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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 August 2009 (20.08.2009)

PCT

(10) International Publication Number
WO 2009/101639 A1

(51) **International Patent Classification**:
C12Q 1/68 (2006.01)

(21) **International Application Number**:
PCT/IS2009/000001

(22) **International Filing Date**:
16 February 2009 (16.02.2009)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
8716 14 February 2008 (14.02.2008) IS

(71) **Applicant** *(for all designated States except US)*: **DECODE GENETICS EHF.** [IS/IS]; Sturlugata 8, 101 Reykjavik (IS).

(72) **Inventors; and**
(75) **Inventors/Applicants** *(for US only)*: **RAFNAR, Thorunn** [IS/IS]; Kvistalandi 24, 108 Reykjavik (IS). **THORGEIRSSON Thorgeir** [IS/IS]; Vesturgata 5a, 101 Reykjavik (IS). **SULEM Patrick** [FR/IS]; Eskihlid 22, 107 Reykjavik (IS). **GELLER, Frank** [DE/IS]; Tjarnastigur 6, 170 Seltjarnames (IS).

(74) **Agent**: **JONSSON, Thorlakur**; deCODE GENETICS EHF., Sturlugata 8, IS-101 Reykjavik (IS).

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

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

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

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

— *with sequence listing part of description (Rule 5.2(a))*

WO 2009/101639 A1

(54) **Title**: SUSCEPTIBILITY VARIANTS FOR LUNG CANCER

(57) **Abstract**: The present invention discloses certain genetic variants as susceptibility variants for lung cancer. The invention relates to methods of risk assessment using such variants. The invention further relates to kits for use in risk assessment of lung cancer.

SUSCEPTIBILITY VARIANTS FOR LUNG CANCER

BACKGROUND OF THE INVENTION

Lung cancer causes more deaths from cancer worldwide than any other form of cancer (Goodman, G.E., *Thorax 57*:994-999 (2002)). In the United States, lung cancer is the primary cause of cancer death among both men and women. In 2007, the death rate from lung cancer was an estimated 160,390 deaths, exceeding the combined total for breast, prostate and colon cancer (America Cancer Society, www.cancer.org). Lung cancer is also the leading cause of cancer death in all European countries and is rapidly increasing in developing countries. While environmental factors, such as lifestyle factors (*e.g.*, smoking) and dietary factors, play an important role in lung cancer, genetic factors also contribute to the disease. For example, a family of enzymes responsible for carcinogen activation, degradation and subsequent DNA repair have been implicated in susceptibility to lung cancer. In addition, an increased risk to familial members outside of the nuclear family has been shown by deCODE geneticists by analysing all lung cancer cases diagnosed in Iceland over 48 years. This increased risk could not be entirely accounted for by smoking indicating that genetic variants may predispose certain individuals to lung cancer (Jonsson *et.al.*, *JAMA 292(24)*:2977-83 (2004); Amundadottir *et.al.*, *PLoS Med. 1(3)*:e65 (2004)).

The five-year survival rate among all lung cancer patients, regardless of the stage of disease at diagnosis, is only 13%. This contrasts with a five-year survival rate of 46% among cases detected while the disease is still localized. However, only 16% of lung cancers are discovered before the disease has spread. Early detection is difficult as clinical symptoms are often not observed until the disease has reached an advanced stage. Currently, diagnosis is aided by the use of chest x-rays, analysis of the type of cells contained in sputum and fiberoptic examination of the bronchial passages. Treatment regimens are determined by the type and stage of the cancer, and include surgery, radiation therapy and/or chemotherapy. In spite of considerable research into therapies for this and other cancers, lung cancer remains difficult to diagnose and treat effectively. Accordingly, there is a great need in the art for improved methods for detecting and treating such cancers.

Environmental risk factors for Lung Cancer:

Smoking of tobacco products, and in particular cigarettes, is the largest known risk factor lung cancer with a global attributable proportion estimated to be approximately 90% in men and

80% in women. Although the risk of lung cancer associated with tobacco smoking is strongly related to duration of smoking, and declines with increasing time from cessation, the estimated lifetime risk of lung cancer among former smokers remains high, ranging from approximately 6% in smokers who give up at the age of 50, to 10% for smokers who give up at age 60, compared to 15% for lifelong smokers and less than 1% in never-smokers (Peto et al. 2000 BMJ, 321, 323-32, Brennan, et al. 2006 Am J Epidemiol 164, 1233-1241). In populations where the large majority of smokers have quit smoking, such as men in the US and UK, the majority of lung cancer cases now occurs among ex-smokers (Doll et al. 1994 BMJ 309, 901-911, Zhu et al. 2001 Cancer Res, 61, 7825-7829). This emphasizes the importance of developing alternative prevention measures for lung cancer including the identification of high risk subgroups.

Genetic risk factors for Lung Cancer:

Notably, only about 15% of lifelong smokers will develop lung cancer by the age of 75, and approximately 5 to 10% of lifetime smokers will develop another tobacco-related cancer (Kjaerheim et al. 1998 Cancer Causes Control 9, 99-108). A possible explanation for this large differences in risk for individuals with similar level of tobacco exposures could be that genetic factors play a determining role in lung cancer susceptibility (Spitz et al. 2005 J Clin Oncol 23, 267-275). Identifying genes, which influence the risk of lung cancer could be important for several aspects of management of the disease.

Segregation analyses predict that the majority of genetic risk for lung cancer is most likely to be polygenic in nature, with multiple risk alleles that confer low to moderate risk and which may interact with each other and with environmental risk factors. Many studies have investigated lung cancer susceptibility based on the presence of low-penetrance, high-frequency single nucleotide polymorphisms in candidate genes belonging to specific metabolic pathways. Genetic polymorphisms of xenobiotic metabolism, DNA repair, cell-cycle control, immunity, addiction and nutritional status have been described as promising candidates but have in many cases proven difficult to confirm (Hung et al. 2005 J Natl Cancer Inst 97, 567-576, Hung et al. 2006 Cancer Res 66, 8280-8286, Landi et al. 2006 Carcinogenesis, in press, Brennan et al.2005 Lancet 366, 1558-60, Hung et al. 2007 Carcinogenesis 28,1334-40, Campa et al. 2007 Cancer Causes Control 18, 449-455, Gemignani et al. 2007 Carcinogenesis 28(6), 1287-93, Hall et al. 2007 Carcinogenesis 28, 665-671, Campa et al. 2005 Cancer Epidemiol Biomarkers Prev 14, 2457-2458, Campa et al. 2005 Cancer Epidemiol Biomarkers Prev 14, 538-539, Hashibe et al. 2006 Cancer Epidemiol Biomarkers Prev 15, 696-703).

For cancers that show a familial risk of around two-fold such as lung cancer (Jonsson et al. 2004 JAMA 292, 2977-2983, Li and Hemminki 2005 Lung Cancer 47, 301-307, Goldgar et al. 1994 J Natl Cancer Inst 86, 1600-1608), the majority of cases will arise from approximately 10%-15% of the population at greatest risk (Pharoah et al. 2002 NatGenet 31, 33-36). The identification of common genetic variants that affect the risk of lung cancer may enable the identification of individuals who are at a very high risk because of their increased genetic susceptibility or, in the case of genes related to nicotine metabolism, because of their inability to quit smoking. Such findings could potentially lead to chemoprevention programs for high risk individuals, and are especially of importance given the high residual risk that remains among ex-smokers, among whom the majority of lung cancers in the US and Europe now occur.

SUMMARY OF THE INVENTION

The present invention relates to methods for risk assessment of lung cancer. Thus, the invention relates to methods of determining a susceptibility to lung cancer in human individuals, including methods of determining an increased susceptibility to, or increased risk of developing, lung cancer, as well as methods of determining a decreased susceptibility or decreased risk of lung cancer or determining a protection against lung cancer, by evaluating certain markers and haplotypes that have been found to be associated with susceptibility of lung cancer. The method also pertains to methods of assessing response to therapeutic methods and/or therapeutic agents using the markers of the invention, as well as to methods for monitoring response to therapeutic agents and/or methods, using the markers of the invention, and to kits and apparati for use in the methods described herein.

In one aspect, the present invention pertains to a method for determining a susceptibility to lung cancer in a human individual, 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, or in a genotype dataset from the individual, wherein the at least one polymorphic marker is a marker within the C15 LD block (SEQ ID NO:1) that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein determination of the presence of the at least one allele is indicative of a susceptibility to lung cancer.

In another aspect, the invention relates to a method of determining a susceptibility to lung cancer in a human individual, the method comprising determining the presence or absence of at least one allele of at least one polymorphic marker is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein determination of the presence of the at least one allele of the at least one polymorphic marker is indicative of a susceptibility to lung cancer in the individual.

In another aspect, the invention relates to a method of determining a susceptibility to lung cancer in a human individual, comprising determining whether at least one at-risk allele in at least one polymorphic marker is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein determination of the presence of the at least one at-risk allele in at least one polymorphic marker is indicative of increased susceptibility to lung cancer in the individual.

In certain embodiments, the at least one polymorphic marker is selected from the group consisting of the markers set forth in Table 4 and Table 6. In some embodiments, the at least one polymorphic marker is selected from the group consisting of rs1051730, and markers in linkage disequilibrium therewith. In some embodiments, the at least one polymorphic marker is selected from the group consisting of rs55853698, rs55781567, rs8192482, ss107794645 and the markers set forth in Table 4. In some embodiments, the at least one polymorphic marker is selected from the group consisting the markers set forth in Table 4. In one preferred embodiment, the at least one polymorphic marker is rs1051730 (SEQ ID NO:1). In another preferred embodiment, the at least one polymorphic marker is rs16969968 (SEQ ID NO:3).

In another aspect, the invention relates to a method for determining a susceptibility to lung cancer in a human individual, 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, or in a genotype dataset from the individual, wherein the at least one polymorphic marker is associated with a gene selected from *CHRNA5*, *CHRNA3* and *CHRNB4*, and wherein determination of the presence of the at least one allele is indicative of a susceptibility to lung cancer.

Being "associated with", in this context, means that the at least one marker is in linkage disequilibrium with at least one of the *CHRNA5, CHRNA3* and *CHRNB4* genes or their regulatory regions. Such markers can be located within the gene, or its regulatory regions, or they can be in linkage disequilibrium with at least one marker within the gene or its regulatory region that has a direct impact on the function of the gene. The functional consequence of the susceptibility variants can be on the expression level of the gene, the stability of its transcript or through amino acid alterations at the protein level, as described in more detail herein.

In one embodiment, the at least one polymorphic marker is selected from the group consisting of rs1051730, rs680244, rs1948 and rs569207, and markers in linkage disequilibrium therewith. In one preferred embodiment, the at least one polymorphic marker is selected from the group consisting of marker rs1051730 (SEQ ID NO:2), and markers in linkage disequilibrium therewith. In one embodiment, the at least one polymorphic marker is selected from the markers set forth in Table 4. In another embodiment, the presence of allele

T in the polymorphic marker rs1051730 (SEQ ID NO:2) is indicative of increased susceptibility of lung cancer for the individual. Certain embodiments include a further step comprising assessing the frequency of at least one haplotype comprising at least two polymorphic markers.

The susceptibility of lung cancer to which the polymorphic markers as described herein relate to may be in the form of an increased or a decreased susceptibility of lung cancer. In some embodiments, the susceptibility is increased susceptibility of lung cancer characterized by an odds ratio (OR) of at least 1.20. In another embodiment, the susceptibility is increased susceptibility characterized by an odds ratio (OR) or relative risk (RR) of at least 1.25. In yet another embodiment, the susceptibility is increased susceptibility characterized by an odds ratio (OR) of at least 1.30. In another embodiment, the susceptibility is increased susceptibility characterized by an odds ratio (OR) of at least 1.35. In other embodiments, the characteristic odds ratio is any other non-integer value between 1.0 and 5.0.

In certain embodiments, the at least one allele or haplotype predictive of increased susceptibility of lung cancer is selected from the group consisting of rs1051730 allele T, rs680244 allele G, rs1948 allele C, rs8034191 allele C, rs2036534 allele T, rs11638372 allele T, rs4887077 allele T, rs6495314 allele C, and rs1996371 allele G.

In some embodiments, the susceptibility is decreased susceptibility of lung cancer characterized by an odds ratio (OR) of less than 1.0. In certain embodiments, the susceptibility is increased susceptibility characterized by an odds ratio (OR) or relative risk (RR)of less than 0.9. In certain other embodiments, the susceptibility is decreased susceptibility characterized by an odds ratio (OR) of less than 0.8. In another embodiment, the susceptibility is decreased susceptibility characterized by an odds ratio (OR) of less than 0.75. In other embodiments, the characteristic odds ratio is any other non-integer value between 0.1 and 1.0.

In certain embodiments, the at least one allele or haplotype predictive of decreased susceptibility of lung cancer is selected from the group consisting of rs1051730 allele C and rs55787222 allele -8.

Another aspect of the invention relates to a method of assessing susceptibility to lung cancer in a human individual, comprising screening a nucleic acid from the individual for at least one polymorphic marker allele or haplotype within SEQ ID NO:1 (C15 LD Block) that correlates with increased occurrence of lung cancer in a human population; wherein determination of the presence of an at-risk allele in the at least one polymorphism or an at-risk haplotype in the nucleic acid identifies the individual as having elevated susceptibility to lung cancer, and wherein the absence of the at least one at-risk allele or at-risk haplotype in the nucleic acid identifies the individual as not having the elevated susceptibility.

In one embodiment, the at least one polymorphic marker is selected from the markers set forth in Table 4, and markers in linkage disequilibrium therewith. In another embodiment, the at least one polymorphic marker is rs1051730 (SEQ ID NO:2). In one such embodiment, the presence of allele T in marker rs1051730 (SEQ ID NO:2) is indicative of increased susceptibility of lung cancer in the individual.

Another aspect of the invention relates to a method of determining a susceptibility to lung cancer, the method comprising: (i) obtaining sequence data about a human individual identifying at least one allele of at least one polymorphic marker, wherein different alleles of the at least one polymorphic marker are associated with different susceptibilities to lung cancer in humans; and (ii) determining a susceptibility to lung cancer from the nucleic acid sequence data, wherein the at least one polymorphic marker is a marker associated with the C15 LD block, or a marker in linkage disequilibirium therewith. In one embodiment, the method comprises obtaining sequence data about at least two polymorphic markers.

In a general sense, genetic markers lead to alternate sequences at the nucleic acid level. If the nucleic acid marker changes the codon of a polypeptide encoded by the nucleic acid, then the marker will also result in alternate sequence at the amino acid level of the encoded polypeptide (polypeptide markers). Determination of the identity of particular alleles at polymorphic markers in a nucleic acid or particular alleles at polypeptide markers comprises whether particular alleles are present at a certain position in the sequence. Sequence data identifying a particular allele at a marker comprises sufficient sequence to detect the particular allele. For single nucleotide polymorphisms (SNPs) or amino acid polymorphisms described herein, sequence data can comprise sequence at a single position, i.e. the identity of a nucleotide or amino acid at a single position within a sequence. Alternatively, the allele can be the allele of the complementary strand of DNA, such that the nucleic acid data includes the identification of at least one allele which is complementary to the allele at the opposite strand.

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

In certain embodiments, sequence data about both alleles of polymorphic markers associated with the C15 LD block are obtained, and the identity of at least one haplotype based on the sequence data is determined, and a susceptibility to the condition is determined from the haplotype data.

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

Obtaining nucleic acid sequence data can in certain embodiments comprise obtaining a biological sample from the human individual and analyzing sequence of the at least one polymorphic marker in nucleic acid in the sample. Analyzing sequence can comprise determining the presence or absence of at least one allele of the at least one polymorphic marker. Determination of the presence of a particular susceptibility allele (*e.g.*, an at-risk allele) is indicative of susceptibility to lung cancer in the human individual. Determination of the absence of a particular suscepbility allele is indicative that the particular susceptibility due to the at least one polymorphic marker is not present in the individual.

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

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

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

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

one polymorphic marker. Obtaining nucleic acid sequence data can also comprise obtaining nucleic acid sequence information from a preexisting record.

Certain embodiments of the invention relate to obtaining nucleic acid sequence data about at least two polymorphic markers associated with the C15 LD block. Other embodiments may relate to obtaining sequence data about more than two polymorphic markers, including three, four, five, six, seven, eight, nine or ten or more polymorphic markers.

The markers associated with the C15 LD block are in certain embodiments markers within the genomic segment with sequence as set forth in SEQ ID NO:1 herein. In some embodiments, the markers are markers associated with one or more of the *CHRNA3*, *CHRNA5* and *CHRNB4* genes. In some embodiments, the markers are selected from the group consisting of rs1051730, and markers in linkage disequilibrium therewith. In one embodiment, the markers are selected from the group consisting of the markers set forth in Table 4 and Table 6. In one embodiment, the marker is rs1051730. In another embodiment, the marker is rs16969968.

Obtaining sequence data may in certain embodiments relate to 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, or in a genotype dataset from the individual. Obtaining information about the absence or presence of particular marker alleles represents sequence information for the marker, identifying particular marker alleles.

In certain embodiments of the invention, a further step of assessing the frequency of at least one haplotype in the individual is performed. In such embodiments, two or more markers, including three, four, five, six, seven, eight, nine or ten or more markers can be included in the haplotype. In certain embodiments, the at least one haplotype comprises markers that are all in LD with rs1051730 and/or rs16969968.

Yet another aspect of the invention relates to a method of identification of a marker for use in assessing susceptibility to lung cancer in human individuals, the method comprising (a) identifying at least one polymorphic marker within the C15 LD block, or at least one polymorphic marker in linkage disequilibrium therewith; (b) determining the genotype status of a sample of individuals diagnosed with lung 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 lung 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 lung cancer. In one embodiment, an increase in frequency of the at least one allele in the at least one polymorphism in individuals diagnosed with lung 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 lung cancer. In another embodiment, a decrease in frequency of the at least one allele in the at least one polymorphism in individuals diagnosed with lung 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, lung cancer. In another embodiment, the significant difference in frequency is characterized by a statistical measure. Obviously, a decrease in frequency of a polymorphism in individuals diagnosed with lung cancer is indicative of the polymorphism being useful for assessing decreased susceptibility, or a protection against, lung cancer. Likewise, an increase in frequency of a polymorphism in individuals diagnosed with lung cancer is indicative of the polymorphism being useful for assessing increased susceptibility of lung cancer. In one embodiment, the at least one polymorphic marker is in linkage disequilibrium with at least one marker selected from the group consisting of rs1051730, rs680244, rs1948, rs8192475 and rs569207.

The invention also relates to a method of genotyping a nucleic acid sample obtained from a human individual, comprising determining the presence or absence of at least one allele of at least one polymorphic marker predictive of increased risk of lung cancer in the sample, wherein the at least one marker is selected from the markers set forth in Table 4, and markers in linkage disequilibrium therewith, and wherein determination of the presence or absence of the at least one allele of the at least one polymorphic marker is predictive of increased risk of lung 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, and single-stranded conformation analysis. In one embodiment, the process comprises allele-specific probe hybridization. In another embodiment, the process comprises allele-specific primer extension. In one preferred embodiment, the process 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 is capable of specifically hybridizing to a first segment of the nucleic acid whose nucleotide sequence is given by SEQ ID NO:1 that comprises at least one polymorphic site; (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. In one embodiment, the copies of the nucleic acid are provided by amplification by Polymerase Chain Reaction (PCR)

Yet another aspect of the invention relates to a method of determining a susceptibility to lung cancer in a human individual, the method comprising determining the identity 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 marker is selected from markers associated with the *CHRNA3* gene, the *CHRNA5* gene, and/or the *CHRNB4* gene, wherein the presence of the at least one allele is indicative of a susceptibility to lung cancer in the individual.

In certain embodiments, the polymorphic markers associated with susceptbility of lung cancer are indicative of a different response rate of the subject to a particular treatment modality for lung cancer. In one embodiment, the treatment modality is selected from the group consisting of surgical treatment, radiation treatment, targeted drug therapy and chemotherapy.

The invention furthermore relates, in another aspect, to a method of assessing an individual for probability of response to a therapeutic agent or method for preventing or ameliorating symptoms associated with lung cancer, 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 a marker within the C15 LD block that is associated with risk of lung cancer in humans, and wherein determination of 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 or method.

Another aspect relates to a method of predicting prognosis of an individual diagnosed with, lung cancer, the method comprising determining the presence or absence of at least one allele of at least one polymorphic is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein determination of the presence of the at least one allele is indicative of a worse prognosis of lung cancer in the individual.

A further aspect relates to a method of monitoring progress of a treatment of an individual undergoing treatment for lung 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 a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein determination of the presence of the at least one allele is indicative of the treatment outcome of the individual.

The treatment for lung cancer can in certain embodiments be selected from surgical treatment (surgical removal of tumor), radiation therapy and chemotherapy. In certain embodiments, the radiation therapy is brachytherapy. The therapeutic agent useful for chemotherapy may be any chemical agent commonly used, or in development, as a chemotherapy agent. In one embodiment, the agent targets an epidermal growth factor receptor. In certain such embodiments, the agent is gefitinib (Iressa) or erlotinib (Tarceva). In certain other embodiments, the agent is an angiogenesis inhibitor. Such inhibitors can for example be antibodies that inhibit the vascular endotheliar growth factor, such as Bevacizumab.

The methods of the invention can, in certain embodiments, further include steps of assessing at least one biomarker in the individual. Such biomarkers are biochemical molecules that are descriptive of the health status of the individual, and whose measurements are useful for aiding in, or use in, determination of a susceptibility to lung cancer. Certain other embodiments may further comprise analyzing non-genetic information to make risk assessment, diagnosis, or prognosis of the individual. The non-genetic information is in one embodiment selected from age, age at onset of the disease, gender, ethnicity, socioeconomic status, previous disease diagnosis, medical history of subject, family history of lung cancer, biochemical- and clinical measurements. Analysis of combined genetic and biomarker and/or non-genetic information can be performed using analysis methods known to the skilled person. In one embodiment, overall risk is calculated by logistic regression.

The invention also relates to a kit for assessing susceptibility to lung cancer in a human individual, the kit comprising (i) 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 a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and (ii) a collection of data comprising correlation data between the polymorphic markers assessed by the kit and susceptibility to lung cancer. In one embodiment, the at least one polymorphic marker is selected from the group consisting of rs1051730, and markers in linkage disequilibrium therewith. In one embodiment, the at least one polymorphic marker is selected from the markers set forth in Table 4. In another embodiment, the at least one polymorphic marker is marker rs1051730 (SEQ ID NO:2).

