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 randombly 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 occurrance 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 occurence 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 $\square\Delta^2$) 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 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). Measuring LD across a region is not straightforward, but one approach is to use the measure r, which was developed in population genetics. 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 genetic segments (such as SNP markers) can be at least 0.1 such as at least 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99 or 1.0. In one preferred embodiment, the significant r^2 value can be at least 0.2. Alternatively, linkage disequilibrium as described herein, refers to linkage disequilibrium 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, 0.99. Thus, linkage disequilibrium represents a correlation between alleles of distinct markers. It is measured by correlation coefficient or |D'| (r^2 up to 1.0 and |D'| up to 1.0). 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, japanese, chinese), as defined (http://www.hapmap.org). In one such embodiment, LD is determined in the CEU population of the HapMap samples. In another embodiment, LD is determined in the YRI population. In another embodiment, LD is determined in a European population. In yet another embodiment, LD is determined in 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., 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 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 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

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.

It has thus become apparent that for any given observed association to a polymorphic marker in the genome, it is likely 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 and/or haplotypes of the invention, as described herein, may be used as surrogate markers. The surrogate markers have in one embodiment relative risk (RR) and/or odds ratio (OR) values smaller than for the markers or haplotypes initially found to be associating with the disease, as described herein. 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, as described herein. 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, such as the variants described herein. Identifying and using such markers for detecting the association discovered by the inventors as described herein 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

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

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 suceptibility region, for example, within an LD block region, association of all possible combinations of genotyped 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 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 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 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 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 NEMO.

For single marker association to a disease, the Fisher exact test can be used to calculate two-sided p-values for each individual allele. Usually, all p-values are presented unadjusted for multiple comparisons unless specifically indicated. The presented frequencies (for microsatellites, SNPs and haplotypes) are allelic frequencies as opposed to carrier frequencies. To minimize any bias due the relatedness of the patients who were recruited as families, first and second-degree relatives can be eliminated from the patient list. Furthermore, the test can be repeated for association correcting for any remaining relatedness among the patients, by extending a variance adjustment procedure previously described (Risch, N. & Teng, J. (*Genome Res.*,

8:1273-1288 (1998)) for sibships so that it can be applied to general familial relationships, and present both adjusted and unadjusted p-values for comparison. 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. The differences are in general very small as expected. To assess the significance of single-marker association corrected for multiple testing we can carry out a randomization test using the same genotype data. Cohorts of patients and controls can be randomized and the association analysis redone multiple times (*e.g.*, up to 500,000 times) and the p-value is the fraction of replications that produced a p-value for some marker allele that is lower than or equal to the p-value we observed using the original patient and control cohorts.

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, $risk(h_i)/risk(h_j) = (f_i/p_i)/(f_j/p_j)$, 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.

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, and hence the less stringent the statistical measure that is 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 are a measure of a true genetic effect, and replication in additional cohorts is not necessarily from a statistical point of view. However, 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, a measured by the OR or RR, is the same in all populations, while the frequency of the variant may differ between the poplations. 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.

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.

As described herein, certain polymorphic markers and haplotypes comprising such markers are 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 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 within haplotype blocks or LD blocks comprising at-risk markers, such as the markers of the present invention, can be used as surrogates for other markers and/or haplotypes within the haplotype block or LD block. Markers with values of r^2 equal to 1 are perfect surrogates for the at-risk 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 represent variants with relative risk values as high or possibly even higher than the at-risk variant. 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 for example be a tandem repeat, such as a minisatellite or a microsatellite, a transposable element (*e.g.*, an *Alu* element), or a structural 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 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 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 for detecting association to breast cancer, or a susceptibility to breast cancer, in an individual. These tagging or surrogate markers that are in LD with the markers of the present invention can also include other markers that distinguish among haplotypes, as these similarly have predictive value for detecting susceptibility to breast cancer.

The present invention can in certain embodiments be practiced by assessing a sample comprising genomic DNA from an individual for the presence of 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 look-up table that comprises correlations between at least one allele of the polymorphism and breast cancer. In some embodiments, the table comprises a correlation for one polymorhpism. In other embodiments, the table comprises a correlation for a plurality of polymorhpisms. 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 a risk measure, such as a relative risk (RR), an absolute risk (AR) or an odds ratio (OR).

The markers of the invention, *e.g.*, polymorphic markers on Chromosome 5p12 and Chromosome 10q26, *e.g.*, the markers presented in Tables 12, 13 and 14, *e.g.*, markers marker rs7703618, rs4415084, rs2067980, rs10035564, rs11743392, rs7716600, rs10941679, rs1219648, may be useful for risk assessment and diagnostic purposes for, either alone or in combination. Thus, even in cases where the increase in risk by individual markers is relatively modest, i.e. on the order of 10-30%, the association may have significant implications. 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 markers on chromosome 5p12 and chromosome 10q26, such as marker rs10941679 and marker rs1219648. Alternatively, markers in LD with either 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, 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).

Thus, in one 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. Results for any two or more markers can be combined in such analysis, such as results for three markers, four markers, five markers, six markers, seven markers, eight markers, nine markers, or ten or more markers. In such embodiments, the genotype status of a plurality of markers and/or 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, may subsequently be used to determine the overall risk conferred based on the genotype status at the multiple loci. Assessment of risk based on such analysis may subsequently be used in the methods and kits of the invention, as described herein.

As described in the above, the haplotype block structure of the human genome has the effect that a large number of variants (markers and/or haplotypes) in linkage disequilibrium with the variant originally associated with a disease or trait may be used as surrogate markers for assessing association to the disease or trait. The number of such surrogate markers will depend on factors such as the historical recombination rate in the region, the mutational frequency in the region (i.e., the number of polymorphic sites or markers in the region), and the extent of LD (size of the LD block) in the region. These markers are usually located within the physical boundaries of the LD block or haplotype block in question as defined using the methods described herein, or by other methods known to the person skilled in the art. However, sometimes marker and haplotype association is found to extend beyond the physical boundaries of the haplotype block as defined. Such markers and/or haplotypes may in those cases be also used as surrogate markers and/or haplotypes for the markers and/or haplotypes physically residing within the haplotype block as defined. As a consequence, markers and haplotypes in LD (typically characterized by r^2 greater than 0.1, such as r^2 greater than 0.2, including r^2 greater than 0.3, also including r^2 greater than 0.4) with the markers and haplotypes of the present invention are also within the scope of the invention, even if they are physically located beyond the boundaries of the haplotype block as defined. This includes markers that are described herein (*e.g.*, Table 1 and Table 3, *e.g.* rs4415084, rs10941679, rs1219648), but may also include other markers that are in strong LD (characterized by r^2 greater than 0.1 or 0.2 and/or $|D'| > 0.8$) with one or more of the markers listed in Table 1 and Table 3.

For the SNP markers described herein, the opposite allele to the allele found to be in excess in patients (at-risk allele) is found in decreased frequency in breast cancer. These markers and haplotypes in LD and/or comprising such markers, are thus protective for breast cancer, i.e. they confer a decreased risk or susceptibility of individuals carrying these markers and/or haplotypes developing breast cancer.

Certain variants of the present invention, including certain haplotypes comprise, in some cases, a combination of various genetic markers, *e.g.*, SNPs and microsatellites. Detecting haplotypes can be accomplished by methods known in the art and/or described herein for detecting sequences at polymorphic sites. Furthermore, correlation between certain haplotypes or sets of markers and disease phenotype can be verified using standard techniques. A representative example of a simple test for correlation would be a Fisher-exact test on a two by two table.

In specific embodiments, a marker allele or haplotype found to be associated with breast cancer, (*e.g.*, marker alleles as listed in Table 1 and Table 3, the markers as listed in Tables 12, 13 and 14, SEQ ID NO:1-237) is one in which the marker allele or haplotype is more frequently present in an individual at risk for breast cancer (affected), compared to the frequency of its presence in a healthy individual (control), wherein the presence of the marker allele or haplotype is indicative of breast cancer or a susceptibility to breast cancer. In other embodiments, at-risk markers in linkage disequilibrium with one or more markers found to be associated with breast cancer are tagging markers that are more frequently present in an individual at risk for breast cancer (affected), compared to the frequency of their presence in a healthy individual (control), wherein the presence of the tagging markers is indicative of increased susceptibility to breast cancer. In a further embodiment, at-risk markers alleles (i.e. conferring increased susceptibility) in linkage disequilibrium with one or more markers found to be associated with breast cancer, are markers comprising one or more allele that is more frequently present in an individual at risk for breast cancer, compared to the frequency of their presence in a healthy individual (control), wherein the presence of the markers is indicative of increased susceptibility to breast cancer.

Study population

In a general sense, the methods and kits of the invention can be utilized from samples containing genomic DNA from any source, i.e. any individual. In preferred embodiments, the individual is a human individual. The individual can be an adult, child, or fetus. The present invention also provides for assessing markers and/or haplotypes in individuals who are members of a target population. 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).

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 the disease 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 sex, males or females. In some embodiments, it relates to assessment of male subjects. In preferred embodiments, it relates to assessment of female subjects.

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 (Styrkarsdottir, U., *et al. N Engl J Med* Apr 29 2008 (Epub ahead of print); 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); Helgadottir, A., *et al., Science* 316:1491-93 (2007); Steinthorsdottir, 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); Amundadottir, 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.

The markers of the present invention found to be associated with breast cancer are believed to 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, and 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, Portugues, Italian, Polish, Bulgarian, Slavic, Serbian, Bosnian, Czech, Greek and Turkish populations. The invention furthermore in other embodiments can be practiced in specific human populations that include Bantu, Mandenk, Yoruba, San, Mbuti Pygmy, Orcadian, Adygei, Russian, Sardinian, Tuscan, Mozabite, Bedouin, Druze, Palestinian, Balochi, Brahui, Makrani, Sindhi, Pathan, Burusho, Hazara, Uygur, Kalash, Han, Dai, Daur, Hezhen, Lahu, Miao, Oroqen, She, Tujia, Tu, Xibo, Yi, Mongolan, Naxi, Cambodian, Japanese, Yakut, Melanesian, Papuan, Karitianan, Surui, Colmbian, Maya and Pima.

In certain embodiments, the invention relates to populations that include black African ancestry such as populations comprising persons of African descent or lineage. Black African ancestry may be determined by self reporting as African-Americans, Afro-Americans, Black Americans,

being a member of the black race or being a member of the negro race. For example, African Americans or Black Americans are those persons living in North America and having origins in any of the black racial groups of Africa. In another example, self-reported persons of black African ancestry may have at least one parent of black African ancestry or at least one grandparent of black African ancestry.

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 as 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 tought 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.

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.

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 risks 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 breat cancer (*e.g.*, the FGF10, MRPS30 and/or FGFR2 genes). 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 FGF10, MRPS30 and/or FGFR2 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.

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 is 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.

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.

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 Galen 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; Elfel, 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 FGF10, MRPS30 and/or FGFR2 genes.

Cancer chemotherapy has well known, dose-limiting side effects on normal tissues particularly the highly proliferative hemopoetic 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 pneumatitis. 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 FGF10, MRPS30 and/or FGFR2 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: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 metastatsis, 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 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).

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 OF THE INVENTION

Diagnostic and screening methods

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 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 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 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 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 litterature and scientific publications. The diagnostic 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 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.

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 (such as breast cancer). 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.

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 (*e.g.*, markers on Chromosome 5p12 and 10q26, *e.g.* the markers and haplotypes as listed in Tables 12, 13 and 14, *e.g.* markers rs4415084, rs10941679 and rs1219648, and markers in linkage disequilibrium therewith) are associated with 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 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 in a nucleic acid sample obtained from the individual, wherein the at least one polymorphic marker is selected from the group consisting of the polymorphic markers listed in Tables 12, 13 and 14, and markers in linkage disequilibrium therewith. In another embodiment, the invention pertains to methods of diagnosing a susceptibility to breast cancer in a human individual, by screening for at least one marker allele or haplotype as listed in Table 12, 13 or 14, or markers in linkage disequilibrium therewith. 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, 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 haplotypes described herein include combinations of alleles at various genetic markers (e.g., SNPs, microsatellites, or other genetic variants). 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

of the present invention 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 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 diagnose a susceptibility to breast cancer, a hybridization sample can be formed by contacting the test sample containing a breast cancer-associated nucleic acid, 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. For example, the nucleic acid probe can comprise all or a portion of a nucleotide sequence comprising the markers set forth in Tables 12, 13 and 14 (SEQ ID NO:1-237), or a nucleotide sequence comprising the FGF10, MRPS30,

HCN1 or FGFR2 genes or fragments thereof, 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 a nucleotide sequence comprising the markers listed in any one of Tables 12, 13 and 14 (SEQ ID NO:1-237), or a nucleotide sequence comprising the FGF10, MRPS30, HCN1 and FGFR2 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, or 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. Hybridization of the PNA probe is thus diagnostic for breast cancer or a susceptibility to 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 ore 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 associated with breast cancer (*e.g.* the polymorphic markers of Tables 12, 13 and 14 (SEQ ID NO:1-237) and markers in linkage disequilibrium therewith). 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.

Allele-specific oligonucleotides can also be used to detect the presence of a particular allele in a nucleic acid associated with breast cancer (e.g. the polymorphic markers of Tables 12, 13 and 14, and markers in linkage disequilibrium therewith), through the use of dot-blot hybridization of amplified oligonucleotides with allele-specific oligonucleotide (ASO) probes (see, for example, Saiki, R. *et al.*, *Nature*, 324:163-166 (1986)). An "allele-specific oligonucleotide" (also referred to herein as an "allele-specific oligonucleotide probe") is an oligonucleotide of approximately 10-50 base pairs or approximately 15-30 base pairs, that specifically hybridizes to a nucleic acid associated with breast cancer, and which contains a specific allele at a polymorphic site (e.g., a marker or haplotype as described herein). An allele-specific oligonucleotide probe that is specific for one or more particular a nucleic acid associated with breast cancer can be prepared using standard methods (see, e.g., Current Protocols in Molecular Biology, *supra*). PCR can be used to amplify the desired region. The DNA containing the amplified region can be dot-blotted using standard methods (*see, e.g.*, Current Protocols in Molecular Biology, *supra*), and the blot can be contacted with the oligonucleotide probe. The presence of specific hybridization of the probe to the amplified region can then be detected. Specific hybridization of an allele-specific oligonucleotide probe to DNA from the subject is indicative of a specific allele at a polymorphic

site associated with cancer, including breast cancer (see, e.g., Gibbs, R. *et al.*, *Nucleic Acids Res.*, *17*:2437-2448 (1989) and WO 93/22456).

With the addition of such analogs as locked nucleic acids (LNAs), the size of primers and probes can be reduced to as few as 8 bases. LNAs are a novel class of bicyclic DNA analogs in which the 2' and 4' positions in the furanose ring are joined via an O-methylene (oxy-LNA), S-methylene (thio-LNA), or amino methylene (amino-LNA) moiety. Common to all of these LNA variants is an affinity toward complementary nucleic acids, which is by far the highest reported for a DNA analog. For example, particular all oxy-LNA nonamers have been shown to have melting temperatures (T_m) of 64°C and 74°C when in complex with complementary DNA or RNA, respectively, as opposed to 28°C for both DNA and RNA for the corresponding DNA nonamer. Substantial increases in T_m are also obtained when LNA monomers are used in combination with standard DNA or RNA monomers. For primers and probes, depending on where the LNA monomers are included (e.g., the 3' end, the 5' end, or in the middle), the T_m could be increased considerably.

In another embodiment, arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from a subject, can be used to identify polymorphisms in a nucleic acid associated with breast cancer (*e.g.* the polymorphic markers of Table 12, 13 and 14 (SEQ ID NO:1-237), and markers in linkage disequilibrium therewith). 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 oligonucleotide arrays, also described as "Genechips™," have been generally described in the art (see, e.g., U.S. Patent No. 5,143,854, PCT Patent Publication Nos. WO 90/15070 and 92/10092). 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); Raqoussis, 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 (1985)); 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.

In another embodiment of the invention, diagnosis of breast cancer or 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 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 FGF10, MRPS30, HCN1 and FGFR2 genes). The haplotypes and markers of the present invention that show association to breast cancer may play a role through their effect on one or more of these nearby genes. Possible mechanisms affecting these genes include, e.g., effects on transcription, effects on RNA 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.

Thus, in another embodiment, the variants (markers or haplotypes) of the invention showing association to breast cancer affect the expression of a nearby gene. It is well known that regulatory element affecting gene expression may be located tenths or even 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. It is thus contemplated that the detection of the markers or haplotypes of the present invention can be used for assessing expression for one or more of the FGF10, MRPS30, HCN1 and FGFR2 genes.

A variety of methods can be used for detecting protein expression levels, including enzyme linked immunosorbent assays (ELISA), Western blots, immunoprecipitations and 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 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 FGF10, MRPS30, and HCN1 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., FGF10, MRPS30, and HCN1) 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, the diagnosis of a susceptibility to breast cancer is made by detecting at least one marker or haplotypes of the present invention (*e.g.*, associated alleles of the markers listed in Tables 12, 13 and 14 (SEQ ID NO:1-237), and markers in linkage disequilibrium therewith), in combination with an additional protein-based, RNA-based or DNA-based assay. The methods of the invention can also be used in combination with an analysis of a subject's family history and risk factors (e.g., environmental risk factors, lifestyle risk factors).

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 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 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 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 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 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 Table 12, 13 and 14 (SEQ ID NO:1-237), and polymorphic markers in linkage disequilibrium therewith. In yet another embodiment the 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 12, 13 and 14. In another embodiment, the marker to be detected is selected from marker rs10941679, rs7703618, rs4415084, rs2067980, rs10035564, rs11743392, rs7716600 and rs1219648. 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 12, 13 and 14. In yet another embodiment, the marker or haplotype to be detected comprises at least one marker selected from the group of markers consisting of markers rs10941679, rs7703618, rs4415084, rs2067980, rs10035564, rs11743392, rs7716600 and rs1219648, and markers in linkage disequilibrium therewith.

In one preferred embodiment, the kit for detecting the markers of the invention comprises a detection oligonucleotide probe, that hybridizes to a segment of template DNA containing a SNP polymorphisms to be detected, an enhancer oligonucleotide probe and an endonuclease. The detection oligonucleotide probe comprises a fluorescent molety 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 molety can be Gig Harbor Green or Yakima Yellow, or other suitable fluorescent moleties. 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, 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.

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.

In one of such embodiments, the presence of the marker or haplotype is indicative of a susceptibility (increased susceptibility or decreased susceptibility) to breast cancer. In another embodiment, the presence of the marker or haplotype is indicative of response to a breast cancer therapeutic agent. In another embodiment, the presence of the marker or haplotype is indicative of breast cancer prognosis. In yet another embodiment, the presence of the marker 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).

Therapeutic agents

Variants of the present invention (*e.g.*, the markers and/or haplotypes of the invention, *e.g.*, the markers listed in Tables 12, 13 and 14, *e.g.*, rs4415084, rs10941679, rs1219648) can be used to identify 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 FGF10, MRPS30, HCN1 and FGFR2 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 of the invention, nucleic acids comprising one or more variant of the invention (*e.g.*, nucleic acids with sequence as set forth in any one of SEQ ID NO:1-237, or fragments thereof) 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 described and reviewed in *AntisenseDrug Technology: Principles, Strategies, and Applications*, Crooke, ed., Marcel Dekker Inc., New York (2001). In general, antisense nucleic acid molecules are designed to be complementary to a region of mRNA expressed by a gene, so that the antisense molecule hybridizes to the mRNA, thus blocking translation of the mRNA into protein. 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 Ter.* 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)

The variants described herein can 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 (markers 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 to treat a disease or disorder, such as breast cancer. 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 of the present invention (*e.g.*, the markers and haplotypes set forth in Tables 12, 13 and 14) 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 concompass 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 adminstered 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 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 a 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 inihibitors. 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 treatement 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

The present invention also 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 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 The third party may be the individual from which the genotype data is derived. The third party may also be a service provider for analyzing the genotype information, for example a service provider who calculates genetic risk based on the genotype of the individual at particular

genetic markers. In one such embodiment, the service provider receives genotype information from a genotype service provider, and stores the genotype information on a readable medium for subsequent analysis. In another embodiment, the genotype provider is also the service provider, i.e. the same party generates genotypes from a DNA sample from an individual, stores the genotype data on a readable medium, and providers service relating to the risk assessment or other interpretation of the genotype data. The additional interpretation may for example include assessment or prediction of the ancestry of the individual, or the genealogical relationship between the individual and a reference individual. The reference individual may for example be a friend, relative or any other person to whom the individual wishes to compare his/her genotypes to. In one particular embodiment, the genotype data is used to derive information about genetic risk factors contributing to increased susceptibility to breast cancer, and report results based on such comparison.

In one aspect, the invention relates to computer-readable media. In general terms, such medium has capabilities of storing (i) identifer information for at least one polymorphic marker or a haplotye; (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 breast cancer; 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 representativ 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.

Additional information about the individual can be stored on the medium, such as ancestry information, information about sex, physical attributes or characterisitics (including height and weight), biochemical measurements (such as blood pressure, blood lipid levels, etc.), or other useful information that is desirable to store or manipulate in the context of the genotype status of a particular individual.

The invention furthermore relates to an apparatus that is suitable for determination or manipulation of genetic data useful for determining a susceptibility to breast cancer in a human individual. Such an apparatus can include a computer-readable memory, a routine for manipulating data stored on the computer-readable memory, and a routine for generating an output that includes a measure of the genetic data. Such measure can include values such as allelic or haplotype frequencies, genotype counts, sex, age, phenotype information, values for odds ratio (OR) or relative risk (RR), population attributable risk (PAR), or other useful information that is either a direct statistic of the original genotype data or based on calculations based on the genetic data.

The markers and haplotypes shown herein to be associated with increased susceptibility (*e.g.*, increased risk) of breast cancer, are in certain embodiments useful for interpretation and/or analysis of genotype data. Thus in certain embodiments, an identification of an at-risk allele for breast cancer, as shown herein, or an allele at a polymorphic marker in LD with any one of the markers shown herein to be associated with breast cancer, 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 the individual from whom the data originates, for example via a user interface accessable 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 ration (OR)) for the disease (*e.g.*, breast cancer). 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 varians in the dataset are made available to the individual, 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. In particular embodiments, the results of risk assessment is made available to a third party, *e.g.*, a physician, other healthcare worker or genetic counselor.

Markers useful in various aspects of the invention

The above-described applications can all be practiced with the markers and haplotypes of the invention that have in more detail been described with respect to methods of assessing susceptibility to breast cancer. Thus, these applications can in general be reduced to practice using markers within the Chr5p12 and Chr10q26 genomic regions as defined herein, including markers as listed in Tables 12, 13 and 14, and markers in linkage disequilibrium therewith. In one embodiment, a marker useful in the various aspects and embodiments of the invention is seleced from the markers set forth in Tables 12, 13 and 14 (SEQ ID NO:1-237). In one embodiment, the marker is selected from marker rs10941679, rs7703618, rs4415084, rs2067980, rs10035564, rs11743392, rs7716600, and rs1219648, and markers in linkage disequilibrium therewith. In another embodiment, the marker is selected from marker rs10941679, rs7703618, rs4415084, rs2067980, rs10035564, rs11743392, rs7716600 and rs1219648. In another embodiment, the marker is selected from rs10941679, and markers in linkage disequilibrium therewith. In one embodiment, the marker is selected from the markers set forth in Table 13. In another embodiment, the marker is selected from marker rs4415084, and markers in linkage disequilibrium therewith. In another embodiment, the marker is selected

from the markers set forth in Table 12. In another embodiment, the marker is selected from marker rs1219648, and markers in linkage disequilibrium therewith. In another embodiment, the marker is selected from the markers set forth in Table 14. In another embodiment, the marker is rs4415084. In another embodiment, the marker is rs10941679. In another embodiment, the marker is rs1219648. In another embodiment, the marker is rs4415084 or rs10941679. In another embodiment, marker alleles conferring increased risk or susceptibility of breast cancer are selected from rs10941679 allele G, rs7703618 allele T, rs4415084 allele G, rs2067980 allele G, rs10035564 allele G, rs11743392 allele T, rs7716600 allele A, and rs1219648 allele G.

Nucleic Acids and Polypeptides

The nucleic acids and polypeptides described herein can be used in methods and kits of the 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 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 60% 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).

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 the polymorphic markers listed in Table 12, Table 13 and Table 14 (SEQ ID NO:1-237), and the nucleotide sequence of the FGF10, MRPS30, HCN1 and FGFR2 genes; or a nucleotide sequence comprising, or consisting of, the complement of the nucleotide sequence of a nucleotide sequence comprising the polymorphic markers listed in Table 12, Table 13 and Table 14 (SEQ ID NO:1-237), and the nucleotide sequence of the FGF10, MRPS30, HCN1 and FGFR2 genes, wherein the nucleotide sequence comprises at least one polymorphic allele contained in the markers and haplotypes described herein. 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 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.

In general, the isolated nucleic acid sequences of the invention can be used as molecular weight markers on Southern gels, and as chromosome markers that are labeled to map related gene positions. The nucleic acid sequences can also be used to compare with endogenous DNA sequences in patients to identify breast cancer or a susceptibility to breast cancer, and as probes, such as to hybridize and discover related DNA sequences or to subtract out known sequences from a sample (e.g., subtractive hybridization). The nucleic acid sequences can further be used to derive primers for genetic fingerprinting, to raise anti-polypeptide antibodies using immunization techniques, and/or as an antigen to raise anti-DNA antibodies or elicit immune responses.

Antibodies

Polyclonal antibodies and/or monoclonal antibodies that specifically bind one form of the gene product but not to the other form of the gene product are also provided. Antibodies are also provided which bind a portion of either the variant or the reference gene product that contains the polymorphic site or sites. 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).

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.

In general, antibodies of the invention (*e.g.*, a monoclonal antibody) can be used to isolate a polypeptide of the invention 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. 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 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; 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 ^{3}H.

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 polymorpic 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 labelled 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 examples.

EXAMPLE 1. *Identification of variants on chromosome 5p12 that associate with risk of breast cancer*

INTRODUCTION

Mutations in breast cancer susceptibility genes *BRCA1* and *BRCA2* account for 15-25% of the familial component of breast cancer risk [Easton, (1999), Breast Cancer Res, **1**, 14-7; Balmain, et al., (2003), Nat Genet, **33 Suppl**, 238-44]. Much of the genetic component of risk of breast cancer remains uncharacterized and is thought to arise from combinations of less penetrant

variants that, individually, may be quite common [Pharoah, et al., (2002), Nat Genet, **31**, 33-6]. Many searches for less penetrant breast cancer risk variants have been carried out using a candidate gene, case-control association approach. Findings from these studies have often proven difficult to replicate [Breast Cancer Association, (2006), J Natl Cancer Inst, **98**, 1382-96]. Recently common missense variants in two genes, *CASP8* and *TGFB1*, have been shown to be associated with breast cancer risk by using well-powered, multi-center analyses [Cox, et al., (2007), Nat Genet, **39**, 352-8]. These reports emphasize the importance of large scale studies with adequate replication when the goal is to identify common variants conferring modest increases in the risk of breast cancer.

RESULTS

Numerous Illumina SNPs in a region on chromosome 5p12 are associated with an increased risk for breast cancer in Iceland

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 1,600 Icelandic breast cancer patients and 11,563 controls. This discovery sample set was designated "Iceland 1". 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 the same area on chromosome 5p12 occupied 39 of the top 50 ranks. The highest ranks occupied by SNPs located in 5p12 were ranks 5 through 9. The region of interest containing these SNPs extended from approximately chromosome 5 co-ordinate 44,094,392bp (position of marker rs7704166; all co-ordinates herein are from NCBI Build 34) to the position of the last Illumina SNP before the centromere; namely rs10941803 at 46,393,984bp. Results from genotyping of the Illumina SNPs in this region are presented in Table 1 and are presented graphically in Figure 1.

In order to further investigate the signals related to one of the highly ranked SNPs, marker rs7703618 on chr5p12, we generated and validated Centaurus assays for this SNP. The Centaurus assay was designated SG05S3065.c1. This SNP assay was used to genotype an independent sample of approximately 591 Icelandic Breast Cancer patients and 1314 controls. This independent sample was designated Iceland 2. As shown in Table 2, the SNP showed a significant association with breast cancer in the Iceland 2 sample, confirming the original observations with Iceland 1. We have thus replicated the original finding observed in the Iceland 1 sample in the independent Iceland 2 sample. The joint P-value for Iceland 1 & Iceland 2

approached a level that would be considered genome-wide significant after applying the conservative Bonferroni correction for the 311,524 SNPs tested [Skol, et al., (2006), Nat Genet, **38**, 209-13].

Association to chromosome 5p12, confirmed in CGEMS data, is genome-wide significant following joint analysis with CGEMS

The Cancer Genetics Markers of Susceptibility (CGEMS) project of the U.S. National Cancer Institute 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/. The CGEMS project found no genome-wide significant signals in the 5p12 region. However, we noted that one SNP, namely rs4415084, had a P-value of 1.38E-07 when data from Iceland (Iceland 1 cohort) and the CGEMS data set was analyzed jointly, that is genome-wide significant after Bonferroni correction. This SNP had an unremarkable P-value of 2.21E-03 in the CGEMS data set, the genome wide significance of the joint value being mostly carried by the Iceland 1 P-value of 9.02E-06. Thus, the CGEMS data, while nowhere significant on its own, provides confirmation of our original observation of association to chromosome 5p12.

Numerous HapMap (non Illumina) SNP markers could show BC risk associations through their correlation with the Illumina SNPs that showed an association in the 5p12 region

We contemplated that there may be allelic heterogeneity at this locus, that is there may be more than one underlying at-risk variant present in the 5p12 region that is correlated to different degrees with the Illumina SNP set tested and exist at different frequencies in different populations. This is based on the observation that there appear to be significant signals distal to the cluster of most significant markers (see FIG 1). Furthermore, we have noted that in some cases association signals are very strong in the Iceland 1 material but are not strong in the CGEMS data: for example the rs4415084 SNP described above gave a very strong signal in Iceland 1 but not in the CGEMS data. Similarly, the SNP that was tested successfully for replication in Iceland, rs7703618, gave a P-value of 6.93E-06 and only 2.37E-02 in the CGEMS data set. Given that there may be allelelc heterogeneity in the 5p12 region, we define a set of HapMap SNPs that, through their correlations with signals that we observed in the Iceland 1 data set, could be used to detect all pathogenic mutations in the 5p12 region. In order to identify such a set of HapMap SNPs, we first identified a class of SNPs that gave P-values of 10E-3 or less in the Iceland 1 data set. We then split this set into equivalence classes, membership of a particular equivalence class being defined as two SNPs that have an r^2 value of >0.8 between

them. This resulted in a set of 6 equivalence classes which we designated A to F. For each equivalence class we started with the SNP that gave the most significant signal in the class (which we denoted the "key" SNP), then by reference to HapMap data we identified all HapMap SNPs that were correlated with the key SNP by an r^2 value of 0.2 or greater and were not themselves represented on the Illumina Hap300 chip. Thus, using the Iceland 1 data we observed signals in several different equivalence classes and that fact, in itself, provides evidence for allelic heterogeneity. These HapMap SNPs could, through their correlations with SNPs in one or more of the equivalence classes we identified, also be used to detect the same breast cancer risk associations we originally observed. The list of the key SNPs and their correlations with HapMap SNPs is shown in Table 4.

FGF10 and MRPS30 are the most likely candidate genes in the 5p12 region

Figure 1 shows a plot of the association signals obtained in the 5p12 region superimposed on a map of recombination hotspots, chromosome bands, known genes, and recombination rates. Recombination hotspots and recombination rates were determined as described by McVean et al. [McVean, et al., (2004), Science, **304**, 581-4]. A representation of r^2 values between HapMap SNPs in the region is also shown. It can be seen that there are three known genes of note in the region; FGF10, MRPS30, and HCN1, along with one poorly characterized gene LOC441070. Two of these, FGF10 and MRPS30 are compelling candidates for an involvement in breast cancer predisposition.

As reviewed by Howard and Ashworth [Howard and Ashworth, (2006), PLoS Genet, **2**, e112], FGF10 is required for normal embryonic development of the breast. FGF10 has been implicated as an oncogene in mouse models of breast cancer by MMTV insertional mutagenesis and FGF10 is over expressed in around 10% of human breast cancers [Theodorou, et al., (2004), Oncogene, **23**, 6047-55]. As can be seen in Figure 1, the FGF10 gene is separated from the main clusters of association signals by a recombination hotspot. However key elements controlling regulation of FGF10 may be present in the region where the strong association signals occur. Alternatively, the association signals may be in linkage disequilibrium with pathogenic mutations within the FGF10 gene itself.

MRPS30 encodes the mitochondrial 28S ribosomal subunit. It is also known as programmed cell death protein 9 (PDCD9). This is the mammalian counterpart of the Gallus gallus pro-apoptotic protein p52. It has been shown to induce apoptosis and activate the stress-responsive JNK1 pathway in mammalian cells. The protein appears to function in apoptosis at least in part through the Bcl-2 pathway [Sun, et al., (1998), Gene, **208**, 157-66; Carim, et al., (1999), Cytogenet Cell Genet, **87**, 85-8; Cavdar Koc, et al., (2001), FEBS Lett, **492**, 166-70]. Although

it has not been implicated previously in breast cancer, its involvement in the above pathways suggest that genetic variants in MRPS30 may be involved in modifying breast cancer risk.

METHODS

Patient and control selection

Approval for this study was granted by the National Bioethics Committee of Iceland and the Icelandic Data Protection Authority. Records of breast cancer diagnoses were obtained from the Icelandic Cancer Registry (ICR). The ICR contains all cases of invasive breast tumors and ductal or lobular carcinoma in-situ diagnosed in Iceland from January 1st 1955. All people living in Iceland who had a diagnosis entered into the ICR up to the end of December 2005 were eligible to participate in the study. The ICR contained records of 4603 individuals diagnosed during this period. A prevalence cohort comprised of all living patients (approximately 2840) were eligible for recruitment into the study. We obtained informed consent, a blood sample, and diagnostic information from 2210 patients, a participation rate of approximately 78%. Genotyping was successful on a total of 2190 patients for rs7703618 . Further details of the recruitment of this patient group have been reported previously [Stacey, et al., (2006), PLoS Med, **3**, e217].

The 12,904 Icelandic controls consisted of 846 individuals randomly selected from the Icelandic Genealogical Database and 12,058 individuals from other ongoing genome-wide association studies at deCODE. Individuals with a diagnosis of breast cancer in the ICR were excluded. Both male and female genders were included.

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 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

A Centaurus assay [Kutyavin, et al., (2006), Nucleic Acids Res, **34**, e128] was designed for rs7703618 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 5 shows the sequence context for the key SNPs discussed herein. Table 6 shows the description of the Centaurus Assay for marker rs7703618 that was developed for genotyping in this study.

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 0.675^2. 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: Association results for Illumina SNPs in the 5p12 region:

SNP	Allele	Position b!d34	P-value	OR	Cases	Frq.Cases	Controls	Frq.Controls
rs7704166	A	44094392	1.15E-01	1.063	1660	0.492	11561	0.477
rs6879107	A	44096606	6.07E-01	1.029	1659	0.856	11554	0.852
rs4334895	G	44105438	4.12E-01	1.037	1658	0.273	11555	0.266
rs6859263	G	44111594	9.65E-01	1.002	1628	0.170	10803	0.169
rs4242104	T	44122673	6.22E-01	1.028	1660	0.858	11563	0.854
rs4502833	C	44144266	6.53E-01	1.020	1625	0.304	11379	0.300
rs6871975	T	44166026	5.11E-01	1.029	1618	0.295	11263	0.289
rs4242107	A	44174584	7.26E-01	1.015	1660	0.278	11562	0.275
rs4242108	T	44174787	6.53E-01	1.025	1660	0.858	11563	0.854
rs4492117	C	44174878	3.54E-01	1.055	1660	0.136	11563	0.130
rs4596368	G	44194286	3.91E-01	1.053	1660	0.121	11562	0.116
rs4866869	G	44195892	7.44E-01	1.014	1660	0.309	11555	0.306
rs4296809	A	44232169	6.18E-01	1.026	1660	0.152	11552	0.149
rs4866880	A	44237263	8.77E-01	1.007	1660	0.302	11562	0.301
rs4866773	A	44264014	7.86E-01	1.015	1659	0.850	11537	0.848
rs1550939	G	44267213	7.20E-01	1.015	1660	0.299	11563	0.296
rs4643965	C	44270309	8.24E-01	1.010	1660	0.259	11562	0.257
rs10512836	T	44273926	6.14E-01	1.025	1660	0.794	11563	0.790
rs726941	C	44279279	9.02E-02	1.078	1660	0.271	11563	0.257
rs2053784	C	44286984	6.31E-01	1.021	1660	0.287	11562	0.283
rs7713769	G	44314707	3.23E-01	1.082	1660	0.684	11561	0.678
rs1011814	G	44381321	2.23E-01	1.051	1660	0.660	11563	0.649
rs11743802	T	44396855	2.16E-02	1.134	1656	0.857	11539	0.840
rs2121875	T	44411048	2.14E-01	1.052	1660	0.660	11563	0.649
rs1384449	A	44422561	7.95E-01	1.012	1649	0.763	11202	0.761
rs2973644	C	44429684	9.54E-01	1.003	1660	0.232	11562	0.232
rs10512852	C	44439070	1.82E-01	1.101	1660	0.923	11562	0.915
rs723166	C	44441516	8.00E-02	1.081	1658	0.740	11537	0.725
rs16901843	T	44448618	4.81E-01	1.036	1660	0.819	11561	0.814
rs4856898	A	44449132	1.00E+00	1.000	1660	0.146	11553	0.146
rs13357659	G	44478386	1.53E-01	1.059	1660	0.406	11562	0.392
rs1351637	G	44497204	8.87E-01	1.007	1660	0.197	11563	0.196
rs922853	G	44497553	9.48E-01	1.004	1660	0.108	11562	0.108
rs1120718	T	44512079	6.10E-02	1.089	1660	0.761	11563	0.746
rs1384450	C	44525645	3.55E-01	1.037	1660	0.585	11562	0.576
rs2082140	T	44541916	1.84E-01	1.112	1660	0.068	11563	0.060
rs17320222	A	44544792	1.78E-01	1.142	1657	0.960	11552	0.955
rs6856565	C	44584014	3.26E-01	1.052	1659	0.176	11553	0.169
rs4463187	G	44614156	1.43E-02	1.100	1660	0.532	11561	0.508
rs7706449	A	44614727	5.00E-02	1.084	1660	0.355	11561	0.337
rs6889804	T	44618310	3.31E-01	1.051	1660	0.176	11561	0.169
rs4642379	T	44620489	3.27E-01	1.052	1660	0.176	11562	0.169
rs6896299	G	44655907	3.37E-01	1.051	1658	0.176	11552	0.168
rs4529201	C	44658472	1.09E-02	1.105	1659	0.531	11539	0.506
rs4415084	T	44706016	9.02E-06	1.194	1660	0.415	11562	0.373
rs2218080	G	44758831	3.37E-05	1.181	1660	0.405	11562	0.366
rs11747159	T	44763211	7.62E-06	1.197	1658	0.397	11561	0.354
rs2330572	C	44766490	2.38E-05	1.184	1660	0.405	11561	0.365
rs994783	G	44768748	2.19E-05	1.185	1658	0.405	11560	0.365

SNP	Allele	Position bld34	P-value	OR	Cases	Frq.Cases	Controls	Frq.Controls
rs6885754	A	44811554	9.68E-01	1.007	1660	0.014	11560	0.013
rs7712949	T	44815846	1.19E-05	1.193	1659	0.392	11547	0.350
rs11746980	A	44823379	1.98E-05	1.186	1660	0.405	11561	0.365
rs16901964	T	44826756	1.97E-05	1.188	1659	0.390	11560	0.350
rs727305	C	44841543	1.51E-05	1.191	1660	0.390	11516	0.349
rs10462081	A	44846166	2.13E-05	1.187	1657	0.390	11557	0.350
rs13163209	A	44849250	1.92E-05	1.188	1660	0.390	11555	0.350
rs13159598	G	44851427	3.31E-05	1.182	1644	0.402	11513	0.363
rs3751648	G	44853580	8.57E-06	1.202	1589	0.387	10844	0.344
rs13174122	C	44856241	3.63E-05	1.181	1658	0.390	11498	0.351
rs11746506	T	44858067	2.12E-05	1.187	1660	0.389	11561	0.350
rs12168871	A	44859505	1.50E-05	1.191	1657	0.390	11532	0.349
rs9637783	G	44865147	1.74E-05	1.190	1655	0.389	11520	0.349
rs4457089	T	44867237	2.06E-05	1.187	1660	0.389	11560	0.349
rs6867533	T	44872793	6.23E-06	1.200	1641	0.398	11355	0.355
rs6896350	C	44878072	2.09E-05	1.187	1660	0.389	11559	0.350
rs1371025	C	44879734	2.08E-05	1.187	1660	0.389	11557	0.349
rs6451775	G	44882269	2.21E-05	1.187	1660	0.389	11561	0.350
rs729599	C	44887761	2.21E-05	1.187	1660	0.389	11561	0.350
rs987394	T	44891679	2.16E-05	1.187	1660	0.389	11558	0.349
rs4440370	A	44898853	2.17E-05	1.187	1659	0.389	11559	0.350
rs4492119	A	44901115	7.16E-06	1.200	1645	0.387	11340	0.345
rs7703497	A	44902529	2.12E-05	1.187	1659	0.389	11559	0.349
rs4395640	T	44914601	8.63E-06	1.197	1652	0.395	11497	0.353
rs7716600	A	44920506	3.12E-05	1.214	1660	0.241	11560	0.208
rs4412123	T	44921769	1.10E-05	1.192	1660	0.409	11560	0.367
rs7705343	G	44925078	1.31E-05	1.190	1658	0.409	11558	0.368
rs4129642	G	44943630	1.11E-05	1.194	1655	0.396	11518	0.355
rs9790879	C	44945386	1.27E-05	1.191	1659	0.409	11561	0.368
rs10462084	G	44946850	2.31E-01	1.087	1659	0.155	11562	0.147
rs9791058	T	44949392	7.66E-06	1.197	1660	0.396	11562	0.354
rs6880275	T	44954436	7.76E-06	1.198	1648	0.395	11484	0.353
rs6870136	G	44956163	7.51E-06	1.197	1660	0.396	11559	0.354
rs6881563	C	44958354	7.28E-06	1.198	1660	0.396	11561	0.354
rs7703818	G	44960080	6.93E-06	1.198	1658	0.396	11557	0.354
rs10077814	C	44962290	1.24E-05	1.191	1659	0.407	11561	0.366
rs6451783	G	44963794	9.41E-06	1.195	1660	0.395	11563	0.353
rs4298259	G	44966212	9.35E-06	1.195	1660	0.395	11562	0.353
rs7728431	T	44968180	9.96E-06	1.195	1660	0.394	11559	0.353
rs12517690	A	44984794	9.58E-06	1.195	1660	0.395	11562	0.353
rs3935213	A	45006945	3.69E-02	1.112	1659	0.182	11561	0.167
rs6866995	C	45022348	3.77E-02	1.112	1660	0.182	11563	0.167
rs2067980	G	45027818	9.89E-04	1.200	1660	0.155	11555	0.132
rs3923826	C	45118278	2.49E-02	1.137	1660	0.870	11562	0.855
rs11743309	A	45167689	2.15E-01	1.067	1660	0.636	11559	0.626
rs12654948	T	45211215	1.28E-01	1.094	1651	0.878	11260	0.686
rs12515620	G	45236970	4.07E-02	1.126	1660	0.135	11558	0.122
rs12515179	C	45292596	4.66E-02	1.120	1660	0.138	11562	0.125
rs10512876	G	45295105	7.73E-01	1.025	1660	0.055	11561	0.054
rs10035564	G	45298001	1.79E-04	1.176	1650	0.291	11477	0.258
rs13180087	C	45311269	5.26E-02	1.127	1659	0.117	11563	0.105
rs4866929	A	45312090	9.34E-02	1.068	1660	0.484	11560	0.468
rs981782	T	45331219	2.28E-04	1.159	1601	0.458	10706	0.421

SNP	Allele	Position bid34	P-value	OR	Cases	Frq.Cases	Controls	Frq.Controls
rs981782	T	45331219	1.05E-01	1.066	1660	0.470	11562	0.454
rs9790873	C	45337015	3.61E-02	1.133	1660	0.128	11554	0.115
rs9292918	G	45346536	4.21E-03	1.161	1656	0.176	11546	0.155
rs6895055	A	45366909	2.34E-03	1.172	1660	0.177	11551	0.155
rs6888352	A	45371416	2.72E-03	1.169	1659	0.177	11551	0.155
rs994092	G	45381260	2.49E-03	1.171	1659	0.177	11557	0.155
rs10473384	A	45389311	2.34E-03	1.172	1659	0.177	11560	0.155
rs1501357	G	45410376	2.61E-03	1.170	1660	0.177	11563	0.155
rs12517615	C	45412289	6.70E-02	1.190	1647	0.047	11288	0.040
rs1501362	T	45423708	2.85E-03	1.167	1660	0.179	11558	0.157
rs6451796	T	45433355	3.11E-03	1.166	1660	0.179	11560	0.157
rs6414906	C	45451822	9.05E-03	1.112	1658	0.375	11560	0.351
rs13162651	C	45455401	5.54E-02	1.081	1655	0.375	11363	0.357
rs1483310	G	45459859	7.27E-02	1.178	1655	0.051	11413	0.044
rs12559024	T	45463191	9.87E-01	1.001	1658	0.080	11549	0.080
rs6892290	G	45472638	1.08E-02	1.109	1660	0.375	11560	0.351
rs6451801	A	45484934	1.02E-02	1.110	1660	0.375	11560	0.351
rs13354786	C	45502578	1.16E-02	1.108	1660	0.377	11563	0.353
rs12651887	T	45515924	1.19E-01	1.076	1657	0.225	11548	0.212
rs1471683	A	45524926	7.07E-02	1.086	1660	0.232	11562	0.217
rs2337414	A	45614170	1.12E-01	1.100	1660	0.884	11562	0.874
rs1852598	G	45648477	2.93E-01	1.049	1660	0.245	11562	0.237
rs11743392	T	45660476	4.43E-04	1.150	1614	0.490	10826	0.455
rs1534391	T	45714539	8.74E-02	1.117	1652	0.884	11396	0.873
rs2878074	C	45761710	1.49E-02	1.103	1660	0.384	11560	0.361
rs4380674	C	45814898	2.50E-01	1.053	1660	0.259	11563	0.249
rs7447717	A	45815753	1.29E-02	1.105	1660	0.384	11559	0.361
rs4388219	A	45826251	1.41E-01	1.063	1660	0.328	11555	0.315
rs10941703	C	45827373	1.02E-01	1.069	1660	0.384	11561	0.349
rs10941704	G	45829007	2.99E-01	1.049	1657	0.245	11531	0.236
rs4551074	T	45838254	2.15E-01	1.057	1660	0.260	11561	0.249
rs13155321	G	45842047	9.36E-02	1.071	1660	0.367	11559	0.351
rs7733616	C	45844224	1.59E-01	1.063	1630	0.290	11372	0.278
rs10069793	G	45848652	6.09E-02	1.079	1650	0.367	11485	0.349
rs13176359	T	45859776	2.72E-01	1.050	1660	0.267	11552	0.256
rs4866973	A	45862187	2.03E-01	1.059	1657	0.259	11530	0.246
rs4242126	A	45871874	2.60E-01	1.051	1660	0.245	11561	0.236
rs6885429	G	45876009	9.87E-02	1.069	1659	0.368	11561	0.351
rs10461763	G	45884708	5.07E-01	1.075	1659	0.034	11557	0.032
rs7730617	A	45885804	1.53E-01	1.063	1659	0.294	11557	0.281
rs13175559	C	45900128	1.10E-01	1.067	1660	0.368	11560	0.353
rs11951003	A	45902630	1.09E-01	1.067	1659	0.368	11559	0.353
rs4331911	G	45904876	1.55E-01	1.060	1627	0.360	11297	0.347
rs13340341	T	45906437	1.53E-01	1.060	1660	0.363	11556	0.349
rs12109205	G	45922179	2.53E-01	1.051	1660	0.288	11559	0.278
rs4368738	T	45936289	1.57E-01	1.061	1659	0.294	11561	0.281
rs9637799	G	45946109	5.03E-01	1.075	1660	0.034	11560	0.032
rs6862657	A	45951498	1.03E-01	1.069	1656	0.366	11533	0.351
rs7443189	A	45958468	6.09E-01	1.057	1654	0.034	11519	0.032
rs11948152	C	46000128	1.69E-01	1.054	1644	0.580	11377	0.567
rs7443384	G	46006394	1.01E-01	1.070	1660	0.345	11562	0.330
rs7701444	C	46014363	1.05E-01	1.068	1660	0.369	11561	0.354
rs13352566	T	46042403	2.03E-01	1.059	1648	0.260	11420	0.249

SNP	Allele	Position bld34	P-value	OR	Cases	Frq.Cases	Controls	Frq.Controls
rs4370277	G	46094169	2.40E-01	1.048	1636	0.402	11316	0.390
rs12173206	A	46121458	5.02E-01	1.078	1659	0.034	11551	0.032
rs10086479	G	46142988	1.85E-01	1.054	1660	0.401	11550	0.388
rs12515804	C	46144107	1.69E-01	1.056	1660	0.401	11561	0.388
rs13175755	T	46145261	2.58E-01	1.083	1660	0.086	11563	0.080
rs7720482	T	46151684	1.66E-01	1.057	1659	0.402	11559	0.388
rs4975890	C	46159139	2.75E-01	1.047	1656	0.330	11552	0.320
rs12690679	C	46206288	3.19E-01	1.040	1659	0.423	11555	0.414
rs12697527	T	46247605	2.98E-01	1.042	1658	0.419	11533	0.409
rs13168297	T	46273834	5.02E-01	1.028	1649	0.348	11436	0.342
rs4975957	A	46310803	2.97E-01	1.042	1660	0.420	11549	0.410
rs12659648	C	46330355	4.96E-01	1.028	1652	0.349	11541	0.343
rs13355128	T	46332613	3.32E-01	1.041	1614	0.649	10822	0.639
rs10941803	C	46393984	3.56E-01	1.038	1613	0.343	11166	0.335

Shown are the SNP names, the identity of the risk allele, the location (in NCBI Build 34 coordinates, the P-value and Odds Ratio (OR) for Breast Cancer association, the numbers of individuals tested and the allele frequencies in the Breast Cancer Case and the Control groups respectively.

Table 2. Replication of signal from SNP rs7703618 in an independent Icelandic Breast Cancer Case/Control sample:

rs7703618 (G)	Frequency			
Cohort (Cases/Controls)	Cases	Controls	OR	P
Iceland 1 (1599/11558)	0.396	0.354	1.20	1.1E-05
Iceland 2 (591/1314)	0.392	0.353	1.18	2.9E-02
Iceland combined (2190/12872)	0.395	0.354	1.19	5.3E-07

Table 3. HapMap SNPs with r^2 values >0.2 in relation to key SNPs in equivalence classes A-F.

SNP 1 is the Illumina SNP with the lowest P-value in each of equivalence classes A-F. SNP 2 is the HapMap SNP that is correlated to SNP 1. D´ is the mean D´ value across all combinations of the alleles of SNP 1 and SNP 2. R2 is the square of the correlation coefficient between the two SNPs. p-min is the P-value corresponding to the strongest linkage disequilibrium observed between alleles of SNP 1 and SNP 2.