In one embodiment, the reagents comprise at least one contiguous oligonucleotide that is capable of hybridizing to a fragment of the genome of the individual comprising the at least one polymorphic marker, a buffer and a detectable label. In one embodiment, the fragment of the genome to which the oligonucleotide is capable of hybridizing has a sequence as set forth in SEQ ID NO:1. In one preferred embodiment, 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. Ideally, the at least one oligonucleotide is completely complementary to the genome of the individual. Mismatches can however be tolerated, as is well known to the skilled person and further described herein. Thus, the at least one oligonucleotide is in certain embodiments not completely complementary to the genome sequence of the individual. In such embodiments, the oligonucleotide can be about 99%, about 98%, about 95%, about 90%, about 85% or about 80% identically to the genomic sequence of the individual. In other embodiments, the oligonucleotide comprises one mismatch, two mismatches, three mismatches or four or more mismatches, compared with the genomic sequence of the individual. In certain embodiments, the oligonucleotide is about 18 to about 50 nucleotides in length. In other embodiments, the oligonucleotide is 20-30 nucleotides in length.

In one preferred embodiment, 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 is capable of specifically hybridizing to a first segment of the nucleic acid whose nucleotide sequence is given by SEQ ID NO: 1 that comprises at least one polymorphic site; 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.

Another aspect of the invention relates to the use of an oligonucleotide probe in the manufacture of a diagnostic reagent for diagnosing and/or assessing susceptibility to lung cancer in a human individual, wherein the probe is capable of hybridizing to a segment of a nucleic acid whose nucleotide sequence is given by SEQ ID NO:1 that comprises at least one

polymorphic site, wherein the fragment is 15-500 nucleotides in length. In one embodiment, the polymorphic site is selected from the polymorphic markers set forth in Table 4, and polymorphisms in linkage disequilibrium therewith. In a preferred embodiment, the at least one polymorphic site is rs1051730 (SEQ ID NO:2).

Computer-implemented aspects are also provided. In one such aspect, the invention provides a computer-readable medium having computer executable instructions for determining susceptibility to lung cancer, the computer readable medium comprising:data indicative of at least one polymorphic marker;a routine stored on the computer readable medium and adapted to be executed by a processor to determine risk of developing the at least one condition for the at least one polymorphic marker; wherein the at least one polymorphic marker is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith. In certain embodiments, the computer readable medium contains data indicative of at least two polymorphic markers. The data may be indicative of at least one polymorphic marker comprises parameters indicative of susceptibility to lung cancer for the at least one polymorphic marker, and wherein risk of developing lung cancer in an individual is based on the allelic status for the at least one polymorphic marker in the individual. In certain embodiments, the data indicative of at least one polymorphic marker comprises data indicative of the allelic status of said at least one polymorphic marker in the individual. The data may further be indicative of at least one haplotype comprising two or more polymorphic markers. The routine may also be adapted to receive input data indicative of the allelic status of said at least one polymorphic marker in said individual.

Another computer-implemented aspect provides an apparatus for determining a genetic indicator in a human individual for lung cancer, comprising:a processor; a computer readable memory having computer executable instructions adapted to be executed on the processor to analyze marker and/or haplotype information for at least one human individual with respect to at least one polymorphic marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and generate an output based on the marker or haplotype information, wherein the output comprises a risk measure of the at least one marker or haplotype as a genetic indicator of lung cancer for the human individual.

In one embodiment, the computer readable memory further comprises data indicative of the frequency of at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of individuals diagnosed with lung cancer, and data indicative of the frequency of at the least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference individuals, and wherein a risk measure is based on a comparison of the at least one marker and/or haplotype status for the human individual to the data indicative of the frequency of the at least one marker and/or haplotype information for the plurality of

individuals diagnosed with lung cancer. In another embodiment, the computer readable memory further comprises data indicative of the risk of developing lung cancer associated with at least one allele of at least one polymorphic marker or at least one haplotype, and wherein a risk measure for the human individual is based on a comparison of the at least one marker and/or haplotype status for the human individual to the risk of lung cancer associated with the at least one allele of the at least one polymorphic marker or the at least one haplotype. In yet another embodiment, the computer readable memory further comprises data indicative of the frequency of at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of individuals diagnosed with lung cancer, and data indicative of the frequency of at the least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference individuals, and wherein risk of developing lung cancer is based on a comparison of the frequency of the at least one allele or haplotype in individuals diagnosed with lung cancer and reference individuals. In certain embodiments, the risk measure is characterized by an odds ratio (OR), a risk ratio (RR) or an absolute risk (AR).

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

In other particular other embodiments of the methods, uses, apparatus or kits of the invention, the presence of at least one at-risk variant, i.e. an at-risk allele in at least one polymorphic marker or an at-risk haplotype, is indicative of lung cancer at an early age, i.e. lung cancer with an early occurrence or onset. Early onset is in some embodiments categorized as onset before age 75. In other embodiments, early onset is categorized as onset before age 70, before age 65, before age 60, before age 55, before age 50, before age 45, or before age 40. Other values for categorization of age at onset are also contemplated,

including, but not limited to, all integer values of age, and such age categories are also within scope of the invention.

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

BRIEF DESCRIPTION OF THE DRAWINGS

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

FIG 2 illustrates the risk for nicotine dependence observed for rs1051730 (1) and rs578776 (2) based on the comparison of 2161 nicotine dependent individuals and 865 low quantity smokers. The frequencies for cases and controls are given in parentheses below the alleles/haplotypes,and the arrows point towards the allele/haplotype for which increased risk is observed. (A) displays the odds ratios observed for the two SNPs and the linkage disequilibrium between them, (B) shows the odds ratios between the three observed haplotypes. There is no significant odds ratio for the haplotype with the protective C_1 allele at rs1051730 and the risk C_2 allele at rs578776 compared with the haplotype with both protective alleles. The comparison of the haplotype with the protective allele at rs1051730 and the risk allele at rs578776 against the haplotype with both high risk alleles shows a significant odds ratio due to rs1051730 allele T.

DETAILED DESCRIPTION OF THE INVENTION

A description of preferred embodiments of the invention follows.

Defintions

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 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.

Sequence conucleotide ambiguity as described herein including the accompanying sequence listing 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 "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).

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 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. Haplotypes are described herein in the context of the marker name and the allele of the marker in that haplotype, *e.g.*, "4 rs1051730" refers to the 4 allele of marker rs1051730 being in the haplotype, and is equivalent to "rs1051730 allele 4". 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), 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 lung 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 lung 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", as described herein, refers to 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" as described herein, refers to 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 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 "lung cancer therapeutic agent" refers to an agent that can be used to ameliorate or prevent symptoms associated with lung cancer.

The term "lung cancer-associated nucleic acid", as described herein, refers to a nucleic acid that has been found to be associated with lung cancer. This includes, but is not limited to, the markers and haplotypes described herein and markers and haplotypes in strong linkage disequilibrium (LD) therewith. In one embodiment, a lung cancer-associated nucleic acid refers to an LD-block found to be associated with lung cancer through at least one polymorphic marker located within the LD block.

The term "antisense agent" or "antisense oligonucleotide" refers, as described herein, to molecules, or compositions comprising molecules, which include a sequence of purine an pyrimidine heterocyclic bases, supported by a backbone, which are effective to hydrogen bond to a corrresponding contiguous bases in a target nucleic acid sequence. The backbone is composed of subunit backbone moieties supporting the purine an pyrimidine hetercyclic bases at positions which allow such hydrogen bonding. These backbone moieties are cyclic moieties of 5 to 7 atoms in size, linked together by phosphorous-containing linkage units of one to three atoms in length. In certain preferred embodiments, the antisense agent comprises an oligonucleotide molecule.

"Lung cancer", in the present context, refers to clinically diagnosed lung cancer. The term encompasses all subclassification of lung cancer, including non-small cell lung carcinoma (CSCLC), squamous cell carcinoma, adenocarcinoma, bronchoalveolar carcinoma, large cell carcinoma, small cell lung carcinoma (SCLC), and combined patterns or subphenotypes of lung cancer.

The "C15 LD block", as defined herein, refers to the genomic region flanked by the SNP markers rs4436747 and rs1383636. This genomic region corresponds to a region of the genome with extensive linkage disequilibrium (LD), as described herein, and within which variants in linkage disequilibrium with rs1051730, also called surrogate variants, can be found (*e.g.*, as set forth in Table 4). The region is located between position 76,501,063 and 76,893,275 in NCBI Build 36, and has the sequence as set forth in SEQ ID NO:1.

Association of genetic variants to lung cancer

A genome-wide association study of SNP markers on a chip containing approximately 317,000 such SNPs has resulted in identification of significant association to markers on Chromosome

15, within the nicotinic acetylcholine receptor gene cluster. As shown in Table 2, marker rs1051730 has been found to associate with increased risk of developing lung. The marker, and markers in linkage disequilibrium therewith, are thus useful for detecting an increased risk, or increased susceptibility, of lung cancer. Any one of these markers, alone or in combination, are useful in the methods, kits, apparatus, uses and media described herein. Exemplary markers in LD with rs1051730 are shown in Table 3 herein, and additional markers useful for practicing the invention as described herein are listed in the tables presented herein, including Tables 4 and 6.

Further variants have been identified through sequencing of the CHRNA5, CHRNA3 and CHRNB4 genes, as described further herein. These additional variants, including rs16969968, can also be useful for diagnostic applications for lung cancer, as described herein.

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 nucleotides. 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 Nucleotide Polymorphisms ("SNPs"). These SNPs are believed to have occurred in a single mutational event, and therefore there are usually two possible alleles possible at each SNPsite; the original allele and the mutated 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 microsatellites, insertions, deletions, inversions and copy number variations. 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. These sequence variants can all 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 lung 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 realise 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. Additional 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 methodologies available for SNP genotyping include, but are not limited to, TaqMan genotyping assays and SNPlex platforms (Applied Biosystems), mass spectrometry (*e.g.*, MassARRAY system from Sequenom), minisequencing methods, real-time PCR, Bio-Plex system (BioRad), CEQ and SNPstream systems (Beckman), Molecular Inversion Probe array technology (*e.g.*, Affymetrix GeneChip), and BeadArray Technologies (*e.g.*, Illumina GoldenGate and Infinium assays). By these or other methods available to the person skilled in the art, one or more alleles at polymorphic markers, including microsatellites, SNPs or other types of polymorphic markers, can be identified.

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

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

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

population allele frequency. The probability of the genotypes of the case's relatives can then be computed by:

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

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

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

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

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

In the present context, an individual who is at an increased susceptibility (i.e., increased risk) for lung cancer, is an individual in whom at least one specific allele at one or more polymorphic marker or haplotype conferring increased susceptibility for lung 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 lung 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 1.2, including but not limited to: at least 1.2, at least 1.3, at least 1.4, at least 1.5, at least 1.6, at least 1.7, at least 1.8, at least 1.9, and at least 2.0. In

a particular embodiment, a risk (relative risk and/or odds ratio) of at least 1.2 is significant. In another particular embodiment, a risk of at least 1.3 is significant. In yet another embodiment, a risk of at least 1.4 is significant. In a further embodiment, a relative risk of at least about 1.5 is significant. In another further embodiment, a significant increase in risk is at least about 1.7 is significant. However, other numerical values bridging the above for defining risk measures are also contemplated, *e.g.* at least 1.15, 1.25, 1.35, and so on, and such cutoffs are also within scope of the present invention. In other embodiments, a significant increase in risk is at least about 20%, including but not limited to about 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 100%, 150%, 200%, 300%, and 500%. In one particular embodiment, a significant increase in risk is at least 20%. In other embodiments, a significant increase in risk is at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90% and at least 100%. Other cutoffs or ranges as deemed suitable by the person skilled in the art to characterize the invention are however also contemplated, and those are also within scope of the present invention. In certain embodiments, a significant increase in risk is characterized by a p-value, such as a p-value of less than 0.05, less than 0.01, less than 0.001, less than 0.0001, less than 0.00001, less than 0.000001, less than 0.0000001, less than 0.00000001, or less than 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 lung cancer (affected), compared to the frequency of its presence in a comparison group (control), and wherein the presence of the marker or haplotype is indicative of susceptibility to lung 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. 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 risk factors for lung cancer (*e.g.,* non-smokers). 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 smoking history, 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.

In other embodiments of the invention, an individual who is at a decreased susceptibility (i.e., at a decreased risk) for lung cancer is an individual in whom at least one specific allele at one or more polymorphic marker or haplotype conferring decreased susceptibility for lung cancer 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 0.95, including but not limited to less than 0.9, less than 0.8, 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 0.8. In another embodiment, significant decreased risk is less than 0.7. In yet another embodiment, significant decreased risk is less than 0.6. In another embodiment, the decrease in risk (or susceptibility) is at least 20%, including but not limited to 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 75%, 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 20%. In another embodiment, a significant decrease in risk is at least about 30%. In another embodiment, the decrease in risk is at least about 40%. 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 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 lung cancer, compared with controls, the other allele of the marker will be found in decreased frequency in the group of individuals with lung cancer, compared with controls. In such a case, one allele of the marker (the one found in increased frequency in individuals with lung cancer) 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.* lung 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., *el 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 lung cancer may be assessed. Such variants may be all genetic, or they may represent a combination of genetic and non-genetic risk variants.

Linkage Disequilibrium

The natural phenomenon of recombination, which occurs on average once for each chromosomal pair during each meiotic event, represents one way in which nature provides variations in sequence (and biological function by consequence). It has been discovered that recombination does not occur randomly in the genome; rather, there are large variations in the frequency of recombination rates, resulting in small regions of high recombination

frequency (also called recombination hotspots) and larger regions of low recombination frequency, which are commonly referred to as Linkage Disequilibrium (LD) blocks (Myers, S. *et al., Biochem Soc Trans* 34:526-530 (2006); Jeffreys, A.J., *et al.,Nature Genet* 29:217-222 (2001); May, C.A., *et al., Nature Genet* 31:272-275(2002)).

Linkage Disequilibrium (LD) refers to a non-random assortment of two genetic elements. For example, if a particular genetic element (*e.g.*, an allele of a polymorphic marker, or a haplotype) occurs in a population at a frequency of 0.50 (50%) and another element occurs at a frequency of 0.50 (50%), then the predicted occurrence of a person's having both elements is 0.25 (25%), assuming a random distribution of the elements. However, if it is discovered that the two elements occur together at a frequency higher than 0.25, 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). Most capture the strength of association between pairs of biallelic sites. Two important pairwise measures of LD are r^2 (sometimes denoted Δ^2) and |D'|. 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 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 indicative of markers being in linkage disequilibrium 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 is 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 r^2 coefficient or |D'| (r^2 up to 1.0 and |D'| up to 1.0). In certain embodiments, linkage disequilibrium is defined in terms of values for both the r^2 and |D'| measures. In one such embodiment, a significant linkage disequilibrium is defined as $r^2 > 0.1$ and |D'| >0.8. In another embodiment, a significant linkage disequilibrium is defined as $r^2 > 0.2$ and |D'| >0.9. Other combinations and permutations of values of r^2 and |D'|for determining linkage disequilibrium are also possible, and within the scope of the invention. Linkage disequilibrium can be determined in a single human population, as defined herein, or it can be determined in a collection of samples comprising individuals from more than one human population. In one embodiment of the invention, LD is determined in a sample from one or more of the HapMap populations (caucasian, african, 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 yet another embodiment, LD is determined in samples from the Icelandic population.

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

Genomic LD maps have been generated across the genome, and such LD maps have been proposed to serve as framework for mapping disease-genes (Risch, N. & Merkiangas, 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.

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

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, and as such are useful for use in the methods and kits of the present invention. One or more causative (functional) variants or mutations may reside within the region found to be associating to the disease or trait. Such variants 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. One 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 linkage region, for example, association of all possible combinations of genotyped markers is studied, provided those markers span a practical region. 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 a 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 for the linkage analysis, 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 described in Risch, N. & Teng, J. (*Genome Res.*, 8:1273-1288 (1998)), DNA pooling (*ibid*) for sibships so that it can be applied to general familial relationships, and present both adjusted and unadjusted p-values for comparison. 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, $\text{risk}(h_i)/\text{risk}(h_j) = (f_i/p_i)/(f_j/p_j)$, 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 is usually quite small, and hence the less stringent the statistical measure that needs to be applied. For example, for a genome-wide search for susceptibility variants for a particular disease or trait using 300,000 SNPs, a correction for the 300,000 tests performed (one for each SNP) can be performed. Since many SNPs on the arrays typically used are correlated (*i.e.,* in LD), they are not independent. Thus, the correction is conservative. Nevertheless, applying this correction factor requires an observed P-value of less than $0.05/300{,}000 = 1.7 \times 10^{-7}$ for the signal to be considered significant applying this conservative test on results from a single study cohort. Obviously, signals found in a genome-wide association study with P-values less than this conservative threshold are a measure of a true genetic effect, and replication in additional cohorts is not necessarily from a statistical point of view. Importantly, however, signals with P-values that are greater than this threshold may also be due to a true genetic effect. Thus, 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.

Risk Calculations

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

Deriving risk from odds-ratios

Most gene discovery studies for complex diseases that have been published to date in authoritative journals have employed a case-control design because of their retrospective setup. These studies sample and genotype a selected set of cases (people who have the specified disease condition) and control individuals. The interest is in genetic variants (alleles) which frequency in cases and controls differ significantly.

The results are typically reported in odds-ratios, that is the ratio between the fraction (probability) with the risk variant (carriers) versus the non-risk variant (non-carriers) in the groups of affected versus the controls, i.e. expressed in terms of probabilities conditional on the affection status:

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

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

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

exactly, but the risk estimates are usually robust to moderate deviations from this assumption.

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

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

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

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

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

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

The risk relative to the average population risk

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

$RR(aa) = Pr(A|aa)/Pr(A) = (Pr(A|aa)/Pr(A|bb))/(Pr(A)/Pr(A|bb)) = r^2/(Pr(aa)\, r^2 + Pr(ab)\, r + Pr(bb)) = r^2/(p^2 r^2 + 2pq\, r + q^2) = r^2/R$

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

As an example, in type-2 diabetes risk, allele T of the disease associated marker rs7903146 in the TCF7L2 gene on chromosome 10 has an allelic OR of 1.37 and a frequency (p) around

0.28 in non-Hispanic white populations. The genotype relative risk compared to genotype CC are estimated based on the multiplicative model.

For TT it is 1.37×1.37 = 1.88; for CT it is simply the OR 1.37, and for CC it is 1.0 by definition.

The frequency of allele C is q = 1 - p = 1 - 0.28 = 0.72. Population frequency of each of the three possible genotypes at this marker is:

$Pr(TT) = p^2 = 0.08$, $Pr(CT) = 2pq = 0.40$, and $Pr(CC) = q^2 = 0.52$

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

$R = 0.08 \times 1.88 + 0.40 \times 1.37 + 0.52 \times 1 = 1.22$

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

RR(TT) = 1.88/1.22 = 1.54, RR(CT) = 1.37/1.22 = 1.12, RR(CC) = 1/1.22 = 0.82.

Combining the risk from multiple markers

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

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

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

Pr(A|g1,g2) = Pr(A|g1)Pr(A|g2)/Pr(A) and Pr(g1,g2) = Pr(g1)Pr(g2)

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

As is in most situations where a statistical model is utilized, the model applied is not expected to be exactly true since it is not based on an underlying bio-physical model. However, the multiplicative model has so far been found to fit the data adequately, i.e. no significant deviations are detected for many common diseases for which many risk variants have been discovered.

As an example, an individual who has the following genotypes at 4 markers associated with risk of type-2 diabetes along with the risk relative to the population at each marker:

Chromo 3 PPARG CC Calculated risk: RR(CC) = 1.03

Chromo 6 CDKAL1 GG Calculated risk: RR(GG) = 1.30

Chromo 9 CDKN2A AG Calculated risk: RR(AG) = 0.88

Chromo 11 TCF7L2 TT Calculated risk: RR(TT) = 1.54

Combined, the overall risk relative to the population for this individual is:
$1.03 \times 1.30 \times 0.88 \times 1.54 = 1.81$

Adjusted life-time risk

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

For example, for a phenotype, if the overall genetic risk relative to the population is 1.8 for a white male, and if the average life-time risk of the phenotype for individuals of his demographic is 20%, then the adjusted lifetime risk for him is 20% x 1.8 = 36%.

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

Risk assessment for lung cancer

As described herein, certain polymorphic markers and haplotypes comprising such markers are found to be useful for risk assessment of lung cancer. Risk assessment can involve the use of the markers for diagnosing a susceptibility to lung cancer. Particular alleles of polymorphic markers are found more frequently in individuals with lung cancer, than in individuals without diagnosis of lung cancer. Therefore, these marker alleles have predictive value for detecting lung cancer, or a susceptibility to lung cancer, in an individual. Tagging markers in linkage disequilibrium with at-risk variants (or protective variants) described herein can be used as surrogates for these markers (and/or haplotypes). Such surrogate markers can be located within a particular haplotype block or LD block. Such surrogate markers can also sometimes be located outside the physical boundaries of such a haplotype block or LD block, either in close vicinity of the LD block/haplotype block, but possibly also located in a more distant genomic location.

Long-distance LD can for example arise if particular genomic regions (*e.g.*, genes) are in a functional relationship. For example, if two genes encode proteins that play a role in a shared metabolic pathway, then particular variants in one gene may have a direct impact on observed variants for the other gene. Without intending to be bound by theory, let us consider the case where a variant in one gene leads to increased expression of the gene product. To counteract this effect and preserve overall flux of the particular pathway, this variant may have led to selection of one (or more) variants at a second gene that conferes decreased expression levels of that gene. These two genes may be located in different genomic locations, possibly even on different chromosomes, but variants within the genes are in apparent LD, not because of their shared physical location within a region of high LD, but rather due to evolutionary forces. Such LD is also contemplated and within scope of the present invention. The skilled person will appreciate that many other scenarios of functional gene-gene interaction are possible, and the particular example discussed here represents only one such possible scenario.

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 as or possibly even higher than the originally identified 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 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 detected at-risk variants, also have predictive value for detecting association to lung cancer, or a susceptibility to lung 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 lung cancer.