SNP A = rs4415084				
SNP 1	SNP 2	D'	R2	p-min
rs4415084	rs7735881	1.000	1.000	3.35E-36
rs4415084	rs7723539	1.000	1.000	3.35E-36
rs4415084	rs4492118	1.000	1.000	5.45E-36
rs4415084	rs4463168	1.000	1.000	1.87E-35
rs4415084	rs920329	1.000	1.000	7.12E-36
rs4415084	rs7720551	1.000	1.000	3.35E-36
rs4415084	rs714130	1.000	1.000	3.35E-36
rs4415084	rs6874055	1.000	1.000	3.37E-36
rs4415084	rs6861560	1.000	1.000	3.35E-36
rs4415084	rs6451770	1.000	1.000	3.35E-36
rs4415084	rs4571480	1.000	1.000	1.43E-35
rs4415084	rs4419600	1.000	1.000	3.35E-36
rs4415084	rs4415085	1.000	1.000	3.35E-36
rs4415084	rs2218081	1.000	1.000	3.35E-36
rs4415084	rs2165010	1.000	1.000	3.37E-36
rs4415084	rs2165009	1.000	1.000	5.45E-36
rs4415084	rs2013513	1.000	1.000	1.43E-35
rs4415084	rs1821936	1.000	1.000	5.45E-36
rs4415084	rs1436825	1.000	1.000	3.35E-36
rs4415084	rs13156930	1.000	1.000	3.35E-36
rs4415084	rs12522626	1.000	1.000	8.86E-36
rs4415084	rs12515012	1.000	1.000	3.35E-36
rs4415084	rs12187196	1.000	1.000	3.35E-36
rs4415084	rs10941678	1.000	1.000	5.45E-36
rs4415084	rs4321755	1.000	1.000	3.35E-36
rs4415084	rs10941677	1.000	1.000	1.43E-35
rs4415084	rs10805885	1.000	1.000	5.45E-36
rs4415084	rs16901937	1.000	0.965	4.27E-34
rs4415084	rs920328	1.000	0.931	3.70E-32
rs4415084	rs7380559	0.923	0.768	1.13E-23
rs4415084	rs4518409	0.923	0.766	1.13E-23
rs4415084	rs1436821	0.923	0.766	1.13E-23
rs4415084	rs1436820	0.923	0.766	1.13E-23
rs4415084	rs13362132	0.923	0.766	1.13E-23
rs4415084	rs13160259	0.923	0.766	1.13E-23
rs4415084	rs11958808	0.923	0.766	1.13E-23
rs4415084	rs1061310	0.923	0.766	1.13E-23
rs4415084	rs10512665	0.923	0.766	1.13E-23
rs4415084	rs1048758	0.923	0.766	1.13E-23
rs4415084	rs10044096	0.923	0.766	1.13E-23
rs4415084	rs9292913	0.922	0.765	1.72E-23
rs4415084	rs13177711	0.922	0.765	1.72E-23
rs4415084	rs11949847	0.921	0.764	2.62E-23
rs4415084	rs4329026	0.923	0.763	2.84E-23
rs4415084	rs7716571	0.923	0.763	2.33E-23
rs4415084	rs7711697	0.922	0.762	1.14E-22
rs4415084	rs7380878	0.921	0.762	4.30E-23
rs4415084	rs10043344	0.921	0.761	6.56E-23
rs4415084	rs10040082	0.957	0.738	2.38E-22
rs4415084	rs7717459	0.919	0.708	1.84E-21
rs4415084	rs6893319	0.919	0.706	1.84E-21
rs4415084	rs6872254	0.919	0.706	1.84E-21
rs4415084	rs6451778	0.919	0.706	1.84E-21
rs4415084	rs4373287	0.919	0.706	1.84E-21

rs4415084	rs1868405	0.919	0.706	1.84E-21
rs4415084	rs1438822	0.919	0.706	1.84E-21
rs4415084	rs1438819	0.919	0.706	1.84E-21
rs4415084	rs13169120	0.919	0.706	1.84E-21
rs4415084	rs13155698	0.919	0.706	1.84E-21
rs4415084	rs13154781	0.919	0.706	1.84E-21
rs4415084	rs10462060	0.919	0.706	1.84E-21
rs4415084	rs10065638	0.919	0.706	1.84E-21
rs4415084	rs10059086	0.919	0.706	1.84E-21
rs4415084	rs10057521	0.919	0.706	1.84E-21
rs4415084	rs10053247	0.919	0.706	1.84E-21
rs4415084	rs10041518	0.919	0.706	1.84E-21
rs4415084	rs10040488	0.919	0.706	1.84E-21
rs4415084	rs10039866	0.919	0.706	1.84E-21
rs4415084	rs12513749	0.919	0.706	3.69E-21
rs4415084	rs6875933	0.919	0.706	3.67E-21
rs4415084	rs7736092	0.918	0.705	2.73E-21
rs4415084	rs10070037	0.918	0.705	2.73E-21
rs4415084	rs6871052	0.918	0.705	5.47E-21
rs4415084	rs7708506	0.917	0.704	4.05E-21
rs4415084	rs4642377	0.920	0.704	1.85E-21
rs4415084	rs4457088	0.919	0.703	2.72E-21
rs4415084	rs10038554	0.918	0.703	2.72E-21
rs4415084	rs3747479	0.919	0.701	3.77E-21
rs4415084	rs6894324	0.918	0.701	6.83E-21
rs4415084	rs9790896	0.880	0.700	8.38E-21
rs4415084	rs6875267	0.915	0.699	4.58E-20
rs4415084	rs6868232	0.918	0.697	1.16E-20
rs4415084	rs11741772	0.918	0.697	2.05E-20
rs4415084	rs9292914	0.904	0.694	6.03E-18
rs4415084	rs11951760	0.912	0.694	6.63E-20
rs4415084	rs12518851	0.908	0.685	2.68E-18
rs4415084	rs7715731	0.915	0.685	1.60E-19
rs4415084	rs1438827	0.881	0.675	2.82E-20
rs4415084	rs12651949	0.911	0.665	4.20E-16
rs4415084	rs11948186	0.914	0.649	1.74E-19
rs4415084	rs10051592	0.914	0.649	1.74E-19
rs4415084	rs16902086	0.802	0.559	2.08E-16
rs4415084	rs3935088	0.905	0.537	5.09E-16
rs4415084	rs10512875	0.901	0.517	6.18E-15
rs4415084	rs10941679	1.000	0.513	5.36E-17
rs4415084	rs4613718	1.000	0.454	3.93E-17
rs4415084	rs830395	1.000	0.402	1.56E-13
rs4415084	rs10044408	1.000	0.330	6.94E-11
rs4415084	rs6869488	0.856	0.287	6.68E-09
rs4415084	rs4460145	0.856	0.287	6.68E-09
rs4415084	rs7709262	0.847	0.275	5.02E-08
rs4415084	rs6874127	0.847	0.273	8.18E-08
rs4415084	rs13183434	1.000	0.268	2.69E-09
rs4415084	rs7716101	0.843	0.264	9.59E-08
rs4415084	rs7709681	0.843	0.264	9.59E-08
rs4415084	rs6894974	0.843	0.264	9.59E-08
rs4415084	rs6865307	0.843	0.264	9.59E-08
rs4415084	rs4533894	0.843	0.264	9.59E-08
rs4415084	rs12521639	0.843	0.264	9.59E-08
rs4415084	rs12054976	0.843	0.264	9.59E-08
rs4415084	rs7731099	0.839	0.264	1.12E-07
rs4415084	rs7701679	0.841	0.263	1.15E-07
rs4415084	rs6862655	0.739	0.262	3.65E-08
rs4415084	rs10059745	0.739	0.262	3.65E-08
rs4415084	rs6451796	0.840	0.255	1.80E-07
rs4415084	rs3923055	0.840	0.255	1.80E-07
rs4415084	rs1501361	0.840	0.255	1.80E-07
rs4415084	rs1392973	0.840	0.255	1.80E-07

rs4415084	rs5866354	0.734	0.254	5.71E-08
rs4415084	rs4639238	0.736	0.254	5.70E-08
rs4415084	rs12374507	0.736	0.254	5.70E-08
rs4415084	rs10066953	0.736	0.254	5.70E-08
rs4415084	rs4371761	0.839	0.252	2.37E-07
rs4415084	rs10054521	0.733	0.249	8.38E-08
rs4415084	rs4502832	0.838	0.246	1.74E-07
rs4415084	rs4485937	0.838	0.246	1.74E-07
rs4415084	rs4369695	0.838	0.246	1.74E-07
rs4415084	rs4286810	0.838	0.246	1.74E-07
rs4415084	rs10941692	0.838	0.246	1.74E-07
rs4415084	rs12522398	0.832	0.243	2.89E-07
rs4415084	rs4866900	0.923	0.232	4.31E-08
rs4415084	rs4493682	0.816	0.212	3.37E-06
rs4415084	rs4306490	0.811	0.210	4.58E-06
rs4415084	rs6893494	0.814	0.206	3.96E-06
rs4415084	rs12523157	0.814	0.206	3.96E-06
rs4415084	rs11954598	0.814	0.206	3.96E-06
rs4415084	rs7720104	0.809	0.205	5.34E-06
rs4415084	rs6884149	0.810	0.204	4.47E-06
rs4415084	rs983940	1.000	0.204	4.35E-09
rs4415084	rs6451767	1.000	0.204	4.35E-09
rs4415084	rs1482669	1.000	0.204	4.35E-09
rs4415084	rs1351633	1.000	0.204	4.35E-09
rs4415084	rs10079222	1.000	0.204	4.35E-09
rs4415084	rs12514414	0.804	0.202	7.18E-06
rs4415084	rs6876773	0.802	0.201	0.00001
rs4415084	rs7711445	0.802	0.201	0.000012
SNP B = rs7703618				
SNP 1	SNP 2	D'	R2	p-min
rs7703618	rs9292914	1.000	1.000	1.09E-30
rs7703618	rs7717459	1.000	1.000	2.19E-35
rs7703618	rs7715731	1.000	1.000	3.63E-33
rs7703618	rs7736092	1.000	1.000	3.33E-35
rs7703618	rs7708505	1.000	1.000	3.33E-35
rs7703618	rs6875267	1.000	1.000	2.26E-34
rs7703618	rs10041518	1.000	1.000	2.19E-35
rs7703618	rs10039866	1.000	1.000	2.19E-35
rs7703618	rs10038554	1.000	1.000	9.62E-35
rs7703618	rs6894324	1.000	1.000	9.62E-35
rs7703618	rs6893319	1.000	1.000	2.19E-35
rs7703618	rs6875933	1.000	1.000	2.19E-35
rs7703618	rs6872254	1.000	1.000	2.19E-35
rs7703618	rs6871052	1.000	1.000	3.33E-35
rs7703618	rs6868232	1.000	1.000	1.90E-34
rs7703618	rs6451778	1.000	1.000	2.19E-35
rs7703618	rs4642377	1.000	1.000	6.45E-35
rs7703618	rs4457098	1.000	1.000	9.62E-35
rs7703618	rs4373287	1.000	1.000	2.19E-35
rs7703618	rs3747479	1.000	1.000	1.44E-34
rs7703618	rs1866406	1.000	1.000	2.19E-35
rs7703618	rs1438822	1.000	1.000	2.19E-35
rs7703618	rs1438819	1.000	1.000	2.19E-35
rs7703618	rs13189120	1.000	1.000	2.19E-35
rs7703618	rs13155698	1.000	1.000	2.19E-35
rs7703618	rs1315478I	1.000	1.000	2.19E-35
rs7703618	rs12651949	1.000	1.000	1.69E-31
rs7703618	rs12518851	1.000	1.000	5.31E-33
rs7703618	rs12513749	1.000	1.000	2.19E-35
rs7703618	rs11951760	1.000	1.000	1.16E-33
rs7703618	rs11741772	1.000	1.000	3.34E-34
rs7703618	rs10462080	1.000	1.000	2.19E-35
rs7703618	rs10070037	1.000	1.000	3.33E-35

rs7703618	rs10065638	1.000	1.000	2.19E-35
rs7703618	rs10059086	1.000	1.000	2.19E-35
rs7703618	rs10057521	1.000	1.000	2.19E-35
rs7703618	rs10053247	1.000	1.000	2.19E-35
rs7703618	rs10040488	1.000	1.000	2.19E-35
rs7703618	rs10040082	1.000	1.000	1.50E-34
rs7703618	rs1438827	1.000	0.964	2.55E-33
rs7703618	rs7711697	1.000	0.929	2.12E-31
rs7703618	rs11958808	1.000	0.929	7.53E-32
rs7703618	rs10044096	1.000	0.929	7.53E-32
rs7703618	rs7380559	1.000	0.929	7.53E-32
rs7703618	rs4518409	1.000	0.929	7.53E-32
rs7703618	rs4329028	1.000	0.929	2.12E-31
rs7703618	rs1438821	1.000	0.929	7.53E-32
rs7703618	rs1438820	1.000	0.929	7.53E-32
rs7703618	rs13362132	1.000	0.929	7.53E-32
rs7703618	rs13160259	1.000	0.929	7.53E-32
rs7703618	rs1081310	1.000	0.929	7.53E-32
rs7703618	rs10512865	1.000	0.929	7.53E-32
rs7703618	rs1048758	1.000	0.929	7.53E-32
rs7703618	rs7716571	1.000	0.929	4.57E-31
rs7703618	rs9292913	1.000	0.928	1.15E-31
rs7703618	rs7380878	1.000	0.928	3.22E-31
rs7703618	rs13177711	1.000	0.928	1.15E-31
rs7703618	rs10043344	1.000	0.928	4.91E-31
rs7703618	rs11949847	1.000	0.928	1.75E-31
rs7703618	rs9790896	0.862	0.856	7.84E-27
rs7703618	rs920328	0.923	0.764	1.31E-23
rs7703618	rs4571460	0.922	0.731	2.33E-22
rs7703618	rs11948186	0.883	0.723	1.25E-21
rs7703618	rs10051592	0.883	0.723	1.25E-21
rs7703618	rs7735881	0.921	0.709	7.34E-22
rs7703618	rs7723539	0.921	0.709	7.34E-22
rs7703618	rs714130	0.921	0.709	7.34E-22
rs7703618	rs6861560	0.921	0.709	7.34E-22
rs7703618	rs5451770	0.921	0.709	7.34E-22
rs7703618	rs4419600	0.921	0.709	7.34E-22
rs7703618	rs4415085	0.921	0.709	7.34E-22
rs7703618	rs4321755	0.921	0.709	7.34E-22
rs7703618	rs2218081	0.921	0.709	7.34E-22
rs7703618	rs1438825	0.921	0.709	7.34E-22
rs7703618	rs13156930	0.921	0.708	7.34E-22
rs7703618	rs12515012	0.921	0.708	7.34E-22
rs7703618	rs12187196	0.921	0.708	7.34E-22
rs7703618	rs7720551	0.920	0.707	1.46E-21
rs7703618	rs6874055	0.919	0.706	1.84E-21
rs7703618	rs2165010	0.919	0.706	1.84E-21
rs7703618	rs920329	0.920	0.706	2.91E-21
rs7703618	rs10805885	0.921	0.706	1.09E-21
rs7703618	rs1821936	0.921	0.705	2.18E-21
rs7703618	rs10941678	0.921	0.705	2.18E-21
rs7703618	rs4463189	0.918	0.704	3.64E-21
rs7703618	rs2013513	0.919	0.704	2.72E-21
rs7703618	rs10941677	0.919	0.704	2.72E-21
rs7703618	rs4492118	0.921	0.703	1.62E-21
rs7703618	rs2165009	0.921	0.703	1.62E-21
rs7703618	rs12522626	0.921	0.703	1.62E-21
rs7703618	rs16901937	0.920	0.662	4.51E-21
rs7703618	rs16902086	0.812	0.635	1.69E-18
rs7703618	rs3935086	0.865	0.570	1.75E-16
rs7703618	rs10512875	0.865	0.570	1.75E-16
rs7703618	rs930395	1.000	0.482	9.85E-16
rs7703618	rs10941679	0.842	0.435	2.26E-12
rs7703618	rs4813718	1.000	0.384	5.94E-15

rs7703618	rs4502632	0.925	0.349	3.00E-10
rs7703618	rs4485937	0.925	0.349	3.00E-10
rs7703618	rs4389695	0.925	0.349	3.00E-10
rs7703618	rs10941692	0.925	0.349	3.00E-10
rs7703618	rs12522398	0.923	0.349	5.27E-10
rs7703618	rs6869468	0.865	0.342	3.96E-10
rs7703618	rs4460145	0.865	0.342	3.96E-10
rs7703618	rs10044408	0.917	0.334	7.15E-09
rs7703618	rs7731099	0.856	0.327	2.53E-09
rs7703618	rs7716101	0.856	0.317	3.51E-09
rs7703618	rs7709661	0.856	0.317	3.51E-09
rs7703618	rs6894974	0.856	0.317	3.51E-09
rs7703618	rs6885307	0.856	0.317	3.51E-09
rs7703618	rs4533894	0.856	0.317	3.51E-09
rs7703618	rs12521639	0.856	0.317	3.51E-09
rs7703618	rs12054976	0.856	0.317	3.51E-09
rs7703618	rs7701679	0.852	0.315	8.33E-09
rs7703618	rs4371761	0.852	0.305	9.59E-09
rs7703618	rs4296810	0.851	0.296	7.62E-09
rs7703618	rs1909937	0.911	0.282	2.53E-08
rs7703618	rs16902069	0.911	0.282	2.53E-08
rs7703618	rs1472584	0.911	0.282	2.53E-08
rs7703618	rs1392970	0.911	0.282	2.53E-08
rs7703618	rs12523398	0.911	0.282	2.53E-08
rs7703618	rs12521953	0.911	0.282	2.53E-08
rs7703618	rs12516488	0.911	0.282	2.53E-08
rs7703618	rs12514615	0.911	0.282	2.53E-08
rs7703618	rs12153189	0.911	0.282	2.53E-08
rs7703618	rs12153053	0.911	0.282	2.53E-08
rs7703618	rs11953498	0.911	0.282	2.53E-08
rs7703618	rs10941693	0.911	0.282	2.53E-08
rs7703618	rs4357042	0.910	0.282	2.95E-08
rs7703618	rs12523359	0.910	0.282	2.95E-08
rs7703618	rs12522305	0.910	0.282	2.95E-08
rs7703618	rs4533895	0.909	0.281	3.44E-08
rs7703618	rs6898476	0.907	0.281	4.74E-08
rs7703618	rs6451795	0.786	0.267	1.07E-07
rs7703618	rs3923055	0.786	0.267	1.07E-07
rs7703618	rs1501361	0.786	0.267	1.07E-07
rs7703618	rs1392973	0.786	0.267	1.07E-07
rs7703618	rs4566805	0.903	0.268	1.31E-07
rs7703618	rs12520430	0.906	0.266	5.16E-08
rs7703618	rs1405918	0.636	0.265	6.95E-08
rs7703618	rs13183434	0.907	0.262	5.17E-08
rs7703618	rs7446090	0.835	0.259	1.68E-07
rs7703618	rs4493662	0.833	0.259	1.97E-07
rs7703618	rs7720104	0.831	0.256	2.29E-07
rs7703618	rs4308490	0.831	0.256	2.75E-07
rs7703618	rs7711444	0.901	0.254	4.16E-07
rs7703618	rs6893494	0.831	0.250	2.47E-07
rs7703618	rs12523157	0.831	0.250	2.47E-07
rs7703618	rs11954598	0.831	0.250	2.47E-07
rs7703618	rs6864149	0.827	0.249	2.79E-07
rs7703618	rs12514414	0.823	0.247	4.51E-07
rs7703618	rs12520124	0.591	0.247	6.18E-07
rs7703618	rs6876773	0.822	0.246	6.39E-07
rs7703618	rs13187565	0.603	0.246	1.02E-06
rs7703618	rs7711446	0.821	0.246	7.62E-07
rs7703618	rs2337483	0.623	0.246	2.67E-07
rs7703618	rs7709262	0.727	0.242	5.48E-07
rs7703618	rs2580260	0.598	0.241	3.14E-07
rs7703618	rs6874127	0.726	0.240	8.96E-07
rs7703618	rs2569162	0.617	0.238	6.61E-07
rs7703618	rs5890289	0.823	0.237	7.69E-07

rs7703618	rs6892627	0.593	0.231	5.31E-07
rs7703618	rs6451814	0.593	0.231	5.31E-07
rs7703618	rs2337952	0.593	0.231	5.31E-07
rs7703618	rs2049656	0.593	0.231	5.31E-07
rs7703618	rs1483303	0.593	0.231	5.31E-07
rs7703618	rs17343002	1.000	0.229	2.29E-09
rs7703618	rs6893773	0.584	0.225	9.49E-07
rs7703618	rs2337951	0.584	0.225	9.49E-07
rs7703618	rs10078625	0.584	0.225	9.49E-07
rs7703618	rs10036065	0.581	0.224	7.95E-07
rs7703618	rs7705696	0.571	0.218	1.42E-06
rs7703618	rs2625494	0.568	0.218	1.15E-06
rs7703618	rs2580258	0.586	0.216	1.15E-06
rs7703618	rs1351720	0.568	0.218	1.15E-06
rs7703618	rs12110137	0.586	0.218	1.15E-06
rs7703618	rs10073636	0.586	0.218	1.15E-06
rs7703618	rs10043792	0.586	0.218	1.15E-06
rs7703618	rs1364732	0.573	0.215	2.06E-06
rs7703618	rs6451810	0.584	0.214	1.52E-06
rs7703618	rs6860200	0.560	0.212	2.47E-06
rs7703618	rs756048	0.567	0.211	2.41E-06
rs7703618	rs7732970	0.554	0.208	2.28E-06
rs7703618	rs6892594	0.554	0.208	2.28E-06
rs7703618	rs6451804	0.554	0.208	2.28E-06
rs7703618	rs7444176	0.572	0.208	2.55E-06
rs7703618	rs9687260	0.550	0.205	2.05E-06
rs7703618	rs7708116	0.550	0.205	2.05E-06
rs7703618	rs1501356	0.750	0.204	2.77E-06
rs7703618	rs12656983	0.750	0.204	2.77E-06
SNP C = rs2067980				
SNP 1	SNP 2	D'	R2	p-min
rs2067980	rs13163434	0.931	0.863	1.68E-17
rs2067980	rs10044408	0.767	0.444	2.19E-09
rs2067980	rs1501358	0.713	0.436	1.93E-09
rs2067980	rs12656983	0.713	0.436	1.93E-09
rs2067980	rs6451796	0.779	0.434	7.64E-10
rs2067980	rs3923055	0.779	0.434	7.64E-10
rs2067980	rs1501361	0.779	0.434	7.64E-10
rs2067980	rs1392973	0.779	0.434	7.64E-10
rs2067980	rs6874127	0.778	0.430	9.74E-10
rs2067980	rs7709262	0.776	0.408	1.92E-09
rs2067980	rs930395	0.774	0.383	4.51E-09
rs2067980	rs6451795	0.640	0.350	8.08E-08
rs2067980	rs11948186	1.000	0.329	8.10E-11
rs2067980	rs10051592	1.000	0.329	8.10E-11
rs2067980	rs6861150	0.577	0.324	3.93E-07
rs2067980	rs10473387	0.743	0.296	4.63E-07
rs2067980	rs16902086	1.000	0.294	4.00E-10
rs2067980	rs7711697	1.000	0.291	4.85E-10
rs2067980	rs7716571	1.000	0.288	1.57E-09
rs2067980	rs10941679	0.759	0.288	1.55E-07
rs2067980	rs10043344	1.000	0.286	8.31E-10
rs2067980	rs7360559	1.000	0.283	6.59E-10
rs2067980	rs4518409	1.000	0.283	6.59E-10
rs2067980	rs4329028	1.000	0.283	1.69E-09
rs2067980	rs1438821	1.000	0.283	6.59E-10
rs2067980	rs1438820	1.000	0.283	6.59E-10
rs2067980	rs13362132	1.000	0.283	6.59E-10
rs2067980	rs13160259	1.000	0.283	6.59E-10
rs2067980	rs11958608	1.000	0.283	6.59E-10
rs2067980	rs1061310	1.000	0.283	6.59E-10
rs2067980	rs10512665	1.000	0.283	6.59E-10
rs2067980	rs1048758	1.000	0.283	6.59E-10

rs2067980	rs10044096	1.000	0.283	6.59E-10
rs2067980	rs9292913	1.000	0.280	7.66E-10
rs2067980	rs7380878	1.000	0.280	2.18E-09
rs2067980	rs13177711	1.000	0.280	7.66E-10
rs2067980	rs11949647	1.000	0.278	8.92E-10
rs2067980	rs7721731	0.736	0.276	1.54E-06
rs2067980	rs9790896	1.000	0.275	1.04E-09
rs2067980	rs1483309	0.753	0.270	2.69E-07
rs2067980	rs1483306	0.753	0.270	2.69E-07
rs2067980	rs13358718	0.753	0.270	2.69E-07
rs2067980	rs10073055	0.753	0.270	2.69E-07
rs2067980	rs10472404	0.750	0.268	5.26E-07
rs2067980	rs12697498	0.741	0.268	2.58E-06
rs2067980	rs4463188	1.000	0.259	3.60E-09
rs2067980	rs4571480	1.000	0.258	2.32E-09
rs2067980	rs2013513	1.000	0.258	2.32E-09
rs2067980	rs10941677	1.000	0.258	2.32E-09
rs2067980	rs4532370	0.750	0.256	4.73E-07
rs2067980	rs1483312	0.750	0.256	4.73E-07
rs2067980	rs1351719	0.750	0.256	4.73E-07
rs2067980	rs12656485	0.750	0.256	4.73E-07
rs2067980	rs6877477	0.802	0.255	1.27E-06
rs2067980	rs6868232	0.900	0.254	3.11E-07
rs2067980	rs7715731	0.893	0.254	8.48E-07
rs2067980	rs7735881	1.000	0.254	2.71E-09
rs2067980	rs7723539	1.000	0.254	2.71E-09
rs2067980	rs7720551	1.000	0.254	2.71E-09
rs2067980	rs714130	1.000	0.254	2.71E-09
rs2067980	rs6874055	1.000	0.254	7.25E-09
rs2067980	rs6861560	1.000	0.254	2.71E-09
rs2067980	rs6451770	1.000	0.254	2.71E-09
rs2067980	rs4419600	1.000	0.254	2.71E-09
rs2067980	rs4415085	1.000	0.254	2.71E-09
rs2067980	rs4321755	1.000	0.254	2.71E-09
rs2067980	rs2218081	1.000	0.254	2.71E-09
rs2067980	rs2165010	1.000	0.254	7.25E-09
rs2067980	rs1438825	1.000	0.254	2.71E-09
rs2067980	rs13158830	1.000	0.254	2.71E-09
rs2067980	rs12515012	1.000	0.254	2.71E-09
rs2067980	rs12187196	1.000	0.254	2.71E-09
rs2067980	rs4452566	0.813	0.254	5.19E-07
rs2067980	rs10040082	0.906	0.252	1.42E-07
rs2067980	rs3747479	0.900	0.252	3.55E-07
rs2067980	rs920329	1.000	0.251	3.15E-09
rs2067980	rs1821938	1.000	0.251	3.15E-09
rs2067980	rs10941678	1.000	0.251	3.15E-09
rs2067980	rs10805685	1.000	0.251	3.15E-09
rs2067980	rs7717459	0.905	0.249	1.51E-07
rs2067980	rs6893319	0.905	0.249	1.51E-07
rs2067980	rs6875933	0.905	0.249	1.51E-07
rs2067980	rs6872254	0.905	0.249	1.51E-07
rs2067980	rs6451776	0.905	0.249	1.51E-07
rs2067980	rs4373287	0.905	0.249	1.51E-07
rs2067980	rs1866406	0.905	0.249	1.51E-07
rs2067980	rs1438822	0.905	0.249	1.51E-07
rs2067980	rs1438819	0.905	0.249	1.51E-07
rs2067980	rs13189120	0.905	0.249	1.51E-07
rs2067980	rs13155698	0.905	0.249	1.51E-07
rs2067980	rs13154781	0.905	0.249	1.51E-07
rs2067980	rs10462080	0.905	0.249	1.51E-07
rs2067980	rs10065636	0.905	0.249	1.51E-07
rs2067980	rs10059086	0.905	0.249	1.51E-07
rs2067980	rs10057521	0.905	0.249	1.51E-07
rs2067980	rs10053247	0.905	0.249	1.51E-07

rs2067980	rs10041518	0.905	0.249	1.51E-07
rs2067980	rs10040488	0.905	0.249	1.51E-07
rs2067980	rs10039866	0.905	0.249	1.51E-07
rs2067980	rs6875287	0.904	0.249	2.50E-07
rs2067980	rs4492118	1.000	0.249	3.66E-09
rs2067980	rs2165009	1.000	0.249	3.66E-09
rs2067980	rs12522626	1.000	0.249	3.66E-09
rs2067980	rs12513749	0.903	0.248	3.02E-07
rs2067980	rs7736092	0.905	0.247	1.73E-07
rs2067980	rs6871052	0.905	0.247	1.73E-07
rs2067980	rs10070037	0.905	0.247	1.73E-07
rs2067980	rs11741772	0.899	0.247	4.65E-07
rs2067980	rs4642377	0.899	0.247	4.29E-07
rs2067980	rs6451806	0.734	0.245	1.74E-06
rs2067980	rs16901937	1.000	0.245	4.22E-09
rs2067980	rs7708506	0.904	0.244	1.99E-07
rs2067980	rs6894324	0.899	0.244	4.89E-07
rs2067980	rs4457088	0.899	0.244	4.89E-07
rs2067980	rs10038554	0.899	0.244	4.89E-07
rs2067980	rs2678967	0.747	0.243	8.09E-07
rs2067980	rs1564684	0.747	0.243	8.09E-07
rs2067980	rs12651949	0.864	0.242	0.000056
rs2067980	rs12518851	0.896	0.240	1.34E-06
rs2067980	rs1438827	0.904	0.240	2.40E-07
rs2067980	rs9292914	0.887	0.237	2.40E-06
rs2067980	rs11951760	0.892	0.230	1.31E-06
rs2067980	rs7447532	1.000	0.222	0.000037
rs2067980	rs5004228	1.000	0.222	0.000037
rs2067980	rs11750364	1.000	0.222	0.000038
rs2067980	rs13357090	1.000	0.222	0.000037
rs2067980	rs920328	0.901	0.221	5.82E-07
rs2067980	rs10462095	0.737	0.221	3.73E-06
rs2067980	rs16902199	0.736	0.219	2.22E-06
rs2067980	rs7724971	0.740	0.219	2.21E-06
rs2067980	rs7719703	0.740	0.219	2.21E-06
rs2067980	rs7700252	0.740	0.219	2.21E-06
rs2067980	rs6898648	0.740	0.219	2.21E-06
rs2067980	rs6894784	0.740	0.219	2.21E-06
rs2067980	rs6894270	0.740	0.219	2.21E-06
rs2067980	rs6886950	0.740	0.219	2.21E-06
rs2067980	rs4437383	0.740	0.219	2.21E-06
rs2067980	rs2337954	0.740	0.219	2.21E-06
rs2067980	rs16902221	0.740	0.219	2.21E-06
rs2067980	rs16902217	0.740	0.219	2.21E-06
rs2067980	rs1483308	0.740	0.219	2.21E-06
rs2067980	rs13359915	0.740	0.219	2.21E-06
rs2067980	rs13357427	0.740	0.219	2.21E-06
rs2067980	rs12109155	0.740	0.219	2.21E-06
rs2067980	rs10473389	0.740	0.219	2.21E-06
rs2067980	rs10074312	0.740	0.219	2.21E-06
rs2067980	rs10068821	0.740	0.219	2.21E-06
rs2067980	rs1852595	0.739	0.217	2.48E-06
rs2067980	rs13159362	0.739	0.217	2.48E-06
rs2067980	rs10214369	0.739	0.217	2.48E-06
rs2067980	rs13361609	0.739	0.215	2.79E-06
rs2067980	rs7445730	0.723	0.211	6.63E-06
rs2067980	rs4339358	0.723	0.211	6.63E-06
rs2067980	rs4242125	0.723	0.211	6.63E-06
rs2067980	rs4560554	0.728	0.209	0.000011
rs2067980	rs4626346	0.737	0.208	3.54E-06
rs2067980	rs10052977	0.737	0.208	3.54E-06
rs2067980	rs3935066	0.730	0.208	7.29E-06
rs2067980	rs4132311	0.697	0.204	0.000013
rs2067980	rs4283798	0.730	0.201	6.64E-06

SNP D = rs10035564				
SNP 1	SNP 2	D'	R2	p-min
rs10035564	rs11948185	1.000	1.000	1.31E-34
rs10035564	rs10051592	1.000	1.000	1.31E-34
rs10035564	rs16902085	1.000	0.894	6.99E-30
rs10035564	rs9292914	0.954	0.874	8.22E-24
rs10035564	rs7711697	0.959	0.822	8.14E-25
rs10035564	rs2935088	1.000	0.819	2.75E-26
rs10035564	rs10512875	1.000	0.819	2.75E-26
rs10035564	rs7380559	0.959	0.794	1.31E-24
rs10035564	rs4518409	0.959	0.794	1.31E-24
rs10035564	rs1438821	0.959	0.794	1.31E-24
rs10035564	rs1438820	0.959	0.794	1.31E-24
rs10035564	rs13362132	0.959	0.794	1.31E-24
rs10035564	rs13160259	0.959	0.794	1.31E-24
rs10035564	rs11958806	0.959	0.794	1.31E-24
rs10035564	rs1061310	0.959	0.794	1.31E-24
rs10035564	rs10512865	0.959	0.794	1.31E-24
rs10035564	rs1048758	0.959	0.794	1.31E-24
rs10035564	rs10044096	0.959	0.794	1.31E-24
rs10035564	rs4328028	0.958	0.793	3.59E-24
rs10035564	rs7716571	0.958	0.792	4.90E-24
rs10035564	rs13177711	0.959	0.792	1.92E-24
rs10035564	rs9292913	0.959	0.792	3.86E-24
rs10035564	rs7380878	0.958	0.791	5.25E-24
rs10035564	rs11949847	0.959	0.790	2.82E-24
rs10035564	rs9790895	0.959	0.790	4.42E-24
rs10035564	rs10043344	0.958	0.789	7.72E-24
rs10035564	rs11741772	0.880	0.745	1.24E-21
rs10035564	rs5875267	0.882	0.745	1.24E-21
rs10035564	rs7717459	0.880	0.721	1.72E-21
rs10035564	rs5893319	0.880	0.721	1.72E-21
rs10035564	rs8872254	0.880	0.721	1.72E-21
rs10035564	rs6451778	0.880	0.721	1.72E-21
rs10035564	rs4373287	0.880	0.721	1.72E-21
rs10035564	rs1866406	0.880	0.721	1.72E-21
rs10035564	rs1438822	0.880	0.721	1.72E-21
rs10035564	rs1438819	0.880	0.721	1.72E-21
rs10035564	rs13169120	0.880	0.721	1.72E-21
rs10035564	rs13155696	0.880	0.721	1.72E-21
rs10035564	rs13154781	0.880	0.721	1.72E-21
rs10035564	rs10462080	0.880	0.721	1.72E-21
rs10035564	rs10065636	0.880	0.721	1.72E-21
rs10035564	rs10059086	0.880	0.721	1.72E-21
rs10035564	rs10057521	0.880	0.721	1.72E-21
rs10035564	rs10053247	0.880	0.721	1.72E-21
rs10035564	rs10041518	0.880	0.721	1.72E-21
rs10035564	rs10040488	0.880	0.721	1.72E-21
rs10035564	rs10039866	0.880	0.721	1.72E-21
rs10035564	rs12513749	0.880	0.721	3.45E-21
rs10035564	rs6875933	0.880	0.720	3.43E-21
rs10035564	rs7736092	0.880	0.719	2.46E-21
rs10035564	rs6871052	0.880	0.719	2.46E-21
rs10035564	rs10070037	0.880	0.719	2.46E-21
rs10035564	rs7706506	0.880	0.719	3.76E-21
rs10035564	rs4542377	0.877	0.718	4.69E-21
rs10035564	rs12651949	0.859	0.718	3.09E-17
rs10035564	rs3747479	0.877	0.717	6.57E-21
rs10035564	rs6894324	0.877	0.716	6.70E-21
rs10035564	rs4457088	0.877	0.716	6.70E-21
rs10035564	rs10038554	0.877	0.716	6.70E-21
rs10035564	rs6868232	0.874	0.715	1.26E-20
rs10035564	rs10040082	0.877	0.714	9.58E-21

rs10035564	rs11951760	0.875	0.706	1.16E-19
rs10035564	rs7715731	0.866	0.704	3.86E-19
rs10035564	rs12518851	0.871	0.700	1.86E-18
rs10035564	rs1438827	0.879	0.692	1.28E-20
rs10035564	rs10941677	0.915	0.669	3.88E-20
rs10035564	rs7735881	0.914	0.649	9.79E-20
rs10035564	rs7723539	0.914	0.649	9.79E-20
rs10035564	rs714130	0.914	0.649	9.79E-20
rs10035564	rs6861560	0.914	0.649	9.79E-20
rs10035564	rs6451770	0.914	0.649	9.79E-20
rs10035564	rs4419600	0.914	0.649	9.79E-20
rs10035564	rs4415085	0.914	0.649	9.79E-20
rs10035564	rs4321755	0.914	0.649	9.79E-20
rs10035564	rs2216081	0.914	0.649	9.79E-20
rs10035564	rs1436825	0.914	0.649	9.79E-20
rs10035564	rs13156030	0.914	0.649	9.79E-20
rs10035564	rs12515012	0.914	0.649	9.79E-20
rs10035564	rs12187196	0.914	0.649	9.79E-20
rs10035564	rs7720551	0.914	0.648	1.95E-19
rs10035564	rs6874055	0.912	0.647	2.45E-19
rs10035564	rs2165010	0.912	0.647	2.45E-19
rs10035564	rs920329	0.913	0.647	1.91E-19
rs10035564	rs10941678	0.914	0.646	1.42E-19
rs10035564	rs10805685	0.914	0.646	1.42E-19
rs10035564	rs1821936	0.914	0.646	2.84E-19
rs10035564	rs4463188	0.911	0.645	4.78E-19
rs10035564	rs4571480	0.912	0.644	3.53E-19
rs10035564	rs2013513	0.912	0.644	3.53E-19
rs10035564	rs4492118	0.914	0.643	2.06E-19
rs10035564	rs2165009	0.914	0.643	2.06E-19
rs10035564	rs12522626	0.914	0.643	2.06E-19
rs10035564	rs16901937	0.913	0.625	4.75E-19
rs10035564	rs920328	0.834	0.580	4.14E-17
rs10035564	rs6889488	1.000	0.489	1.22E-15
rs10035564	rs4460145	1.000	0.489	1.22E-15
rs10035564	rs7731099	1.000	0.478	4.71E-15
rs10035564	rs7716101	1.000	0.463	7.32E-15
rs10035564	rs7709661	1.000	0.463	7.32E-15
rs10035564	rs7701679	1.000	0.463	6.99E-15
rs10035564	rs6884974	1.000	0.463	7.32E-15
rs10035564	rs6885307	1.000	0.463	7.32E-15
rs10035564	rs4533694	1.000	0.463	7.32E-15
rs10035564	rs12521639	1.000	0.463	7.32E-15
rs10035564	rs12054978	1.000	0.463	7.32E-15
rs10035564	rs4371761	1.000	0.452	2.30E-14
rs10035564	rs4502832	1.000	0.438	4.20E-14
rs10035564	rs4485937	1.000	0.438	4.20E-14
rs10035564	rs4389695	1.000	0.438	4.20E-14
rs10035564	rs4296810	1.000	0.438	4.20E-14
rs10035564	rs12522398	1.000	0.438	6.21E-14
rs10035564	rs10941692	1.000	0.438	4.20E-14
rs10035564	rs10044408	1.000	0.425	2.83E-13
rs10035564	rs7720104	1.000	0.400	1.04E-12
rs10035564	rs4493682	1.000	0.400	8.68E-13
rs10035564	rs4308490	1.000	0.400	1.04E-12
rs10035564	rs5876773	1.000	0.390	1.45E-12
rs10035564	rs7711446	1.000	0.388	2.05E-12
rs10035564	rs6893494	1.000	0.388	1.21E-12
rs10035564	rs12523157	1.000	0.388	1.21E-12
rs10035564	rs12514414	1.000	0.388	2.44E-12
rs10035564	rs11954598	1.000	0.388	1.21E-12
rs10035564	rs6864149	1.000	0.388	1.44E-12
rs10035564	rs6896476	1.000	0.375	5.36E-12
rs10035564	rs6890289	1.000	0.375	4.52E-12

rs10035564	rs4533695	1.000	0.375	4.52E-12
rs10035564	rs7709262	0.869	0.370	1.13E-10
rs10035564	rs1405918	0.768	0.368	6.67E-11
rs10035564	rs4357042	1.000	0.364	7.23E-12
rs10035564	rs1909937	1.000	0.364	6.12E-12
rs10035564	rs1472584	1.000	0.364	6.12E-12
rs10035564	rs1392970	1.000	0.364	6.12E-12
rs10035564	rs12523398	1.000	0.364	6.12E-12
rs10035564	rs12523359	1.000	0.364	7.23E-12
rs10035564	rs12522305	1.000	0.364	7.23E-12
rs10035564	rs12521953	1.000	0.364	6.12E-12
rs10035564	rs12516488	1.000	0.364	6.12E-12
rs10035564	rs12153189	1.000	0.364	6.12E-12
rs10035564	rs12153053	1.000	0.364	6.12E-12
rs10035564	rs11953498	1.000	0.364	6.12E-12
rs10035564	rs10941593	1.000	0.364	6.12E-12
rs10035564	rs16902068	1.000	0.364	6.12E-12
rs10035564	rs12514615	1.000	0.364	6.12E-12
rs10035564	rs6874127	0.865	0.357	3.82E-10
rs10035564	rs13187565	0.747	0.357	1.02E-09
rs10035564	rs4613718	1.000	0.356	6.07E-14
rs10035564	rs2625494	0.766	0.354	9.14E-11
rs10035564	rs2580258	0.766	0.354	9.14E-11
rs10035564	rs1951720	0.766	0.354	9.14E-11
rs10035564	rs12110137	0.766	0.354	9.14E-11
rs10035564	rs10073636	0.766	0.354	9.14E-11
rs10035564	rs10043792	0.766	0.354	9.14E-11
rs10035564	rs12520124	0.722	0.351	7.29E-10
rs10035564	rs4566805	1.000	0.350	3.68E-11
rs10035564	rs12520430	1.000	0.349	2.49E-11
rs10035564	rs6451796	0.863	0.345	5.58E-10
rs10035564	rs3923055	0.863	0.345	5.58E-10
rs10035564	rs1501361	0.863	0.345	5.58E-10
rs10035564	rs1392973	0.863	0.345	5.58E-10
rs10035564	rs2589162	0.756	0.340	6.65E-10
rs10035564	rs13180434	1.000	0.340	2.99E-11
rs10035564	rs7446090	0.922	0.339	6.66E-10
rs10035564	rs2580260	0.727	0.338	3.66E-10
rs10035564	rs7711444	1.000	0.337	6.16E-11
rs10035564	rs7446182	0.724	0.326	7.25E-10
rs10035564	rs6892627	0.724	0.326	7.25E-10
rs10035564	rs6451814	0.724	0.326	7.25E-10
rs10035564	rs2337952	0.724	0.326	7.25E-10
rs10035564	rs2049656	0.724	0.326	7.25E-10
rs10035564	rs1483303	0.724	0.326	7.25E-10
rs10035564	rs12654213	0.721	0.324	6.66E-10
rs10035564	rs7717787	0.715	0.323	1.43E-09
rs10035564	rs7705696	0.715	0.323	1.43E-09
rs10035564	rs6893773	0.717	0.322	1.41E-09
rs10035564	rs2337951	0.717	0.322	1.41E-09
rs10035564	rs10078625	0.717	0.322	1.41E-09
rs10035564	rs2337483	0.721	0.322	1.76E-09
rs10035564	rs10035065	0.713	0.320	1.23E-09
rs10035564	rs755046	0.712	0.315	2.82E-09
rs10035564	rs1384732	0.710	0.314	3.47E-09
rs10035564	rs6451810	0.718	0.308	2.46E-09
rs10035564	rs8860200	0.678	0.307	5.17E-09
rs10035564	rs7732970	0.682	0.300	4.88E-09
rs10035564	rs6892594	0.682	0.300	4.88E-09
rs10035564	rs8451804	0.682	0.300	4.88E-09
rs10035564	rs7711528	0.838	0.299	3.91E-08
rs10035564	rs9667260	0.679	0.298	4.57E-09
rs10035564	rs7706116	0.679	0.298	4.57E-09
rs10035564	rs6451802	0.617	0.293	5.15E-08

rs10035564	rs4455566	0.672	0.293	8.79E-09
rs10035564	rs6451793	0.784	0.285	8.16E-08
rs10035564	rs10039263	0.674	0.280	1.09E-08
rs10035564	rs6895191	0.672	0.279	1.47E-08
rs10035564	rs6878425	0.672	0.279	1.47E-08
rs10035564	rs10462097	0.672	0.279	1.47E-08
rs10035564	rs7444176	0.672	0.279	2.05E-08
rs10035564	rs16902084	0.760	0.273	1.87E-07
rs10035564	rs1501358	0.838	0.272	4.94E-08
rs10035564	rs12855983	0.838	0.272	4.94E-08
rs10035564	rs16902083	0.778	0.265	1.49E-07
rs10035564	rs10941679	0.634	0.263	1.41E-07
rs10035564	rs6882139	0.632	0.258	6.10E-08
rs10035564	rs6451843	0.624	0.252	1.14E-07
rs10035564	rs6862655	0.815	0.250	3.35E-08
rs10035564	rs10059745	0.815	0.250	3.35E-08
rs10035564	rs7716785	0.521	0.246	2.76E-07
rs10035564	rs7444405	0.626	0.244	1.36E-07
rs10035564	rs4639238	0.812	0.243	4.87E-08
rs10035564	rs12374507	0.812	0.243	4.87E-08
rs10035564	rs10066963	0.812	0.243	4.87E-08
rs10035564	rs930395	0.686	0.242	4.25E-07
rs10035564	rs5866354	0.807	0.241	8.01E-08
rs10035564	rs10054521	0.811	0.239	6.85E-08
rs10035564	rs12520838	0.589	0.235	3.01E-07
rs10035564	rs7709131	0.592	0.234	3.01E-07
rs10035564	rs7445572	0.592	0.234	3.01E-07
rs10035564	rs6861150	0.818	0.227	8.14E-07
rs10035564	rs6451795	0.757	0.222	1.20E-06
rs10035564	rs13156198	0.585	0.220	6.49E-07
rs10035564	rs4569881	0.553	0.211	1.33E-06
rs10035564	rs13361919	0.553	0.211	1.33E-06
rs10035564	rs13185201	0.553	0.211	1.33E-06
rs10035564	rs10941740	0.561	0.210	4.73E-08
rs10035564	rs12697517	0.603	0.209	2.07E-06
rs10035564	rs12697503	0.600	0.209	2.90E-08
rs10035564	rs7443976	0.578	0.208	1.36E-06
SNP E = rs11743392				
SNP 1	SNP 2	D'	R2	p-min
rs11743392	rs13179816	0.927	0.631	4.07E-20
rs11743392	rs2625494	1.000	0.527	1.80E-19
rs11743392	rs2580258	1.000	0.527	1.80E-19
rs11743392	rs1351720	1.000	0.527	1.80E-19
rs11743392	rs6451810	1.000	0.527	3.30E-19
rs11743392	rs12110137	1.000	0.527	1.80E-19
rs11743392	rs10073636	1.000	0.527	1.80E-19
rs11743392	rs10043792	1.000	0.527	1.80E-19
rs11743392	rs1384732	1.000	0.517	4.95E-19
rs11743392	rs755048	1.000	0.513	8.25E-19
rs11743392	rs7446162	1.000	0.509	6.24E-19
rs11743392	rs2337483	1.000	0.509	1.13E-18
rs11743392	rs2049656	1.000	0.509	6.24E-19
rs11743392	rs1483303	1.000	0.509	6.24E-19
rs11743392	rs6892627	1.000	0.509	6.24E-19
rs11743392	rs6451814	1.000	0.509	6.24E-19
rs11743392	rs2569162	1.000	0.509	2.03E-18
rs11743392	rs2337952	1.000	0.509	6.24E-19
rs11743392	rs1405916	1.000	0.509	6.24E-19
rs11743392	rs12654213	1.000	0.509	6.24E-19
rs11743392	rs7444176	1.000	0.505	1.08E-17
rs11743392	rs6893773	1.000	0.505	1.03E-18
rs11743392	rs2580260	1.000	0.505	1.03E-18
rs11743392	rs2337951	1.000	0.505	1.03E-18

rs11743392	rs10078625	1.000	0.505	1.03E-18
rs11743392	rs10036065	1.000	0.505	1.03E-18
rs11743392	rs7717787	1.000	0.500	1.72E-18
rs11743392	rs7705696	1.000	0.500	1.72E-18
rs11743392	rs7732970	1.000	0.492	2.09E-18
rs11743392	rs6892594	1.000	0.492	2.09E-18
rs11743392	rs6451804	1.000	0.492	2.09E-18
rs11743392	rs9687260	1.000	0.492	2.09E-18
rs11743392	rs7706116	1.000	0.492	2.09E-18
rs11743392	rs4455566	1.000	0.487	3.45E-18
rs11743392	rs6860200	1.000	0.483	5.74E-18
rs11743392	rs13187565	1.000	0.479	1.77E-16
rs11743392	rs12520124	1.000	0.475	5.18E-17
rs11743392	rs7444405	0.951	0.461	1.77E-15
rs11743392	rs10039283	0.951	0.461	1.77E-15
rs11743392	rs7443976	0.951	0.460	3.54E-15
rs11743392	rs6895191	0.950	0.460	3.12E-15
rs11743392	rs6878425	0.950	0.460	3.12E-15
rs11743392	rs10462097	0.950	0.460	3.12E-15
rs11743392	rs6882139	0.950	0.444	5.53E-15
rs11743392	rs13156198	0.949	0.444	1.11E-14
rs11743392	rs10041478	0.948	0.438	3.12E-14
rs11743392	rs6451843	0.948	0.436	3.57E-14
rs11743392	rs10042199	0.947	0.433	8.75E-14
rs11743392	rs12520938	0.948	0.428	4.12E-14
rs11743392	rs7709131	0.948	0.428	3.36E-14
rs11743392	rs7445572	0.948	0.428	3.36E-14
rs11743392	rs6884718	0.948	0.428	3.36E-14
rs11743392	rs4302598	0.948	0.428	3.36E-14
rs11743392	rs4277924	0.948	0.428	3.36E-14
rs11743392	rs13361118	0.948	0.428	3.36E-14
rs11743392	rs13155231	0.948	0.428	3.36E-14
rs11743392	rs12654375	0.948	0.428	3.36E-14
rs11743392	rs12652235	0.948	0.428	3.36E-14
rs11743392	rs12523291	0.948	0.428	3.36E-14
rs11743392	rs12188166	0.948	0.428	3.36E-14
rs11743392	rs10941727	0.948	0.428	3.36E-14
rs11743392	rs6451802	1.000	0.428	9.62E-16
rs11743392	rs10805706	0.948	0.428	8.73E-14
rs11743392	rs4569881	0.947	0.413	9.91E-14
rs11743392	rs13361919	0.947	0.413	9.91E-14
rs11743392	rs13185201	0.947	0.413	9.91E-14
rs11743392	rs10941740	0.942	0.410	1.20E-12
rs11743392	rs12697503	0.856	0.402	3.42E-12
rs11743392	rs13166830	0.944	0.395	1.42E-12
rs11743392	rs12697517	0.848	0.390	1.27E-11
rs11743392	rs7713759	0.852	0.383	7.69E-12
rs11743392	rs13164722	0.852	0.383	7.69E-12
rs11743392	rs12153540	0.852	0.383	7.69E-12
rs11743392	rs11958566	0.849	0.381	1.35E-11
rs11743392	rs12697523	0.851	0.378	1.23E-11
rs11743392	rs12690678	0.851	0.378	1.23E-11
rs11743392	rs12656953	0.851	0.378	1.23E-11
rs11743392	rs7718785	1.000	0.362	3.18E-14
rs11743392	rs7719500	1.000	0.276	1.59E-11
rs11743392	rs2589181	1.000	0.276	1.59E-11
rs11743392	rs4367308	1.000	0.276	1.59E-11
rs11743392	rs4262323	1.000	0.276	1.59E-11
rs11743392	rs10041772	1.000	0.276	1.59E-11
rs11743392	rs10041767	1.000	0.276	1.59E-11
rs11743392	rs4283798	1.000	0.274	2.54E-11
rs11743392	rs13188585	1.000	0.270	6.52E-11
rs11743392	rs4626346	1.000	0.265	3.64E-11
rs11743392	rs13361609	1.000	0.265	4.42E-11

rs11743392	rs4975889	1.000	0.260	6.07E-11
rs11743392	rs1852595	1.000	0.260	6.07E-11
rs11743392	rs10214369	1.000	0.260	6.07E-11
rs11743392	rs7700252	1.000	0.255	8.24E-11
rs11743392	rs1483308	1.000	0.255	8.24E-11
rs11743392	rs9686560	1.000	0.255	8.24E-11
rs11743392	rs7724971	1.000	0.255	8.24E-11
rs11743392	rs7719703	1.000	0.255	8.24E-11
rs11743392	rs7714713	1.000	0.255	8.24E-11
rs11743392	rs7703405	1.000	0.255	8.24E-11
rs11743392	rs7447232	1.000	0.255	8.24E-11
rs11743392	rs6898646	1.000	0.255	8.24E-11
rs11743392	rs6894784	1.000	0.255	8.24E-11
rs11743392	rs6894273	1.000	0.255	8.24E-11
rs11743392	rs6886950	1.000	0.255	8.24E-11
rs11743392	rs4560554	1.000	0.255	1.63E-10
rs11743392	rs4452566	1.000	0.255	1.16E-10
rs11743392	rs4437383	1.000	0.255	8.24E-11
rs11743392	rs4407637	1.000	0.255	8.24E-11
rs11743392	rs2337954	1.000	0.255	8.24E-11
rs11743392	rs16902221	1.000	0.255	8.24E-11
rs11743392	rs16902217	1.000	0.255	8.24E-11
rs11743392	rs13359915	1.000	0.255	8.24E-11
rs11743392	rs13357427	1.000	0.255	8.24E-11
rs11743392	rs13159362	1.000	0.255	1.16E-10
rs11743392	rs12109155	1.000	0.255	8.24E-11
rs11743392	rs10074312	1.000	0.255	8.24E-11
rs11743392	rs10066021	1.000	0.255	8.24E-11
rs11743392	rs10473389	1.000	0.255	8.24E-11
rs11743392	rs6414908	1.000	0.253	2.47E-10
rs11743392	rs10462095	1.000	0.252	1.87E-10
rs11743392	rs7445730	1.000	0.249	1.37E-10
rs11743392	rs4339358	1.000	0.249	1.37E-10
rs11743392	rs4242125	1.000	0.249	1.37E-10
rs11743392	rs16902199	1.000	0.249	1.37E-10
rs11743392	rs12655230	1.000	0.249	1.37E-10
rs11743392	rs6877477	1.000	0.243	3.19E-10
rs11743392	rs7706959	1.000	0.239	3.06E-10
rs11743392	rs7446602	1.000	0.239	3.06E-10
rs11743392	rs13356124	1.000	0.239	3.06E-10
rs11743392	rs4132311	1.000	0.237	3.66E-10
rs11743392	rs2878967	1.000	0.234	4.04E-10
rs11743392	rs16902186	1.000	0.234	4.04E-10
rs11743392	rs1564664	1.000	0.234	4.04E-10
rs11743392	rs1405916	1.000	0.234	4.04E-10
rs11743392	rs12653475	1.000	0.234	4.04E-10
rs11743392	rs12697498	1.000	0.232	1.15E-09
rs11743392	rs7722360	1.000	0.228	6.74E-10
rs11743392	rs4532370	1.000	0.224	6.76E-10
rs11743392	rs1483312	1.000	0.224	6.76E-10
rs11743392	rs1351719	1.000	0.224	6.76E-10
rs11743392	rs12656485	1.000	0.224	6.76E-10
rs11743392	rs10052977	0.919	0.224	7.24E-08
rs11743392	rs6451606	1.000	0.219	1.98E-09
rs11743392	rs1483309	1.000	0.215	1.89E-09
rs11743392	rs1483306	1.000	0.215	1.89E-09
rs11743392	rs13358716	1.000	0.215	1.89E-09
rs11743392	rs10472404	1.000	0.215	1.89E-09
rs11743392	rs10073055	1.000	0.215	1.89E-09
rs11743392	rs12518113	0.846	0.209	6.42E-07
rs11743392	rs4288123	1.000	0.209	3.15E-09
rs11743392	rs17268008	0.767	0.205	1.34E-06
rs11743392	rs4975924	0.844	0.204	8.60E-07
rs11743392	rs4126583	0.844	0.204	8.60E-07