The presence of certain alleles at certain polymorphic markers (*e.g.*, allele T in marker rs1051730) is indicative of increased risk of developing lung cancer. In general, homozygous carriers of an at-risk allele (*e.g.*, individuals who carry two copies of the T allele of marker rs1051730) are of particularly high risk or susceptibility of developing lung cancer. Thus, in certain embodiments of the invention, the presence of two copies of an at-risk allele is indicative of increased susceptibility or risk of lung cancer. In other embodiments, heterozygous individuals carrying one copy of the at-risk allele are at increased risk or susceptibility of lung 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 lung 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 lung cancer. Detecting particular alleles of polymorphic markers can in certain embodiments be done by obtaining nucleic acid sequence data about a particular human individual, that identifies at least one allele of at least one polymorphic marker. Different alleles of the at least one marker are associated with different susceptibility to the disease in humans. Obtaining nucleic acid sequence data can comprise nucleic acid sequence at a single nucleotide position, which is sufficient to identify alleles at SNPs. The nucleic acid sequence data can also comprise sequence at any other number of nucleotide positions, in particular for genetic markers that comprise multiple nuclotide positions, and can be anywhere from two to hundreds of thousands, possibly even millions, of nucleotides (in particular, in the case of copy number variations (CNVs)).

In certain embodiments, the invention can be practiced utilizing a dataset comprising information about the genotype status of at least one polymorphic marker associated with lung cancer (or markers in linkage disequilibrium with at least one marker associated with lung 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 the disease. A positive result for a variant (*e.g.*, marker allele) associated with lung cancer, is indicative of the individual from which the dataset is derived is at increased susceptibility (increased risk) of lung cancer.

In certain embodiments of the invention, a polymorphic marker is correlated to lung 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 lung 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 disease (*e.g.*, lung cancer), a risk for the disease, or a susceptibility to the disease, 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 disclosed to be predictive of susceptibility to lung cancer, as disclosed herein, *e.g.*, the markers presented in Table 4, may be useful for risk assessment and diagnostic purposes for, either alone or in combination. Even in cases where the increase in risk by individual risk factors is relatively modest, *e.g.* on the order of 10-30%, the association may have significant implications. Thus, relatively common genetic 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 lung cancer.

Thus, in one embodiment of the invention, a plurality of variants (genetic markers, biomarkers and/or haplotypes) is used for overall risk assessment of lung cancer. 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 lung cancer. 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.

A variety of biomarkers have been associated with lung cancer, and can all be useful in combination with the genetic variants disclosed herein. These include, but are not limited to, protein markers, non-protein small molecule markers, mRNA markers and DNA markers. Protein markers include, but are not limited to, epidermal growth factor recpetors (EGFR), cytokeratins, MUC1, LUNX, KS1/4, telomerase, cell cycle proteins, such as cyclins (*e.g.*, cyclin D1, cyclin A, cyclin B1) and cyclin-dependent kinases, G1-S transition proteins, such as Rb (retinoblastoma susceptibility protein), apoptosis proteins, death receptors, caspases, death-associated protein, Bcl-2 family, p-53, angiogenesis growth factors (*e.g.*, VEGF, VEGFR-1, VEGFR-2, PDGF, bFGF, IL-8, collagen XVIII), inhibitors of angiogenesis, and markers of angiogenetic activity. These biomarkers may be used alone or in combination for risk assessment of lung cancer, in combination with at least one genetic variant as described herein. The skilled person will appreciate that expression of these protein factors can be made at the protein level, such as by monoclonal or multiclonal antibodies, or by other methods known to the skilled person. Alternatively, determination of mRNA levels of the corresponding mRNA precursor can be used as a measure of expression levels.

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 (*e.g.*, lung cancer) 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 4), but may also include other markers that are in strong LD (*e.g.*, 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 4.

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 patients with lung cancer. These markers and haplotypes in LD and/or comprising such markers, are thus protective for lung cancer, i.e. they confer a decreased risk or susceptibility of individuals carrying these markers and/or haplotypes for developing lung 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 lung cancer, is one in which the marker allele or haplotype is more frequently present in an individual who is at risk for lung 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 lung cancer or a susceptibility to lung cancer. In other embodiments, at-risk markers in linkage disequilibrium with one or more markers found to be associated with lung cancer (*e.g.*, markers as listed in Tables 4 and 6) are tagging markers that are more frequently present in an individual at risk for lung cancer (affected), compared to the frequency of their presence in a non-affected or healthy individual (control), wherein the presence of the tagging markers is indicative of increased susceptibility to lung cancer . In a further embodiment, at-risk markers alleles (i.e. conferring increased susceptibility) in linkage disequilibrium with one or more markers shown herein to be associated with lung cancer, are markers comprising one or more allele that is more frequently present in an individual at risk for lung cancer, compared to the frequency of their presence in a non-affected or healthy individual (control), wherein the presence of the markers is indicative of increased susceptibility to lung cancer.

Study population

In a general sense, the methods, uses and kits of the invention as described herein can be utilized from samples containing genomic DNA from any source. In preferred embodiments, the individual from whom the sample is derived 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 human individuals who are members of a target population. Such a target population is in one embodiment a population or group of individuals at particular 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 lung cancer or related diseases, smoking history, family history of disease).

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 lung cancer in any age range described in the above. It is also contemplated that a range of ages may be relevant in certain embodiments, such as age at onset at more than age 45 but less than age 60. Other age ranges are however also contemplated, including all age ranges bracketed by the age values listed in the above. The invention furthermore relates to individuals of either gender, males or females.

The Icelandic population is a Caucasian population of Northern European ancestry. A large number of studies reporting results of genetic linkage and association in the Icelandic population have been published in the last few years. Many of those studies show replication of variants, originally identified in the Icelandic population as being associating with a particular disease, in other populations (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.

It is thus believed that the markers of the present invention described herein to be associated with lung 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, Portuguese, Italian, Polish, Bulgarian, Slavic, Serbian, Bosnian, Chech, 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, Adygel, 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. In another embodiment, the invention relates to individuals of Caucasian origin.

Ancestry is in certain embodiment based on self-reported ancestry. The ancestry or racial contribution in individual subjects may also be determined by genetic analysis. Genetic analysis of ancestry may for example 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, have different population frequencies in different populations, or are polymorphic in one population but not in another. The person skilled in the art will however apply the methods available and as thought herein to practice the present invention in any given human population. This may include assessment of polymorphic markers in the LD region of the present invention, so as to identify those markers that give strongest association within the specific population. Thus, the at-risk variants of the present invention may reside on different haplotype background and in different frequencies in various human populations. However, utilizing methods known in the art and the markers of the present invention, the invention can be practiced in any given human population.

Utility of Genetic Testing

The person skilled in the art will appreciate and understand that the variants described herein in general do not, by themselves, provide an absolute identification of individuals who will develop lung cancer. The variants described herein do however indicate increased and/or decreased likelihood that individuals carrying the at-risk variants disclosed herein will develop lung cancer. This information is however extremely valuable in itself, as outlined in more detail in the following, as it can be used, for example, to initiate preventive measures at an early stage, perform regular physical exams to monitor the development, progress and/or appearance of symptoms of lung cancer, or to schedule exams at a regular interval to identify lung cancer in its early stages, so as to be able to apply treatment at an early stage which is often critical for successful lung cancer therapy.

The knowledge about a genetic variant that confers a risk of developing lung cancer offers the opportunity to apply a genetic test to distinguish between individuals with increased risk of developing lung cancer (i.e. carriers of the at-risk variants disclosed herein) and those with decreased risk of developing lung cancer (i.e. carriers of protective variants, and/or non-carriers of at-risk variants). The core value of genetic testing is the possibility of being able to diagnose disease, or a predisposition to disease, at an early stage and provide information to the clinician about prognosis/aggressiveness of the disease in order to be able to apply the most appropriate treatment.

Individuals with a family history of lung cancer and carriers of at-risk variants may benefit from genetic testing since the knowledge of the presence of a genetic risk factor, or evidence for increased risk of being a carrier of one or more risk factors, may provide incentive for implementing a healthier lifestyle, by avoiding or minimizing known environmental risk factors for lung cancer. For example, an individual who is a current smoker and is identified as a carrier of one or more of the variants shown herein to be associated with increased risk of lung cancer , may, due to his/her increased risk of developing the disease, choose to quit smoking.

Integration of Genetic Risk Models into Clinical Management of Lung Cancer:

Management of lung cancer currently relies on a combination of primary prevention (most importantly abstinence from smoking), early diagnosis and appropriate treatments. There are clear clinical imperatives for integrating genetic testing into several 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.

Primary prevention

Primary prevention options currently focus on avoiding exposure to tobacco smoke or other environmental toxins that have been associated with the development of lung cancer.

Early Diagnosis

Patients who are identified as being at high risk for lung cancer may be referred to have chest X-rays or sputum cytology examination. In addition, a spiral CT scan is a newly-developed procedure for lung cancer screening. Numerous lung cancer screening trials are currently taking place but presently, the U.S. Preventive Services Task Force (USPSTF) concludes that evidence is insufficient to recommend for or against screening asymptomatic persons for lung cancer with either low dose computerized tomography (LDCT), chest x-ray, sputum cytology, or a combination of these tests.

Many of the screening protocols being evaluated involve some form of radiation or and invasive procedure such as bronchoscopy. These protocols carry certain risks and may prove hard to implement due to the considerable costs involved. In light of the fact that only about 15% of lifetime smokers develop lung cancer, it is clear that the great majority of individuals at risk would be needlessly subjected to repeated screening tests with the associated costs and negative side-effects. The identification of genetic biomarkers that affect the the risk of developing lung cancer could be used to help identify individuals should be offered extreme help in risk reduction programs such as smoking termination. In the case of failure to stop smoking, or in the case of ex-smokers, such genetic biomarkers could help in defining the subpopulation of individuals that would benefit the most from screening.

Less than 10% of lung cancer cases arise in individuals that have never smoked. Genetic biomarkers that predict the risk of lung cancer would be particularly useful in this group. The genetic component of this form of the diease is likely to be even stronger than in tobacco-related lung cancer. If genetic variants that affect the risk of non-smoking lung cancer were known, it might be possible to identify individuals at high risk for this disease and subject them to regular screening tests.

Variants over the CHRNA3/CHRNA5/CHRNB4 gene cluster have previously been reported as potentially associated with risk of nicotine dependence (Saccone, *et al., Hum Mol Genet* **16:**36-49 (2007)). However, the evidence for association reported was weak, and a large number of other genomic locations were also reported as potentially associated with nicotine

dependence. The present inventors have confirmed the suggested association of variants in the region to smoking phenotypes. The present inventors have also surprisingly found that the rs1051730 marker, and markers in linkage disequilibrium therewith, show strong association to lung cancer. While smoking is a known risk factor for this disease, the effect of the rs1051730 variant on lung cancer cannot be explained by the commonly used phenotypes for nicotine dependence (ND), such as smoking quantity (SQ) (which is correlated with the Fagerström score and nicotine dependence according to the DSM-IV criteria). This will be further described in the following.

The rs1051730 marker is associated with SQ as shown in Table 1. Of the 13,945 smokers studied by the inventors, 501 are known to have developed lung cancer. The SQ levels 1, 2 and 3 have a calculated relative risk of 2.1, 2.4 and 2.9 for lung cancer, respectively, compared with SQ level 0 (1-10 cigarettes/day). If it is assumed that only smokers developed lung cancer, the frequency of the rs1051730 allele T variant can be calculated as a weighted average, using these relative risk estimates. Then, the predicted frequency of the variant in lung cancer is [(0.305x0.260) + (0.350x0.459x2.1) + (0.380x0.214x2.4)+(0.391x0.067x2.9)] divided by [0.260 + (0.459x2.1) + (0.214x2.4) + (0.067x2.9)], or 35.6% (see Table 2). It should be noted that this is an overestimate, since non-smokers are given a weight of zero in this calculation. Still, compared to the population frequency of 34.4% for the variant, the odds ratio for lung cancer based on this calculation is only 1.05, which is much smaller than the observed value of 1.31 (Table 2). It should be noted that even if the relative risks for SQ levels 2 and 3 were doubled, the calculated frequency and the corresponding OR value for lung cancer would only increase to 36.3% and 1.09, respectively. In other words, the SQ measure only explains a small proportion of the increased risk for lung cancer that is observed for rs1051730 allele T. The same conclusion will be reached using nicotine dependence phenotypes such as Fagerström score and DSM-IV criteria, since the frequency of the variant for these phenotypes is comparable to SQ.

These surprising observations show that the risk conferred by rs1051730 for lung cancer cannot be explained by their effect on the smoking quantity phenotype. Thus, there is an unexpected and surprising additional risk for lung cancer conferred by rs1051730 and correlated variants.

METHODS

Methods for risk assessment of lung cancer are described herein and are encompassed by the invention. The invention also encompasses methods of assessing an individual for probability of response to a therapeutic agent for lung cancer, as well as methods for predicting the effectiveness of a therapeutic agent for lung cancer. Kits for assaying a sample from a

subject to detect susceptibility to lung cancer are also encompassed by the invention.

Diagnostic and screening methods

In certain embodiments, the present invention pertains to methods of determining a susceptibility to lung cancer, by detecting particular alleles at genetic markers that appear more frequently in lung cancer subjects or subjects who are susceptible to lung cancer. In a particular embodiment, the invention is a method of diagnosing a susceptibility to lung cancer by detecting at least one allele of at least one polymorphic marker (*e.g.*, the markers described herein). The present invention describes methods whereby detection of particular alleles of particular markers or haplotypes is indicative of a susceptibility to lung cancer. Such prognostic or predictive assays can also be used to determine prophylactic treatment of a subject prior to the onset of symptoms or prior to diagnosis of lung 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 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 simultaneously assessed for up to one million SNPs. The resulting genotype information, made available to the customer can be compared to information from the public literature about disease or trait risk associated with various SNPs. Methods for generating complete sequence information about the genomic sequence of individuals, which can be used for establishing genotype information (sequence identity at polymorphic sites), are also being developed. The diagnostic application of disease-associated alleles as described herein, can thus be performed either by the individual, through analysis of his/her genotype data, or by a health professional based on results of a clinical test. In other words, the diagnosis or assessment of a susceptibility based on genetic risk can be made by health professionals, genetic counselors or by the layman, based on information about his/her genotype and publications on various risk factors. In the present context, the term "diagnosing", "diagnose susceptibility", and "determine susceptibility", is meant to refer to any available method for such determination, 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. Genotype and/or sequence results 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 lung 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 determining a decreased susceptibility to lung cancer, by detecting particular genetic marker alleles or haplotypes that appear less frequently in individuals diagnosed with lung cancer than in individual not diagnosed with lung cancer or in the general population. Such variants confer a decreased risk of, or protection against, lung cancer. Exemplary variants include the alternate allele of the SNP markers shown herein to be associated with increased risk of lung cancer. In one embodiment, the protective variant for lung cancer is selected from the group consisting of rs1051730 allele C, or marker alleles in linkage disequilibrium therewith. In another embodiment, the protective variant for lung cancer is rs55787222 allele -8 (containing 2 copies of the microsatellite repeat).

As described and exemplified herein, particular marker alleles or haplotypes are associated with risk of lung cancer. In one embodiment, the marker allele or haplotype is one that confers a significant risk or susceptibility to lung cancer. In another embodiment, the invention relates to a method of determining a susceptibility to lung 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 Table 4 and Table 6, and markers in linkage disequilibrium (*e.g.*, defined by numerical values for $r^2 > 0.2$) therewith. In another embodiment, the invention pertains to methods of determining a susceptibility to lung cancer in human individual, by screening for at least one marker allele or haplotype as listed in Table 4, or markers in linkage disequilibrium therewith. In another embodiment, the invention pertains to methods of determining a suscepbility to lung cancer by identifying particular alleles at polymorphic markers associated with at least one of the *CHRNA3*, *CHRNA5* and *CHRNB4* genes. In one embodiment, the marker allele is more frequently present in a subject having, or who is susceptible to, lung 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 lung cancer. These diagnostic methods involve detecting the presence or absence of at least one marker allele or haplotype that is associated with lung cancer. Haplotypes include combinations of alleles at various genetic markers (*e.g.*, SNPs, microsatellites). 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 nucleic acid associated with lung cancer (*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 lung 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 numerical value of r^2 greater than 0.2 and/or |D'| > 0.8).

In one embodiment, determination of a susceptibility to lung cancer can be accomplished using hybridization methods, such as Southern analysis, Northern analysis, and/or *in situ* hybridizations (see Current Protocols in Molecular Biology, Ausubel, F. *et al.*, eds., John Wiley & Sons, including all supplements). 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.

To determine a susceptibility to lung cancer, a hybridization sample is formed by contacting the test sample containing a nucleic acid associated with lung cancer, 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 the nucleotide sequence of the C15 LD Block (SEQ ID NO:1), as described herein, optionally comprising at least one allele of at least one marker 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 C15 LD Block (SEQ ID NO:1), 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 allelic combination (*i.e.*, a haplotype) and therefore is susceptible to lung 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, or is capable of hybridizing, 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.

In another hybridization method, Northern analysis (see Current Protocols in Molecular Biology, Ausubel, F. *et al.*, eds., John Wiley & Sons, *supra*) is used to identify the presence of a polymorphism associated with lung cancer. For Northern analysis, a test sample of RNA is obtained from the subject by appropriate means. As described herein, specific hybridization of a nucleic acid probe to RNA from the subject is indicative of a particular allele complementary to the probe. For representative examples of use of nucleic acid probes, see, for example, U.S. Patent Nos. 5,288,611 and 4,851,330.

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 lung cancer. Hybridization of the PNA probe is thus diagnostic for lung cancer or a susceptibility to lung 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 lung 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 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 lung cancer. Further, the expression of the variant(s) can be quantified as physically or functionally different.

In another embodiment of the methods 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 lung cancer. 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 lung 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. 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 lung 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 can be prepared using standard methods (see, *e.g.*, Current Protocols in Molecular Biology, *supra*). PCR can be used to amplify the desired region. Standard techniques can be used to detect hybridization of the allele-specific oligonucleotide to the nucleic acid sample. Specific hybridization of an allele-specific oligonucleotide probe to DNA from the subject is indicative of a specific allele at a polymorphic site (see, *e.g.*, Gibbs, R. *et al.*, *Nucleic Acids Res.*, *17*:2437-2448 (1989) and WO 93/22456).

In another embodiment, arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from a subject, can be used to identify particular alleles at polymorphic sites. For example, an oligonucleotide array can be used. Oligonucleotide arrays typically comprise a plurality of different oligonucleotide probes that are coupled to a surface

of a substrate in different known locations. These arrays can generally be produced using mechanical synthesis methods or light directed synthesis methods that incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis methods, or by other methods known to the person skilled in the art (see, *e.g.*, Bier, F.F., *et al. Adv Biochem Eng Biotechnol* 109:433-53 (2008); Hoheisel, J.D., *Nat Rev Genet* 7:200-10 (2006); Fan, J.B., *et al. Methods Enzymol* 410:57-73 (2006); 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 lung 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 lung cancer or determination of a susceptibility to lung cancer, can be made by examining expression and/or composition of a polypeptide encoded by a nucleic acid associated with lung 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 lung 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 lung 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. The haplotypes and markers of the present invention that show association to lung cancer may play a role through their effect on one or more of these nearby genes. In one embodiment, the gene is selected from the group consisting of *CHRNA3, CHRNA5* and *CHRNB4*. 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 lung cancer affect the expression of a nearby gene, such as one or more of the *CHRNA3*, *CHRNA5* and *CHRNB4* genes. It is well known that regulatory element affecting gene expression may be located far away, even as far as tenths or hundreds of kilobases away, from the promoter region of a gene. By assaying for the presence or absence of at least one allele of at least one polymorphic marker of the present invention, it is thus possible to assess the expression level of such nearby genes. 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 these 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 lung cancer. An alteration in expression of a polypeptide encoded by a nucleic acid associated with lung 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 lung 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 lung cancer is made by detecting a particular splicing variant encoded by a nucleic acid associated with lung cancer, or a particular pattern of splicing variants.

Quantitative or qualitative alterations can be present in the polypeptide. 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, lung cancer. In one embodiment, the control sample is from a subject that does not possess a variant shown herein to be associated with lung cancer. 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 lung 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 lung 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 lung cancer 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 lung cancer is made by detecting at least one marker or haplotypes of the present invention (*e.g.*, associated alleles of the markers listed in Table 4, 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 lung cancer, means for analyzing the nucleic acid sequence of a nucleic acid associated with lung cancer, means for analyzing the amino acid sequence of a polypeptide encoded by a nucleic acid associated with lung cancer, etc. The kits can for example include necessary buffers, nucleic acid primers for amplifying nucleic acids of the invention (*e.g.*, a nucleic acid segment comprising 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 other diagnostic assays for lung cancer.

In one embodiment, the invention is a kit for assaying a sample from a subject to detect the presence of a susceptibility to lung 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 (*e.g.*, the markers set forth in Table 4, and markers in linkage disequilibrium therewith). 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 4, 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 lung 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 lung 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 set forth in Table 4 and Table 6. In some embodiments, 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 listed in Table 4 and Table 6. In another embodiment, the marker or haplotype to be detected comprises at least one marker from markers in linkage disequilibrium to at least one of the markers listed in Table 4. In another embodiment, the marker to be detected is rs1051730.

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. As explained in the above, the detection oligonucleotide probe comprises 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, 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 (*e.g.*, a particular marker allele) or haplotype is indicative of a susceptibility (increased susceptibility or decreased susceptibility) to lung cancer. In another embodiment, the presence of the marker or haplotype is indicative of response to a therapeutic agent for lung cancer. In another embodiment, the presence of the marker or haplotype is indicative of prognosis of lung cancer. In yet another embodiment, the presence of the marker or haplotype is indicative of progress of treatment of lung cancer. Such treatment may include intervention by surgery, medication or by other means (*e.g.*, lifestyle changes).

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

In a further aspect of the present invention, a pharmaceutical pack (kit) is provided, the pack comprising a therapeutic agent and a set of instructions for administration of the therapeutic agent to humans diagnostically tested for one or more variants of the present invention, as disclosed herein. The therapeutic agent can be a small molecule drug, an antibody, a peptide, an antisense or rnai molecule, or other therapeutic molecules. In one embodiment, an individual identified as a carrier of at least one variant of the present invention is instructed to take a prescribed dose of the therapeutic agent. In one such embodiment, an individual identified as a homozygous carrier of at least one variant of the present invention is instructed to take a prescribed dose of the therapeutic agent. In another embodiment, an individual identified as a non-carrier of at least one variant of the present invention is instructed to take a prescribed dose of the therapeutic agent.

Treatment

Treatment for lung cancer can involve surgical removal of tumor, chemotherapy, or radiation therapy, as well as combinations of these methods. The decision about which treatments will

be appropriate for a given individual take into account the histological type (small cell lung carcinoma (SCLC) or non-small cell lung carcinoma (NSCLC), localization and extent of the tumor as well as the overall health status of the patient.