rs11743392	rs12697524	0.844	0.204	8.60E-07
rs11743392	rs12523279	0.844	0.204	8.60E-07
rs11743392	rs12522090	0.844	0.204	8.60E-07
rs11743392	rs12019302	0.844	0.204	8.60E-07
rs11743392	rs11949184	0.844	0.204	8.60E-07
rs11743392	rs10941798	0.844	0.204	8.60E-07
rs11743392	rs10452111	0.844	0.204	8.60E-07
rs11743392	rs10941748	0.844	0.204	1.01E-06
rs11743392	rs4975948	0.836	0.202	1.67E-06
rs11743392	rs7721731	1.000	0.201	9.28E-09
SNP F = rs7716600				
SNP 1	SNP 2	D'	R2	p-min
rs7716600	rs930395	1.000	1.000	3.05E-27
rs7716600	rs10941679	1.000	0.777	1.77E-21
rs7716600	rs12651949	1.000	0.505	4.24E-14
rs7716600	rs7715731	1.000	0.482	2.21E-14
rs7716600	rs10040082	1.000	0.480	2.37E-15
rs7716600	rs6858232	1.000	0.473	8.95E-15
rs7716600	rs7717459	1.000	0.471	2.94E-15
rs7716600	rs4642377	1.000	0.471	8.77E-15
rs7716600	rs10041518	1.000	0.471	2.94E-15
rs7716600	rs10040488	1.000	0.471	2.94E-15
rs7716600	rs10039866	1.000	0.471	2.94E-15
rs7716600	rs6893319	1.000	0.471	2.94E-15
rs7716600	rs6875933	1.000	0.471	2.94E-15
rs7716600	rs6872254	1.000	0.471	2.94E-15
rs7716600	rs6451778	1.000	0.471	2.94E-15
rs7716600	rs4373287	1.000	0.471	2.94E-15
rs7716600	rs1866406	1.000	0.471	2.94E-15
rs7716600	rs1438822	1.000	0.471	2.94E-15
rs7716600	rs1438819	1.000	0.471	2.94E-15
rs7716600	rs13189120	1.000	0.471	2.94E-15
rs7716600	rs13155698	1.000	0.471	2.94E-15
rs7716600	rs13154781	1.000	0.471	2.94E-15
rs7716600	rs12513749	1.000	0.471	2.94E-15
rs7716600	rs10462080	1.000	0.471	2.94E-15
rs7716600	rs10065638	1.000	0.471	2.94E-15
rs7716600	rs10059086	1.000	0.471	2.94E-15
rs7716600	rs10057521	1.000	0.471	2.94E-15
rs7716600	rs10053247	1.000	0.471	2.94E-15
rs7716600	rs7736092	1.000	0.468	3.68E-15
rs7716600	rs4457088	1.000	0.468	1.09E-14
rs7716600	rs10038554	1.000	0.468	1.09E-14
rs7716600	rs6894324	1.000	0.468	1.09E-14
rs7716600	rs6871052	1.000	0.468	3.68E-15
rs7716600	rs3747479	1.000	0.468	1.09E-14
rs7716600	rs10070037	1.000	0.468	3.68E-15
rs7716600	rs7708506	1.000	0.465	4.80E-15
rs7716600	rs6875267	1.000	0.461	1.69E-14
rs7716600	rs11741772	1.000	0.461	4.98E-14
rs7716600	rs12518851	1.000	0.459	2.75E-14
rs7716600	rs1438827	1.000	0.454	8.91E-15
rs7716600	rs11951760	1.000	0.451	3.30E-14
rs7716600	rs10043344	1.000	0.444	1.35E-14
rs7716600	rs11958808	1.000	0.437	1.57E-14
rs7716600	rs10044096	1.000	0.437	1.57E-14
rs7716600	rs7711697	1.000	0.437	4.46E-14
rs7716600	rs7380559	1.000	0.437	1.57E-14
rs7716600	rs4518409	1.000	0.437	1.57E-14
rs7716600	rs4329028	1.000	0.437	4.46E-14
rs7716600	rs1438821	1.000	0.437	1.57E-14
rs7716600	rs1438820	1.000	0.437	1.57E-14
rs7716600	rs13362132	1.000	0.437	1.57E-14

rs7716600	rs13160259	1.000	0.437	1.57E-14
rs7716600	rs1061310	1.000	0.437	1.57E-14
rs7716600	rs10512685	1.000	0.437	1.57E-14
rs7716600	rs1046758	1.000	0.437	1.57E-14
rs7716600	rs9292913	1.000	0.434	1.96E-14
rs7716600	rs7380878	1.000	0.434	5.55E-14
rs7716600	rs7716571	1.000	0.434	5.55E-14
rs7716600	rs13177711	1.000	0.434	1.96E-14
rs7716600	rs11949847	1.000	0.431	2.45E-14
rs7716600	rs920326	1.000	0.421	3.47E-14
rs7716600	rs9292914	1.000	0.404	7.68E-11
rs7716600	rs4571460	1.000	0.400	1.17E-13
rs7716600	rs2013513	1.000	0.400	1.17E-13
rs7716600	rs7735881	1.000	0.392	1.56E-13
rs7716600	rs7723539	1.000	0.392	1.56E-13
rs7716600	rs7720551	1.000	0.392	1.56E-13
rs7716600	rs714130	1.000	0.392	1.56E-13
rs7716600	rs8874055	1.000	0.392	4.15E-13
rs7716600	rs6861560	1.000	0.392	1.56E-13
rs7716600	rs6451770	1.000	0.392	1.56E-13
rs7716600	rs4419800	1.000	0.392	1.56E-13
rs7716600	rs4415085	1.000	0.392	1.56E-13
rs7716600	rs4321755	1.000	0.392	1.56E-13
rs7716600	rs2218081	1.000	0.392	1.56E-13
rs7716600	rs2165010	1.000	0.392	4.15E-13
rs7716600	rs1436825	1.000	0.392	1.56E-13
rs7716600	rs13156930	1.000	0.392	1.56E-13
rs7716600	rs12515012	1.000	0.392	1.56E-13
rs7716600	rs12187196	1.000	0.392	1.56E-13
rs7716600	rs4463188	1.000	0.390	7.67E-13
rs7716600	rs920329	1.000	0.389	1.95E-13
rs7716600	rs1621936	1.000	0.389	1.95E-13
rs7716600	rs10941678	1.000	0.389	1.95E-13
rs7716600	rs10941677	1.000	0.389	5.16E-13
rs7716600	rs10805685	1.000	0.389	1.95E-13
rs7716600	rs4492118	1.000	0.385	2.44E-13
rs7716600	rs2165009	1.000	0.385	2.44E-13
rs7716600	rs12522626	1.000	0.385	2.44E-13
rs7716600	rs16901937	1.000	0.378	3.20E-13
rs7716600	rs9790896	0.931	0.370	6.39E-11
rs7716600	rs1482698	0.926	0.328	2.69E-10
rs7716600	rs13183434	0.685	0.314	1.52E-07
rs7716600	rs1482685	0.848	0.256	3.68E-08
rs7716600	rs1384451	0.848	0.256	3.68E-08
rs7716600	rs2200123	0.757	0.252	8.51E-07
rs7716600	rs11749658	1.000	0.248	4.20E-06
rs7716600	rs1482667	0.779	0.234	2.90E-07
rs7716600	rs11948186	0.679	0.233	9.05E-07
rs7716600	rs10051592	0.679	0.233	9.05E-07
rs7716600	rs10473355	0.770	0.231	5.67E-07
rs7716600	rs10472394	0.778	0.231	3.47E-07
rs7716600	rs2877162	0.777	0.229	3.52E-07
rs7716600	rs2330551	0.777	0.229	3.52E-07
rs7716600	rs10055789	0.777	0.229	3.52E-07
rs7716600	rs10055953	0.776	0.228	4.20E-07
rs7716600	rs987852	0.775	0.221	5.02E-07
rs7716600	rs4242112	0.775	0.221	5.02E-07
rs7716600	rs2330553	0.775	0.221	5.02E-07
rs7716600	rs12054807	0.775	0.221	5.02E-07
rs7716600	rs2877163	0.774	0.219	5.99E-07
rs7716600	rs10941665	0.774	0.219	5.99E-07
rs7716600	rs7356597	0.763	0.214	1.35E-06
rs7716600	rs10473354	0.760	0.207	1.67E-06

Table 4: Sequence Contexts of Key SNPs.

Key SNP A: rs4415084 deCODE Name: SG05S3092 (SEQ ID NO:235)

caggttatgctacttccctggaggacctctcaaaaggaagctgtttgttctatttcttttctcatctgtcccaggact
aggtattgcattaggagatcccttgcttcccactgctgcttttaaatcatttcatttccttcttcccttcattcttc
ccaaatgcaaggtctttcaactttcatttcgtgctacactctgccctttattgctgctctctggaatttgtggtcac
tgtccctcatacactgaaaactcacatacctctacctctagccctgttgtattcctgatgacttgagca[C/T]cca
agggagtgatacatacagcactggtcaatcatttctttacctgccacacatacagcaatctttaatttcaatagcct
tagccactcattcccaaataatgcttggatcatgcacattatcatgagtaaatacacccatgtctgaaatcctgatt
tcaagtacttcccaatttttctgtcttttctttactttcagctcacagaaacaattcttccaccatattaaaaactc
taatccaattcacttgttccaccacttttttattcattattctctcctgtctttactttcttcct

Key SNP B: rs7703618 deCODE Name: SG05S3065 (SEQ ID NO:140)

gtgaggacacagagccaaaccatttcaccagagggctgagtaactctaatctggcaggatgattatcctacacaggt
tgcaatggcccctgaaatttggacgcactttgtgagagaccagtgtctagataactaggaactaggtaaatgttgga
gagctgcttcccttcatttctgtcattgtctgtttcatttcctttgcattgtttgttgatctgtattaaacaaaaat
gaaagcaaaccttgtatctgagtctccatttttaccaatcctcacatttatggttcagtgtcttagtct[A/G]gtt
tcgaataacaagaaccttttgtacttggaagtataaaacttgatagcagcaacattattgatatttagagctcagta
cctgtctaattacaggcaggcagaaagaagtgtcaaggtattcttgcttatcaggtcacaggtaatttcttcctcta
agaattcataaactgatagactaatattggagaaagaaatgcaatttaattgctgaaagtctgtttcagtttactgg
tcttgtaatagaggtaaaattctaaacaacttggggagctttggtgagaattaaaataggtgggtg

Key SNP C: rs2067980 deCODE Name: SG05S3114 (SEQ ID NO:160)

gaatatgacgtcatataggcattaatttccatgttatgaattcaccagtaaaattgtttaaacagagaagtaaacaa
gacggtaatgttattcaggtaaaagtagagagggaaaagaaatattggaaccagttcagcaaccaaaatggtgccag
agcccaagcatgagttattaaaggctggtggttcctctctcctgacccattaccattcttatctctgatgctccagg
ctgtcagtttctttcttttttgaccatatacaggtaaggaaagcccatttatgagctattttatttcca[A/G]gtt
ttaaaaatgtcaattgatataggctatgatctacagtaatgcttaatctattgaagttttgcatcaaattccatct
taagatgcaagcctgaagcccatttaatgccaaatgtaaatacaagtgctagtttcaaagggcaagattcaaagaaa
gacaaacagaagaaaagtattttaattgctatctaaaagaaggctgtgttcttgggtgaatactttgttgatgtatt
tggggtagaaacagagggagaaataattatgtaatgttaagctgttttctaaaattccagggctcc

Key SNP D: rs10035564 deCODE Name: SG05S3104

acaataagtttttagtgatattagatttttttcatttttggaagaagaacagaaaaagtgtaaaaagatggaataat
atagaaaatggtagctggaggattcaaagaagaactcacttttatcatgtcaaagctaaaatataaattgtagattt
tgcatatgtacaatgagcagaaacacatagctgaagaaagaagtgtgctaaataaataaatgaagtattaatgattg

agcagagtcttagaaagttggacatgttaagagcattgaatctatttagcctttcatgccatgcccaaa[A/G]tca
gaattttaacctatactaggactttaagacaaaaaataggcaaacaaaatcacaaagtgttacaattgacatatgca
gtgaattgtttcccttaaaaacaacatttttttttagttatatcactactataaaatttattcttcaacaggcaac
taaacgtaatctggtttaatctttttttataaaggaacattttaaagtaattcttttctcctaacagaccatctatt
ttcctctaaatctctttagctttaatatctatttagcgataaacagtgcatgaaataaacagctc

Key SNP E: rs11743392 deCODE Name: SG05S3093
aatctttcaaatatattcatctctcacttatttagggactgattatccaatttgtgaactatccctgtggcttctcc
tcttttctctaatgatttctcctctgcctatttccttaaatcgttttaatactaaatgagctgcatgaaaacagaaa
agaagctaaagcagcaaaatttgatacatataaacagtactgcaaaagaatttcatttgtgctcatatgtttttgaa
ttttcaattttctgttaccccacttccatatttcacactccagattatgtcacccccacccaactcccaa[C/T]aat
ttgaaattcaaatttggaaattcatctattggttcatttagttggaaactgcatattcacaggtggagagtggaata
tatttcaaaaccacagagaaaaaaaaaaaaaacgtaattcaacttcgttaatttgtttttaattttccaaagctggaa
attgtctctatatctcaattgatgagtttctgagctaaaaacaaaacaaaacaaaacaaaacatcatttcctgtaac
cagatttcactgctttcattctaagcaagatgatataaataacaatgagtagtcaagtatttattc

Key SNP F: rs7716600 (SEQ ID NO:125)

SG05S3097
aggcctaatggttgtatatatatatttttttatttggtagcagaaaagactttaaaatatgttgatgtttgcgaggt
aaagcatctatgtagggcattactatcaaggctttttttttctgcttgagtctatattacaaacattttattatgtc
tctgctgagattaatttaaatgtgcaaattttcaattcctaatataaagataaaatgtaaagttgatccaaaaatac
aaaaaaagtgataaaacttagtttgtaatatagactcatatatcatatttttagttctatttcaatgct[A/G]tct
agaatttttatcattgcttttacctgaagattcaaattgttttggcatcagtcgggaaatcagtttgtttagctag
caaaaatagacattaataaataaacccagaatacttagaagagatagatagggacccagatctctcaagaaatacgg
ctacagctaattgctatttctacacaaattaacaagcaagctataaactggcatgtgggatttttttttttttttt
tctctgagacaaggtttcactctctctcccagacgggagtgcagtggtaccatcttggttcagggc

Table 5: SNP and Centaurus Assay Description:

SNP name: SG05S3065 or rs7703618
Mapping Information (Build 34)
chr5:44.960080+
Assay SG05S3065.c1 of type CENTAURUS, status Verified
Forward Primer: GCAAACCTTGTATCTGAGTCTCCAT
Reverse Primer: GTGACCTGATAAGCAAGAATACCT
Vic probe: CTTA*GTCTGGT
Fam probe: TCT*T*A*GTCTAGT
Enhancer: TCGAATAACAAGAACC
* Indicates a modified base as described in [Kutyavin, et al., (2006), Nucleic Acids Res, **34**, e128].

EXAMPLE 2. *Refinement of association signal on chromosome 5p12, correlation with clinical variables and investigation of FGFR2 locus on chromosome 10q26.*

The signals we have identified on chromosome 5p12 localize to a large stretch of chromosome 5p12-11 exhibiting a low recombination. From this region we selected 10 SNPs from the Illumina Hap300 chip set, generated Centaurus SNP assays[Kutyavin, et al., (2006), Nucleic Acids Res, **34**, e128] and typed them in additional samples from Iceland, and in replication samples from Sweden, Holland, Spain and the United States. In total, 5028 cases and 32090 controls of European ancestry were studied. The most strongly associated Illumina SNP in the region was rs4415084, the T allele giving a combined odds ratio (OR) of 1.16 and a *P* value of $6.4x10^{-10}$, which meets the Bonferroni criteria for genome-wide significance (Table 6). In the replication samples alone, rs4415084 gave an OR of 1.14 ($P = 7.5 \times 10^{-5}$). To refine the signal, we typed a further 11 SNPs that were not on the Illumina chip, but were in LD with Hap300 SNPs giving a substantial signal. Data from these SNPs is presented in Tables 6 and 7. The strongest overall signal (OR 1.19, $P = 2.9x10^{-11}$) originated from the G allele of rs10941679, a non-Illumina SNP that is correlated to rs4415084 ($D' = 0.99$, $r^2 = 0.51$ in the Icelandic population, Table 8).

Allele G-rs10941679 is less common than T-rs4415084 and is almost completely contained on the T-rs4415084 background. However, in a multivariate analysis, T-rs4415084 retained nominal significance after correction for G-rs10941679 ($P = 0.042$), and *vice versa* ($P = 0.0017$, Table 9). Therefore, despite being highly correlated, neither SNP accounts completely for the observed signal at 5p12. We concluded that both T-rs4415084 and G-rs10941679 confer risk of breast cancer. Multivariate analysis revealed that the signal from SNP rs7703618 could be accounted for entirely by either T-rs4415084 or G-rs10941679 (Table 9).

We reviewed the medical records of the patients, if they were available, and analyzed the combined data from the Icelandic and replication sample sets for the two risk variants at 5p12 and marker rs1219648 at the *FGFR2* locus on chromosome 10. All three variants conferred significantly greater risk of estrogen receptor (ER) positive breast cancer than of ER negative tumours (Tables 6 and 10). A similar preferential risk was seen for progesterone receptor positive tumours for the 5p12 variants (Table 10). We previously reported that susceptibility variants on 2q35 and 16q12 are particularly associated with ER positive tumours[Stacey, et al., (2007), Nat Genet]. The present findings add further support to the notion that ER positive and ER negative tumours have different genetic components to their risks.

The 5p12 SNPs also showed associations with lower histological grade, which was explained by the association with ER status in multivariate analysis. The *FGFR2* SNP was more frequent in node positive than node negative tumours. There were no significant associations with tumour stage or histopathology (Table 10). No variant showed a significant association with age at diagnosis. The *FGFR2* SNP was associated with a family history of breast cancer, in line with

previous reports [Huijts, et al., (2007), Breast Cancer Res, 9, R78; Easton, et al., (2007), Nature 447:1087-93]. Similar tendencies, though not statistically significant, were observed for the 5p12 SNPs (Table 11).

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 2277 patients. Of these, genotypes were derived from Illumina Hap300 chips for 1791 patients and from Centaurus assays for 486 patients. The 26,199 Icelandic controls consisted of individuals selected from ongoing Illumina-based genome-wide association studies at deCODE. 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 642 patients. The 1540 successfully genotyped controls 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 counselling 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.

U.S. Multiethnic Cohort: The Multiethnic Cohort study (MEC) consists of over 215,000 men and women in Hawaii and Los Angeles (with additional African-Americans from elsewhere in California). The cohort is comprised predominantly of African Americans, Native Hawaiians, Japanese Americans, Latinos and European Americans who entered the study between 1993 and 1996 by completing a 26-page self-administered questionnaire that asked detailed information about dietary habits, demographic factors, personal behaviors, history of prior medical conditions, family history of common cancers, and for women, reproductive history and exogenous hormone use. The participants were between the ages 45 and 75 at enrolment. Incident cancers in the MEC are identified by cohort linkage to population-based cancer Surveillance, Epidemiology and End Results (SEER) registries covering Hawaii and Los Angeles County, and to the California State cancer registry covering all of California. Beginning in 1994, blood samples were collected from incident breast cancer cases and a random sample of MEC participants to serve as a control pool for genetic analyses in the cohort. Eligible cases in the nested breast cancer case-control study consisted of women with incident invasive cancer diagnosed after enrolment in the MEC through December 31, 2002. Controls were participants without breast cancer prior to entry into the cohort and without a diagnosis up to December 31, 2002. Controls were frequency matched to cases based on race/ethnicity and age (in 5-year intervals).

Nigeria: We obtained genotypes from 689 incident breast cancer cases and 469 controls from Ibadan, Nigeria. Cases were consecutively recruited at presentation and later histologically confirmed in the Departments of Surgery and Radiotherapy, University College Hospital, Ibadan, Nigeria. This hospital serves a catchment area of 3 million people and is an oncology referral centre for other hospitals in the region. Population-based controls were recruited randomly from the community adjoining the hospital. After a community consultation process, control subjects were invited to attend a clinic set up for the purposes of the study. Controls were cancer-free at recruitment and over 18 years of age.

Genotyping

Approximately 1791 Icelandic patients and 26199 controls were genotyped on Illumina Hap300 SNP arrays, as described previously[Stacey, et al., (2007), Nat Genet **39**:865-9]. All other genotyping was carried out using Nanogen Centaurus assays [Kutyavin, et al., (2006), Nucleic Acids Res, **34**, e128] that were generated for SNPs shown in Table 7. Primer sequences are available on request. 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 Nanogen platforms and the observed mismatch rate was lower than 0.5%. All genotyping was carried out at the deCODE Genetics facility. All physical coordinates are given according to NCBI Build 35.

Clinical Parameters

Estrogen and progesterone receptor status was derived from immunohistochemical or immunometric assay results reported in medical records. A receptor level of $\geq$10 fmol/mg or an immunohistochemical observation of $\geq$ 10% positive nuclei was considered to be positive. Stage was determined according to the American Joint Committee on Cancer, 6th Edition. Histological subtype was determined from SNOMED-M (or equivalent ICDO) codes as follows: "Invasive Ductal Carcinoma": 8500/3, 8521/3; "DCIS" (Ductal Carcinoma In-Situ and related in-situ carcinomas): 8500/2, 8050/2, 8201/2, 8501/2, 8503/2, 8507/2, 8522/2; "Invasive Lobular Carcinoma": 8520/3; "LCIS" (Lobular Carcinoma In-Situ): 8520/2; "Tubular or Mucinous": 8211/3, 8480/3, 8481/3; "Medullary Carcinoma": 8510/3, 8512/3; "Mixed Invasive": 8522/3, 8523/3, 8524/3, 8541/3, 8543/3; "Other Invasive": 8050/3, 8141/3, 8200/3, 8260/3, 8323/3, 8401/3, 8490/3, 8501/3, 8503/3, 8504/3, 8530/3, 8540/3. Tumours with the following non-specific codes were excluded from analysis of histopathological types: 8000/3, 8010/2, 8010/3, 8010/6, 8020/3, 8140/2, 8140/3, 8230/3. Histological Grade was specified according to the Nottingham (Elston-Ellis modification of the Scarff-Bloom-Richardson) system. Node status was

analyzed for stages I to IIIB and was based on pathological staging obtained by axillary lymph node dissection and/or sentinel node biopsy. The Sweden Familial sample set was not used in analysis of clinical parameters.

Statistical Analyses

We calculated the OR for each SNP allele assuming the multiplicative model; *i.e.* assuming that the relative risk of the two alleles that a person carries multiplies. Allelic frequencies and OR are presented for the markers. The associated *P* values were calculated with the standard likelihood ratio χ^2 statistic as implemented in the NEMO software package (Gretarsdottir S., *et al.*, *Nat. Genet.* 35:131-38 (2003)). Confidence intervals were calculated assuming that the estimate of OR has a log-normal distribution. For SNPs that were in strong LD, whenever the genotype of one SNP was missing for an individual, the genotype of the correlated SNPs were used to impute genotypes through a likelihood approach as previously described (Gretarsdottir S., *et al.*, *Nat. Genet.* 35:131-38 (2003)). This ensured that results presented for different SNPs were based on the same number of individuals, allowing meaningful comparisons of OR and *P*-values. Joint analyses of multiple case-control replication groups were carried out using a Mantel-Haenszel model in which the groups were allowed to have different population frequencies for alleles or genotypes but were assumed to have common relative risks. The tests of heterogeneity were performed by assuming that the allele frequencies were the same in all groups under the null hypothesis, but each group had a different allele frequency under the alternative hypothesis. Joint analyses of multiple groups of cases were performed using an extended Mantel-Haenszel model that corresponds to a polytomous logistic regression using the group indicator as a covariate. There was no evidence of heterogeneity between the replication sample sets for any of the SNPs tested. Association of risk variants with age at diagnosis and with histological grade were tested by linear regression between the parameter value and the number of copies of the risk allele carried by each individual.

For analysis of family history we calculated for each genotype a familial relative risk for first degree relatives by adapting our previously described method (Amundadottir L.T., *et al.*, *PLoS Med.* 1:e65 (2004)) to accommodate genotype-specific familial relative risks ($gfRR_{gt}$). For each SNP genotype we determined a $gfRR_{gt}$ as:

$$gfRR_{gt} = \frac{a/r}{x/n}$$

where *r* is the number of first-degree relatives of breast cancer patients with genotype *gt* (counting multiple times those individuals who are related to more than one patient with

genotype *gt* and *a* is the number of first-degree relatives of breast cancer patients with genotype *gt* who are themselves affected with breast cancer. In the denominator, *n* is the size of the population and *x* is the number of people in the population affected with the disease (from ICR records). In order to compare the observed $gfRR_{gt}$ of one genotype with another, we calculated the ratio $gfRR_{gt1}/gfRR_{gt2}$. The significance of these latter ratios was determined by simulation: Controls groups for each $gfRR_{gt}$ were drawn randomly from the set of breast cancer patients genotyped for the SNP in question. The control groups were the same size as each corresponding observed $gfRR_{gt}$ group and were matched on the numbers of parents listed in the Icelandic Genealogical Database (0, 1, or 2). Ten thousand iterations of control group $gfRR_{gt1}/gfRR_{gt2}$ ratios were calculated and the *P* value determined by counting how often the ratio for the control groups matched or exceed the observed $gfRR_{gt1}/gfRR_{gt2}$.

We calculated genotype specific ORs, by estimating the genotype frequencies in the population assuming Hardy-Weinberg equilibrium. No significant deviations from multiplicity were observed. Potential interactions between loci were examined using correlation tests of allele counts and by case-control association of carriers and non-carriers. No significant interactions were observed.

Some of the Icelandic patients and controls are related to each other, both within and between groups, causing the χ^2 statistic to have a mean >1. We estimated the inflation factor by simulating genotypes through the Icelandic genealogy, as described previously(Grant S.F., *et al.*, *Nat Genet* 38:320-3 (2006)), and corrected the χ^2 statistics for Icelandic OR's accordingly. The estimated inflation factor was 1.08 for the complete set of Icelanders (cases and controls) and smaller, but ≥ 1, for all the other subsets used in the analysis of the clinical phenotypes.

Table 6. Association of SNPs in 5p12 and 10q26 loci with risk for breast cancer

				Number		Frequency				
Location	**SNP**	**Allele**	**Sample Set**	**Cases**	**Controls**	**Cases**	**Controls**	**OR[a]**	**95%CI**	***P*[b]**
5p12	rs4415084	T	Iceland[c]	2277	26199	0.409	0.372	1.17	(1.10,1.25)	1.9×10^{-6}
			Sweden	833	1750	0.443	0.417	1.11	(0.99,1.25)	8.0×10^{-2}
			Holland	744	2034	0.433	0.402	1.13	(1.01,1.28)	3.9×10^{-2}
			Spain	642	1540	0.396	0.362	1.16	(1.01,1.33)	3.3×10^{-2}
			MEC European Americans	532	567	0.471	0.424	1.21	(1.01,1.43)	3.5×10^{-2}
			Non-Icelanders[d]	2751	5891	0.436	0.401	1.14	(1.07,1.22)	7.5×10^{-5}
			All samples[d]	5028	32090	0.431	0.386	1.16	(1.10,1.21)	6.4×10^{-10}
			CGEMS[e]	*1141*	*1140*	*0.437*	*0.395*	*1.19*		2.2×10^{-3}
			All ER Positive[d]	2729	32090	0.444	0.398	1.23	(1.18,1.30)	1.8×10^{-11}
			All ER Negative[d]	744	32090	0.391	0.395	0.98	(0.88,1.10)	7.7×10^{-1}
			All ER Positive vs Negative[d]	2729	744	0.444	0.391	1.25	(1.11,1.41)	2.0×10^{-4}
5p12	rs10941679	G	Iceland[c]	2277	26199	0.269	0.235	1.20	(1.11,1.29)	2.2×10^{-6}
			Sweden	833	1750	0.312	0.273	1.21	(1.06,1.37)	3.8×10^{-3}
			Holland	744	2034	0.298	0.258	1.22	(1.07,1.39)	3.2×10^{-3}
			Spain	642	1540	0.214	0.198	1.10	(0.94,1.30)	2.3×10^{-1}
			MEC European Americans	532	567	0.293	0.253	1.23	(1.02,1.48)	3.4×10^{-2}
			Non-Icelanders[d]	2751	5891	0.279	0.245	1.19	(1.11,1.28)	2.9×10^{-6}
			All samples[d]	5028	32090	0.277	0.243	1.19	(1.13,1.26)	2.9×10^{-11}
			All ER Positive[d]	2736	32090	0.286	0.243	1.27	(1.19,1.35)	2.5×10^{-13}
			All ER Negative[d]	744	32090	0.254	0.243	1.05	(0.92,1.18)	4.8×10^{-1}
			All ER Positive vs Negative[d]	2736	744	0.286	0.254	1.21	(1.08,1.36)	4.2×10^{-3}
5p12	rs7703618	G	All samples[d]	5028	32090	0.369	0.338	1.13	(1.08,1.18)	3.3×10^{-7}
5p12	rs10035564	G	All samples[d]	5028	32090	0.312	0.301	1.10	(1.04,1.15)	5.3×10^{-4}
5p12	rs4866929	A	All samples[d]	5028	32090	0.527	0.519	1.04	(1.00,1.09)	6.7×10^{-2}
5p12	rs981782	T	All samples[d]	5028	32090	0.507	0.500	1.04	(0.99,1.09)	1.0×10^{-1}

Location	SNP	Allele	Sample Set	Number		Frequency		OR[a]	95%CI	P[b]
				Cases	Controls	Cases	Controls			
10q26	rs1219648	G	Iceland[c]	2270	28180	0.492	0.453	1.17	(1.10,1.25)	1.2×10^{-6}
			Sweden	822	1725	0.466	0.381	1.37	(1.21,1.54)	3.0×10^{-7}
			Holland	741	2001	0.455	0.389	1.31	(1.17,1.48)	8.7×10^{-6}
			Spain	635	1493	0.477	0.424	1.24	(1.09,1.41)	1.5×10^{-3}
			Non-Icelanders[d]	2198	5219	0.463	0.398	1.31	(1.22,1.41)	1.2×10^{-13}
			All samples[d]	4468	31409	0.470	0.412	1.23	(1.17,1.29)	1.3×10^{-17}
			All ER Positive[d]	2354	31409	0.481	0.412	1.29	(1.22,1.38)	3.4×10^{-16}
			All ER Negative[d]	657	31409	0.413	0.412	0.99	(0.86,1.10)	8.3×10^{-1}
			All ER Positive vs Negative[d]	2354	657	0.481	0.413	1.30	(1.15,1.47)	2.9×10^{-5}

[a] Allelic Odds Ratios calculated under the multiplicative model [b]All *P* values are two sided and have been adjusted for relatedness and other potential stratification of the Icelandic cases and controls. [c]Icelandic data are combined Illumina and Centaurus assay-derived replication data sets. [d]For analyses of combined data for the "Non-Icelanders", "All Samples" and ER groups, the OR and *P* values were calculated using the Mantel-Haenszel method, and the frequencies as simple (arithmetic) means of the frequencies of individual groups. [e]CGEMS data are displayed for comparative purposes only and were not included in any of the calculations.

Table 7. Association with breast cancer for all variants tested in 5p12

	Number		Frequency						
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.409	0.372	1.17	(1.10,1.25)	1.9E-06	T	rs4415084
Sweden	833	1750	0.443	0.417	1.11	(0.99,1.25)	8.0E-02		
Holland	744	2034	0.433	0.402	1.13	(1.01,1.28)	3.9E-02		
Spain	642	1540	0.396	0.362	1.16	(1.01,1.33)	3.3E-02		
MEC European American	532	567	0.471	0.424	1.21	(1.01,1.43)	3.5E-02		
Non-Icelanders	2751	5891	0.436	0.401	1.14	(1.07,1.22)	7.5E-05		
All European Ancestry	5028	32090	0.431	0.396	1.16	(1.10,1.21)	6.4E-10		
MEC African American	428	457	0.630	0.641	0.95	(0.78,1.16)	6.5E-01		
Nigeria	689	469	0.689	0.648	1.20	(1.00,1.44)	4.6E-02		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.269	0.235	1.20	(1.11,1.29)	2.2E-06	G	rs10941679
Sweden	833	1750	0.312	0.273	1.21	(1.06,1.37)	3.8E-03		
Holland	744	2034	0.298	0.256	1.22	(1.07,1.39)	3.2E-03		
Spain	642	1540	0.214	0.198	1.10	(0.94,1.30)	2.3E-01		
MEC European American	532	567	0.293	0.253	1.23	(1.02,1.48)	3.4E-02		
Non-Icelanders	2751	5891	0.279	0.245	1.19	(1.11,1.28)	2.9E-06		
All European Ancestry	5028	32090	0.277	0.243	1.19	(1.13,1.26)	2.9E-11		
MEC African American	428	457	0.218	0.213	1.03	(0.82,1.29)	8.0E-01		
Nigeria	689	469	0.175	0.191	0.90	(0.72,1.12)	3.3E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.393	0.356	1.18	(1.10,1.25)	9.8E-07	G	rs7703618
Sweden	833	1750	0.405	0.386	1.08	(0.96,1.22)	1.9E-01		
Holland	744	2034	0.398	0.383	1.06	(0.94,1.20)	3.3E-01		
Spain	642	1540	0.324	0.313	1.05	(0.91,1.21)	4.8E-01		
MEC European	532	567	0.427	0.391	1.16	(0.98,1.38)	9.2E-02		

American									
Non-Icelanders	2751	5891	0.388	0.368	1.08	(1.01,1.16)	2.3E-02		
All European Ancestry	5028	32090	0.389	0.366	1.13	(1.08,1.18)	3.3E-07		
MEC African American	428	457	0.349	0.348	1.01	(0.83,1.22)	9.4E-01		
Nigeria	689	469	0.327	0.335	0.96	(0.81,1.15)	6.9E-01		

Sample Set	Cases	Controls	Cases	Controls	OR	95%CI	P	Allele	SNP
Iceland	2277	26199	0.288	0.261	1.14	(1.07,1.23)	1.8E-04	G	rs10035564
Sweden	833	1750	0.322	0.313	1.04	(0.92,1.18)	5.1E-01		
Holland	744	2034	0.331	0.319	1.06	(0.93,1.20)	3.9E-01		
Spain	642	1540	0.309	0.310	0.99	(0.86,1.14)	9.1E-01		
Non-Icelanders	2751	5891	0.321	0.314	1.03	(0.96,1.09)	4.0E-01		
All European Ancestry	5028	32090	0.312	0.301	1.10	(1.04,1.15)	5.3E-04		
Nigeria	689	469	0.767	0.739	1.17	(0.95,1.43)	1.3E-01		

Sample Set	Cases	Controls	Cases	Controls	OR	95%CI	P	Allele	SNP
Iceland	2277	26199	0.484	0.468	1.07	(1.00,1.14)	4.6E-02	A	rs4866929
Sweden	833	1750	0.515	0.514	1.01	(0.90,1.13)	9.3E-01		
Holland	744	2034	0.549	0.551	0.99	(0.88,1.12)	8.8E-01		
Spain	642	1540	0.502	0.489	1.05	(0.93,1.20)	4.2E-01		
MEC European American	532	567	0.585	0.572	1.05	(0.89,1.25)	5.4E-01		
Non-Icelanders	2751	5891	0.538	0.531	1.02	(0.96,1.09)	5.6E-01		
All European Ancestry	5028	32090	0.527	0.519	1.04	(1.00,1.09)	6.7E-02		
MEC African American	428	457	0.884	0.893	0.92	(0.68,1.24)	5.7E-01		
Nigeria	689	469	0.999	0.999	NA	NA	NA		

Sample Set	Cases	Controls	Cases	Controls	OR	95%CI	P	Allele	SNP
Iceland	2277	26199	0.472	0.458	1.06	(0.99,1.13)	8.4E-02	T	rs981782
Sweden	833	1750	0.512	0.509	1.01	(0.90,1.14)	8.4E-01		
Holland	744	2034	0.544	0.543	1.00	(0.89,1.13)	9.5E-01		
Spain	642	1540	0.501	0.489	1.05	(0.92,1.19)	5.0E-01		
Non-Icelanders	2751	5891	0.519	0.514	1.02	(0.95,1.10)	6.2E-01		
All European	5028	32090	0.507	0.500	1.04	(0.99,1.09)	1.0E-01		

Ancestry									
Nigeria	689	469	0.999	0.999	NA	NA	NA		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.618	0.591	1.12	(1.04,1.21)	3.5E-03	T	rs4613718
Sweden	833	1750	0.616	0.594	1.10	(0.97,1.24)	1.3E-01		
Holland	744	2034	0.637	0.593	1.20	(1.06,1.36)	3.0E-03		
Spain	642	1540	0.635	0.609	1.12	(0.98,1.28)	9.8E-02		
Non-Icelanders	2751	5891	0.629	0.599	1.16	(1.08,1.24)	6.4E-05		
All European Ancestry	5028	32090	0.627	0.597	1.14	(1.08,1.20)	1.3E-06		
Nigeria	689	469	0.763	0.755	1.05	(0.86,1.28)	6.4E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.402	0.367	1.16	(1.09,1.24)	5.2E-06	G	rs994793
Sweden	833	1750	0.434	0.414	1.09	(0.97,1.22)	1.7E-01		
Holland	744	2034	0.438	0.411	1.12	(0.99,1.26)	7.2E-02		
Spain	642	1540	0.374	0.352	1.10	(0.96,1.26)	1.6E-01		
Non-Icelanders	2751	5891	0.415	0.392	1.11	(1.03,1.19)	4.9E-03		
All European Ancestry	5028	32090	0.412	0.386	1.14	(1.09,1.20)	6.0E-08		
Nigeria	689	469	0.623	0.611	1.05	(0.88,1.25)	5.9E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.398	0.363	1.16	(1.09,1.24)	4.7E-06	T	rs6867533
Sweden	833	1750	0.431	0.412	1.08	(0.96,1.22)	1.9E-01		
Holland	744	2034	0.430	0.408	1.09	(0.97,1.23)	1.4E-01		
Spain	642	1540	0.370	0.349	1.10	(0.96,1.26)	1.8E-01		
Non-Icelanders	2751	5891	0.410	0.390	1.10	(1.02,1.18)	9.8E-03		
All European Ancestry	5028	32090	0.407	0.383	1.14	(1.08,1.19)	1.1E-07		
Nigeria	689	469	0.545	0.528	1.07	(0.90,1.27)	4.3E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.236	0.209	1.17	(1.08,1.26)	5.7E-05	A	rs7716600
Sweden	833	1750	0.265	0.239	1.15	(1.01,1.32)	3.8E-02		
Holland	744	2034	0.254	0.235	1.11	(0.97,1.27)	1.4E-01		
Spain	642	1540	0.177	0.177	1.00	(0.84,1.19)	9.8E-01		
Non-Icelanders	2751	5891	0.232	0.217	1.11	(1.02,1.20)	1.8E-02		

All European Ancestry	5028	32090	0.233	0.215	1.15	(1.08,1.21)	1.8E-06		
Nigeria	689	469	0.169	0.194	0.85	(0.68,1.06)	1.4E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.312	0.283	1.15	(1.07,1.24)	8.0E-05	A	rs3935086
Sweden	833	1750	0.316	0.311	1.02	(0.90,1.16)	7.5E-01		
Holland	744	2034	0.321	0.321	1.00	(0.88,1.14)	9.8E-01		
Spain	642	1540	0.308	0.310	0.99	(0.86,1.14)	9.0E-01		
MEC European American	532	567	0.380	0.349	1.15	(0.94,1.40)	1.9E-01		
Non-Icelanders	2751	5891	0.331	0.323	1.02	(0.95,1.10)	5.6E-01		
All European Ancestry	5028	32090	0.327	0.315	1.09	(1.03,1.14)	1.3E-03		
CGEMS	NA	NA	NA	NA	NA		NA		
MEC African American	428	457	0.706	0.702	1.02	(0.84,1.24)	8.3E-01		
Nigeria	689	469	0.831	0.799	1.24	(1.00,1.54)	5.5E-02		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.151	0.131	1.17	(1.07,1.28)	4.7E-04	G	rs2067980
Sweden	833	1750	0.165	0.148	1.14	(0.97,1.34)	1.2E-01		
Holland	744	2034	0.159	0.165	0.96	(0.82,1.13)	6.0E-01		
Spain	642	1540	0.126	0.139	0.89	(0.73,1.08)	2.3E-01		
MEC European American	NA	NA	NA	NA	NA		NA		
Non-Icelanders	2751	5891	0.150	0.151	1.00	(0.91,1.09)	9.5E-01		
All European Ancestry	5028	32090	0.150	0.146	1.10	(1.03,1.17)	6.3E-03		
Nigeria	689	469	0.085	0.089	0.96	(0.71,1.30)	7.7E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.199	0.182	1.12	(1.03,1.22)	8.2E-03	A	rs7731099
Sweden	833	1750	0.188	0.196	0.95	(0.82,1.10)	5.0E-01		
Holland	744	2034	0.200	0.198	1.02	(0.88,1.18)	8.4E-01		
Spain	642	1540	0.189	0.186	1.02	(0.86,1.20)	8.2E-01		
MEC European American	NA	NA	NA	NA	NA		NA		
Non-Icelanders	2751	5891	0.192	0.193	0.99	(0.90,1.10)	8.9E-01		
All European	5028	32090	0.194	0.190	1.06	(1.00,1.13)	5.9E-02		

Ancestry									
Nigeria	689	469	0.494	0.476	1.07	(0.90,1.28)	4.1E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.123	0.105	1.20	(1.08,1.33)	6.3E-04	A	rs13183434
Sweden	833	1750	0.130	0.116	1.14	(0.96,1.37)	1.4E-01		
Holland	744	2034	0.130	0.136	0.95	(0.80,1.13)	5.9E-01		
Spain	642	1540	0.108	0.117	0.91	(0.74,1.12)	3.8E-01		
Non-Icelanders	2751	5891	0.123	0.123	1.00	(0.85,1.19)	9.6E-01		
All European Ancestry	5028	32090	0.123	0.118	1.12	(1.04,1.20)	2.7E-03		
Nigeria	689	469	0.072	0.072	1.00	(0.73,1.38)	9.9E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.254	0.233	1.12	(1.04,1.21)	2.1E-03	G	rs10512875
Sweden	833	1750	0.282	0.279	1.01	(0.89,1.15)	8.4E-01		
Holland	744	2034	0.297	0.283	1.07	(0.94,1.22)	3.0E-01		
Spain	642	1540	0.277	0.280	0.99	(0.85,1.14)	8.7E-01		
Non-Icelanders	2751	5891	0.286	0.281	1.02	(0.96,1.09)	5.3E-01		
All European Ancestry	5028	32090	0.278	0.269	1.08	(1.02,1.14)	5.2E-03		
Nigeria	689	469	0.569	0.536	1.15	(0.96,1.36)	1.3E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.285	0.262	1.12	(1.05,1.21)	1.3E-03	G	rs16902086
Sweden	833	1750	0.340	0.325	1.07	(0.95,1.21)	2.8E-01		
Holland	744	2034	0.335	0.329	1.03	(0.91,1.17)	6.6E-01		
Spain	642	1540	0.312	0.311	1.01	(0.88,1.16)	9.1E-01		
MEC European American	532	567	0.381	0.378	1.02	(0.85,1.21)	8.7E-01		
Non-Icelanders	2751	5891	0.342	0.335	1.03	(0.96,1.11)	3.5E-01		
All European Ancestry	5028	32090	0.331	0.321	1.08	(1.02,1.13)	3.5E-03		
MEC African American	428	457	0.689	0.687	1.01	(0.83,1.23)	9.3E-01		
Nigeria	689	469	0.791	0.747	1.28	(1.05,1.57)	1.4E-02		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.122	0.107	1.16	(1.04,1.30)	9.9E-03	A	rs6861150

Sweden	833	1750	0.123	0.123	1.00	(0.83,1.19)	9.7E-01		
Holland	744	2034	0.122	0.121	1.01	(0.84,1.21)	9.1E-01		
Spain	642	1540	0.110	0.119	0.91	(0.74,1.12)	3.7E-01		
Non-Icelanders	2751	5891	0.118	0.121	0.98	(0.91,1.07)	7.0E-01		
All European Ancestry	5028	32090	0.119	0.118	1.07	(0.98,1.15)	1.2E-01		
Nigeria	689	469	0.070	0.068	1.02	(0.74,1.40)	9.2E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.127	0.111	1.17	(1.04,1.30)	6.7E-03	C	rs6451795
Sweden	833	1750	0.133	0.132	1.01	(0.85,1.20)	9.1E-01		
Holland	744	2034	0.130	0.130	1.00	(0.84,1.20)	9.8E-01		
Spain	642	1540	0.116	0.129	0.89	(0.73,1.09)	2.6E-01		
Non-Icelanders	2751	5891	0.127	0.130	0.97	(0.87,1.08)	5.8E-01		
All European Ancestry	5028	32090	0.127	0.125	1.06	(0.99,1.15)	1.1E-01		
Nigeria	689	469	0.474	0.463	1.05	(0.88,1.25)	6.0E-01		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.497	0.475	1.09	(1.02,1.16)	6.7E-03	T	rs11743392
Sweden	833	1750	0.499	0.513	0.95	(0.84,1.06)	3.6E-01		
Holland	744	2034	0.542	0.554	0.95	(0.85,1.07)	4.2E-01		
Spain	642	1540	0.498	0.485	1.05	(0.93,1.20)	4.2E-01		
Non-Icelanders	2751	5891	0.513	0.517	0.98	(0.91,1.05)	5.6E-01		
All European Ancestry	5028	32090	0.509	0.507	1.04	(0.99,1.09)	9.5E-02		
Nigeria	689	469	0.999	0.999	NA	NA	NA		
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.408	0.372	1.17	(1.09,1.25)	1.8E-05	T	rs7718785
Sweden	833	1750	0.382	0.406	0.90	(0.80,1.02)	9.9E-02		
Holland	744	2034	0.424	0.438	0.94	(0.84,1.06)	3.5E-01		
Spain	642	1540	0.408	0.400	1.03	(0.90,1.18)	6.5E-01		
MEC European American	532	567	0.472	0.463	1.04	(0.86,1.26)	7.0E-01		
Non-Icelanders	2751	5891	0.422	0.427	0.96	(0.90,1.03)	2.8E-01		
All European Ancestry	5028	32090	0.419	0.416	1.05	(1.00,1.11)	3.1E-02		
MEC African	428	457	0.659	0.675	0.93	(0.70,1.23)	6.1E-01		

American									
Nigeria	689	469	0.710	0.663	1.24	(1.03,1.49)	2.0E-02		
=====									
Sample Set	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Iceland	2277	26199	0.505	0.530	0.90	(0.85,0.97)	2.9E-03	G	rs13179818
Sweden	833	1750	0.507	0.492	1.06	(0.95,1.20)	3.0E-01		
Holland	744	2034	0.435	0.436	1.00	(0.88,1.12)	9.4E-01		
Spain	642	1540	0.495	0.537	0.84	(0.74,0.96)	1.1E-02		
Non-Icelanders	2751	5891	0.479	0.488	0.97	(0.90,1.04)	4.3E-01		
All European Ancestry	5028	32090	0.485	0.499	0.93	(0.89,0.98)	6.0E-03		
Nigeria	689	469	0.001	0.014	0.07	(0.02,0.28)	2.1E-04		
=====									

Allelic Odds Ratios calculated under the multiplicative model. All P values are two sided and have been adjusted for relatedness and other potential stratification of the Icelandic cases and controls. Icelandic data are combined Illumina and Centaurus assay-derived replication data sets. For analyses of combined data for the "Non-Icelanders" and "All European Ancestry" the OR and P values were calculated using the Mantel-Haenszel method, and the frequencies as simple (arithmetic) means of the frequencies of individual groups. CGEMS data are displayed for comparative purposes only and were not included in any of the calculations.

Table 8. LD relations between 5p12 SNPs in Iceland

	rs981782	rs4866929	rs7703618	rs10035564	rs4415084	rs10941679	
rs981782	NA	0.94	0.10	0.38	0.06	0.01	r2
rs4866929	0.99	NA	0.11	0.37	0.07	0.02	
rs7703618	0.19	0.45	NA	0.46	0.81	0.45	
rs10035564	0.95	0.96	0.85	NA	0.37	0.13	
rs4415084	0.29	0.32	0.94	0.79	NA	0.51	
rs10941679	0.19	0.24	0.90	0.59	0.09	NA	
	D'						

Table 9. Multivariate analysis of SNPs in Sp12

Adjusted for >	rs4415084		rs10941679		rs7703618		rs10035564		rs4866929		rs981782	
Tested Variant	P	OR	P	OR	P	OR	P	OR	P	OR	P	OR
rs4415084	NA	NA	4.2E-02	1.07	3.8E-04	1.19	2.8E-07	1.16	3.6E-09	1.16	2.6E-09	1.16
rs10941679	1.7E-03	1.13	NA	NA	1.8E-05	1.17	1.3E-08	1.18	1.4E-10	1.19	1.0E-10	1.19
rs7703618	4.9E-01	0.97	5.2E-01	1.02	NA	NA	1.8E-04	1.13	1.9E-06	1.13	1.4E-06	1.13
rs10035564	6.9E-01	0.99	4.6E-01	1.02	8.9E-01	0.99	NA	NA	3.6E-03	1.10	1.9E-03	1.11
rs4866929	9.8E-01	1.00	6.2E-01	1.01	9.2E-01	1.00	9.5E-01	0.99	NA	NA	4.0E-01	1.06
rs981782	9.7E-01	1.00	6.3E-01	1.01	8.8E-01	1.00	5.6E-01	0.98	7.6E-01	0.98	NA	NA

Table 10. Clinical Correlations

	Number		Frequency						
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Estrogen Receptor									
test:ER positive vs control									
Iceland	1129	26199	0.428	0.372	1.26	(1.16,1.38)	8.9E-08	4	rs4415084
Sweden	377	1750	0.443	0.417	1.11	(0.95,1.30)	1.9E-01	4	rs4415084
Holland	541	2034	0.441	0.403	1.17	(1.02,1.34)	2.4E-02	4	rs4415084
Spain	320	1540	0.420	0.362	1.28	(1.07,1.52)	6.4E-03	4	rs4415084
MEC European American	362	567	0.486	0.424	1.28	(1.06,1.56)	1.1E-02	4	rs4415084
All European Ancestry	2729	32090	0.444	0.396	1.23	(1.16,1.30)	1.8E-11	4	rs4415084
test:ER negative vs control									
Iceland	361	26199	0.373	0.372	1.00	(0.00, inf)	1.0E+00	4	rs4415084
Sweden	77	1750	0.384	0.417	0.87	(0.62,1.21)	4.0E-01	4	rs4415084
Holland	125	2034	0.388	0.402	0.94	(0.72,1.22)	6.4E-01	4	rs4415084
Spain	98	1540	0.343	0.361	0.92	(0.68,1.25)	5.9E-01	4	rs4415084
MEC European American	83	567	0.470	0.424	1.20	(0.86,1.68)	2.9E-01	4	rs4415084
All European Ancestry	744	32090	0.391	0.396	0.98	(0.88,1.10)	7.7E-01	4	rs4415084
test:ER positive vs negative									
Iceland	1129	361	0.428	0.373	1.26	(1.06,1.50)	8.2E-03	4	rs4415084
Sweden	377	77	0.443	0.384	1.28	(0.90,1.82)	1.7E-01	4	rs4415084
Holland	541	125	0.441	0.388	1.25	(0.94,1.65)	1.3E-01	4	rs4415084
Spain	320	98	0.420	0.343	1.39	(1.00,1.94)	5.2E-02	4	rs4415084
MEC European American	362	83	0.486	0.470	1.07	(0.75,1.51)	7.2E-01	4	rs4415084
All European Ancestry	2729	744	0.444	0.391	1.25	(1.11,1.41)	2.0E-04	4	rs4415084
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
test:ER positive vs control									
Iceland	1134	26199	0.284	0.236	1.29	(1.17,1.42)	3.1E-07	3	rs10941679
Sweden	377	1750	0.307	0.273	1.18	(0.99,1.40)	6.4E-02	3	rs10941679
Holland	541	2034	0.304	0.256	1.26	(1.09,1.46)	2.4E-03	3	rs10941679
Spain	320	1540	0.244	0.197	1.31	(1.07,1.61)	9.5E-03	3	rs10941679
MEC European American	364	567	0.302	0.253	1.28	(1.04,1.58)	2.0E-02	3	rs10941679
All European Ancestry	2736	32090	0.288	0.243	1.27	(1.19,1.35)	2.5E-12	3	rs10941679

test:ER negative vs control									
Iceland	361	26199	0.242	0.236	1.03	(0.86,1.24)	7.2E-01	3	rs10941679
Sweden	77	1750	0.266	0.273	0.97	(0.67,1.39)	8.5E-01	3	rs10941679
Holland	125	2034	0.267	0.258	1.05	(0.78,1.40)	7.5E-01	3	rs10941679
Spain	98	1540	0.184	0.197	0.91	(0.63,1.32)	6.3E-01	3	rs10941679
MEC European American	83	567	0.313	0.253	1.35	(0.94,1.93)	1.0E-01	3	rs10941679
All European Ancestry	744	32090	0.254	0.243	1.05	(0.92,1.18)	4.8E-01	3	rs10941679
test:ER positive vs negative									
Iceland	1134	361	0.284	0.242	1.25	(1.03,1.51)	2.5E-02	3	rs10941679
Sweden	377	77	0.307	0.266	1.22	(0.83,1.79)	3.1E-01	3	rs10941679
Holland	541	125	0.304	0.267	1.20	(0.89,1.64)	2.4E-01	3	rs10941679
Spain	320	98	0.244	0.184	1.43	(0.96,2.13)	7.5E-02	3	rs10941679
MEC European American	364	83	0.302	0.313	0.95	(0.66,1.37)	7.7E-01	3	rs10941679
All European Ancestry	2736	744	0.288	0.254	1.21	(1.06,1.38)	4.2E-03	3	rs10941679
==========									
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
test:ER positive vs control									
Iceland	1126	26190	0.504	0.453	1.23	(1.13,1.34)	1.6E-06	3	rs1219648
Sweden	372	1725	0.466	0.381	1.42	(1.21,1.67)	1.6E-05	3	rs1219648
Holland	539	2001	0.468	0.389	1.39	(1.21,1.59)	2.5E-06	3	rs1219648
Spain	317	1493	0.484	0.424	1.27	(1.07,1.51)	5.8E-03	3	rs1219648
MEC European American	NA	NA	NA	NA	NA	(0.00,0.00)	NA	3	rs1219648
All European Ancestry	2354	31409	0.481	0.412	1.29	(1.22,1.38)	3.4E-16	3	rs1219648
test:ER negative vs control									
Iceland	360	26190	0.440	0.453	0.95	(0.82,1.10)	5.0E-01	3	rs1219648
Sweden	76	1725	0.349	0.381	0.87	(0.62,1.22)	4.3E-01	3	rs1219648
Holland	124	2001	0.399	0.389	1.04	(0.80,1.36)	7.5E-01	3	rs1219648
Spain	97	1493	0.464	0.424	1.17	(0.88,1.57)	2.8E-01	3	rs1219648
MEC European American	NA	NA	NA	NA	NA	(0.00,0.00)	NA	3	rs1219648
All European Ancestry	657	31409	0.413	0.412	0.99	(0.88,1.10)	8.3E-01	3	rs1219648
test:ER positive vs negative									
Iceland	1126	360	0.504	0.440	1.29	(1.09,1.53)	2.7E-03	3	rs1219648
Sweden	372	76	0.466	0.349	1.63	(1.14,2.34)	7.3E-03	3	rs1219648
Holland	539	124	0.468	0.399	1.33	(1.00,1.75)	4.7E-02	3	rs1219648
Spain	317	97	0.484	0.464	1.08	(0.79,1.50)	6.2E-01	3	rs1219648