Surgical removal of the tumor is generally performed for early-stage (stage I or sometimes stage II) NSCLC and is the treatment of choice for cancer that has not spread beyond the lung. About 10%-35% of lung cancers can be removed surgically, but removal does not always result in a cure since the tumors may already have spread and can recur at a later time. Among people who have an isolated, slow-growing lung cancer removed, 25%-40% are alive five years after diagnosis. Surgery is less often performed with SCLC because these tumors are more likely to have spread at the time of diagnosis.

Radiation therapy may be employed as a treatment for both NSCLC and SCLC. Radiation therapy may be given as curative therapy, palliative therapy (using lower doses of radiation than with curative regimens) or as adjuvant therapy to surgery or chemotherapy. Brachytherapy is a term used to describe the use of a small pellet of radioactive material placed directly into the cancer or into the airway next to the cancer. Radiation therapy generally only shrinks a tumor or limits its growth when given as a sole therapy, yet in 10%-15% of people it leads to long-term remission and palliation of the cancer. Combining radiation therapy with chemotherapy can further increase the chances of survival.

Both NSCLC and SCLC may be treated with chemotherapy. Chemotherapy may be given alone, as an adjuvant to surgical therapy, or in combination with radiotherapy. While a number of chemotherapeutic drugs have been developed, the platinum-based drugs have been the most effective in treatment of lung cancers.

Chemotherapy is the treatment of choice for most SCLC, since these tumors are generally widespread in the body when they are diagnosed. Only half of people who have SCLC survive for four months without chemotherapy. With chemotherapy, their survival time is increased up to four- to fivefold. Chemotherapy alone is not particularly effective in treating NSCLC, but when NSCLC have metastasized, it can prolong survival in many cases.

In recent years, new drugs have been developed that target specifically defined genetic changes in the tumor, also called targeted therapy. The most notable of these drugs are compounds that target the epidermal growth factor receptor (gefitinib (Iressa), erlotinib (Tarceva)) which is amplified in a subset of lung cancer tumors. Inhibitors of angiogenesis such as antibodies that inhibit the vascular endotheliar growth factor (notably Bevacizumab) have also been shown to prolong survival in advanced lung cancer when added to the standard chemotherapy regimen. Bevacizumab may be adminstered in combination with paclitaxel and/or carboplatin.

A number of agents are in early stages of clinical research, including cyclo-oxygenase 2 inhibitors, proteasome inhibitors, bexaroteine and the apopotosis promoter exisulind. Current developmental strategies include proto-oncogene inhibition, phosphoinositide 3-kinase inhibition, histone deacylase inhibition, and tumor suppressor gene replacement. The methods of the invention are also applicable in the context of treatement by any of such therapeutic agents.

However, despite these recent advances in lung cancer treatment, the prognosis of the great majority of cases remains poor and there is a great need for development of more effective drugs.

Direct effect of nicotine on lung tissue through the nAChRs

Although traditionally not considered to be carcinogenic in itself, accumulating evidence suggests that nicotine contributes directly to lung carcinogenesis through stimulation of nAChRs in non-neuronal cells. nACh receptors are expressed in normal lung epithelial cells and respond to stimulation by nicotine by increased cellular proliferation and attenuation of apoptosis (Conti-Fine et al. 2000 Eur J Pharmacol 393(1-3):279-94, West et al. 2003 J Clin Invest 111(1):81-90). Furthermore, stimulation with nicotine or the nicotine-derived nitrosamine NNK induces a partially transformed phenotype in these cells, suggesting that nicotine may have a role as a tumor promoter in the lung (Ho et al. 2005 Toxicol and Appl Pharmaco 205(2):133-48). nAChRs are present on lung cancer cells of various histological types and stimulation with nicotine can promote tumor cell proliferation, tumor cell migration, invasion and reduce apoptosis of cells under hypoxia (Tsurutani et al. 2005 Carcinogenesis 26(7):1182-95, Xu and Deng 2006 J Biol Chem 281(7):4457-66, West et al. 2003 J Clin Invest 111(1):81-90, Heeschen et al. 2007 Nat Med 7(7):833-9, Maneckjee and Minna 1990 Proc Natl Acad Sci U S A 87(9):3294-8, Maneckjee and Minna 2004 Cell Growth Differ 5(10):1033-40). Furthermore, nicotine stimulation has been shown to induce tumor angiogenesis both through a cholinergic pathway, independently from the angiogenic pathways mediated by growth factor receptors, and by promoting VEGF expression through nAChRs (Heeschen et al. 2001 Nat Med 7(7):833-9, Cooke 2007 Life Sci 80(24-25):2347-51, Zhang et al. 2007 Clin Cancer Res 13(16):4686-94). In particular, nAChRs containing the α-3, α -5 and α -4 subunits are expressed on SCLC cells where they serve as receptors for autocrine growth (Song et al. 2003 Cancer Res 63(1):214-21).

In light of the evidence above, it is conceivable that sequence polymorphisms in the *CHRNA3/CHRNA5/CHRNB4* gene cluster might directly modulate the vulnerability to lung cancer thorugh direct effect of nicotine on the lung tissue. If this is true, it is also possible

that agents that modulate the activity of the nAChR in the lung may have a chemopreventive or even therapeutic potential.

Variants of the present invention (*e.g.*, the markers and/or haplotypes of the invention, *e.g.*, the markers listed in Table 4, *e.g.*, the markers listed in Table 6, *e.g.*, marker rs1051730 and/or marker rs16969968) can be used to identify novel therapeutic targets for lung cancer. For example, genes containing, or in linkage disequilibrium with, variants (markers and/or haplotypes) associated with lung cancer (*e.g.*, the *CHRNA3/CHRNA5/CHRNB4* 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 lung cancer, or prevent or delay onset of symptoms associated with lung 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, 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. By binding the appropriate target sequence, an RNA-RNA, DNA-DNA or RNA-DNA duplex is formed. The antisense oligonucleotides are complementary to the sense or coding strand of a gene. It is also possible to form a triple helix, where the antisense oligonucleotide binds to duplex DNA.

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

In certain embodiments, the antisense agent is an oligonucleotide that is capable of binding to a nucleotide segment of a gene selectet from the group consisting of *CHRNA3*, *CHRNA5* and *CHRNB4*. Antisense nucleotides can be from 5-500 nucleotides in length, including 5-200 nucleotides, 5-100 nucleotides, 10-50 nucleotides, and 10-30 nucleotides. In certain preferred embodiments, the antisense nucleotides is from 14-50 nucleotides in length, includign 14-40 nucleotides and 14-30 nucleotides. In certain embodiments, the antisense nucleotide is capable of binding to a nucleotide segment of a gene within the genomic segment with sequence as set forth in SEQ ID NO:1.

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 lung 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 set forth in Table 4) 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 a disease, including lung cancer, or a defect causing the disease, 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 lung cancer. Thus, the variants of the invention are useful as targets for the identification and/or development of therapeutic agents. Such methods may 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 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, *e.g.* a product of one or more of the *CHRNA3*, *CHRNA5* and *CHRNB4* genes. 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 lung 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 of treatment

As is known in the art, individuals can have differential responses to a particular therapy (*e.g.*, a therapeutic agent or therapeutic method). 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, *e.g.* a different response rate, to a particular treatment modality. This means that a patient diagnosed with lung 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 lung cancer. Therefore, the presence or absence of the marker allele or haplotype could aid in deciding what treatment should be used for the patient. For example, for a newly diagnosed patient, the presence of a marker or haplotype of the present invention may be assessed (*e.g.*, through testing DNA derived from a blood sample, as described herein). If the patient is positive for a marker allele or haplotype at (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. 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 lung cancer, or a susceptibility to lung cancer, at an early stage, to select the most appropriate treatment and/or preventive measure, and provide information to the clinician about prognosis/aggressiveness of the disease in order to be able to apply the most appropriate treatment.

The treatment for lung cancer can in certain embodiments be selected from surgical treatment (surgical removal of tumor), radiation therapy and chemotherapy. It is contemplated that the markers described herein to be associated with lung cancer can be used to predict the efficacy of any of these particular treatment modules. In certain embodiments, the markers of the inventions, as described herein may be used to determine an appropriate combination of therapy, which can include any one, two or three of these treatement modules. In certain embodiments, the radiation therapy is brachytherapy. The

agent useful for chemotherapy may be any chemical agent commonly used, or in development, as a chemotherapy agent, including, but not limited to, cisplatin, carboplatin, gemcitabine (4-amino-1-[3,3-difluoro-4-hydroxy-5- (hydroxymethyl) a tetrahydrofuran-2-yl]-1H-pyrimidin- 2-one), paclitaxel ((2a,4a,5β,7β,10β,13a)-4,10-bis(acetyloxy)-13-{[(2*R*,3*S*)-3-(benzoylamino)-2-hydroxy-3-phenylpropanoyl]oxy}-1,7-dihydroxy-9-oxo-5,20-epoxytax-11-en-2-yl benzoate), docetaxel ((2*R*,3*S*)-*N*-carboxy-3-phenylisoserine, *N*-tert-butyl ester, 13-ester with 5, 20-epoxy-1, 2, 4, 7, 10, 13-hexahydroxytax-11-en-9-one 4-acetate 2-benzoate), etoposide (4'-demethyl-epipodophyllotoxin 9-[4,6-O-(*R*)-ethylidene-beta-D-glucopyranoside], 4' -(dihydrogen phosphate)), vinorelbine (4-(acetyloxy)-6,7-didehydro-15-((2*R*,6*R*,8*S*)-4-ethyl-1,3,6,7,8,9-hexahydro- 8-(methoxycarbonyl)-2,6-methano- 2*H*-azecino(4,3-*b*)indol-8-yl)-3-hydroxy- 16-methoxy-1methyl,methylester, (2beta,3beta,4beta,5alpha,12*R*,19alpha) - aspidospermidine-3-carboxylic acid), and etoposide (4'-demethyl-epipodophyllotoxin 9-[4,6-O-(*R*)-ethylidene-beta-D-glucopyranoside] 4' -(dihydrogen phosphate)). Chemotherapy agents may be used alone or in combination. In one embodiment, the agent targets an epidermal growth factor receptor. In certain such embodiments, the agent is gefitinib (Iressa; *N*-(3-chloro-4-fluoro-phenyl)-7-methoxy-6-(3-morpholin-4-ylpropoxy)quinazolin-4-amine) or erlotinib (Tarceva; *N*-(3-ethynylphenyl)-6,7-bis(2-methoxyethoxy)quinazolin-4-amine). In certain other embodiments, the agent is angiogenesis inhibitor. Such inhibitors can for example be antibodies that inhibit the vascular endotheliar growth factor, such as Bevacizumab (Avastin).

The present invention also relates to methods of monitoring progress or effectiveness of a treatment for a lung cancer, as described in the above. 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 lung 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 (*e.g.*, the nicotinic acetylcholine receptor family), in samples taken before and during treatment. Alternatively, metabolites belonging to the biological network or metabolic pathway (*e.g.*, nicotine metabolites) 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 at least one at-risk variant 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 lung 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. This application can improve the safety of clinical trials, but can also enhance the chance that a clinical trial will demonstrate statistically significant efficacy, which may be limited to a certain sub-group of the population. Thus, one possible outcome of such a trial is that carriers of certain genetic variants, *e.g.*, the markers and haplotypes of the present invention, are statistically significantly likely to show positive response to the therapeutic agent, i.e. experience alleviation of symptoms associated with lung cancer when taking the therapeutic agent or drug as prescribed.

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

Computer-implemented aspects

As understood by those of ordinary skill in the art, the methods and information described herein may be implemented, in all or in part, as computer executable instructions on known computer readable media. For example, the methods described herein may be implemented in hardware. Alternatively, the method may be implemented in software stored in, for

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

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

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

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

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

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

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

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

acoustic, RF, infrared and other wireless media. Combinations of the any of the above should also be included within the scope of computer readable media.

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

The computer 110 may also include other removable/non-removable, volatile/nonvolatile computer storage media. By way of example only, Fig. 1 illustrates a hard disk drive 140 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 151 that reads from or writes to a removable, nonvolatile magnetic disk 152, and an optical disk drive 155 that reads from or writes to a removable, nonvolatile optical disk 156 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 141 is typically connected to the system bus 121 through a non-removable memory interface such as interface 140, and magnetic disk drive 151 and optical disk drive 155 are typically connected to the system bus 121 by a removable memory interface, such as interface 150.

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

interface 190. In addition to the monitor, computers may also include other peripheral output devices such as speakers 197 and printer 196, which may be connected through an output peripheral interface 190.

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

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

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

While the risk evaluation system and method, and other elements, have been described as preferably being implemented in software, they may be implemented in hardware, firmware, etc., and may be implemented by any other processor. Thus, the elements described herein may be implemented in a standard multi-purpose CPU or on specifically designed hardware or firmware such as an application-specific integrated circuit (ASIC) or other hard-wired device as desired, including, but not limited to, the computer 110 of Fig. 1. When implemented in

software, the software routine may be stored in any computer readable memory such as on a magnetic disk, a laser disk, or other storage medium, in a RAM or ROM of a computer or processor, in any database, etc. Likewise, this software may be delivered to a user or a diagnostic system via any known or desired delivery method including, for example, on a computer readable disk or other transportable computer storage mechanism or over a communication channel such as a telephone line, the internet, wireless communication, etc. (which are viewed as being the same as or interchangeable with providing such software via a transportable storage medium).

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

Accordingly, the invention relates to computer-implemented applications using the polymorphic markers and haplotypes described herein, and genotype and/or disease-association data derived therefrom. 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 and/or sequence information derived from an individual on readable media, so as to be able to provide the genotype information to a third party (*e.g.*, the individual, a guardian of the individual, a health care provider or genetic analysis service provider), or for deriving information from the genotype data, *e.g.*, by comparing the genotype data to information about genetic risk factors contributing to increased susceptibility to lung cancer, and reporting results based on such comparison.

Nucleic acids and polypeptides

The nucleic acids and polypeptides described herein can be used in methods and kits of the present invention. 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 80% or at least about 90% (on a molar basis) of all macromolecular species present. With regard to genomic DNA, the term "isolated" also can refer to nucleic acid molecules that are separated from the chromosome with which the genomic DNA is naturally associated. For example, the isolated nucleic acid molecule can contain less than about 250 kb, 200 kb, 150 kb, 100 kb, 75 kb, 50 kb, 25 kb, 10 kb, 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of the nucleotides that flank the nucleic acid molecule in the genomic DNA of the cell from which the nucleic acid molecule is derived.

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

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

The percent identity of two nucleotide or amino acid sequences can be determined by aligning the sequences for optimal comparison purposes (*e.g.*, gaps can be introduced in the sequence of a first sequence). The nucleotides or amino acids at corresponding positions are then compared, and the percent identity between the two sequences is a function of the number of identical positions shared by the sequences (i.e., % identity = # of identical positions/total # of positions x 100). In certain embodiments, the length of a sequence aligned for comparison purposes is at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, at least 90%, or at least 95%, of the length of the reference sequence. The actual

comparison of the two sequences can be accomplished by well-known methods, for example, using a mathematical algorithm. A non-limiting example of such a mathematical algorithm is described in Karlin, S. and Altschul, S., *Proc. Natl. Acad. Sci. USA, 90*:5873-5877 (1993). Such an algorithm is incorporated into the NBLAST and XBLAST programs (version 2.0), as described in Altschul, S. *et al.*, *Nucleic Acids Res., 25*:3389-3402 (1997). When utilizing BLAST and Gapped BLAST programs, the default parameters of the respective programs (*e.g.*, NBLAST) can be used. See the website on the world wide web at ncbi.nlm.nih.gov. In one embodiment, parameters for sequence comparison can be set at score=100, wordlength=12, or can be varied (*e.g.*, W=5 or W=20).

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, the nucleotide sequence of C15 LD Block (SEQ ID NO:1), or a nucleotide sequence comprising, or consisting of, the complement of the nucleotide sequence of C15 LD Block (SEQ ID NO:1), 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 labelled (*e.g.*, radiolabelled) 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 a susceptibility to lung 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 *Surf*ZAP™ 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 polymorphic marker of the invention, can be used to identify individuals that require modified treatment modalities.

Antibodies can furthermore be useful for assessing expression of variant proteins in disease states, such as in active stages of a disease, or in an individual with a predisposition to a disease related to the function of the protein, in particular lung 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 lung 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 (*e.g.*, CHRNA3, CHRNA5 and/or CHRNB4 proteins), 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.

EXEMPLIFICATION

Example 1.

The following contains description of the identification of susceptibility factors found to be associated with lung cancer (LC) through single-point analysis of SNP markers.

METHODS

Subjects

For all studies involving Icelandic subjects, the study protocols were approved by the National Bioethics Committee (NBC) and the Data Protection Authority (DPA) of Iceland. The DPA encrypted all personal identifiers associated with information or blood samples using the third-party encryption system (Gulcher, J.R., *et al., Eur J Hum Genet* **8:**739-42 (2000)). Overall the Icelandic study involves 10,995 subjects with information on SQ available in the GWA, an additional 2,950 subjects with information on SQ, and 4,203 never-smokers. In the LC study, 665 patients and 28,752 population controls were used (see Table 2 for details).

Smoking. All Icelandic subjects in the study of smoking-related phenotypes, including Icelandic population controls, were originally recruited for different genetic studies conducted over eleven years (1996-2007) at deCODE genetics and information on the number of cigarettes smoked per day (cpd) was available from various questionnaires. The cpd information was categorised into SQ level and used as a quantitative variable. Detailed information on SQ was also available for the foreign LC populations.

Nicotine Dependence. For a subset of the Icelandic smokers, information on the criteria used to diagnose ND was available from ongoing studies of ND and Anxiety/Depression (Thorgeirsson, T.E., *et al., Am J Hum Genet* **72:**1221-30 (2003)). We excluded individuals with diagnoses of other substance dependence or abuse giving a total of 2,394 ND subjects. A score of 4 or higher on the FTND (Heatherton, T.F. *et al., Br J Addict* **86:**1119-27 (1991)), or endorsement of at least three of the DSM criteria were used to assign affected status for ND.

Additional information on the Icelandic smoking and ND study group is available in the Supplementary Information.

Lung Cancer. Case control groups from three European populations were used in the studies on LC (Iceland, Spain and the Netherlands). Iceland: Recruitment was initiated in the year 1998 using a nationwide list from the Icelandic Cancer Registry (ICR). Approximately 1,265 LC patients were alive during the period of recruitment and of those 665 participated in the project. Information in the ICR includes year and age at diagnosis, year of death, SNOMED code and ICD-10 classification. Histological and cytological verification was available for 647 cases; the remaining 18 cases were diagnosed clinically. The Netherlands: The patients and controls were identified retrospectively through three different ongoing studies on genetic risk factors of disease. All three study protocols were approved by the Institutional Review Board of the Radboud University Nijmegen Medical Centre (RUNMC). A total of 90 patients and 2,018 controls are included in this study. Spain: The patients were recruited from the Oncology Department of Zaragoza Hospital in Zaragoza, from June 2006 to June 2007. During the 12 month interval of recruitment, 330 patients were invited to participate, and 292 enrolled (88% participation rate). Clinical information including age at onset and histology were collected from medical records. The 1,474 control individuals were approached at the University Hospital in Zaragoza. Study protocols were approved by the Institutional Review Board of Zaragoza University Hospital. All subjects gave written informed consent.

Genotyping

All 10,995 samples in the GWA study of SQ were genotyped using genotyping systems and specialised software from Illumina (Human Hap300 and Human Hap300-duo+ Bead Arrays, Illumina) (Barrett, J.C. & Cardon, L.R., *Nat Genet* **38:**659-62 (2006). Marker rs1051730 was genotyped using a Centaurus assay (Nanogen) for 8,566 Icelandic samples and all samples in the foreign study groups.

Statistical Analysis

Adjustment for relatedness in the Icelandic studies

Evaluation of statistical significance took the relatedness of the Icelandic individuals into account by dividing the test-statistic with a correction factor. For the GWA this was done by the method of genomic control (Devlin, B., *et al, Nat Genet* **36:**1129-30 (2004) using all 306,207 SNPs passing quality control. In all other comparisons genotype information for the total number of tested individuals was only available for SNP rs1051730, and the correction factor for the $\chi 2$ test-statistic was determined applying a simulation procedure using the

known genealogy which we had previously employed (Grant, S.F., *et al., Nat Genet* **38**:320-23 (2006)). We simulated 100,000 sets of genotypes for the SNP through the Icelandic genealogy of 739,000 individuals. The simulated genotypes were used in the applied tests resulting in 100,000 tests under the null hypothesis and the mean of the respective $\chi 2$ test-statistics gives the correction factor.

Regression Analysis

The year of birth had been rounded to the nearest 5. When year of birth adjustment was applied to study the effect of the variant, year of birth was treated as a categorical variable with four levels: ≤ 1930, 1935 to 1945, 1950 to 1960, and ≥ 1965. This divided the 13,945 smokers studied in groups of 3774, 3416, 4027 and 2728, which was the closest we could get to having four groups of equal size. The same categories were applied when analysing the data from Spain and the Netherlands.

Genotypic Odds Ratios

In general, the odds ratios for rs1051730 were calculated assuming a multiplicative model, i.e. the risks of the two alleles a person carries are expected to multiply. For example, if OR is the risk of T relative to C, then the risk of a homozygote TT individual will be OR times that of a heterozygote CT, and OR^2 times that of a homozygote CC. Additionally, genotypic ORs were calculated under the assumption of Hardy-Weinberg equilibrium in the controls (no control population showed a deviation from Hardy-Weinberg equilibrium).

Results

We performed a genome-wide association (GWA) study of smoking quantity (SQ), utilizing questionnaire data limited to basic questions on smoking behaviour that were available for a large number of lifetime smokers. The GWA scan comprises 10,995 Icelandic smokers who had been assayed with Infinium HumanHap300 SNP chips (Illumina). A set of 306,207 single nucleotide polymorphisms (SNPs), fulfilling our quality criteria, was tested. We focussed on cigarette smoking, with SQ reported as cigarettes per day (cpd). All SQ data were clustered into categories and we refer to them as "SQ levels", the SQ levels are: 0 (1-10 cpd), 1 (11-20 cpd), 2 (21-30 cpd), and 3 (31+ cpd). Each increment represents an increase in SQ of 10 cpd. Allele T of rs1051730 was most strongly associated with SQ, and the association was highly significant ($P=5\times10^{-16}$). The SNP is within the CHRNA3 gene in a linkage disequilibrium block (C15 LD block, SEQ ID NO:1) also containing two other nicotinic acetylcholine receptor

(nAChR) genes, CHRNA5 and CHRNB4. Six other SNPs on chromosome 15q24 passed the threshold of genome-wide significance ($P<2\times10^{-7}$), but they are all correlated with rs1051730 (r^2 = 0.14-0.93). An additional 2,950 smokers from Iceland were genotyped for rs1051730 giving a total of 13,945 smokers (Table 1) with mean variant frequency of 34.7%, which is not significantly different from the frequency of 34.4% observed in 4,203 individuals who were genotyped and reported never having smoked (OR=1.01, 95% CI:0.96-1.07, P=0.60). Indeed, the frequency of the variant in the 3,627 low quantity smokers (≤ 10 cpd), is significantly less than the frequency in those who do not smoke (OR=0.83, 95%CI:0.78-0.90, $P=4.5\times10^{-7}$). The increase in frequency between levels varies, and the largest increase (4.5%) is observed between the lowest levels (0 and 1), whereas the increase between the highest levels (2 and 3) is just 1.1%.