MEC European American	NA	NA	NA	NA	NA	(0.00,0.00)	NA	3	rs1219648
All European Ancestry	2354	657	0.481	0.413	1.30	(1.15,1.47)	2.9E-05	3	rs1219648
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Progesterone Receptor									
test:PR positive vs control									
Iceland	1049	26199	0.422	0.372	1.23	(1.13,1.35)	3.8E-06	4	rs4415084
Sweden	300	1750	0.445	0.417	1.12	(0.94,1.34)	2.0E-01	4	rs4415084
Holland	404	2034	0.442	0.403	1.17	(1.01,1.37)	4.0E-02	4	rs4415084
Spain	269	1540	0.424	0.361	1.30	(1.08,1.57)	6.0E-03	4	rs4415084
MEC European American	294	567	0.490	0.424	1.30	(1.06,1.60)	1.2E-02	4	rs4415084
All European Ancestry	2316	32090	0.445	0.396	1.22	(1.15,1.30)	7.3E-10	4	rs4415084
test:PR negative vs control									
Iceland	424	26199	0.393	0.372	1.09	(0.95,1.25)	2.2E-01	4	rs4415084
Sweden	98	1750	0.393	0.417	0.90	(0.67,1.21)	5.0E-01	4	rs4415084
Holland	260	2034	0.415	0.402	1.05	(0.88,1.27)	5.8E-01	4	rs4415084
Spain	144	1540	0.352	0.362	0.96	(0.74,1.23)	7.4E-01	4	rs4415084
MEC European American	126	567	0.468	0.424	1.19	(0.89,1.59)	2.3E-01	4	rs4415084
All European Ancestry	1052	32090	0.404	0.396	1.05	(0.96,1.15)	2.7E-01	4	rs4415084
test:PR positive vs negative									
Iceland	1049	424	0.422	0.393	1.13	(0.96,1.33)	1.4E-01	4	rs4415084
Sweden	300	98	0.445	0.393	1.24	(0.89,1.72)	2.0E-01	4	rs4415084
Holland	404	260	0.442	0.415	1.11	(0.89,1.39)	3.4E-01	4	rs4415084
Spain	269	144	0.424	0.351	1.36	(1.01,1.84)	4.2E-02	4	rs4415084
MEC European American	294	126	0.489	0.467	1.09	(0.80,1.49)	5.8E-01	4	rs4415084
All European Ancestry	2316	1052	0.445	0.404	1.16	(1.04,1.29)	6.2E-03	4	rs4415084
test:PR positive vs control									
Iceland	1054	26199	0.284	0.235	1.29	(1.17,1.43)	5.8E-07	3	rs10941679
Sweden	300	1750	0.307	0.273	1.18	(0.98,1.43)	8.9E-02	3	rs10941679
Holland	404	2034	0.299	0.258	1.23	(1.04,1.46)	1.5E-02	3	rs10941679
Spain	269	1540	0.240	0.197	1.28	(1.03,1.60)	2.7E-02	3	rs10941679
MEC European American	296	567	0.307	0.253	1.31	(1.05,1.64)	1.7E-02	3	rs10941679
All European Ancestry	2323	32090	0.288	0.243	1.27	(1.18,1.36)	7.2E-11	3	rs10941679

test:PR negative vs control									
Iceland	424	26199	0.243	0.236	1.04	(0.88,1.22)	6.3E-01	3	rs10941679
Sweden	98	1750	0.265	0.273	0.96	(0.69,1.33)	8.1E-01	3	rs10941679
Holland	260	2034	0.294	0.258	1.20	(0.98,1.47)	8.2E-02	3	rs10941679
Spain	144	1540	0.208	0.197	1.07	(0.79,1.44)	6.6E-01	3	rs10941679
MEC European American	126	567	0.313	0.253	1.35	(1.00,1.83)	5.0E-02	3	rs10941679
All European Ancestry	1052	32090	0.265	0.243	1.11	(1.00,1.23)	5.4E-02	3	rs10941679
test:PR positive vs negative									
Iceland	1054	424	0.284	0.243	1.24	(1.03,1.48)	2.3E-02	3	rs10941679
Sweden	300	98	0.307	0.265	1.23	(0.86,1.76)	2.6E-01	3	rs10941679
Holland	404	260	0.300	0.294	1.03	(0.81,1.31)	8.1E-01	3	rs10941679
Spain	269	144	0.240	0.208	1.20	(0.85,1.69)	3.0E-01	3	rs10941679
MEC European American	296	126	0.307	0.313	0.97	(0.71,1.33)	8.5E-01	3	rs10941679
All European Ancestry	2323	1052	0.288	0.265	1.14	(1.02,1.28)	2.7E-02	3	rs10941679

==========

Sample / Comparison	Cases	Controls	Cases	Controls	OR	95%CI	P	Allele	SNP
test:PR positive vs control									
Iceland	1047	26190	0.492	0.453	1.17	(1.07,1.28)	4.6E-04	3	rs1219648
Sweden	295	1725	0.456	0.381	1.36	(1.14,1.63)	5.8E-04	3	rs1219648
Holland	403	2001	0.457	0.389	1.32	(1.13,1.54)	3.7E-04	3	rs1219648
Spain	266	1493	0.477	0.424	1.24	(1.03,1.49)	2.3E-02	3	rs1219648
MEC European American	NA	NA	NA	NA	NA	(0.00,0.00)	NA	3	rs1219648
All European Ancestry	2011	31409	0.470	0.412	1.23	(1.15,1.32)	7.1E-10	3	rs1219648
test:PR negative vs control									
Iceland	423	26190	0.470	0.453	1.07	(0.94,1.23)	3.1E-01	3	rs1219648
Sweden	97	1725	0.407	0.381	1.12	(0.83,1.50)	4.6E-01	3	rs1219648
Holland	258	2001	0.448	0.389	1.27	(1.06,1.53)	1.0E-02	3	rs1219648
Spain	143	1493	0.490	0.424	1.30	(1.02,1.66)	3.4E-02	3	rs1219648
MEC European American	NA	NA	NA	NA	NA	(0.00,0.00)	NA	3	rs1219648
All European Ancestry	921	31409	0.454	0.412	1.16	(1.06,1.28)	2.1E-03	3	rs1219648
test:PR positive vs negative									
Iceland	1047	423	0.492	0.470	1.09	(0.93,1.28)	2.9E-01	3	rs1219648
Sweden	295	97	0.456	0.407	1.22	(0.88,1.69)	2.4E-01	3	rs1219648
Holland	403	258	0.457	0.448	1.04	(0.83,1.29)	7.5E-01	3	rs1219648
Spain	266	143	0.477	0.490	0.95	(0.71,1.27)	7.4E-01	3	rs1219648
MEC European American	NA	NA	NA	NA	NA	(0.00,0.00)	NA	3	rs1219648

All European Ancestry	2011	921	0.470	0.454	1.07	(0.96,1.19)	2.5E-01	3	rs1219648
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Histopathology:									
test:Invasive Ductal vs Control									
All European Ancestry	2897	32090	0.431	0.396	1.17	(1.10,1.24)	1.6E-07	4	rs4415084
All European Ancestry	2899	32090	0.276	0.243	1.18	(1.11,1.26)	4.7E-07	3	rs10941679
All European Ancestry	2512	31409	0.465	0.412	1.21	(1.14,1.29)	2.8E-10	3	rs1219648
test:Invasive Lobular vs Control									
All European Ancestry	419	32090	0.422	0.396	1.07	(0.93,1.23)	3.4E-01	4	rs4415084
All European Ancestry	419	32090	0.264	0.243	1.13	(0.97,1.33)	1.2E-01	3	rs10941679
All European Ancestry	363	31409	0.519	0.412	1.38	(1.19,1.60)	2.2E-05	3	rs1219648
test:Tubular vs Control									
All European Ancestry	187	32090	0.444	0.396	1.20	(0.98,1.48)	7.8E-02	4	rs4415084
All European Ancestry	187	32090	0.321	0.243	1.22	(0.97,1.53)	9.6E-02	3	rs10941679
All European Ancestry	149	31409	0.445	0.412	1.18	(0.94,1.49)	1.5E-01	3	rs1219648
test:Other Invasive vs Control									
All European Ancestry	75	30340	0.367	0.390	0.89	(0.64,1.24)	4.9E-01	4	rs4415084
All European Ancestry	75	30340	0.241	0.236	1.01	(0.00, Inf)	1.0E+00	3	rs10941679
All European Ancestry	58	29684	0.458	0.422	1.24	(0.86,1.79)	2.5E-01	3	rs1219648
test:Mixed Invasive vs Control									
All European Ancestry	192	30550	0.461	0.404	1.31	(1.07,1.61)	8.5E-03	4	rs4415084
All European Ancestry	192	30550	0.323	0.255	1.46	(1.17,1.83)	8.0E-04	3	rs10941679
All European Ancestry	147	29916	0.486	0.407	1.35	(1.07,1.71)	1.1E-02	3	rs1219648
test:Medullary vs Control									
All European Ancestry	43	30340	0.425	0.390	1.15	(0.74,1.77)	5.4E-01	4	rs4415084
All European Ancestry	43	30340	0.313	0.236	1.13	(0.68,1.87)	6.3E-01	3	rs10941679
All European Ancestry	42	29684	0.424	0.422	0.99	(0.64,1.53)	9.7E-01	3	rs1219648
test:DCIS vs Control									
All European Ancestry	275	30340	0.468	0.390	1.25	(1.05,1.49)	1.1E-02	4	rs4415084
All European Ancestry	275	30340	0.268	0.236	1.31	(1.09,1.59)	5.1E-03	3	rs10941679
All European Ancestry	272	29684	0.360	0.422	1.05	(0.88,1.25)	5.9E-01	3	rs1219648

test:LCIS vs Control									
All European Ancestry	28	29773	0.239	0.379	0.72	(0.41,1.27)	2.6E-01	4	rs4415084
All European Ancestry	28	29773	0.169	0.230	0.90	(0.46,1.78)	7.7E-01	3	rs10941679
All European Ancestry	28	29684	0.492	0.422	1.18	(0.69,2.00)	5.4E-01	3	rs1219648
test:Other Non-invasive vs Control									
All European Ancestry	12	28233	0.429	0.387	1.15	(0.50,2.65)	7.5E-01	4	rs4415084
All European Ancestry	12	28233	0.340	0.247	1.55	(0.64,3.77)	3.4E-01	3	rs10941679
All European Ancestry	12	28191	0.493	0.421	1.17	(0.51,2.64)	7.1E-01	3	rs1219648
test:Heterogeneity, All Types									
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	1.9E-01	4	rs4415084
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	5.8E-01	3	rs10941679
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	5.8E-01	3	rs1219648
test:Heterogeneity, Invasive Types									
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	4.4E-01	4	rs4415084
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	5.1E-01	3	rs10941679
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	6.1E-01	3	rs1219648
==========									
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Stage									
test:Stage 0 (in-situ) vs Control									
All European Ancestry	267	29773	0.391	0.379	1.21	(1.02,1.45)	2.9E-02	4	rs4415084
All European Ancestry	267	29773	0.294	0.230	1.27	(1.05,1.55)	1.5E-02	3	rs10941679
All European Ancestry	265	29684	0.399	0.422	1.02	(0.85,1.21)	8.5E-01	3	rs1219648
test:Stage1 vs Control									
All European Ancestry	1394	31523	0.412	0.388	1.19	(1.10,1.29)	2.1E-05	4	rs4415084
All European Ancestry	1394	31523	0.273	0.241	1.19	(1.09,1.30)	1.5E-04	3	rs10941679
All European Ancestry	1385	31409	0.473	0.412	1.23	(1.13,1.33)	4.1E-07	3	rs1219648
test:Stage2 vs Control									
All European Ancestry	1161	31523	0.408	0.388	1.10	(1.01,1.21)	2.5E-02	4	rs4415084
All European Ancestry	1161	31523	0.272	0.241	1.22	(1.11,1.35)	4.9E-05	3	rs10941679
All European Ancestry	1156	31409	0.468	0.412	1.22	(1.12,1.33)	5.3E-06	3	rs1219648

test:Stage3&4 vs Control									
All European Ancestry	438	31523	0.424	0.388	1.12	(0.97,1.29)	1.1E-01	4	rs4415084
All European Ancestry	438	31523	0.271	0.241	1.14	(0.97,1.33)	1.1E-01	3	rs10941679
All European Ancestry	435	31409	0.486	0.412	1.31	(1.14,1.50)	1.1E-04	3	rs1219648
test:Heterogeneity, Stages 1-4									
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	3.9E-01	4	rs4415084
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	6.8E-01	3	rs10941679
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	6.3E-01	3	rs1219648
test:All Invasvie Stages (1-4) vs Control									
All European Ancestry	3233	31523	0.416	0.388	1.15	(1.09,1.22)	4.6E-07	4	rs4415084
All European Ancestry	3233	31523	0.271	0.241	1.19	(1.12,1.27)	1.6E-08	3	rs10941679
All European Ancestry	3216	31409	0.472	0.412	1.24	(1.17,1.31)	7.6E-15	3	rs1219648
test:In-situ (Stage 0) vs Invasive (Stage 1-4)									
All European Ancestry	267	2749	0.391	0.410	1.04	(0.86,1.24)	7.0E-01	4	rs4415084
All European Ancestry	267	2749	0.294	0.260	1.05	(0.86,1.29)	6.0E-01	3	rs10941679
All European Ancestry	265	2739	0.399	0.480	0.84	(0.70,1.00)	5.2E-02	3	rs1219648
Sample / Comparison	**Cases**	**Controls**	**Cases**	**Controls**	**OR**	**95%CI**	**P**	**Allele**	**SNP**
Grade									
test:Grade1 vs Control									
All European Ancestry	471	31523	0.443	0.388	1.26	(1.10,1.44)	6.6E-04	4	rs4415084
All European Ancestry	471	31523	0.295	0.241	1.25	(1.08,1.45)	2.5E-03	3	rs10941679
All European Ancestry	467	31409	0.479	0.412	1.21	(1.06,1.39)	4.4E-03	3	rs1219648
test:Grade2 vs Control									
All European Ancestry	985	31523	0.428	0.388	1.20	(1.09,1.31)	1.8E-04	4	rs4415084
All European Ancestry	985	31523	0.287	0.241	1.27	(1.15,1.41)	5.5E-06	3	rs10941679
All European Ancestry	981	31409	0.476	0.412	1.31	(1.19,1.43)	1.8E-08	3	rs1219648
test:Grade3 vs Control									
All European Ancestry	690	31523	0.402	0.388	1.05	(0.94,1.17)	4.2E-01	4	rs4415084

All European Ancestry	690	31523	0.251	0.241	1.05	(0.92,1.19)	4.7E-01	3	rs10941679
All European Ancestry	683	31409	0.447	0.412	1.13	(1.01,1.26)	2.8E-02	3	rs1219648
test:Trend Test Grade									
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	1.8E-02	4	rs4415084
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	2.0E-02	3	rs10941679
All European Ancestry	NA	NA	NA	NA	NA	(0.00,0.00)	2.9E-01	3	rs1219648
Node Status									
test:Node positive vs control									
All European Ancestry	1120	31523	0.407	0.388	1.10	(1.01,1.21)	2.6E-02	4	rs4415084
All European Ancestry	1122	31523	0.264	0.241	1.16	(1.05,1.28)	3.2E-03	3	rs10941679
All European Ancestry	1113	31409	0.484	0.412	1.32	(1.21,1.44)	2.0E-10	3	rs1219648
test:Node negative vs control									
All European Ancestry	1883	31523	0.421	0.388	1.18	(1.10,1.26)	4.3E-06	4	rs4415084
All European Ancestry	1886	31523	0.276	0.241	1.22	(1.12,1.31)	7.4E-07	3	rs10941679
All European Ancestry	1873	31409	0.470	0.412	1.20	(1.12,1.28)	3.0E-07	3	rs1219648
test:Node positive vs negative									
All European Ancestry	1120	1883	0.406	0.421	0.94	(0.84,1.05)	2.6E-01	4	rs4415084
All European Ancestry	1122	1886	0.264	0.277	0.97	(0.86,1.09)	5.6E-01	3	rs10941679
All European Ancestry	1113	1873	0.484	0.470	1.11	(1.00,1.24)	4.7E-02	3	rs1219648

Table 11. *1° Familial* Relative Risks by SNP Genotype

SNP	Location	Geno type 1	# Affected with Geno type 1	gfRRgt1	# of Affected 1° Relatives for Geno type 1	Geno type 2	# Affected with Genotype 2	gfRRgt 2	# of Affected 1° Relatives for Genotype 2	gfRRgt1/ gfRRgt2	*P*-value
rs4415084	5p12	C/T	1089	1.758	317	C/C	781	1.632	202	1.077	0.0822
		T/T	373	1.932	111	C/C	781	1.632	202	1.184	0.0511
		T/T	373	1.932	111	C/T	1089	1.758	317	1.099	0.2410
rs10941679	5p12	A/G	884	1.832	262	A/A	1152	1.694	311	1.081	0.1314
		G/G	148	2.192	50	A/A	1152	1.694	311	1.294	0.0599
		G/G	148	2.192	50	A/G	884	1.832	262	1.197	0.1581
rs1219648	10q26	A/G	1107	1.709	301	A/A	600	1.532	152	1.115	0.1152
		G/G	563	2.063	186	A/A	600	1.532	152	1.346	0.0019
		G/G	563	2.063	186	A/G	1107	1.709	301	1.207	0.0076

Table 12. Surrogate markers for marker rs4415084. Markers with values of r^2 greater than 0.2 to rs4415084 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 rs4415084, 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 sequnces for the marker.

Anchor SNP	Corr SNP	r^2	D'	P-value	Pos in Bld 36	SEQ ID NO:
rs4415084	rs4866900	0.92417	0.236675	3.02E-08	44480857	1
rs4415084	rs7712213	1	0.20783	4.15E-09	44467026	2
rs4415084	rs1482690	1	0.201693	5.32E-09	44524597	3
rs4415084	rs1482663	1	0.207207	3.23E-09	44578859	4
rs4415084	rs1351633	1	0.207207	3.23E-09	44579608	5
rs4415084	rs983940	1	0.207207	3.23E-09	44579893	6
rs4415084	rs4866905	1	0.207207	3.23E-09	44591624	7
rs4415084	rs10079222	1	0.207207	3.23E-09	44597230	8
rs4415084	rs4463187	0.738181	0.259028	3.59E-08	44604412	9
rs4415084	rs10054521	0.736037	0.254284	5.31E-08	44611928	10
rs4415084	rs10059745	0.741785	0.267307	2.28E-08	44622995	11
rs4415084	rs6882655	0.741785	0.267307	2.28E-08	44626667	12
rs4415084	rs4639238	0.741785	0.267307	2.28E-08	44627752	13
rs4415084	rs10066953	0.738181	0.259028	3.59E-08	44636753	14
rs4415084	rs12074507	0.738181	0.259028	3.59E-08	44640070	15
rs4415084	rs4573006	0.738181	0.259028	3.59E-08	44647407	16
rs4415084	rs4529201	0.738181	0.259028	3.59E-08	44649726	17
rs4415084	rs5866354	0.738181	0.259028	3.59E-08	44662567	18
rs4415084	rs4463188	1	1	7.12E-36	44678427	19
rs4415084	rs4321755	1	1	1.28E-36	44681952	20
rs4415084	rs4492116	1	1	2.08E-36	44682382	21
rs4415084	rs4613716	1	0.459459	2.24E-17	44685701	22
rs4415084	rs7735881	1	1	1.28E-36	44685933	23
rs4415084	rs7723539	1	1	1.28E-36	44695967	24
rs4415084	rs10805685	1	1	2.08E-36	44697715	25
rs4415084	rs10941677	1	1	5.45E-36	44698156	26
rs4415084	rs4415084	1	1	-	44698272	235
rs4415084	rs4415085	1	1	1.28E-36	44698716	27
rs4415084	rs7720551	1	1	1.28E-36	44700234	28
rs4415084	rs6674055	1	1	3.35E-36	44702722	29
rs4415084	rs4419600	1	1	1.28E-36	44714291	30
rs4415084	rs12187196	1	1	1.28E-36	44719576	31
rs4415084	rs12522626	1	1	3.38E-36	44721455	32
rs4415084	rs4571460	1	1	5.45E-36	44722845	33
rs4415084	rs6451770	1	1	1.28E-36	44727152	34
rs4415084	rs12515012	1	1	1.28E-36	44730292	35
rs4415084	rs2165009	1	1	2.08E-36	44733673	36
rs4415084	rs13156930	1	1	1.28E-36	44733792	37
rs4415084	rs920328	1	0.93135	1.41E-32	44734808	38
rs4415084	rs1821936	1	1	2.08E-36	44735239	39
rs4415084	rs714130	1	1	1.28E-36	44737175	40
rs4415084	rs2013513	1	1	5.45E-36	44738063	41
rs4415084	rs920329	1	1	2.71E-36	44738264	42
rs4415084	rs2218081	1	1	1.28E-36	44740897	43
rs4415084	rs10941679	1	0.512661	2.03E-17	44742255	44

Anchor SNP	Corr SNP	r^2	D'	P-value	Pos In Bld 36	SEQ ID NO:
rs4415084	rs2165010	1	1	3.35E-36	44742537	45
rs4415084	rs1438825	1	1	1.28E-36	44742688	46
rs4415084	rs6861560	1	1	1.28E-36	44744135	47
rs4415084	rs16901937	1	0.965497	1.66E-34	44744698	48
rs4415084	rs2216080	0.964747	0.930737	2.74E-31	44750087	49
rs4415084	rs11747159	0.920891	0.708022	7.34E-22	44773467	50
rs4415084	rs2330572	0.889429	0.736776	7.58E-23	44776746	51
rs4415084	rs994793	0.889429	0.736776	7.58E-23	44779004	52
rs4415084	rs1438827	0.884157	0.677153	1.15E-20	44787713	53
rs4415084	rs11949847	0.922578	0.766182	1.05E-23	44787926	54
rs4415084	rs7712949	0.920891	0.708022	7.34E-22	44806102	55
rs4415084	rs13154781	0.920891	0.708022	7.34E-22	44810784	56
rs4415084	rs11746960	0.924573	0.767952	4.51E-24	44813835	57
rs4415084	rs7711697	0.924114	0.764101	4.56E-23	44816160	58
rs4415084	rs16901964	0.920891	0.708022	7.34E-22	44819012	59
rs4415084	rs6875933	0.920497	0.707417	1.46E-21	44822453	60
rs4415084	rs727305	0.920891	0.708022	7.34E-22	44831799	61
rs4415084	rs13177711	0.923589	0.767079	6.68E-24	44832719	62
rs4415084	rs1438820	0.924573	0.767952	4.51E-24	44833527	63
rs4415084	rs1438819	0.920891	0.708022	7.34E-22	44833603	64
rs4415084	rs12651949	0.919999	0.664841	4.20E-18	44833669	65
rs4415084	rs10462080	0.920891	0.708022	7.34E-22	44834809	66
rs4415084	rs10462081	0.920891	0.708022	7.34E-22	44836422	67
rs4415084	rs13183209	0.920891	0.708022	7.34E-22	44839506	68
rs4415084	rs6872254	0.920891	0.708022	7.34E-22	44839541	69
rs4415084	rs7717459	0.920891	0.708022	7.34E-22	44840262	70
rs4415084	rs13159598	0.924573	0.767952	4.51E-24	44841683	71
rs4415084	rs3761548	0.919139	0.701965	5.44E-21	44843835	72
rs4415084	rs3747479	0.919139	0.701965	5.44E-21	44844819	73
rs4415084	rs1866406	0.920891	0.708022	7.34E-22	44845702	74
rs4415084	rs13174122	0.920891	0.708022	7.34E-22	44846497	75
rs4415084	rs11746506	0.920891	0.708022	7.34E-22	44848323	76
rs4415084	rs12188871	0.918911	0.679189	7.59E-21	44849761	77
rs4415084	rs11741772	0.917793	0.698764	8.18E-21	44850354	78
rs4415084	rs7716571	0.923792	0.762262	2.31E-23	44852741	79
rs4415084	rs7720797	0.924573	0.767952	4.51E-24	44853066	80
rs4415084	rs9637783	0.920891	0.708022	7.34E-22	44855403	81
rs4415084	rs1061310	0.924573	0.767952	4.51E-24	44856607	82
rs4415084	rs4457089	0.920891	0.708022	7.34E-22	44857493	83
rs4415084	rs13189120	0.920891	0.708022	7.34E-22	44858040	84
rs4415084	rs930395	1	0.402174	5.87E-14	44858215	85
rs4415084	rs10512685	0.924573	0.767952	4.51E-24	44859124	86
rs4415084	rs6887533	0.924573	0.767952	4.51E-24	44863049	87
rs4415084	rs6868232	0.920271	0.698768	4.83E-21	44863437	88
rs4415084	rs12513749	0.920694	0.707719	1.47E-21	44863960	89
rs4415084	rs12518851	0.910529	0.687491	1.15E-16	44863988	90
rs4415084	rs1048758	0.924573	0.767952	4.51E-24	44864351	91
rs4415084	rs13155698	0.920891	0.708022	7.34E-22	44864438	92
rs4415084	rs13160259	0.924573	0.767952	4.51E-24	44864721	93
rs4415084	rs6896350	0.920891	0.708022	7.34E-22	44868328	94
rs4415084	rs1371025	0.91986	0.707156	1.08E-21	44868990	95
rs4415084	rs4586389	0.920891	0.708022	7.34E-22	44872313	96

Anchor SNP	Corr SNP	r^2	D'	P-value	Pos in Bld 36	SEQ ID NO:
rs4415084	rs6451775	0.920891	0.708022	7.34E-22	44872545	97
rs4415084	rs7380559	0.924573	0.767952	4.51E-24	44872767	98
rs4415084	rs729599	0.920891	0.708022	7.34E-22	44878017	99
rs4415084	rs987394	0.920891	0.708022	7.34E-22	44882135	100
rs4415084	rs7715731	0.917461	0.686969	6.41E-20	44882601	101
rs4415084	rs4440370	0.920891	0.708022	7.34E-22	44889109	102
rs4415084	rs4492119	0.920497	0.707417	1.46E-21	44891371	103
rs4415084	rs7703497	0.920891	0.708022	7.34E-22	44892785	104
rs4415084	rs6451778	0.920891	0.708022	7.34E-22	44893745	105
rs4415084	rs13062132	0.924573	0.767952	4.51E-24	44894017	106
rs4415084	rs1438621	0.922391	0.762606	2.84E-23	44894208	107
rs4415084	rs1438622	0.920891	0.708022	7.34E-22	44894929	108
rs4415084	rs4373287	0.920891	0.708022	7.34E-22	44898641	109
rs4415084	rs6871052	0.919658	0.706847	2.16E-21	44899074	110
rs4415084	rs6893319	0.920891	0.708022	7.34E-22	44899486	111
rs4415084	rs10053247	0.920891	0.708022	7.34E-22	44899716	112
rs4415084	rs10040082	0.957789	0.739047	9.26E-23	44901611	113
rs4415084	rs10057521	0.920891	0.708022	7.34E-22	44901743	114
rs4415084	rs10065838	0.920891	0.708022	7.34E-22	44901919	115
rs4415084	rs6894324	0.919555	0.702623	2.73E-21	44903093	116
rs4415084	rs4395640	0.920891	0.708022	7.34E-22	44904857	117
rs4415084	rs10070037	0.920891	0.708022	7.34E-22	44905994	118
rs4415084	rs4518409	0.924573	0.767952	4.51E-24	44906609	119
rs4415084	rs9292913	0.923589	0.767079	6.86E-24	44906636	120
rs4415084	rs9292914	0.906242	0.69671	2.34E-18	44907136	121
rs4415084	rs10059086	0.920891	0.708022	7.34E-22	44907764	122
rs4415084	rs11951760	0.914383	0.695917	2.65E-20	44907929	123
rs4415084	rs4329028	0.924573	0.767952	4.51E-24	44908110	124
rs4415084	rs7716600	1	0.391985	1.56E-13	44910762	125
rs4415084	rs4412123	0.924573	0.767952	4.51E-24	44912045	126
rs4415084	rs7705343	0.924573	0.767952	4.51E-24	44915334	127
rs4415084	rs10040488	0.920891	0.708022	7.34E-22	44916045	128
rs4415084	rs4642377	0.920589	0.703468	1.84E-21	44920997	129
rs4415084	rs4391175	0.920891	0.708022	7.34E-22	44925813	130
rs4415084	rs4129642	0.920891	0.708022	7.34E-22	44933866	131
rs4415084	rs9790879	0.924573	0.767952	4.51E-24	44935642	132
rs4415084	rs9790896	0.86242	0.702742	3.43E-21	44935848	133
rs4415084	rs4457088	0.919555	0.702623	2.73E-21	44936711	134
rs4415084	rs4866784	0.920891	0.708022	7.34E-22	44936808	135
rs4415084	rs9791056	0.920891	0.708022	7.34E-22	44939648	136
rs4415084	rs6880275	0.920891	0.708022	7.34E-22	44944692	137
rs4415084	rs6870136	0.920891	0.708022	7.34E-22	44946419	138
rs4415084	rs6861563	0.920891	0.708022	7.34E-22	44948610	139
rs4415084	rs7703618	0.920891	0.708022	7.34E-22	44950336	140
rs4415084	rs10077814	0.924573	0.767952	4.51E-24	44952546	141
rs4415084	rs6451763	0.920891	0.708022	7.34E-22	44954050	142
rs4415084	rs4298259	0.920891	0.708022	7.34E-22	44956468	143
rs4415084	rs7736092	0.91986	0.707156	1.09E-21	44956752	144
rs4415084	rs7728431	0.920232	0.705178	1.49E-21	44958436	145
rs4415084	rs7708506	0.918803	0.706267	1.62E-21	44958461	146
rs4415084	rs10039866	0.920891	0.708022	7.34E-22	44960818	147
rs4415084	rs10043344	0.922377	0.762771	2.62E-23	44962275	148