Association of the same variant with ND was previously reported in a candidate gene study involving 3,713 SNPs (Saccone, S.F., *et al., Hum Mol Genet* **16:**36-49 (2007)). We assessed the association with ND, defined as a score of 4 or higher on the FTND or endorsement of at least 3 of the 7 DSM-IV criteria. The variant is associated with ND in Iceland in a subset of 2,394 smokers from the SQ study tested both against 28,455 population controls (OR=1.17, 95%CI:1.10-1.25, $P=3.3\times10^{-6}$), and 3,506 low-quantity smokers (OR=1.40, 95%CI:1.29-1.52, $P=7\times10^{-15}$). Both the FTND and the DSM-IV scales include many items that are not based on SQ and their total scores are measures of ND severity. In our ND group, positive scores on most items in both scales show a trend toward higher frequency of the variant, as does the total score on both the FTND and DSM-IV scales. Thus the frequency of the variant increases with addiction severity, and is 46.8% and 43.8% for the highest decile of FTND, and DSM-IV, respectively.

We studied the effect of the variant on lung cancer risk. The study was based on 1,024 cases and 32,244 controls from Iceland, Spain and the Netherlands. The results (Table 2) represent the overall effect on LC including indirect effects through SQ and ND. Significant association was observed with LC for both the Icelandic data (OR = 1.27, $P = 4.1\times10^{-5}$) and for Spain and the Netherlands combined (OR = 1.39, $P = 6.6\times10^{-5}$). These two estimates are not significantly different from each other (P = 0.34), and combining results from all three groups gave an OR of 1.31 (95%CI:1.19-1.44, $P=1.5\times10^{-8}$). There was no significant difference in frequency of the variant between histological types of LC, which is not surprising given the small number of cases per group.

Genotypic ORs for LC and ND did not deviate significantly from those obtained for the multiplicative model, and no significant differences in the ORs between sexes were observed.

According to our estimates for Icelandic LC patients, the correlation between SQ and LC is consistent with numbers reported in other studies (Haimian, C.A., *et al., N Engl J Med* **354:**333-42 (2006); Stellman, S.D, *et al., Ann Epidemiol* **13:**294-302 (2003)). Combining

these estimates with our estimate of the association of the variant with SQ, the expected OR between the variant and lung cancer is only about 1.05 in Iceland, which is well below the direct OR estimate for LC of 1.27 (95%CI: 1.13-1.43). These results were obtained as follows:

In the 13,945 Icelandic smokers studied, 501 are known lung cancer cases. Using logistic regression adjusted for sex and year of birth, compared to SQ level 0, levels 1 to 3 are estimated to have relative risks of 2.1, 2.4 and 2.9 for lung cancer, respectively. Notably, the biggest jump in relative risks occurred between level 0 and level 1. Assuming these relative risk estimates and applying them to the distribution of smokers in various SQ levels as displayed in Table 1, frequency of the variant in lung cancer patients can be calculated as a weighted average. Specifically, the predicted frequency is

$$[(0.305\times0.260)+(0.350\times0.459\times2.1)+(0.380\times0.214\times2.4)+(0.391\times0.067\times2.9)]$$

divided by

$$[0.260+(0.459\times2.1)+(0.214\times2.4)+(0.067\times2.9)],$$

or 35.6%. Note that this calculation assumes only smokers would have lung cancer (non-smokers are given weight zero) and hence could over-estimate the frequency of the variant. Still, compared to the frequency of the variant in population controls (34.4%), the OR is only around 1.05. Note that since the frequency of the variant in SQ level 1 is only 35%, to increase the predicted frequency requires increasing the weights of SQ levels 2 and 3. However, even if we doubled the relative risks for SQ levels 2 and 3, from 2.4 and 2.9 to 4.8 and 5.8 respectively, the frequency and OR predicted for lung cancer patients would only increase respectively to 36.3% and 1.09

These results mean that the effect of the rs1051730 variant on LC is not explained by its effect on the traditional ND phenotypes (SQ, FTND score or DSM criteria). It is possible that an effect on other aspects of smoking behaviour, smoking duration in particular, may account for the observed difference between the indirect and direct estimates of the LC risks. An alternative possibility is that the variant directly confers risk of LC *e.g.* by increasing the vulnerability to tobacco smoke or through other unknown mechanisms.

Calculating the population attributable risk (PAR) for the variant results in 18% for lung cancer. While this is at best a ballpark figure given the complex interplay between the variant, smoking, and smoking-related diseases, it is likely that the variant accounts for a substantial fraction of LC cases and the associated morbidity and mortality.

Table 1: Genotype Status and Smoking Quantity (SQ) Level of 13,945 Icelandic Smokers.

Cigarettes per day (SQ level)	Genotype of rs1051730 GG	GT	TT	Total n (Freq.)	Frequency of T allele
1 to 10 (0)	1,743	1,558	326	3,627 (0.260)	0.305
11 to 20 (1)	2,727	2,865	810	6,402 (0.459)	0.350
21 to 30 (2)	1,145	1,416	427	2,988 (0.214)	0.380
31 and more (3)	341	448	139	928 (0.067)	0.391
All levels (Frequency)	5,956 (0.427)	6,287 (0.451)	1,702 (0.122)	13,945 (1.000)	0.347
Mean SQ level (SD)	1.01 (0.85)	1.12 (0.86)	1.22 (0.85)	1.09 (0.86)	

Table 2: Association of rs1051730 allele T with Lung Cancer

Study Group	Controls n	freq	Cases n	freq	OR	(95% CI)	*P*
Lung Cancer							
Iceland	28,752	0.342	665	0.398	1.27	(1.13 - 1.43)	4.1×10^{-5}
Spain	1,474	0.390	269	0.483	1.46	(1.22 - 1.76)	5.4×10^{-5}
The Netherlands	2,018	0.314	90	0.350	1.18	(0.86 - 1.61)	0.31
Foreign combined	3,492	-	359	-	1.38	(1.18 - 1.62)	6.6×10^{-5}
All combined	32,244	-	1,024	-	1.31	(1.19 - 1.44)	1.5×10^{-8}

5

Table 3. Association of markers in LD with rs1051730 to Lung Cancer (LC). Shown is the marker name, associating allele, numerical value of the LD measure R^2 to rs1051730 based on HapMap CEU data (http://www.hapmap.org), relative risk of the association (RR), the number of affecteds and controls and the allelic frequencies in those groups.

Lung Cancer								
marker	**allele**	**R^2 to rs1051730**	**p-value**	**RR**	**Naff**	**Aff Freq**	**Ncon**	**Con Freq**
rs1051730	T	-	1.75E-05	1.30	610	0.404098	26229	0.342522
rs8034191	C	0.93	3.49E-05	1.29	610	0.596721	26228	0.656093
rs2036534	T	0.14	0.026071	1.18	609	0.79803	26218	0.77052
rs11638372	T	0.44	0.04754	1.13	608	0.418586	26222	0.389597
rs4887077	T	0.43	0.048422	1.13	610	0.415574	26219	0.386781
rs6495314	C	0.44	0.068595	1.12	610	0.418852	26209	0.392232
rs1996371	G	0.45	0.085206	1.11	610	0.418852	26223	0.393681

Table 4. Surrogate markers for rs1051730. Shown are marker names, position of the polymorphic site in NCBI Build 36, the position of the polymorphic site in SEQ ID NO:1, and values for the LD measures |D'|, R^2, and p-value. Linkage disequilibrium was determined using genotypes from the HapMap Caucasian CEU dataset (http://www.hapmap.org).

A.

marker	position (Build 36)	Position SEQ ID NO:1	|D'|	R^2	p-value
rs4436747	76501063	1	0.784105	0.263895	3.65E-09
rs2568498	76508987	7925	0.788769	0.276514	1.54E-09
rs1394371	76511524	10462	0.804627	0.542323	5.30E-16
rs12899131	76513940	12878	0.784105	0.263895	3.65E-09
rs2568500	76513983	12921	0.777839	0.256775	8.52E-09
rs17483548	76517368	16306	0.858671	0.712237	5.42E-22
rs17405217	76518204	17142	0.858671	0.712237	5.42E-22
rs17483721	76520786	19724	0.858671	0.712237	5.42E-22
rs1847529	76522125	21063	0.788769	0.276514	1.54E-09
rs8041628	76522410	21348	0.784596	0.274417	2.37E-09
rs2656052	76527987	26925	0.858671	0.712237	5.42E-22
rs2568494	76528019	26957	0.856736	0.726896	7.55E-21
rs7181486	76528673	27611	0.858429	0.711836	1.09E-21
rs17483929	76529431	28369	0.858671	0.712237	5.42E-22
rs10519198	76529809	28747	0.788769	0.276514	1.54E-09
rs12909921	76530315	29253	0.812377	0.282282	1.41E-08
rs12910090	76530355	29293	0.788769	0.276514	1.54E-09
rs2656065	76537604	36542	0.852863	0.70688	2.79E-21
rs11639224	76540426	39364	0.821543	0.308926	2.70E-09
rs1964678	76541055	39993	0.813921	0.237809	1.03E-07
rs2009746	76541157	40095	0.858848	0.736891	1.59E-22
rs17484235	76548469	47407	0.858671	0.712237	5.42E-22

marker	position (Build 36)	Position SEQ ID NO:1	|D'|	R^2	p-value
rs4299116	76553249	52187	0.806306	0.233961	2.88E-07
rs1504550	76553305	52243	0.858278	0.731261	2.15E-21
rs12910910	76554905	53843	0.8077	0.230774	2.33E-07
rs8043227	76555926	54864	0.813921	0.237809	1.03E-07
rs17484524	76559731	58669	0.852974	0.705337	2.19E-21
rs8042238	76561326	60264	0.809173	0.240375	2.02E-07
rs8042260	76561429	60367	0.790462	0.222858	1.35E-06
rs12903295	76566027	64965	0.85721	0.236711	3.84E-07
rs12904234	76566439	65377	0.810067	0.236137	1.49E-07
rs965604	76576278	75216	0.813921	0.237809	1.03E-07
rs13180	76576543	75481	0.813921	0.237809	1.03E-07
rs1062980	76579582	78520	0.810734	0.236646	2.16E-07
rs4362358	76583159	82097	0.813921	0.237809	1.03E-07
rs9788721	76589924	88862	1	0.871795	7.70E-31
rs8034191	76593078	92016	1	0.871795	7.70E-31
rs12591557	76598787	97725	1	0.366812	3.19E-14
rs10519203	76601101	100039	1	0.871795	7.70E-31
rs12914694	76601499	100437	1	0.38914	4.69E-14
rs8031948	76603112	102050	1	0.871795	1.79E-30
rs1504545	76605526	104464	1	0.372294	1.21E-14
rs952215	76606208	105146	1	0.372294	1.21E-14
rs952216	76606257	105195	1	0.361233	3.45E-14
rs12902493	76606330	105268	1	0.372294	1.21E-14
rs11636131	76608661	107599	1	0.372294	1.21E-14
rs11632604	76608969	107907	1	0.372294	1.21E-14
rs12910289	76609120	108058	1	0.366812	2.04E-14
rs1504546	76611290	110228	1	0.372294	1.21E-14

marker	position (Build 36)	Position SEQ ID NO:1	|D'|	R^2	p-value
rs12906951	76612617	111555	1	0.366812	2.04E-14
rs3885951	76612972	111910	1	0.247573	2.53E-09
rs931794	76613235	112173	1	0.871795	7.70E-31
rs12916999	76613967	112905	1	0.380531	7.49E-15
rs12915366	76618808	117746	1	0.345992	8.47E-14
rs12916483	76619452	118390	1	0.363636	5.69E-14
rs3813572	76619643	118581	1	0.372294	1.21E-14
rs3813571	76619847	118785	1	0.372294	1.21E-14
rs4886571	76620813	119751	1	0.366812	3.19E-14
rs4243083	76620885	119823	1	0.369369	3.36E-14
rs2292117	76621744	120682	1	0.372294	1.21E-14
rs11858230	76622607	121545	1	0.363707	1.34E-13
rs8025429	76623417	122355	1	0.369369	3.36E-14
rs4887062	76624856	123794	0.943182	0.343351	9.61E-13
rs4887063	76626770	125708	1	0.372294	1.91E-14
rs8053	76628275	127213	1	0.372294	1.21E-14
rs1979907	76629294	128232	1	0.386792	7.82E-15
rs1979906	76629344	128282	1	0.394619	2.68E-15
rs1979905	76629429	128367	1	0.385965	4.45E-15
rs4887064	76629902	128840	1	0.385965	4.45E-15
rs12907966	76630106	129044	1	0.385965	4.45E-15
rs880395	76631411	130349	1	0.385965	4.45E-15
rs905740	76631441	130379	1	0.385965	4.45E-15
rs7164030	76631716	130654	1	0.385965	4.45E-15
rs4275821	76636596	135534	1	0.333333	2.17E-13
rs7173512	76636969	135907	1	0.333333	2.17E-13
rs2036527	76638670	137608	0.963677	0.837627	2.25E-27

marker	position (Build 36)	Position SEQ ID NO:1	|D'|	R^2	p-value
rs588765	76652480	151418	1	0.37788	1.31E-14
rs6495306	76652948	151886	1	0.4	1.60E-15
rs17486278	76654537	153475	0.962446	0.895088	2.47E-28
rs601079	76656634	155572	1	0.4	1.60E-15
rs495956	76656985	155923	1	0.333333	2.17E-13
rs680244	76658343	157281	1	0.4	1.60E-15
rs621849	76659916	158854	1	0.4	1.60E-15
rs7180002	76661048	159986	0.964252	0.867797	1.54E-28
rs692780	76663560	162498	1	0.333333	2.17E-13
rs11637635	76664205	163143	1	0.333333	2.17E-13
rs481134	76664618	163556	1	0.394619	2.68E-15
rs951266	76665596	164534	0.964252	0.867797	1.54E-28
rs555018	76666297	165235	1	0.394619	2.68E-15
rs647041	76667536	166474	1	0.392185	3.91E-15
rs17408276	76668673	167611	1	0.320988	5.47E-13
rs16969968	76669980	168918	1	0.901961	1.21E-31
rs518425	76670868	169806	1	0.226415	1.13E-09
rs514743	76671282	170220	1	0.320988	5.47E-13
rs615470	76673043	171981	1	0.320988	5.47E-13
rs660652	76674887	173825	1	0.320988	5.47E-13
rs472054	76675049	173987	1	0.315353	1.37E-12
rs578776	76675455	174393	1	0.212454	2.37E-09
rs6495307	76677376	176314	1	0.385965	4.45E-15
rs1051730	**76681394**	**180332**	**1**	**1**	-
rs3743077	76681951	180889	1	0.392185	1.00E-14
rs1317286	76683184	182122	1	0.901961	2.87E-31
rs12914385	76685778	184716	1	0.787879	8.22E-27

marker	position (Build 36)	Position SEQ ID NO:1	|D'|	R^2	p-value
rs2869546	76694400	193338	1	0.333333	2.17E-13
rs3743075	76696507	195445	1	0.308943	1.35E-12
rs3743074	76696535	195473	1	0.325598	5.08E-13
rs3743073	76696594	195532	1	0.315353	1.37E-12
rs8040868	76698236	197174	1	0.759036	2.11E-26
rs1878399	76699058	197996	1	0.4	1.60E-15
rs1948	76704454	203392	1	0.242424	2.23E-10
rs7178270	76708132	207070	1	0.347826	9.23E-14
rs17487223	76711042	209980	0.926323	0.748065	5.19E-24
rs950776	76713073	212011	1	0.285714	7.88E-12
rs11636753	76716001	214939	1	0.342105	2.39E-13
rs11637890	76722474	221412	1	0.353448	5.46E-14
rs11633223	76722531	221469	1	0.369369	5.26E-14
rs11634351	76731773	230711	0.747959	0.510482	3.07E-13
rs1021070	76733918	232856	1	0.345992	8.47E-14
rs7181405	76735207	234145	1	0.340426	1.43E-13
rs11638830	76735374	234312	0.768682	0.50015	1.70E-14
rs17487514	76740840	239778	0.577029	0.240648	1.33E-06
rs12899135	76741434	240372	0.729687	0.464109	7.89E-13
rs12910237	76743393	242331	1	0.329004	4.10E-13
rs1996371	76743861	242799	0.739121	0.476262	7.13E-14
rs6495314	76747584	246522	0.739121	0.476262	7.13E-14
rs922691	76751049	249987	0.936182	0.310116	9.53E-11
rs12905641	76751417	250355	1	0.315353	9.23E-13
rs11639372	76753710	252648	0.74855	0.483781	6.43E-13
rs12902602	76754456	253394	0.751972	0.483132	9.37E-13
rs1021071	76755234	254172	0.739121	0.476262	7.13E-14

marker	position (Build 36)	Position SEQ ID NO:1	|D'|	R^2	p-value
rs11072785	76755284	254222	0.736113	0.49926	1.22E-13
rs11857532	76755323	254261	0.724283	0.41331	4.03E-12
rs4886580	76756440	255378	0.721631	0.451338	1.29E-12
rs8038920	76761600	260538	1	0.320988	5.47E-13
rs4887077	76765419	264357	0.694601	0.406611	1.52E-11
rs11638372	76770614	269552	0.694601	0.406611	1.52E-11
rs922692	76771269	270207	0.694601	0.406611	1.52E-11
rs12910627	76781988	280926	0.694601	0.406611	1.52E-11
rs11072791	76784131	283069	0.694601	0.406611	1.52E-11
rs11638490	76795005	293943	0.677562	0.38728	1.73E-10
rs11629637	76806079	305017	0.698873	0.423336	6.02E-12
rs3813565	76806665	305603	0.704707	0.447924	1.02E-12
rs4887082	76812122	311060	0.694601	0.406611	1.52E-11
rs12286	76838814	337752	0.661245	0.381187	9.19E-11
rs1809420	76843824	342762	0.655389	0.361998	3.12E-10
rs7174367	76851722	350660	0.647747	0.356148	1.33E-09
rs7171916	76855006	353944	0.608975	0.302175	1.50E-08
rs1994017	76867361	366299	0.917337	0.204002	9.97E-08
rs12905740	76869419	368357	0.90045	0.204099	2.63E-06
rs2277545	76870646	369584	0.510929	0.214706	4.42E-06
rs1564499	76871863	370801	0.917337	0.204002	9.97E-08
rs12903203	76871988	370926	0.539624	0.225571	1.78E-06
rs3743057	76876062	375000	0.917337	0.204002	9.97E-08
rs8038189	76886081	385019	0.920091	0.214074	4.66E-08
rs922693	76886593	385531	0.919558	0.218072	4.83E-08
rs1383636	76893275	392213	0.922667	0.224377	2.15E-08

B

Example 2.

All exons, promoters, and 5' and 3'UTRs were sequenced for each of the *CHRNA5*, *CHRNA3* and *CHRNB4* genes in the nicotinic acetylcholine receptor subunit cluster in a sample of lung cancer patients (n=184), nicotine dependent smokers (n=176) and low quantity smokers (n=175). The regions that were sequenced are indicated in Table 5. In total, 111 variants were found, 47 of which were not present in dbSNP129. A full description of all variants is found in Table 6, including position, alleles, frequency and possible functional significance. Statistical analysis focused on 50 variants with minor allele frequencies greater than 1%. Results of this analysis are found in Table 7. Given the strong established effect seen with rs1051730, we expect to find significant results for this SNP and correlated SNPs. P-values which include an adjustment for the effect of rs1051730 are thus also included in the table.

We examined linkage disequilibrium (LD) among these polymorphisms in order to define equivalence groups in which all polymorphisms have $r^2>0.8$ to one SNP identified as head of the group (Table 8). Six equivalence groups are formed accounting for all but three of the polymorphisms with frequency greater than 5% (See Table 8). These three polymorphisms had strongest LD to the head of class A (rs1051730; r^2 between 0.64 and 0.79) and are thus reported together with that group.

Genotypes from Illumina Human Hap300 chips are available for all subjects sequenced, as well as for additional subjects in each group. Information on linkage disequilibrium within the sequencing sample was used to identify appropriate tagging variants from the Illumina chip to effectively increase sample size for variants of interest.

rs16969968

The non-synonymous *CHRNA5* variant rs16969968 has previously been highlighted in the literature (Saccone, SF, *et al. Hum Mol Genet* 16:36-49 (2007)). In European Americans, LD is strong between this variant and rs1051730 according to the Hapmap project data (D'=1, r^2=0.9; Table 4). We found these two variants to be equivalent in our sequencing sample.

rs1051730 equivalence group

In addition to rs16969968, several other SNPs were found to be in very strong LD with rs1051730 in Iceland. These include rs55853698, rs55781567 and rs8192482, all with

r^2>0.93 to rs1051730/rs16969968. Because LD is so strong in Iceland, we cannot differentiate between these 5 SNPs. Another SNP, ss107794645, exhibited weaker LD with rs1051730/rs16969968 (D'=0.91, r^2=0.69). Within the sequencing sample this SNP gave a stronger risk than rs1051730 for nicotine dependence (OR=1.65 vs. 1.49) but not lung cancer (OR=1.53 vs 1.58). A single SNP assay was designed to further test this variant in Iceland. After additional subjects were genotyped, the OR of this variant is 1.26 (p=0.006, p=0.8 after adjustment for rs1051730) for lung cancer (n=645), and is 1.18 (p=0.02, p=0.8 after adjustment for rs1051730) for nicotine dependence (n=2068), both tested against low quantity smokers (n=535),. These results indicate that risk associated with ss107794645 is due to its LD with rs1051730.

rs12907519/rs8192475

The results from our sequencing analysis alone indicate a significant protective effect of the C allele of rs12907519, a SNP located in intron 1 of *CHRNA3*. With low quantity smokers as controls, the variant has an OR of 0.34 for nicotine dependence (p=0.007 after adjustment) and 0.21 for lung cancer (p=0.0003 after adjustment). This SNP is within equivalence group D, in strong LD with rs8192475 (r^2=0.93) which is included on the Illumina chip. With all genotypes available for rs8192475, association of this variant is not significant for lung cancer (OR=0.78, p=0.5 after adjustment for rs1051730) or nicotine dependence (OR=0.87, p=0.9 after adjustment) when compared to low quantity smokers (see Table 3). Given the strong LD between these variants, we can rule out association of rs12907519 with either lung cancer or nicotine dependence.