Anchor SNP	Corr SNP	r^2	D'	P-value	Pos in Bld 36	SEQ ID NO:
rs4415084	rs10038554	0.919555	0.702623	2.73E-21	44962864	149
rs4415084	rs10044096	0.924573	0.767952	4.51E-24	44963122	150
rs4415084	rs10041518	0.920891	0.708022	7.34E-22	44963163	151
rs4415084	rs12517690	0.920891	0.708022	7.34E-22	44975050	152
rs4415084	rs6875287	0.917171	0.700475	1.83E-20	44977387	153
rs4415084	rs11958808	0.924573	0.767952	4.51E-24	44980847	154
rs4415084	rs3935086	0.904005	0.519752	1.24E-15	44996680	155
rs4415084	rs3935213	0.850441	0.267659	2.47E-08	44997201	156
rs4415084	rs4460145	0.8543	0.276388	1.01E-08	45004083	157
rs4415084	rs8869488	0.8543	0.276388	1.01E-08	45006273	158
rs4415084	rs5866995	0.847224	0.259219	3.63E-08	45012604	159
rs4415084	rs2067980	1	0.265513	1.01E-09	45016074	160
rs4415084	rs4296810	0.836134	0.238566	2.48E-07	45019819	161
rs4415084	rs7709661	0.834245	0.243662	4.06E-07	45039848	162
rs4415084	rs6894974	0.839281	0.254787	1.68E-07	45056286	163
rs4415084	rs4533894	0.841387	0.255659	1.41E-07	45060926	164
rs4415084	rs4371761	0.83644	0.244535	3.42E-07	45061977	165
rs4415084	rs7716101	0.841387	0.255659	1.41E-07	45065624	166
rs4415084	rs7731099	0.841176	0.262636	1.12E-07	45073783	167
rs4415084	rs7701679	0.828648	0.232155	9.87E-07	45078551	168
rs4415084	rs12522398	0.82443	0.225046	6.11E-07	45085230	169
rs4415084	rs4502832	0.836134	0.238566	2.48E-07	45087138	170
rs4415084	rs11948166	0.916552	0.65068	6.95E-20	45087191	171
rs4415084	rs12054976	0.841387	0.255659	1.41E-07	45093077	172
rs4415084	rs4485937	0.836134	0.238566	2.48E-07	45101400	173
rs4415084	rs4389695	0.836134	0.238566	2.48E-07	45107668	174
rs4415084	rs13183434	1	0.265513	1.01E-09	45110390	175
rs4415084	rs12521639	0.841387	0.255659	1.41E-07	45114238	176
rs4415084	rs10051582	0.916552	0.65068	6.95E-20	45126063	177
rs4415084	rs6885307	0.841387	0.255659	1.41E-07	45130260	178
rs4415084	rs10805692	0.836134	0.238566	2.48E-07	45135215	179
rs4415084	rs10941692	0.836134	0.238566	2.48E-07	45135535	180

Table 13. Surrogate SNP markers for marker rs10941679. Markers with values of r^2 greater than 0.2 to rs10941679 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 rs10941679, 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 sequnces for the marker.

Discovery SNP	Corr SNP	R^2	D'	P-value	Pos in Bld 36	SEQ ID NO:
rs10941679	rs10470354	0.697983	0.234357	3.79E-07	44432110	181
rs10941679	rs12054807	0.712181	0.251051	9.37E-08	44433098	182
rs10941679	rs10941665	0.711093	0.248052	1.16E-07	44434453	183
rs10941679	rs7356597	0.701296	0.242799	2.48E-07	44435967	184
rs10941679	rs2200123	0.700877	0.292463	6.90E-08	44444748	185
rs10941679	rs10472394	0.71647	0.283453	5.64E-08	44445489	186
rs10941679	rs10055789	0.715333	0.260073	5.92E-08	44446940	187
rs10941679	rs10055953	0.714293	0.257042	7.33E-08	44447011	188
rs10941679	rs2330551	0.715333	0.260073	5.92E-08	44448702	189
rs10941679	rs987852	0.712181	0.251051	9.37E-08	44450245	190
rs10941679	rs1482668	0.712181	0.251051	9.37E-08	44450407	191
rs10941679	rs2877162	0.715333	0.260073	5.92E-08	44451149	192
rs10941679	rs2877163	0.711093	0.248052	1.16E-07	44451226	193
rs10941679	rs2330553	0.712181	0.251051	9.37E-08	44451426	194
rs10941679	rs1482667	0.704607	0.251717	1.60E-07	44452403	195
rs10941679	rs4242112	0.712181	0.251051	9.37E-08	44452490	196
rs10941679	rs1384451	0.766758	0.281037	1.07E-08	44455011	197
rs10941679	rs1482665	0.766758	0.281037	1.07E-08	44456232	198
rs10941679	rs13357659	0.766758	0.281037	1.07E-08	44466842	199
rs10941679	rs6893590	0.918347	0.220962	1.03E-07	44467227	200
rs10941679	rs8180484	0.915674	0.20433	2.43E-07	44507720	201
rs10941679	rs1384450	0.915674	0.20433	2.43E-07	44515901	202
rs10941679	rs10941667	0.915899	0.20528	2.44E-07	44530438	203
rs10941679	rs16901890	0.628697	0.216876	3.40E-06	44548272	204
rs10941679	rs2126434	0.913386	0.20415	4.85E-07	44548566	205
rs10941679	rs2126435	0.917481	0.212203	1.43E-07	44552968	206
rs10941679	rs4866777	0.919213	0.220312	8.32E-08	44574747	207
rs10941679	rs1482696	0.830339	0.353461	9.52E-11	44575210	208
rs10941679	rs4866902	0.917008	0.219257	1.66E-07	44580477	209
rs10941679	rs10805664	0.919213	0.220312	8.32E-08	44587002	210
rs10941679	rs7708449	0.664362	0.242856	3.43E-07	44604983	211
rs10941679	rs7713139	0.661866	0.237133	5.11E-07	44605617	212
rs10941679	rs10462078	0.664362	0.242956	3.43E-07	44621291	213
rs10941679	rs7448715	0.664362	0.242956	3.43E-07	44621309	214
rs10941679	rs4866911	0.664362	0.242956	3.43E-07	44622497	215
rs10941679	rs4392631	0.664362	0.242956	3.43E-07	44628924	216
rs10941679	rs4866779	0.668197	0.252735	2.11E-07	44659107	217
rs10941679	rs11952948	0.682849	0.246642	4.69E-07	44663041	218
rs10941679	rs4463168	1	0.510791	1.02E-16	44676427	19
rs10941679	rs4321755	1	0.512661	2.03E-17	44681952	20
rs10941679	rs4492116	1	0.509261	2.69E-17	44682382	21
rs10941679	rs4613718	1	0.235547	4.63E-10	44685701	22
rs10941679	rs7735681	1	0.512661	2.03E-17	44685933	23
rs10941679	rs7723539	1	0.512661	2.03E-17	44695967	24
rs10941679	rs10805685	1	0.509261	2.69E-17	44697715	25
rs10941679	rs10941677	1	0.509261	7.08E-17	44698156	26

Discovery SNP	Corr SNP	R^2	D'	P-value	Pos In Bld 35	SEQ ID NO:
rs10941679	rs4415084	1	0.512661	2.03E-17	44698272	219
rs10941679	rs4415085	1	0.512661	2.03E-17	44698716	27
rs10941679	rs7720551	1	0.512661	2.03E-17	44700234	28
rs10941679	rs6874055	1	0.512661	5.36E-17	44702722	29
rs10941679	rs4419600	1	0.512661	2.03E-17	44714291	30
rs10941679	rs12187196	1	0.512661	2.03E-17	44719576	31
rs10941679	rs12522626	1	0.505814	3.57E-17	44721455	32
rs10941679	rs4571480	1	0.509261	7.08E-17	44722945	33
rs10941679	rs6451770	1	0.512661	2.03E-17	44727152	34
rs10941679	rs12515012	1	0.512661	2.03E-17	44730292	35
rs10941679	rs2165009	1	0.509261	2.69E-17	44733873	36
rs10941679	rs13156930	1	0.512661	2.03E-17	44733792	37
rs10941679	rs920328	1	0.55045	2.49E-16	44734608	38
rs10941679	rs1621936	1	0.509261	2.69E-17	44735239	39
rs10941679	rs714100	1	0.512661	2.03E-17	44737175	40
rs10941679	rs2013513	1	0.509261	7.08E-17	44738063	41
rs10941679	rs920329	1	0.509261	2.69E-17	44738264	42
rs10941679	rs2218081	1	0.512661	2.03E-17	44740897	43
rs10941679	rs10941679	1	1	-	44742255	236
rs10941679	rs2165010	1	0.512661	5.36E-17	44742537	45
rs10941679	rs1438825	1	0.512661	2.03E-17	44742886	46
rs10941679	rs6861560	1	0.512661	2.03E-17	44744135	47
rs10941679	rs16901937	1	0.494973	5.47E-17	44744896	48
rs10941679	rs2218080	0.843571	0.456436	5.90E-14	44750087	49
rs10941679	rs11747159	0.841573	0.434894	2.26E-12	44773467	50
rs10941679	rs2330572	0.834953	0.383744	3.61E-11	44776746	51
rs10941679	rs994793	0.834953	0.383744	3.61E-11	44779004	52
rs10941679	rs1438827	0.839428	0.417031	5.91E-12	44787713	53
rs10941679	rs11949847	0.836361	0.39369	2.42E-11	44787926	54
rs10941679	rs7712949	0.841573	0.434894	2.26E-12	44806102	55
rs10941679	rs13154781	0.841573	0.434894	2.26E-12	44810784	56
rs10941679	rs11746980	0.837222	0.4	1.49E-11	44813635	57
rs10941679	rs7711697	0.831574	0.396173	3.79E-11	44816160	58
rs10941679	rs16901964	0.841573	0.434894	2.26E-12	44819012	59
rs10941679	rs6875933	0.841573	0.434894	2.26E-12	44822453	60
rs10941679	rs727305	0.841573	0.434894	2.26E-12	44831799	61
rs10941679	rs13177711	0.836797	0.395865	1.90E-11	44832719	62
rs10941679	rs1438820	0.837222	0.4	1.49E-11	44833527	63
rs10941679	rs1438819	0.841573	0.434894	2.26E-12	44833603	64
rs10941679	rs12651949	0.797511	0.412615	1.53E-08	44833869	65
rs10941679	rs10462080	0.841573	0.434894	2.26E-12	44834809	66
rs10941679	rs10462081	0.841573	0.434894	2.26E-12	44836422	67
rs10941679	rs13183209	0.841573	0.434894	2.26E-12	44839506	68
rs10941679	rs6872254	0.841573	0.434894	2.26E-12	44839541	69
rs10941679	rs7717459	0.841573	0.434894	2.26E-12	44840262	70
rs10941679	rs13159598	0.837222	0.4	1.49E-11	44841683	71
rs10941679	rs3761648	0.843009	0.44764	1.34E-12	44843836	72
rs10941679	rs3747479	0.843009	0.44764	1.34E-12	44844918	73
rs10941679	rs1866406	0.841573	0.434894	2.26E-12	44845702	74
rs10941679	rs13174122	0.841573	0.434894	2.26E-12	44846497	75
rs10941679	rs11746506	0.841573	0.434894	2.26E-12	44848323	76
rs10941679	rs12188871	0.843661	0.453852	8.25E-13	44849761	77

Discovery SNP	Corr SNP	R^2	D'	P-value	Pos In Bld 36	SEQ ID NO:
rs10941679	rs11741772	0.832079	0.420989	1.87E-11	44850354	78
rs10941679	rs7716571	0.8282	0.392326	7.08E-11	44852741	79
rs10941679	rs7720787	0.837222	0.4	1.49E-11	44853066	80
rs10941679	rs9637763	0.841573	0.434894	2.26E-12	44855403	81
rs10941679	rs1061310	0.837222	0.4	1.49E-11	44856607	82
rs10941679	rs4457069	0.841573	0.434894	2.26E-12	44857493	83
rs10941679	rs13189120	0.841573	0.434894	2.26E-12	44858040	84
rs10941679	rs930395	1	0.784483	4.20E-22	44858215	85
rs10941679	rs10512865	0.837222	0.4	1.49E-11	44859124	86
rs10941679	rs6867533	0.837222	0.4	1.49E-11	44863049	87
rs10941679	rs6868232	0.830185	0.426748	1.62E-11	44863437	88
rs10941679	rs12513749	0.840368	0.43365	4.47E-12	44863960	89
rs10941679	rs12518851	0.830617	0.419161	5.99E-11	44863988	90
rs10941679	rs1048758	0.837222	0.4	1.49E-11	44864351	91
rs10941679	rs13155698	0.841573	0.434894	2.26E-12	44864438	92
rs10941679	rs13160259	0.837222	0.4	1.49E-11	44864721	93
rs10941679	rs6896350	0.841573	0.434894	2.26E-12	44868328	94
rs10941679	rs1371025	0.841219	0.431852	2.88E-12	44869990	95
rs10941679	rs4596389	0.841573	0.434894	2.26E-12	44872313	96
rs10941679	rs6451775	0.841573	0.434894	2.26E-12	44872545	97
rs10941679	rs7380559	0.837222	0.4	1.49E-11	44872767	98
rs10941679	rs729599	0.841573	0.434894	2.26E-12	44876017	99
rs10941679	rs987394	0.841573	0.434894	2.26E-12	44882135	100
rs10941679	rs7715731	0.816216	0.41132	1.82E-10	44882601	101
rs10941679	rs4440370	0.841573	0.434894	2.26E-12	44889109	102
rs10941679	rs4492119	0.841573	0.434894	2.26E-12	44891371	103
rs10941679	rs7703497	0.841573	0.434894	2.26E-12	44892765	104
rs10941679	rs6451778	0.841573	0.434894	2.26E-12	44893745	105
rs10941679	rs13362132	0.837222	0.4	1.49E-11	44894017	106
rs10941679	rs1438821	0.828917	0.387517	5.73E-11	44894208	107
rs10941679	rs1438822	0.841573	0.434894	2.26E-12	44894929	108
rs10941679	rs4373287	0.841573	0.434894	2.26E-12	44898641	109
rs10941679	rs6871052	0.839995	0.430591	5.70E-12	44899074	110
rs10941679	rs6893319	0.841573	0.434894	2.26E-12	44899486	111
rs10941679	rs10053247	0.841573	0.434894	2.26E-12	44899716	112
rs10941679	rs10040082	0.84267	0.444569	1.72E-12	44901611	113
rs10941679	rs10057521	0.841573	0.434894	2.26E-12	44901743	114
rs10941679	rs10065636	0.841573	0.434894	2.26E-12	44901919	115
rs10941679	rs6894324	0.835715	0.427926	7.66E-12	44903093	116
rs10941679	rs4395640	0.841573	0.434894	2.26E-12	44904857	117
rs10941679	rs10070037	0.841573	0.434894	2.26E-12	44905994	118
rs10941679	rs4516409	0.837222	0.4	1.49E-11	44906609	119
rs10941679	rs9292913	0.836797	0.398866	1.90E-11	44906636	120
rs10941679	rs9292914	0.824223	0.446215	9.11E-11	44907138	121
rs10941679	rs10059068	0.841573	0.434894	2.26E-12	44907764	122
rs10941679	rs11951760	0.885323	0.466609	1.29E-12	44907929	123
rs10941679	rs4329028	0.837222	0.4	1.49E-11	44908110	124
rs10941679	rs7716800	1	0.7772	1.77E-21	44910762	125
rs10941679	rs4412123	0.837222	0.4	1.49E-11	44912045	126
rs10941679	rs7705343	0.837222	0.4	1.49E-11	44915334	127
rs10941679	rs10040488	0.841573	0.434894	2.26E-12	44916045	128
rs10941679	rs4642377	0.636082	0.43096	6.03E-12	44920997	129

Discovery SNP	Corr SNP	R^2	D'	P-value	Pos In Bld 36	SEQ ID NO:
rs10941679	rs4391175	0.841573	0.434894	2.26E-12	44925813	130
rs10941679	rs4129842	0.841573	0.434894	2.26E-12	44933886	131
rs10941679	rs9790879	0.837222	0.4	1.49E-11	44935642	132
rs10941679	rs9790896	0.780757	0.340217	9.70E-10	44935848	133
rs10941679	rs4457088	0.835715	0.427926	7.66E-12	44936711	134
rs10941679	rs4666784	0.841573	0.434894	2.26E-12	44936888	135
rs10941679	rs9791056	0.841573	0.434894	2.26E-12	44939648	136
rs10941679	rs6680275	0.841573	0.434894	2.26E-12	44944692	137
rs10941679	rs6670136	0.841573	0.434894	2.26E-12	44946419	138
rs10941679	rs6681563	0.841573	0.434894	2.26E-12	44948610	139
rs10941679	rs7703618	0.841573	0.434894	2.26E-12	44950336	140
rs10941679	rs10077814	0.837222	0.4	1.49E-11	44952545	141
rs10941679	rs6451783	0.841573	0.434894	2.26E-12	44954050	142
rs10941679	rs4298259	0.841573	0.434894	2.26E-12	44956468	143
rs10941679	rs7736092	0.841219	0.431652	2.88E-12	44956752	144
rs10941679	rs7728431	0.838562	0.431223	4.37E-12	44956436	145
rs10941679	rs7708506	0.840855	0.428766	3.87E-12	44958451	146
rs10941679	rs10039666	0.841573	0.434894	2.26E-12	44960818	147
rs10941679	rs10043344	0.83823	0.407626	1.23E-11	44962275	148
rs10941679	rs10038554	0.835715	0.427926	7.66E-12	44962864	149
rs10941679	rs10044096	0.837222	0.4	1.49E-11	44963122	150
rs10941679	rs10041516	0.841573	0.434894	2.26E-12	44963163	151
rs10941679	rs12517690	0.841573	0.434894	2.26E-12	44975050	152
rs10941679	rs6875287	0.833288	0.419966	3.75E-11	44977387	153
rs10941679	rs11958808	0.837222	0.4	1.49E-11	44980847	154
rs10941679	rs2067980	0.767084	0.304748	4.34E-08	45018074	160
rs10941679	rs11948186	0.634593	0.268543	1.36E-07	45087191	171
rs10941679	rs13183434	0.685481	0.243358	9.99E-07	45110390	175
rs10941679	rs10051592	0.634593	0.268543	1.36E-07	45126083	177

Table 14. Surrogate SNP markers for marker rs1219648. Markers with values of r^2 greater than 0.2 to rs1219648In 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 rs1219648, 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 sequnces for the marker.

Discovery SNP	Corr SNP	R^2	D'	P-value	Pos in Bld 36	SEQ ID NO:
rs1219648	rs3750817	1	0.487805	5.60E-18	123322567	220
rs1219648	rs11200014	1	0.964392	9.67E-33	123324920	221
rs1219648	rs2912780	1	0.965418	2.47E-34	123327107	222
rs1219648	rs2981579	1	0.965418	2.47E-34	123327325	223
rs1219648	rs1078806	1	0.986387	3.53E-35	123328965	224
rs1219648	rs2981578	1	0.644156	5.26E-30	123330301	225
rs1219648	rs1219648	1	1	-	123336180	237
rs1219648	rs1219643	1	0.272727	6.24E-10	123338345	226
rs1219648	rs2912774	1	1	2.49E-37	123338652	227
rs1219648	rs2936870	1	1	2.49E-37	123338892	228
rs1219648	rs17102287	1	0.446154	8.77E-16	123340161	229
rs1219648	rs2860197	1	1	7.38E-37	123341292	230
rs1219648	rs2420946	1	1	1.03E-36	123341314	231
rs1219648	rs2981582	1	1	2.49E-37	123342307	232
rs1219648	rs3135715	1	0.426087	3.82E-15	123344716	233
rs1219648	rs1047111	0.90484	0.215627	5.60E-07	123347551	234

CLAIMS

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 obtained from the individual or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237), 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.

2. The method according to Claim 1, wherein the at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14.

3. The method according to Claim 1 or Claim 2, wherein the at least one polymorphic marker is selected from rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237).

4. The method according to any of the preceding Claims, further comprising assessing the frequency of at least one haplotype in the individual.

5. The method of any 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 T in rs4415084 (SEQ ID NO:235), allele G in rs10941679 (SEQ ID NO:236) and/or allele G in rs1219648 (SEQ ID NO:237) 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 1.15.

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 1.16, including at least 1.17, at least 1.18, at least 1.19, at least 1.20, at least 1.21, at least 1.22, at least 1.23, at least 1.24, at least 1.25, at least 1.26, at least 1.27, at least 1.28 at least 1.29 and at least 1.30.

9. The method according to any of the 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 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 any one of the markers set forth in Table 12, Table 13 and Table 14 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 of the 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 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 of the preceding claims, further comprising analyzing non-genetic information to make risk assessment, diagnosis, or prognosis of the individual.

15. The method of Claim 14, wherein the non-genetic information is selected from age, gender, ethnicity, socioeconomic status, previous disease diagnosis, medical history of subject, family history of breast cancer, biochemical measurements, and clinical measurements.

16. The method of any of the Claims 11 – 15, further comprising calculating combined risk.

17. A method of determining risk of developing at least a second primary tumor in an individual previously diagnosed with breast cancer, the method 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, 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.

18. The method according to Claim 17, wherein the at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14.

19. 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 rs10941679 (SEQ ID NO:236), *rs4415084* (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and markers in linkage disequilibrium therewith, and wherein the presence of the at least one allele is indicative of a susceptibility to breast cancer.

20. The kit according to Claim 19, wherein the at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14.

21. The kit according to Claim 19 or Claim 20, wherein the reagents comprise at least one contiguous oligonucleotide that hybridizes to a fragment of the genome of the individual comprising the at least one polymorphic marker, a buffer and a detectable label.

22. The kit according to any of the Claims 19 – 21, wherein the reagents comprise at least one pair of oligonucleotides that hybridize to opposite strands of a genomic nucleic acid 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 polymorphic marker, and wherein the fragment is at least 30 base pairs in size.

23. The kit according to Claim 21 or Claim 22, wherein the at least one oligonucleotide is completely complementary to the genome of the individual.

24. The kit according to any of the Claims 19 - 23, wherein the kit comprises:

a. a detection oligonucleotide probe that is from 5-100 nucleotides in length;

b. an enhancer oligonucleotide probe that is from 5-100 nucleotides in length; and

c. 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 one of SEQ ID NO:1-237, 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 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 label when the detection probe is hybridized to the nucleic acid.

25. An apparatus for determining a genetic indicator for breast cancer in a human individual, comprising:

a computer readable memory; and

a routine stored on the computer readable memory;

wherein the routine is adapted to be executed on a processor to analyze marker and/or haplotype information for at least one human individual with respect to at least one polymorphic marker selected from rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and markers in linkage disequilibrium therewith, and generate an output based on the marker or haplotype information, wherein the output comprises an individual risk measure of the at least one marker or haplotype as a genetic indicator of breast cancer for the human individual.

26. The apparatus of Claim 25, wherein the routine further comprises a risk measure for breast cancer associated with the at least one marker allele and/or haplotype, wherein the risk measure is based on a comparison of the frequency of at least one allele of at least one polymorphic marker and/or haplotype in a plurality of individuals diagnosed with breast cancer and an indicator of the frequency of the at least one allele of at least one polymorphic marker and/or haplotype in a plurality of reference individuals, and wherein the individual risk for the human individual is based on a comparison of the carrier status of the individual for the at least one marker allele and/or haplotype and the risk measure for the at least one marker allele and/or haplotype.

27. The apparatus according to Claim 25 or 26, wherein the at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14.

28. 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 at least one of rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237);

b. determining the genotype status of a sample of individuals diagnosed with, or having a susceptibility to, breast cancer; and

c. 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.

29. The method according to Claim 28, 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.

30. The method according to Claim 28 or Claim 29, 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.

31. 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, 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.

32. The method according to Claim 31, wherein determination of the presence of allele T in rs4415084 (SEQ ID NO:235), allele G in rs10941679 (SEQ ID NO:236) and/or allele G in rs1219648 (SEQ ID NO:237) is indicative of increased susceptibility of breast cancer in the individual.

33. The method according to Claim 31 or 32, wherein 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.

34. The method according to any of the Claims 31 - 33, wherein 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 micro array technology)

35. The method according to Claim 34, wherein the process comprises allele-specific probe hybridization.

36. The method according to Claim 34, wherein the process is microarray technology.

37. The method according to any of the Claims 31 – 35, comprising:

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:1-237;

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 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.

38. A method of assessing an individual for probability of response to a breast cancer therapeutic agent, 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, 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.

39. A method of predicting prognosis of an individual diagnosed with breast cancer, the method 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 rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and 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.

40. 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 obtained from the individual, or in a genotype dataset derived from the individual, wherein the at least one polymorphic marker is selected from rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237), and markers in linkage disequilibrium therewith, wherein the presence of the at least one allele is indicative of the treatment outcome of the individual.

41. The method according to any of the Claims 38 – 40, wherein the at least one polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14.

42. 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:1-237, wherein the probe is 15-500 nucleotides in length.

43. A computer-readable medium on which is stored:

a. an identifier for at least one polymorphic marker;

b. an indicator of the frequency of at least one allele of said at least one polymorphic marker in a plurality of individuals diagnosed with breast cancer; and

c. an indicator of the frequency of the least one allele of said at least one polymorphic markers in a plurality of reference individuals;

wherein the at least one polymorphic marker is selected rs10941679 (SEQ ID NO:236), rs4415084 (SEQ ID NO:235), and rs1219648 (SEQ ID NO:237, and polymorphic markers in linkage disequilibrium therewith.

44. The medium according to Claim 43, wherein the polymorphic marker is selected from the markers set forth in Table 12, Table 13 and Table 14.

45. The medium according to claim 43 or 44, further comprising information about the ancestry of the plurality of individuals.

46. The method, kit, use, medium or apparatus according to any of the preceding Claims, wherein the presence of the at least one allele is predictive of estrogen receptor positive or progesterone receptor positive breast cancer.

47. The method, kit, use, medium or apparatus according to any of the preceding claims, wherein linkage disequilibrium between markers is characterized by particular numerical values of the linkage disequilibrium measures r^2 and/or |D'|.

48. The method, kit, use, medium or apparatus according to any of the preceding claims, wherein linkage disequilibrium between markers is characterized by values of r^2 of at least 0.2.

49. The method, kit, use, medium or appartus according to any of the preceding claims, wherein the human individual is of an ancestry that includes European ancestry.