Equivalence Classes in Illumina samples

Four equivalence classes are headed by a SNP on the Illumina chip. A fifth can be tagged with r^2=0.98 by a haplotype of two SNPs from the chip. Results within the larger chip sample are displayed for all tagged classes in Table 9, with and without adjustment for the effect of rs1051730. One class (A) is headed by rs1051730. Within the chip genotyped sample analyzed here, the T allele is strongly associated with both nicotine dependence (OR=1.4, $p=7.4\times10^{-15}$) and lung cancer (OR=1.52, $p=1.5\times10^{-11}$). Of the SNPs which head the remaining 4 classes tagged by Illumina chips, with and without correction for rs1051730, only rs8192475 displayed significant association in any of the three tests within the sequencing sample In the larger chip-genotyped sample, several SNPs have significant p-values due to correlation with rs1051730. After adjustment for the effect of rs1051730 the SNP rs1948 has a p-value of 0.006. This presents the possibility that a protective effect for lung cancer might

exist for a variant in this equivalence class which occurs primarily on the same background as the risk effect of rs1051730. Any such effect would be small, and is masked by the comparably strong risk associated with rs1051730.

rs578776

The SNP rs578776 has recently been reported to be an independent, second risk variant for nicotine dependence within this LD block (Bierut, LJ, *et al. Am J Psych* 165:1163-71 (2008)).. We genotyped additional nicotine dependent cases and low quantity smokers so that our data set would be large enough to address the relationship of rs578776 to rs1051730/rs16969968. According to Hapmap project data, in European Americans LD between the variants is D'=1, r^2=0.2 (www.hapmap.org). In Iceland we see similar results (D'= 0.99, r^2=0.19, n=3026). The risk allele of rs1051730/rs16969968 is fixated on the background of the major allele of rs578776. Therefore there are only 3 haplotypes possible. We find that all the risk associated with rs578776 is confined to the haplotype which includes the risk variant of rs1051730/rs16969968 (OR=1.34, p=1.56x10^{-4}; Figure 2). The frequency of the haplotype containing the protective allele of rs1051730 and the risk allele of rs578776 occurs at a lower frequency in nicotine dependence (37.6%) compared to low quantity smokers (39.7%). There is no evidence to support an independent risk for nicotine dependence associated with the rs578776 variant.

Rare Variants

Of the variants identified with sequencing, 59 occur at frequencies of less than 1%. Table 10 includes the number of carriers in each phenotype group for each of these variants. Among them are 7 missense mutations and one 20bp exonic deletion. The exonic deletion occurs in *CHRNA3* in one subject from the nicotine dependence group. This individual received a score of 4 on the FTND scale and did not meet DSM criteria for nicotine dependence. None of the rare variants alone can fully account for the signal observed. We cannot however rule out the possibility that among these variants are rare high penetrance variants which might influence risk of one or both conditions.

Three length polymorphisms: rs3841324, rs55787222, and rs60706203

Three length polymorphisms were genotyped directly in additional subjects. These include a 22bp insertion/deletion, rs3841324, in the promoter of *CHRNA5*, identified in a scan for promoter polymorphisms affecting gene expression (Buckland, PR, *et al. Hum Mutat* , rs60706203, a 3bp insertion/deletion in the leader sequence of *CHRNA3*, and rs55787222, a 4bp microsatellite in the promoter region of *CHRNA3*. Additional rare alleles of each of the last 2 variants were identified with additional genotyping (see Table 6).

Results for association analysis of these markers are presented in Table 11. P-values were adjusted to take into account the effect of rs1051730. There is no significant association with lung cancer or nicotine dependence for either rs3841324 or rs60706203. In the case of rs55787222, a 4 bp microsatellite in the promoter region of *CHRNA3*, the allele containing 2 copies of the 4bp sequence is not associated with either condition before correction for rs1051730. After correction, however, the p- value is 0.004 for association with nicotine dependence. Within this sample, the p-value for rs1051730 is 0.001 for nicotine dependence. It appears that the allele of rs55787222 which contains 2 copies (-8 allele with respect to CEPH 1347-02 reference) may be mildly protective against nicotine dependence. The risk allele of rs1051730 is fixated on this background, and the risk contributed by this variant is stronger than the protective effect which may be supplied by this allele of rs55787222. However, the risk for rs1051730 is observed for the comparison of both nicotine dependence and lung cancer against low quantity smokers. The possible protective effect of rs55787222 is only observed for nicotine dependence.

Expression

We measured expression of *CHRNA5* in two tissues to address whether genetic variants in the cluster are associated with expression regulation. In particular, rs3841324 has been reported as a promoter regulatory element in cell culture (Buckland PR, *et al. Hum Mutat* 26:214-23 (2005)). We sought to test the effect of this variant on expression *in vivo*. Expression of *CHRNA5* was strongly associated with rs3841324 genotype, with relative expression levels higher for the short allele in blood ($r=0.72$, $p=4x10^{-71}$) and subcutaneous adipose tissue ($r=0.73$, $p=2x10^{-63}$). Association with expression of *CHRNA5* was also examined for the other SNPs within the LD block, with one marker from each equivalence class tested (see Table 12). All markers were significantly associated with expression. Adjusting for rs3841324 reduces the significance of the association for the other SNPs drastically, in subcutaneous adipose tissue only rs1051730 remains nominally significant ($p=0.018$) and three SNPs show nominally significant association in blood (minimum $p=0.006$ for rs1051730). Overall, expression in blood and subcutaneous adipose tissue is strongly associated with rs3841324.

However, we cannot rule out an additional comparably weak effect of another SNP, which was best captured by rs1051730. Expression of *CHRNA5* was not associated with lifetime regular smoking, or with smoking within the past 24 hours (data not shown).

We have established that there is no risk for nicotine dependence or lung cancer associated with this variant independent of the risk associated with rs1051730 (Table 11). However, there is strong LD between the two variants. The T allele of rs1051730 only appears on the haplotype background including the long, or low expression, allele of rs3841324.

A careful characterization of the CHRNA5/CHRNA3/CHRNB4 cluster does not identify any variants with stronger association to nicotine dependence or lung cancer than rs1051730/rs16969968. Therefore the SNP non-synonymous SNP rs16969968 remains the variant most likely to have functional effects leading to the observed association signals within this region.

Materials and Methods

Subjects for Sequencing

Three groups of subjects were selected for sequencing analysis: (1) lung cancer patients (n=184), (2) nicotine dependent smokers without other addictions (n=176) and (3) low-quantity smokers (n=175) (See Table 13 for demographic information). Low-quantity smokers reported regular smoking for at least one year and reported only social smoking or less than 5 cigarettes per day. Subjects with lung cancer show the highest frequency of the identified risk variant, and generally constitute a population with high lifetime smoking exposure. Our sample of nicotine dependent individuals received the diagnosis based on questionnaire data addressing two systems of classification of nicotine dependence, the Fagerstrom Test for Nicotine Dependence (FTND) (Heatherton, TF, *et al. Br J Addict* 86:1119-27)) and the criteria of the Diagnositic and Statistical Manual, Version IV (DSM). Subjects met criteria under either or both systems (FTND 4+ or DSM 3+). Individuals with other substance dependence or abuse diagnoses were excluded. Our previous analysis indicated that the effect of the risk variant was to increase smoking quantity among smokers, rather than affecting initiation. Therefore, we used smokers with low consumption as a control group for study.

Subjects for additional genotyping

Certain variants of interest were specifically genotyped in additional individuals. For the length variants rs55787222, rs3841324 and rs60706203, the subjects included 567 lung cancer patients, 1623 nicotine dependent smokers and 608 low quantity smokers (See Table 14).

All subjects sequenced have also been genotyped with Human Hap300 or Human Hap300-duo1 Bead Arrays (Illumina; San Diego, CA, USA). Additional subjects from each group have also been genotyped using these chips. LD information obtained in the sequencing cohort was used to identify tagging SNPs for testing in the larger sample, which included 669 lung cancer cases, 1950 nicotine dependent smokers and 4680 low quantity smokers (See Table 14)

The study protocols were approved by the National Bioethics Committee (NBC) and the Data Protection Authority (DPA) of Iceland. The DPA has encrypted all personal identifiers linked to phenotype information or blood samples using a third-party encryption system(15)(15). All subjects are of Icelandic ancestry.

Sequencing

The exons, 5' and 3' UTRs, and flanking sequences 1kb upstream of *CHRNA5*, *CHRNA3*, and *CHRNB4* were sequenced. Sequence for the region was obtained from NCBI build 36. A total of 57 primer pairs were designed. The position of regions sequenced (build 36) can be found in Table 5. PCR amplification and sequencing reactions were set up on Zymark ALH300 workstations, with amplification performed on MJR Tetrads. PCR products were purified using AMPure (Agencourt Bioscience). Dye terminator removal was performed using CleanSEQ (Agencourt) to repurify. Electrophoresis was performed on Applied Biosystems 3730 DNA Analyzers. Sequence editing and analysis were performed using deCODE Genetics Sequence Miner software. SNP calling was done by both manual inspection and automated calling. All SNPs identified through automated calling were then confirmed by manual inspection of the sequence traces. Insertion/deletions and microsatellites were identified by manual inspection. Simple, rare insertion/deletions were called manually.

Genotyping

Additional genotyping of SNPs was done using the Centaurus platform (Nanogen). Three variants, rs55787222, rs3841324, and rs60706203, observed in the sequencing, were genotyped in a larger population. For these markers primers were designed using Primer3. PCR reactions were set up on Zymark ALH300 workstations and amplification performed on MJR Tetrads. PCR products were pooled, an internal size standard added, and then resolved on Applied Biosystems 3730 DNA Analyzers. Primers and PCR conditions are available on request. Genotypes were called and edited using deCODE Allele Caller and deCODE-GT.

Expression Analysis

The variant rs3841324 was identifed as a promoter element with significant effect on transcription of *CHRNA5* in a genome scan for regulatory elements (Buckland, PR, *et al. Hum Mutat* 26:214-23 (2005)). We therefore examined its role in regulating expression of the gene in blood and subcutaneous adipose tissue using an expression cohort previously described (Emilsson, V., *et al. Nature* 452:423-8 (2008)). From this cohort, genotype and expression data were used from 446 individuals with blood samples and 376 individuals with subcutaneous adipose tissue samples.

RNA samples were purified using RNeasy Mini Kit (Quiagen), and integrity analyzed using Agilent 2100 Bioanalyzer. Total RNA was converted to cDNA using the High Capacity cDNA Archive Kit (Applied Biosystems). Two Taqman assays were designed for *CHRNA5*, so that positive results cannot be attributable to the specific assay used. The probes are located at different exon boundaries, one crossing exon 2 and 3, and the other crossing exons 3 and 4. Real-time PCR was carried out according to manufacturer's recommendations on an ABI Prism 7900HT Sequence Detection System. Quantification was performed using the ΔΔCt method (User Bulletin no. 2, Applied Biosystems 2001). A housekeeping gene, in this case *GUSB,* was run in parallel for normalization.

Statistical Analysis

A likelihood ratio test was used for analysis using χ^2 statistics. In all cases p-values are reported both with and without correction for the effect of rs1051730. P-values are reported without correction for multiple testing. In the analysis of the larger samples generated from Illumina genotypes and individual genotyping of length polymorphisms, p-values are

corrected for relatedness among affecteds as described previously using a simulation procedure with the known genealogy (Grant, SF., *et al. Nat Genet*38:320-3 (2006)).

The expression data were log-transformed, adjusted for sex and age with a linear regression model, and the standardized residuals were used as the variable. There were 307 individuals present in both data sets and their residuals for the two tissues tested were highly correlated ($r=0.65$, $p=7x10^{-39}$).

In analysis of equivalence classes in larger cohorts, genotypes for rs569207 are inferred. Allele T is tagged by a haplotype of allele C at rs1051730 and allele G at rs680244 ($r^2=0.98$ in the sequencing data) in the analysis of Illumina data. In the expression analysis genotypes were inferred using a two SNP haplotype based on allele G at rs680244 and allele T at rs578776 ($r^2=0.99$ in the sequencing data).

Table 5. Build 36 positions for regions sequenced

Gene	Region	Build 36 position
CHRNA5	5' Flanking & Exon 1	76643986-76645528
	Exon 2	76659873-76660680
	Exon 3	76665714-76666400
	Exon 4	76667349-76668117
	Exon 5	76668894-76670363
	Exon 6 & 3' Flanking	76672141-76673771
CHRNB4	5' Flanking & Exon 1	76720345-76721584
	Exon 2	76714503-76715286
	Exon 3	76710160-76711022
	Exon 4	76708007-76709537
	Exon 5 & 3' Flanking	76703378-76704963
CHRNA3	5' Flanking & Exon 1	76699749-76701312
	Exon 2 &3	76697716-76698675
	Exon 4	76696032-76696844
	Exon 5	76680349-76682010
	Exon 6 & 3' Flanking	76674343-76676459

Table 6. Descriptive information on all variants from sequencing

A: CHRNA5

Marker	Ref SNP ID	Position (B 36)	Pos in Seq ID No: 1	Major Allele	Minor Allele	Minor Allele Freq	Function	aa change
SG15S363	ss107794609	76644594	143532	G	T	0,2%	near CHRNA5	
DG15S1561	rs3841324	76644868	143806	22bp[1]	-	42,2%	near CHRNA5	
SG15S468	rs56182392	76644934	143872	G	A	1,3%	near CHRNA5	
SG15S364	rs503464	76644951	143889	T	A	21,4%	near CHRNA5	
SG15S365	rs55853698	76644994	143932	T	G	36,5%	utr	
SG15S366	rs55781567	76645041	143979	C	G	36,8%	utr	
SG15S411	ss107794620	76645331	144269	G	A	0,2%	intron	
SG15S412	rs684513	76645455	144393	C	G	19,4%	intron	
SG15S312	rs6495306	76652948	151886	A	G	43,3%	intron	
SG15S151	rs680244	76658343	157281	G	A	43,4%	intron	
SG15S311	rs621849	76659916	158854	A	G	43,3%	intron	
SG15S352	ss107794606	76660070	159008	A	C	0,6%	intron	
SG15S469	ss107794638	76660154	159092	G	A	0,2%	intron	
SG15S353	rs569207	76660174	159112	C	T	21,0%	intron	
SG15S470	ss107794639	76660617	159555	A	G	0,6%	intron	
SG15S344	rs55982512	76666113	165051	C	T	0,4%	intron	
SG15S345	rs555018	76666297	165235	A	G	42,5%	intron	
SG15S346	rs647041	76667536	166474	C	T	43,1%	intron	
CHRNA5_0	ss107794648	76667615	166553	TC	-	0,1%	intron	
SG15S347	rs12898919	76667632	166570	G	C	4,8%	intron	
SG15S348	rs2229961	76667807	166745	G	A	1,1%	non-synon	V->I
SG15S471	rs56201623	76669059	167997	C	T	0,1%	intron	
SG15S349	ss107794603	76669155	168093	T	C	0,4%	intron	
SG15S350	ss107794604	76669481	168419	C	T	0,3%	synon	
SG15S148	rs16969968	76669980	168918	G	A	36,0%	non-synon	D->N
SG15S351	ss107794605	76670141	169079	C	T	0,4%	intron	
SG15S355	ss107794607	76672424	171362	G	C	1,0%	intron	
CHRNA5_1	ss107794649	76672962	171900	ACT	-	0,1%	utr	
SG15S356	rs615470	76673043	171981	C	T	38,2%	utr	
SG15S357	rs8192483	76673204	172142	G	A	0,1%	utr	
SG15S358	rs55783657	76673213	172151	G	A	1,3%	utr	
SG15S359	rs8192482	76673253	172191	C	T	35,7%	utr	
SG15S360	rs564585	76673282	172220	A	G	24,8%	utr	
SG15S361	ss107794608	76673351	172289	G	A	0,1%	utr	

B: CHRNA3

Marker	Ref SNP ID	Position (B 36)	Pos in Seq ID No: 1	Major Allele	Minor Allele	Minor Allele Freq	Function	aa change
SG15S389	rs12899226	76674493	173431	A	C	4,9%	near CHRNA3	
SG15S390	rs55736590	76674550	173488	C	T	0,7%	near CHRNA3	
CHRNA3_1	rs34238957	76674771	173709	-	CTCT	38,3%	utr	
SG15S391	rs660652	76674887	173825	C	T	38,2%	utr	
SG15S445	ss107794646	76675048	173986	T	C	0,2%	utr	
SG15S392	rs472054	76675049	173987	C	T	38,2%	utr	
CHRNA3_2	rs35186448	76675294	174232	-	CCCC	20,9%	utr	
SG15S393	rs56113144	76675406	174344	C	T	0,3%	utr	
SG15S162	rs578776	76675455	174393	G	A	24,4%	utr	
SG15S394	ss107794615	76676128	175066	T	A	0,2%	non-synon	I->N
SG15S382	rs56403513	76680842	179780	C	T	0,1%	synon	
SG15S383	ss107794613	76681061	179999	C	T	0,1%	synon	
SG15S446	ss107794633	76681390	180328	C	T	0,1%	non-synon	H->Y
SG15S149	rs1051730	76681394	180332	G	A	35,9%	synon	
SG15S384	rs55958820	76681412	180350	C	A	1,5%	synon	
SG15S385	rs8192480	76681475	180413	T	C	0,1%	synon	
CHRNA3_4	ss107794647	76681496	180434	20bp[2]	-	0,1%	frameshift	
SG15S386	ss107794614	76681539	180477	C	T	0,1%	non-synon	P->L
SG15S387	rs3743078	76681814	180752	C	G	20,9%	intron	
SG15S388	rs3743077	76681951	180889	G	A	42,8%	intron	
SG15S447	ss107794634	76696119	195057	G	A	0,9%	intron	
SG15S448	rs4887069	76696125	195063	C	C	21,3%	intron	
SG15S376	rs8192479	76696453	195391	G	A	3,1%	synon	
SG15S377	rs3743075	76696507	195445	G	A	37,9%	synon	
SG15S378	rs3743074	76696535	195473	T	C	38,0%	intron	
SG15S379	rs3743073	76696594	195532	A	C	38,1%	intron	
SG15S380	rs41280050	76696612	195550	C	T	1,7%	intron	
SG15S381	ss107794612	76696708	195646	G	A	0,8%	intron	
SG15S449	ss107794635	76696793	195731	G	C	0,4%	intron	
CHRNA3_0	ss107794650	76698094	197032	-	A	4,1%	intron	
CHRNA3_0	ss107794650	76698094	197032	A	-	0,2%	intron	
SG15S367	rs8040868	76698236	197174	A	G	41,4%	synon	
SG15S368	rs8192475	76698285	197223	C	T	5,0%	non-synon	R->H
SG15S374	ss107794610	76698484	197422	C	T	0,1%	intron	
SG15S375	ss107794611	76698488	197426	G	C	0,2%	intron	
SG15S396	rs7170068	76699998	198936	C	T	24,3%	intron	
SG15S397	ss107794616	76700025	198963	G	A	0,1%	intron	
SG15S398	ss107794617	76700033	198971	A	G	0,1%	intron	
SG15S399	rs12907519	76700099	199037	A	G	5,0%	intron	
DG15S1563	rs60706203	76700142	199080	AGC[3]	-	39,6%	non-synon	L/-
SG15S450	ss107794636	76700424	199362	C	A	0,2%	near CHRNA3	
DG15S1568	rs55787222	76700428	199366	(CGCC)2-7[4]			near CHRNA3	
SG15S413	ss107794621	76700491	199429	C	G	0,1%	near CHRNA3	
SG15S414	ss107794622	76700600	199538	C	T	0,8%	near CHRNA3	
SG15S415	ss107794623	76700884	199822	T	C	0,1%	near CHRNA3	
SG15S466	ss107794637	76700888	199826	T	C	0,1%	near CHRNA3	

Marker	Ref SNP ID	Position (B 36)	Pos in Seq ID No: 1	Major Allele	Minor Allele	Minor Allele Freq	Function	aa change
SG15S416	ss107794624	76700993	199931	A	G	0,2%	near CHRNA3	
SG15S417	rs12911814	76701039	199977	T	G	5,1%	near CHRNA3	
SG15S467	rs13329271	76701285	200223	T	G	10,0%	near CHRNA3	

[1] CTATTTCCCTCTGGCCCCGCCC

[2] ATCGATTTTCGCCTTATCGT

[3] The major allele contains 7 copies of AGC, the minor 6. With genotyping of 2798 individuals, one individual was identified with 8 copies.

[4] The marker contains 2-7 repeats of (CGCC); frequencies for alleles of rs55787222 are based on specific genotyping of this variant in 2935 individuals.

Frequences are as follows - 2: 41.6%, 3: 0.07%, 4: 50.4%, 5: 7.4%, 6: 0.02%, 7: 0.6%

C: CHRNB4

Marker	Ref SNP ID	Position (B 36)	Pos in Seq ID No: 1	Major Allele	Minor Allele	Minor Allele Freq	Function	aa change
SG15S402	rs2904130	76703677	202615	C	G	36,0%	near CHRNB4	
SG15S401	ss107794618	76704151	203089	C	G	0,8%	utr	
SG15S400	rs55952530	76704371	203309	G	A	1,5%	utr	
SG15S313	rs1948	76704454	203392	C	T	34,6%	utr	
SG15S476	ss107794644	76704907	203845	C	T	0,1%	intron	
SG15S410	rs7178270	76708132	207070	G	C	40,5%	intron	
SG15S409	rs56317523	76708398	207336	C	T	0,4%	non-synon	A->V
SG15S408	rs56235003	76708657	207595	C	T	0,8%	non-synon	R->C
SG15S407	rs3743072	76708817	207755	C	T	0,1%	synon	
SG15S406	rs55919125	76709249	208187	C	T	4,3%	synon	
SG15S404	rs56218866	76709284	208222	A	G	0,1%	non-synon	S->G
SG15S405	rs56095004	76709295	208233	G	A	0,7%	non-synon	R->Q
SG15S403	ss107794619	76709464	208402	C	T	0,1%	intron	
SG15S472	ss107794640	76710242	209180	A	G	0,3%	intron	
SG15S473	ss107794641	76710250	209188	A	G	0,2%	intron	
SG15S420	rs12914008	76710560	209498	C	T	3,5%	non-synon	T->I
SG15S419	ss107794625	76710751	209689	A	G	0,1%	intron	
SG15S418	rs28534575	76710900	209838	A	C	21,1%	intron	
SG15S474	ss107794642	76710925	209863	G	A	0,1%	intron	
SG15S426	rs12440298	76714644	213582	A	C	0,2%	intron	
SG15S425	ss107794630	76714649	213587	T	G	0,2%	intron	
SG15S424	ss107794629	76714670	213608	A	G	0,1%	intron	
SG15S423	ss107794628	76714717	213655	C	T	0,1%	intron	
SG15S422	ss107794627	76714925	213863	C	T	0,3%	non-synon	R->S
SG15S421	ss107794626	76715163	214101	C	T	0,1%	intron	
SG15S475	ss107794643	76720731	219669	G	C	0,1%	near CHRNB4	
SG15S430	ss107794632	76720780	219718	C	T	0,2%	near CHRNB4	
SG15S429	ss107794631	76721203	220141	A	T	0,2%	near CHRNB4	
SG15S428	ss107794645	76721373	220311	G	C	40,7%	near CHRNB4	

Table 7. Comparison of frequencies for markers with minor allele frequency greater than 1% in Nicotine Dependence (ND), Low Quantity Smokers (LQS) and Lung Cancer (LC). Padj - p-value after adjustment for the effect of rs1051730. Allele number for rs55787222 refers to number of copies of 4bp repeat.

Marker	Ref SNP ID	Allele	ND		LQS		LC		ND against LQS			LC against LQS			LC against ND		
			N	Freq	N	Freq	N	Freq	OR	P	Padj	OR	P	Padj	OR	P	Padj
SG15S149	rs1051730	T	176	0,384	175	0,294	184	0,397	1,49	0,01	-	1,58	0,004	-	1,06	0,7	-
SG15S347	rs12898919	C	175	0,040	172	0,076	178	0,028	0,51	0,04	0,1	0,35	0,004	0,01	0,69	0,4	0,4
SG15S389	rs12899226	G	174	0,040	173	0,081	182	0,027	0,48	0,02	0,06	0,32	0,001	0,005	0,67	0,3	0,4
SG15S399	rs12907519	C	160	0,034	158	0,095	162	0,022	0,34	0,002	0,007	0,21	0,00004	0,0003	0,62	0,3	0,3
SG15S417	rs12911814	C	174	0,040	172	0,084	172	0,029	0,46	0,02	0,04	0,33	0,001	0,006	0,71	0,4	0,4
SG15S420	rs12914008	T	175	0,031	171	0,056	182	0,019	0,55	0,1	0,2	0,33	0,009	0,02	0,60	0,3	0,3
SG15S148	rs16969968	A	171	0,389	174	0,293	184	0,397	1,53	0,008	1,0	1,59	0,004	1,0	1,03	0,8	1,0
SG15S313	rs1948	T	176	0,335	174	0,351	184	0,351	0,93	0,7	0,3	1,00	1,0	0,09	1,07	0,7	0,4
SG15S348	rs2229961	A	176	0,017	174	0,003	183	0,014	6,02	0,05	0,1	4,81	0,1	0,2	0,80	0,7	0,7
SG15S418	rs28534575	C	172	0,183	170	0,244	182	0,206	0,69	0,05	0,3	0,80	0,2	0,9	1,16	0,4	0,3
SG15S402	rs2904130	G	176	0,349	174	0,379	184	0,351	0,88	0,4	0,6	0,88	0,4	0,4	1,00	1,0	0,8
CHRNA3_1	rs34238957	0	169	0,385	172	0,390	180	0,375	0,98	0,9	0,1	0,94	0,7	0,1	0,96	0,8	0,8
CHRNA3_0	rs34844435	0	175	0,963	174	0,934	184	0,973	1,83	0,08	0,2	2,53	0,01	0,04	1,38	0,4	0,5
CHRNA3_0	rs34844435	1	175	0,037	174	0,066	184	0,022	0,55	0,08	0,2	0,31	0,003	0,01	0,58	0,2	0,2
CHRNA3_2	rs35186448	2	175	0,191	174	0,239	181	0,199	0,76	0,1	0,5	0,79	0,2	0,9	1,05	0,8	0,6
SG15S379	rs3743073	G	175	0,386	175	0,383	177	0,376	1,01	0,9	0,08	0,97	0,8	0,08	0,96	0,8	1,0
SG15S378	rs3743074	G	175	0,386	175	0,383	184	0,372	1,01	0,9	0,08	0,96	0,8	0,1	0,94	0,7	0,9
SG15S377	rs3743075	T	176	0,384	175	0,383	184	0,372	1,00	1,0	0,08	0,96	0,8	0,1	0,95	0,8	0,9
SG15S388	rs3743077	T	176	0,426	170	0,465	176	0,395	0,86	0,3	0,5	0,75	0,06	0,9	0,88	0,4	0,6
SG15S387	rs3743078	C	176	0,190	173	0,237	179	0,201	0,76	0,1	0,5	0,81	0,2	1,0	1,07	0,7	0,6
DG15S1561	rs3841324	del	152	0,424	163	0,457	177	0,387	0,88	0,4	0,5	0,75	0,06	1,0	0,86	0,3	0,5
SG15S380	rs41280050	A	175	0,020	175	0,011	180	0,019	1,77	0,4	0,3	1,72	0,4	0,4	0,97	1,0	1,0
SG15S392	rs472054	A	174	0,388	175	0,386	184	0,372	1,01	1,0	0,09	0,94	0,7	0,1	0,94	0,7	0,9
SG15S448	rs4887069	G	175	0,194	174	0,241	181	0,204	0,76	0,1	0,5	0,81	0,2	0,9	1,07	0,7	0,6
SG15S364	rs503464	A	164	0,195	171	0,240	176	0,207	0,77	0,2	0,5	0,83	0,3	1,0	1,08	0,7	0,6
SG15S345	rs555018	G	174	0,422	170	0,462	175	0,391	0,85	0,3	0,5	0,75	0,06	1,0	0,88	0,4	0,5
SG15S366	rs55781567	G	165	0,394	171	0,307	181	0,403	1,47	0,02	0,5	1,53	0,008	0,6	1,04	0,8	1,0
SG15S358	rs55783657	A	176	0,009	173	0,006	183	0,025	1,48	0,7	0,9	4,34	0,03	0,1	2,93	0,09	0,09
DG15S1568	rs557872221	4	162	0,506	166	0,551	164	0,503	0,83	0,2	0,2	0,82	0,2	0,09	0,99	0,9	0,8
DG15S1568	rs55787222	5	162	0,071	166	0,087	164	0,052	0,80	0,4	0,8	0,57	0,07	0,2	0,72	0,3	0,3
DG15S1568	rs55787222	2	162	0,423	166	0,361	164	0,436	1,29	0,1	0,09	1,37	0,05	0,1	1,06	0,7	0,8
SG15S365	rs55853698	G	168	0,390	171	0,304	180	0,400	1,46	0,02	0,9	1,53	0,008	0,6	1,04	0,8	0,6
SG15S406	rs55919125	T	176	0,043	174	0,032	183	0,055	1,36	0,4	0,3	1,77	0,1	0,05	1,30	0,5	0,4

SG15S400	rs55952530	A	170	0,009	171	0,020	178	0,017	0,43	0,2	0,3	0,82	0,7	1,0	1,93	0,3	0,3
SG15S384	rs55958820	T	176	0,017	172	0,012	180	0,017	1,47	0,5	0,8	1,44	0,6	0,9	0,98	1,0	0,9
SG15S468	rs56182392	A	158	0,013	168	0,012	173	0,014	1,06	0,9	0,8	1,22	0,8	0,6	1,14	0,8	0,8
SG15S360	rs564585	G	175	0,223	173	0,301	183	0,221	0,67	0,02	0,2	0,66	0,02	0,2	0,99	1,0	0,8
SG15S353	rs569207	T	176	0,190	175	0,237	182	0,203	0,76	0,1	0,5	0,82	0,3	1,0	1,09	0,7	0,5
SG15S162	rs578776	T	175	0,223	175	0,300	179	0,209	0,67	0,02	0,2	0,62	0,006	0,09	0,92	0,7	0,8
DG15S1563	rs60706203	del	165	0,409	154	0,399	161	0,379	1,04	0,8	0,09	0,92	0,6	0,3	0,88	0,4	0,5
SG15S356	rs615470	T	176	0,386	174	0,388	184	0,372	0,99	1,0	0,09	0,94	0,7	0,1	0,94	0,7	0,8
SG15S311	rs621849	G	176	0,426	175	0,466	168	0,405	0,85	0,3	0,5	0,78	0,1	0,7	0,92	0,6	0,8
SG15S346	rs647041	T	172	0,427	172	0,471	177	0,395	0,84	0,3	0,6	0,74	0,04	0,7	0,88	0,4	0,4
SG15S312	rs6495306	G	176	0,426	175	0,466	168	0,405	0,85	0,3	0,5	0,78	0,1	0,7	0,92	0,6	0,8
SG15S391	rs660652	A	169	0,385	171	0,389	179	0,374	0,98	0,9	0,1	0,94	0,7	0,1	0,96	0,8	0,9
SG15S151	rs680244	A	176	0,426	156	0,474	168	0,405	0,82	0,2	0,7	0,75	0,07	0,9	0,92	0,6	0,8
SG15S412	rs684513	G	172	0,180	156	0,215	168	0,188	0,80	0,3	0,7	0,84	0,4	0,8	1,05	0,8	0,6
SG15S410	rs7178270	C	174	0,402	172	0,445	177	0,370	0,84	0,3	0,8	0,73	0,04	0,6	0,87	0,4	0,4
SG15S367	rs8040868	C	175	0,429	174	0,382	182	0,431	1,21	0,2	0,04	1,23	0,2	0,02	1,01	0,9	0,8
SG15S368	rs8192475	T	176	0,040	175	0,083	184	0,027	0,46	0,02	0,05	0,31	0,001	0,004	0,67	0,3	0,4
SG15S376	rs8192479	T	176	0,028	175	0,026	182	0,038	1,11	0,8	0,7	1,52	0,3	0,8	1,37	0,5	0,5
SG15S359	rs8192482	T	175	0,380	173	0,292	183	0,396	1,49	0,01	1,0	1,59	0,003	1,0	1,07	0,7	1,0
SG15S428	ss107794645	C	172	0,451	161	0,332	183	0,432	1,65	0,002	0,1	1,53	0,007	0,5	0,93	0,6	0,4

Table 8. Equivalence classes for SNPs with minor allele frequency greater than 5%. All variants with frequency greater than 5% were grouped into equivalence classes based on $r^2>0.8$. A lead SNP for each class was chosen, and r2 for each variant to that SNP is listed. Three variants do not fit into these classes. They are listed separately under class A, to which each has the strongest LD.

Class	A		B		C		D		E		F	
Head	**rs1051730**	r^2	**rs680244**	r^2	**rs1948**	r^2	**rs8192475**	r^2	**rs578776**	r^2	**rs569207**	r^2
	rs16969968	1,00	rs34238957	0,82	rs2904130	0,92	rs34844435	0,88	rs564585	0,99	rs35186448	0,99
	rs8192482	1,00	rs3841324	0,91			rs12898919	1,00			rs503464	0,86
	rs55853698	0,93	rs60706203	0,87			rs12899226	1,00			rs3743078	0,99
	rs55781567	0,93	rs621849	1,00			rs12907519	0,93			rs7170068	0,87
			rs6495306	1,00			rs12911814	1,00			rs684513	0,83
			rs555018	1,00							rs28534575	0,83
	rs55787222	0,64	rs647041	0,99							rs4887069	0,96
	rs8040868	0,79	rs615470	0,82							rs13329271	0,90
	ss107794645	0,69	rs3743075	0,81								
			rs3743074	0,81								
			rs3743073	0,81								
			rs3743077	1,00								
			rs660652	0,82								
			rs472054	0,82								
			rs7178270	0,80								

Table 9. Association Results for Equivalences Classes in Larger Samples. P-values include correction for relatedness among groups; Padj1 adjusts the P-value for the effect of rs1051730 (LQS = Low Qantity Smokers; ND = Nicotine Dependence; LC = Lung Cancer)

Marker	Class	Allele	LQS		ND		LC		ND against LQS			LC against LQS			LC against ND		
			N	freq	N	freq	N	freq	OR	Padj	Padj 1	OR	Padj	Padj 1	OR	Padj	Padj 1
rs1051730	A	T	4676	0,309	1950	0,384	669	0,404	1,40	7.1x10-15	-	1,52	1.5x10-11	-	1,09	0,2	-
rs680244	B	G	4680	0,556	1950	0,595	669	0,593	1,18	9.2x10-5	0,3	1,16	0,01	0,03	0,99	0,9	0,2
rs1948	C	C	4674	0,650	1950	0,682	669	0,667	1,15	0,001	0,4	1,08	0,2	0,006	0,93	0,3	0,04
rs8192475	D	C	4674	0,947	1948	0,954	668	0,958	1,15	0,2	0,9	1,28	0,09	0,5	1,11	0,5	0,6
rs569207	F	C	4675	0,751	1950	0,787	669	0,809	1,22	3.0x10-5	0,2	1,41	3.1x10-6	0,03	1,15	0,09	0,2

Table10. All variants with frequency less than 1%.

Markers	Ref SNP ID	Allele	Number of Carriers of Minor Allele		
			LQS	ND	LC
CHRNA3_0	rs55665143	(-A)	0	0	2
CHRNA3_4	ss107794647	(-20bp)[1]	0	1	0
CHRNA5_0	ss107794648	(-TC)	1	0	0
CHRNA5_1	ss107794649	(-ACT)	0	0	1
SG15S344	rs55982512	T	1	1	2
SG15S349	ss107794603	C	1	2	1
SG15S350	ss107794604	T	0	0	3
SG15S351	ss107794605	T	4	0	0
SG15S352	ss107794606	C	1	2	3
SG15S355	ss107794607	C	4	1	5
SG15S357	rs8192483	A	0	1	0
SG15S361	ss107794608	A	1	0	0
SG15S363	ss107794609	T	1	1	0
SG15S374	ss107794610	A	1	0	0
SG15S375	ss107794611	G	1	0	1
SG15S381	ss107794612	T	2	4	3
SG15S382	rs56403513	A	0	0	1
SG15S383	ss107794613	A	1	0	0
SG15S385	rs8192480	G	0	0	1
SG15S390	rs55736590	A	3	1	3
SG15S393	rs56113144	A	0	2	1
SG15S394	ss107794615	T	1	0	1
SG15S397	ss107794616	T	0	1	0
SG15S398	ss107794617	C	1	0	0
SG15S401	ss107794618	G	2	2	4
SG15S403	ss107794619	T	0	0	1
SG15S404	rs56218866	G	0	0	1
SG15S405	rs56095004	A	0	2	5
SG15S407	rs3743072	T	0	0	1
SG15S408	rs56235003	T	4	1	3
SG15S409	rs56317523	T	0	1	3
SG15S411	ss107794620	A	1	0	1
SG15S413	ss107794621	C	1	0	0
SG15S414	ss107794622	A	3	1	2
SG15S415	ss107794623	G	0	0	1
SG15S416	ss107794624	C	1	0	0
SG15S419	ss107794625	G	0	0	1
SG15S421	ss107794626	T	1	0	0
SG15S422	ss107794627	T	0	2	1
SG15S423	ss107794628	T	0	1	0
SG15S424	ss107794629	G	0	1	0
SG15S425	ss107794630	G	0	0	2
SG15S426	rs12440298	C	0	1	1
SG15S429	ss107794631	T	0	2	0
SG15S430	ss107794632	T	1	0	1
SG15S445	ss107794646	C	2	0	0
SG15S446	ss107794633	A	0	0	1
SG15S447	ss107794634	A	6	1	2
SG15S449	ss107794635	G	2	1	1
SG15S450	ss107794636	T	1	0	0
SG15S466	ss107794637	G	0	1	0
SG15S469	ss107794638	A	2	0	0
SG15S470	ss107794639	G	3	2	1
SG15S471	rs56201623	T	0	0	1
SG15S472	ss107794640	G	1	1	0
SG15S473	ss107794641	C	0	2	0
SG15S474	ss107794642	A	0	1	0
SG15S475	ss107794643	C	1	0	0
SG15S476	ss107794644	T	0	1	0

[1] ATCGATTTTCGCCTTATCGT. Other alleles in paranthesis are indel polymorphisms of the respective alleles

Table 11. Association Results for Insertion/Deletions and Microsatellites. The results for the T allele of rs1051730 within the sample genotyped for length variants is included here in the table for comparison. P - includes correction for relatedness among groups.
Padj corresponds to P-value after adjustment for the effect of rs1051730.

Marker	Allele	LQS		ND		LC		ND against LQS			LC against LQS			LC against ND		
		N	freq	N	freq	N	freq	OR	P	Padj	OR	P	Padj	OR	P	Padj
rs1051730	T	608	0,306	1623	0,359	567	0,384	1,27	0,001	0,001	1,42	8.5x10-5	8.5x10-5	0,98	0,8	0,4
rs3841324	del	608	0,433	1623	0,403	567	0,399	0,88	0,08	0,8	0,87	0,1	0,4	0,95	0,5	0,8
rs60706203	ins	608	0,398	1623	0,385	567	0,374	0,95	0,4	0,2	0,90	0,2	0,2	1,15	0,05	0,2
rs55787222	2	608	0,382	1623	0,404	567	0,438	1,10	0,2	0,004	1,26	0,006	0,2	0,89	0,1	0,4
rs55787222	4	608	0,536	1623	0,514	567	0,484	0,91	0,2	0,1	0,81	0,01	0,6	0,94	0,7	0,9
rs55787222	5	608	0,076	1623	0,075	567	0,071	1,00	1,0	0,5	0,94	0,7	0,7	1,05	0,9	0,8
rs55787222	7	608	0,004	1623	0,006	567	0,006	1,43	0,5	0,4	1,50	0,5	0,4	1,11	0,1	0,1

Table 12. Association analysis of the effect of rs3841324 genotype on expression of CHRNA5. Other variants within the block were tested as well by including the head of each equivalence class.

Marker	Allele	Class	Whole Blood						Subcutaneous Adipose					
			r	LCL	UCL	P	adjP1	adjP2	r	LCL	UCL	P	adjP1	adjP2
rs3841324	0	B	0,72	0,67	0,76	3.7x10-71	-	-	0,73	0,68	0,77	1.9x10-63	-	-
rs1051730	4	A	-0,54	-0,60	-0,47	2.0x10-34	0,006	-	-0,53	-0,60	-0,45	3.9x10-28	0,02	-
rs680244	1	B	0,71	0,66	0,75	4.3x10-69	0,03	0,1	0,71	0,65	0,75	3.6x10-58	0,07	0,3
rs1948	4	C	0,58	0,51	0,63	1.3x10-40	0,8	0,5	0,56	0,49	0,63	2.0x10-32	0,8	0,6
rs8192475	4	D	0,26	0,17	0,34	3.0x10-8	0,03	0,04	0,23	0,13	0,33	5.3x10-6	0,5	0,6
rs578776	4	E	-0,14	-0,23	-0,05	0,002	0,06	0,5	-0,10	-0,20	0,00	0,004	0,08	0,7
rs569207[1]	4	F	-0,23	-0,31	-0,14	1.3x10-6	0,06	0,3	-0,19	-0,29	-0,09	1.6x10-4	0,09	0,5

[1]. rs569207 allele T (= allele 4) is tagged by a haplotype using allele G at rs680244 and allele T at rs578776 (r2=0.99)
adjP1 - adjusted for the effect of rs3841324; adjP2 - adjusted for effects of both rs3841324 and rs1051730

Table 13. Demographics: Sequencing Cohort

Cohort	N	Sex (M/F)	Age (yrs)
Low Quantity Smokers	175	57/118	55.8±18.4
Nicotine Dependence	176	79/97	50.6±10.4
Lung Cancer	184	98/86	72.6±10.8

Table 14. Demographics for cohorts used in analyses including additional genotyping

A - Length Polymorphisms	**N**	**Male/Female**	**Age (years)**
Nicotine Dependence	1623	602/1021	50.4±11.2
Lung Cancer	567	291/276	70.6±11.0
Low Quantity Smokers	608	192/416	58.4±18.2
B - rs578776			
Nicotine Dependence	2161	758/1403	50.1±11.3
Low Quantity Smokers	865	283/582	57.7±18.8
C - Illumina			
Nicotine Dependence	1950	689/1261	51.0±11.0
Lung Cancer	669	340/329	70.6±11.0
Low Quantity Smokers	4681	1203/3478	63.9±19.1

Age is in years ± S.D.

CLAIMS

1. A method for determining a susceptibility to lung cancer in a human individual, 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, or in a genotype dataset from the individual, wherein the at least one polymorphic marker is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein determination of the presence of the at least one allele is indicative of a susceptibility to lung cancer.

2. The method of claim 1, wherein the at least one polymorphic marker is selected from the group consisting of the markers set forth in Table 4 and Table 6.

3. The method of claim 1, wherein the at least one polymorphic marker is selected from the group consisting of rs1051730, and markers in linkage disequilibrium therewith.

4. The method of claim 3, wherein the at least one polymorphic marker is selected from the group consisting of rs55853698, rs55781567, rs8192482, ss107794645 and the markers set forth in Table 4.

5. The method of any one of the preceding claims, wherein the at least one polymorphic marker is rs1051730.

6. The method of any one of the claims 1 - 4, wherein the at least one polymorphic marker is rs16969968.

7. The method of any one of the preceding claims, further comprising assessing the frequency of at least one haplotype comprising at least two polymorphic markers.

8. The method of any one of the preceding claims, wherein the susceptibility is increased susceptibility.

9. The method of claim 8, wherein the presence of the at least one allele or haplotype is indicative of increased susceptibility with a relative risk of at least 1.25.

10. The method of claim 8 or claim 9, wherein the presence of the at least one allele or haplotype is indicative of increased susceptibility of lung cancer with a relative risk of at least 1.30.

11. The method of any one of the claims 8 - 10, wherein the at least one allele or haplotype is selected from the group consisting of rs1051730 allele T, rs680244 allele

G, rs1948 allele C, rs8034191 allele C, rs2036534 allele T, rs11638372 allele T, rs4887077 allele T, rs6495314 allele C, and rs1996371 allele G.

12. The method of any one of the claims 1-7, wherein the susceptibility is decreased susceptibility.

13. The method of claim 12, wherein the at least one marker allele or haplotype is indicative of decreased susceptibility of lung cancer with a relative risk of less than 0.8.

14. The method of claim 12 or claim 13, wherein the at least one marker allele or haplotype is selected from the group consisting of rs1051730 allele C and rs55787222 allele -8.

15. The method of any one of the preceding Claims, wherein the presence of the marker or haplotype is indicative of a different response rate of the subject to a particular treatment modality for lung cancer.

16. The method of any one of the preceding claims, wherein the individual is of an ancestry that includes Caucasian ancestry.

17. The method of Claim 16, wherein the ancestry is self-reported.

18. The method of Claim 16, wherein the ancestry is determined by detecting at least one allele of at least one polymorphic marker in a sample from the individual, wherein the presence or absence of the allele is indicative of the ancestry of the individual.

19. A method for determining a susceptibility to lung cancer in a human individual, 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, or in a genotype dataset from the individual, wherein the at least one polymorphic marker is associated with a gene selected from CHRNA5, CHRNA3 and CHRNB4, and wherein determination of the presence of the at least one allele is indicative of a susceptibility to lung cancer.

20. The method of claim 19, wherein the at least one polymorphic marker is selected from the group consisting of rs1051730, rs680244, rs1948 and rs569207, and markers in linkage disequilibrium therewith.

21. The method of claim 19, wherein the at least one polymorphic marker is selected from the group consisting of rs1051730 and markers in linkage disequilibrium therewith.

22. The method of claim 21, wherein the at least one polymorphicmarker is selected from the group consisting of the markers set forth in Table 4.

23. The method of any one of the claims 19 – 22, wherein the polymorphism is rs16969968.

24. The method of any one of the claims 19 – 22, wherein the polymorphism is rs1051730.

25. The method of any of the claims 19 - 24, wherein the susceptibility is increased susceptibility.

26. The method of claim 25, wherein determination of the presence of rs1051730 allele T, rs680244 allele G, rs1948 allele C, rs8034191 allele C, rs2036534 allele T, rs11638372 allele T, rs4887077 allele T, rs6495314 allele C, and/or rs1996371 allele G is indicative of increased susceptibility to lung cancer.

27. The method of any of the preceding claims, wherein determination of the presence of the marker or haplotype is indicative of a different response rate of the subject to a particular treatment modality for lung cancer.

28. The method of claim 27, wherein the treatement modality is at least one of surgical treatment, radiation treatment, targeted drug therapy and chemotherapy.

29. A method of determining a susceptibility to lung cancer, the method comprising:

obtaining sequence data about a human individual identifying at least one allele of at least one polymorphic marker, wherein different alleles of the at least one polymorphic marker are associated with different susceptibilities to lung cancer in humans, and

determining a susceptibility to lung cancer from the nucleic acid sequence data,

wherein the at least one polymorphic marker is selected from the group consisting of rs1051730, and markers in linkage disequilibirium therewith.

30. The method of claim 29, comprising obtaining sequence data about at least two polymorphic markers.

31. The method of claim 29 or claim 30, wherein determination of a susceptibility comprises comparing the sequence data to a database containing correlation data between the polymorphic markers and susceptibility to lung cancer.

32. The method of claim 31, wherein the database comprises at least one risk measure of susceptibility to lung cancer for the at least one polymorphic marker.

33. The method of claim 32, wherein the database comprises a look-up table containing at least one risk measure of lung cancer for the polymorphic markers.

34. The method of any one of the claims 29 - 33, wherein the sequence data is nucleic acid sequence data.

35. The method of claim 34, wherein obtaining nucleic acid sequence data comprises obtaining a biological sample from the human individual and analyzing sequence of the at least one polymorphic marker in nucleic acid in the sample.

36. The method of claim 34 or claim 35, wherein analyzing sequence of the at least one polymorphic marker comprises determining the presence or absence of at least one allele of the at least one polymorphic marker.

37. The method of any one of the claims 29 – 33, wherein the sequence data is amino acid sequence data.

38. The method of any one of claims 29 - 37, wherein the obtaining sequence data comprises obtaining sequence information from a preexisting record.

39. The method of any one of the claims 29 - 38, further comprising reporting the susceptibility to at least one entity selected from the group consisting of the individual, a guardian of the individual, a genetic service provider, a physician, a medical organization, and a medical insurer.

40. The method of any one of the claims 29 – 39, wherein the at least one polymorphic marker is selected from the group consisting of rs1051730, and markers in linkage disequilibrium therewith.

41. The method of claim 40, wherein the at least one polymorphic marker is selected from the group consisting of the markers listed in Table 4.

42. The method of claim 40 or claim 41, wherein the at least one polymorphic marker is rs1051730.

43. The method of claims 40 or claim 41, wherein the at least one polymorphic marker is rs16969968.

44. A method of identification of a marker for use in assessing susceptibility to lung cancer, the method comprising

a. identifying at least one polymorphic marker within the C15 LD block, or at least one polymorphic marker in linkage disequilibrium therewith;

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

45. The method of Claim 44, wherein the at least one polymorphic marker is in linkage disequilibrium with at least one marker selected from the group consisting of rs1051730, rs680244, rs1948, rs8192475 and rs569207.

46. The method of claim 44 or claim 45, wherein the at least one polymorphic marker is in linkage disequilibrium with marker rs1051730.

47. The method of any of the claims 44 - 46, 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, lung 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 lung cancer.

48. The method of any of the claims 44 - 47, 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, lung 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, lung cancer.

49. A method of assessing an individual for probability of response to a therapeutic agent for preventing and/or ameliorating symptoms associated with lung cancer, 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 a marker within the C15 LD block that is associated with risk of

lung cancer in humans, or a marker in linkage disequilibrium therewith, and 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.

50. A method of predicting prognosis of an individual diagnosed with lung 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 a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein the presence of the at least one allele is indicative of a worse prognosis of lung cancer in the individual.

51. A method of monitoring progress of a treatment of an individual undergoing treatment for lung 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 is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and wherein the presence of the at least one allele is indicative of the treatment outcome of the individual.

52. The method of any of the Claims 49 - 51, wherein the at least one polymorphic marker is selected from the group consisting of rs10896450, marker rs11228565, marker rs7947353 and marker rs10943605, and markers in linkage disequilibrium therewith.

53. The method of any one of the preceding Claims, further comprising analyzing non-genetic information from the individual to make risk assessment, diagnosis, or prognosis of the individual.

54. The method of Claim 53, wherein the non-genetic information is selected from age, gender, ethnicity, socioeconomic status, smoking history, medical history, family history of lung cancer, biochemical measurements, and clinical measurements.

55. The method of any one of the preceding claims, further comprising assessing the presence or absence of at least one additional genetic risk factor for lung cancer in the individual.

56. The method of any of the Claims 53 - 55, further comprising calculating overall risk.

57. A kit for assessing susceptibility to lung cancer in a human individual, the kit comprising

(i) 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 a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and

(ii) a collection of data comprising correlation data between the polymorphic markers assessed by the kit and susceptibility to lung cancer.

58. The kit of claim 57, wherein the at least one polymorphic marker is selected from the group consisting of rs1051730, and markers in linkage disequilibrium therewith.

59. The kit of Claim 57 or claim 58, wherein the at least one marker is selected from the group consisting of the markers set forth in Table 4.

60. The kit of any one of the Claims 57 - 59, wherein the reagents comprise at least one contiguous oligonucleotide that is capable of hybridizing to a fragment of the genome of the individual comprising the at least one polymorphic marker, a buffer and a detectable label.

61. The kit of any of the Claims 57 - 60, 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.

62. The kit of claim 60 or 61, wherein the at least one oligonucleotide is completely complementary to the genome of the individual.

63. The kit of any one of the claims 60 - 62, wherein the oligonucleotide is about 18 to about 50 nucleotides in length.

64. The kit of any of the Claims 60 - 63, wherein the oligonucleotide is 20-30 nucleotides in length.

65. Use of an oligonucleotide probe in the manufacture of a diagnostic reagent for assessing susceptibility to lung cancer in a human individual, wherein the probe is capable of hybridizing selectively to a segment of a nucleic acid with sequence as set forth in SEQ ID NO:1 that comprises at least one polymorphic site, and wherein the probe is 15-500 nucleotides in length.

66. The use according to Claim 65, wherein the polymorphic site is selected from the group consisting of marker rs1051730, and markers in linkage disequilibrium therewith.

67. The use of Claim 65 or 66, wherein the polymorphic site is selected from the markers set forth in Table 4.

68. A computer-readable medium having computer executable instructions for determining susceptibility to lung cancer in a human individual, the computer readable medium comprising:

data indicative of at least one polymorphic marker;

a routine stored on the computer readable medium and adapted to be executed by a processor to determine risk of developing lung cancer in an individual for the at least one polymorphic marker;

wherein the at least one polymorphic marker is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith.

69. The computer readable medium according to claim 68, wherein the computer readable medium contains data indicative of at least two polymorphic markers.

70. The computer readable medium according to claim 68 or claim 69, wherein the data indicative of at least one polymorphic marker comprises parameters indicative of susceptibility to lung cancer for the at least one polymorphic marker, and wherein risk of developing lung cancer in an individual is based on the allelic status for the at least one polymorphic marker in the individual.

73. The computer readable medium according to any one of the claims 68 – 70, wherein said data indicative of at least one polymorphic marker comprises data indicative of the allelic status of said at least one polymorphic marker in the individual.

74. The computer readable medium of any one of the claims 68 – 73, wherein said routine is adapted to receive input data indicative of the allelic status of said at least one polymorphic marker in said individual.

75. The computer readable medium of any one of the claims 68 - 74, wherein the at least one polymorphic marker is selected from the group consisting of the markers set forth in Table 4.

76. The computer readable medium of any one of claims 68 - 75, comprising data indicative of at least one haplotype comprising two or more polymorphic markers.

77. An apparatus for determining a genetic indicator for lung cancer in a human individual, comprising:

a processor

a computer readable memory having computer executable instructions adapted to be executed on the processor to analyze marker and/or haplotype information for at least one human individual with respect to at least one polymorphic marker, wherein the polymorphic marker is a marker within the C15 LD block that is associated with risk of lung cancer in humans, or a marker in linkage disequilibrium therewith, and

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

78. The apparatus according to Claim 77, wherein the computer readable memory further comprises data indicative of the frequency of at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of individuals diagnosed with lung cancer, and data indicative of the frequency of at the least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference individuals, and wherein a risk measure is based on a comparison of the at least one marker and/or haplotype status for the human individual to the data indicative of the frequency of the at least one marker and/or haplotype information for the plurality of individuals diagnosed with lung cancer.

79. The apparatus according to Claim 77, wherein the computer readable memory further comprises data indicative of the risk of developing lung cancer associated with at least one allele of at least one polymorphic marker or at least one haplotype, and wherein a risk measure for the human individual is based on a comparison of the at least one marker and/or haplotype status for the human individual to the risk of lung cancer associated with the at least one allele of the at least one polymorphic marker or the at least one haplotype.

80. The apparatus according to Claim 77, wherein the computer readable memory further comprises data indicative of the frequency of at least one allele of at least one polymorphic marker or at least one haplotype in a plurality of individuals diagnosed with lung cancer, and data indicative of the frequency of at the least one allele of at least one polymorphic marker or at least one haplotype in a plurality of reference

individuals, and wherein risk of developing lung cancer is based on a comparison of the frequency of the at least one allele or haplotype in individuals diagnosed with lung cancer and reference individuals.

81. The apparatus according to any one of claims 77 - 80, wherein the at least one marker or haplotype comprises at least one marker selected from the group consisting of the markers set forth in Table 4.

82. The apparatus according to any one of the Claims 77 - 81, wherein the risk measure is characterized by an Odds Ratio (OR) or a Relative Risk (RR).

83. 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'|.

84. 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 0.1.

85. 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.

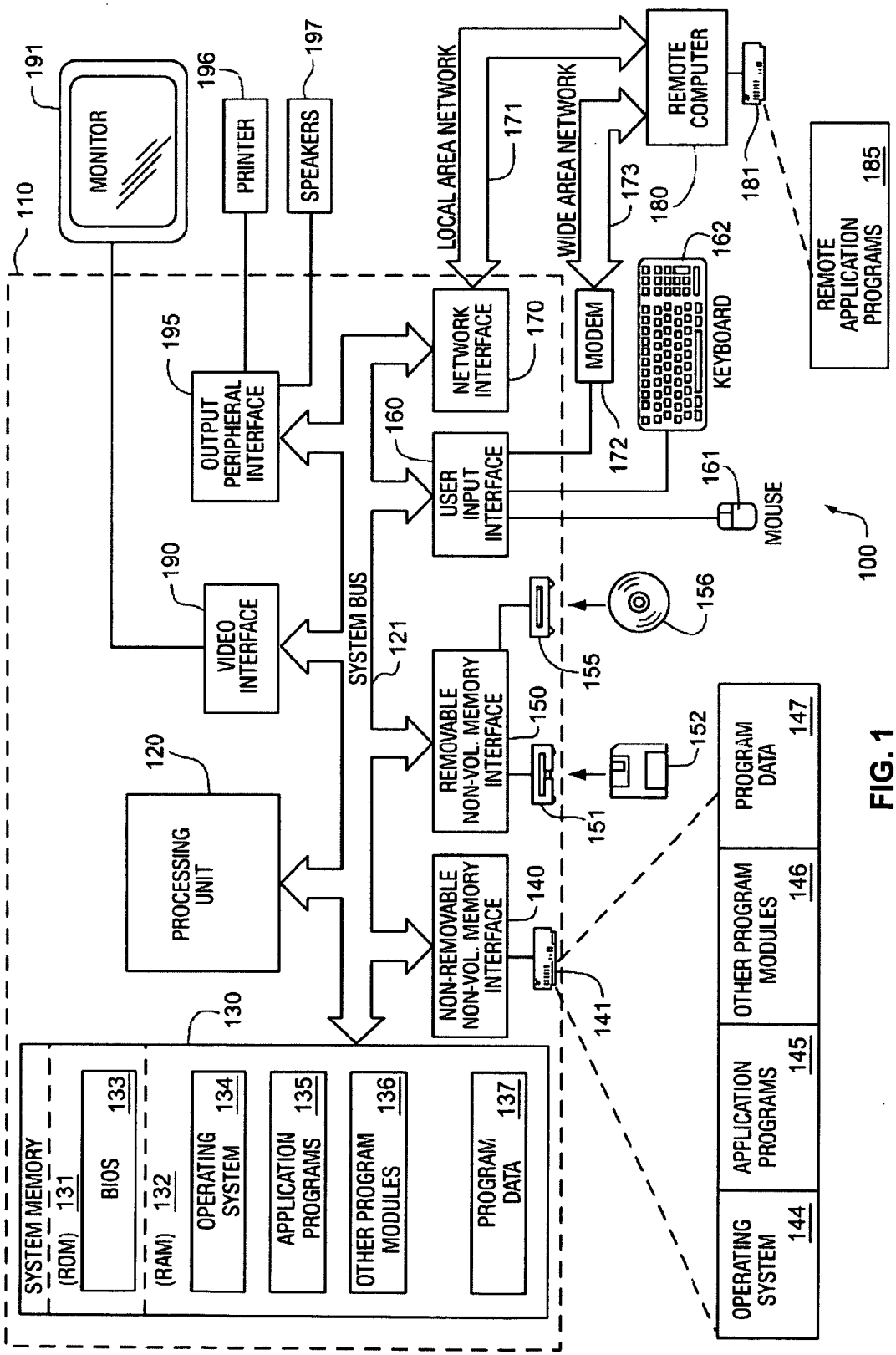
100
110
120
121
130
131
132
133
134
135
136
137
140
141
144
145
146
147
150
151
152
155
156
160
161
162
170
171
172
173
180
181
185
190
191
195
196
197
SYSTEM MEMORY
(ROM)
BIOS
(RAM)
OPERATING SYSTEM
APPLICATION PROGRAMS
OTHER PROGRAM MODULES
PROGRAM DATA
PROCESSING UNIT
NON-REMOVABLE NON-VOL. MEMORY INTERFACE
REMOVABLE NON-VOL. MEMORY INTERFACE
SYSTEM BUS
VIDEO INTERFACE
OUTPUT PERIPHERAL INTERFACE
USER INPUT INTERFACE
NETWORK INTERFACE
MODEM
LOCAL AREA NETWORK
WIDE AREA NETWORK
REMOTE COMPUTER
MONITOR
PRINTER
SPEAKERS
KEYBOARD
MOUSE
REMOTE APPLICATION PROGRAMS
OPERATING SYSTEM
APPLICATION PROGRAMS
OTHER PROGRAM MODULES
PROGRAM DATA

FIG. 1

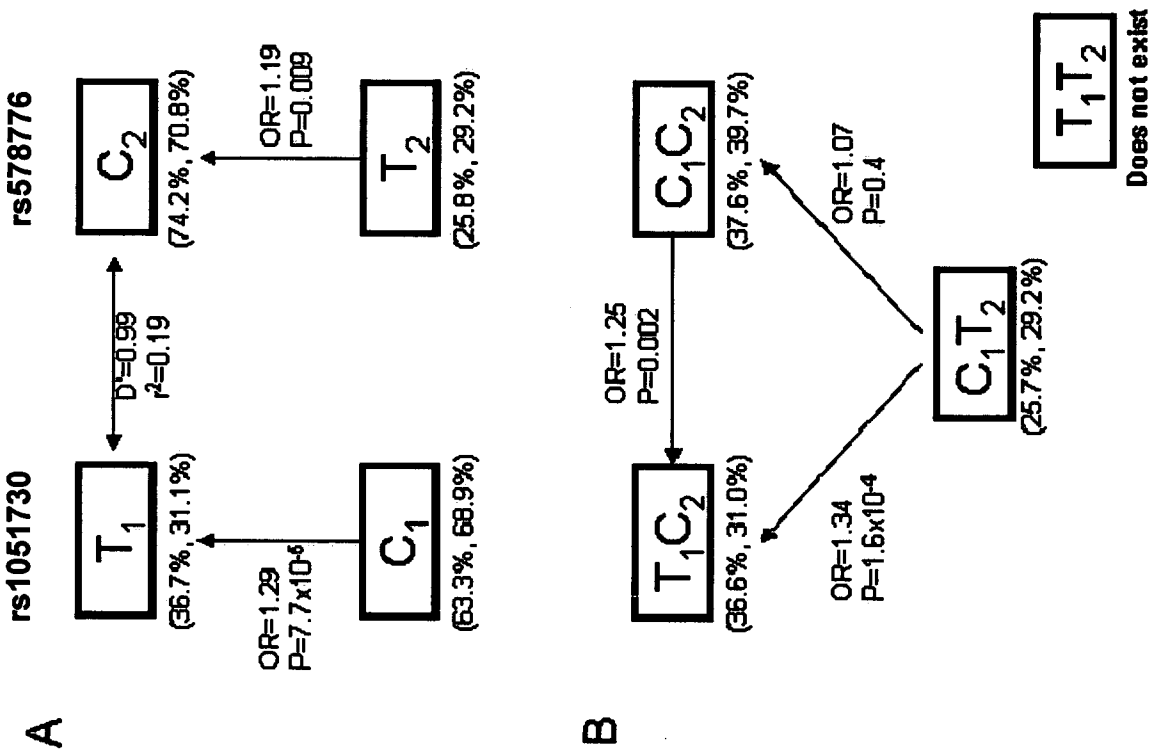
A
rs1051730
rs578776
T1
(36.7%, 31.1%)
C1
(63.3%, 68.9%)
C2
(74.2%, 70.8%)
T2
(25.8%, 29.2%)
D'=0.99
r²=0.19
OR=1.29
P=7.7x10⁻⁶
OR=1.19
P=0.009
B
T1C2
(36.6%, 31.0%)
C1C2
(37.6%, 39.7%)
C1T2
(25.7%, 29.2%)
T1T2
Does not exist
OR=1.25
P=0.002
OR=1.07
P=0.4
OR=1.34
P=1.6x10⁻⁴

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No
PCT/IS2009/000001

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SACCONE SCOTT F ET AL: "Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs" HUMAN MOLECULAR GENETICS, OXFORD UNIVERSITY PRESS, SURREY, vol. 16, no. 1, 1 January 2007 (2007-01-01), pages 36-49, XP002482274 ISSN: 0964-6906 [retrieved on 2006-11-29] table 2 page 41, column 1, paragraph 1 ----- -/--	1-85

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 April 2009	28/04/2009

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Helliot, Bertrand

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IS2009/000001

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.b of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application and necessary to the claimed invention, the international search was carried out on the basis of:

   a. type of material

   [X] a sequence listing

   [ ] table(s) related to the sequence listing

   b. format of material

   [X] on paper

   [X] in electronic form

   c. time of filing/furnishing

   [X] contained in the international application as filed

   [X] filed together with the international application in electronic form

   [ ] furnished subsequently to this Authority for the purpose of search

2. [ ] In addition, in the case that more than one version or copy of a sequence listing and/or table relating thereto has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No
PCT/IS2009/000001

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	LESSOV-SCHLAGGAR ET AL: "Genetics of nicotine dependence and pharmacotherapy" BIOCHEMICAL PHARMACOLOGY, PERGAMON, OXFORD, GB, vol. 75, no. 1, 12 December 2007 (2007-12-12), pages 178-195, XP022387004 ISSN: 0006-2952 page 179, column 1, paragraph 2 the whole document	1-85
P,X	HUNG RAYJEAN J ET AL: "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25" NATURE, NATURE PUBLISHING GROUP, LONDON, UK, vol. 452, no. 7187, 1 April 2008 (2008-04-01), pages 633-637, XP002482276 ISSN: 0028-0836 figure 2 the whole document	1-85
P,X	AMOS CHRISTOPHER I ET AL: "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1" NATURE GENETICS, NATURE PUBLISHING GROUP, US, vol. 40, no. 5, 1 May 2008 (2008-05-01), pages 616-622, XP009100741 ISSN: 1546-1718 table 3 the whole document	1-85
P,X	THORGEIRSSON THORGEIR E ET AL: "A variant associated with nicotine dependence, lung cancer and peripheral arterial disease" NATURE, NATURE PUBLISHING GROUP, LONDON, UK, vol. 452, no. 7187, 3 April 2008 (2008-04-03), pages 638-642, XP002482277 ISSN: 0028-0836 table 4 the whole document	1-85