(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 June 2003 (19.06.2003)

(21) International Application Number: PCT/IB02/04811

(22) International Filing Date: 29 October 2002 (29.10.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:


(72) Inventors; and


(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SI, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: BIALLELIC MARKERS OF D-AMINO ACID OXIDASE AND USES THEREOF

(57) Abstract: The invention concerns the human DAO gene, polynucleotides, and biallelic markers. The invention also concerns the association established between schizophrenia and the biallelic markers. The invention provides means to determine the predisposition of individuals to schizophrenia or related CNS disorder, as well as means for the disease diagnosis and prognosis.
BI Allelic MARKERS OF D-AMINO ACID OXIDASE AND USES THEREOF

FIELD OF THE INVENTION

The present invention is in the field of pharmacogenomics, and is primarily directed to bi allelic markers that are located in or in the vicinity of the D-amino acid oxidase (DAO) gene and the uses of these markers. The present invention encompasses methods of establishing associations between these markers and central nervous system disorders such as schizophrenia and other mood related disorders. The present invention also provides means to determine the predisposition of individuals to said disease as well as means for the diagnosis of such diseases and for the prognosis/detection of an eventual treatment response to agents acting on the leukotriene pathway.

RELATED APPLICATION INFORMATION

This application claims priority on United States provisional patent application Serial No. 60/340,400, filed December 12, 2001, entitled "BI Allelic markers of D-amino acid oxidase and uses thereof".

BACKGROUND OF THE INVENTION

Advances in the technological armamentarium available to basic and clinical investigators have enabled increasingly sophisticated studies of brain and nervous system function in health and disease. Numerous hypotheses both neurobiological and pharmacological have been advanced with respect to the neurochemical and genetic mechanisms involved in central nervous system (CNS) disorders, including psychiatric disorders and neurodegenerative diseases. However, CNS disorders have complex and poorly understood etiologies, as well as symptoms that are overlapping, poorly characterized, and difficult to measure. As a result future treatment regimes and drug development efforts will be required to be more sophisticated and focused on multigenic causes, and will need new assays to segment disease populations, and provide more accurate diagnostic and prognostic information on patients suffering from CNS disorders.

Neurological Basis of CNS Disorders

Neurotransmitters serve as signal transmitters throughout the body. Diseases that affect neurotransmission can therefore have serious consequences. For example, for over 30 years the leading theory to explain the biological basis of many psychiatric disorders such as depression has been the monoamine hypothesis. This theory proposes that depression is partially due to a deficiency in one of the three main biogenic monoamines, namely dopamine, norepinephrine and/or serotonin. In addition to the monoamine hypothesis, numerous arguments tend to show the value in taking into account the overall function of the brain and no longer only considering a single neuronal system. In this context, the value of dual specific actions on the central aminergic systems including second and third messenger systems has now emerged.

Endocrine Basis of CNS Disorders

It is furthermore apparent that the main monoamine systems, namely dopamine,
norepinephrine and serotonin, do not completely explain the pathophysiology of many CNS disorders. In particular, it is clear that CNS disorders may have an endocrine component; the hypothalamic-pituitary-adrenal (HPA) axis, including the effects of corticotrophin-releasing factor and glucocorticoids, plays an important role in the pathophysiology of CNS disorders. In the hypothalamus-pituitary-adrenal (HPA) axis, the hypothalamus lies at the top of the hierarchy regulating hormone secretion. It manufactures and releases peptides (small chains of amino acids) that act on the pituitary, at the base of the brain, stimulating or inhibiting the pituitary's release of various hormones into the blood. These hormones, among them growth hormone, thyroid-stimulating hormone and adrenocorticotrophic hormone (ACTH), control the release of other hormones from target glands. In addition to functioning outside the nervous system, the hormones released in response to pituitary hormones also feed back to the pituitary and hypothalamus. There they deliver inhibitory signals that serve to limit excess hormone biosynthesis.

**CNS Disorders**

Neurotransmitter and hormonal abnormalities are implicated in disorders of movement (e.g. Parkinson's disease, Huntington's disease, motor neuron disease, etc.), disorders of mood (e.g. unipolar depression, bipolar disorder, anxiety, etc.) and diseases involving the intellect (e.g. Alzheimer's disease, Lewy body dementia, schizophrenia, etc.). In addition, these systems have been implicated in many other disorders, such as coma, head injury, cerebral infarction, epilepsy, alcoholism and the mental retardation states of metabolic origin seen particularly in childhood.

**Genetic Analysis of Complex Traits**

Until recently, the identification of genes linked with detectable traits has relied mainly on a statistical approach called linkage analysis. Linkage analysis is based upon establishing a correlation between the transmission of genetic markers and that of a specific trait throughout generations within a family. Linkage analysis involves the study of families with multiple affected individuals and is useful in the detection of inherited traits, which are caused by a single gene, or possibly a very small number of genes. But, linkage studies have proven difficult when applied to complex genetic traits. Most traits of medical relevance do not follow simple Mendelian monogenic inheritance. However, complex diseases often aggregate in families, which suggests that there is a genetic component to be found. Such complex traits are often due to the combined action of multiple genes as well as environmental factors. Such complex trait, include susceptibilities to heart disease, hypertension, diabetes, cancer and inflammatory diseases. Drug efficacy, response and tolerance/toxicity can also be considered as multifactorial traits involving a genetic component in the same way as complex diseases. Linkage analysis cannot be applied to the study of such traits for which no large informative families are available. Moreover, because of their low penetrance, such complex traits do not segregate in a clear cut Mendelian manner as they are passed from one generation to the next. Attempts to map such diseases have been plagued by inconclusive results, demonstrating the need for more sophisticated genetic tools. Knowledge of genetic variation in the
neuronal and endocrine systems is important for understanding why some people are more susceptible to disease or respond differently to treatments. Ways to identify genetic polymorphism and to analyze how they impact and predict disease susceptibility and response to treatment are needed. Although the genes involved in the neuronal and endocrine systems represent major drug targets and are of high relevance to pharmaceutical research, we still have scant knowledge concerning the extent and nature of sequence variation in these genes and their regulatory elements. In the case where polymorphisms have been identified the relevance of the variation is rarely understood. While polymorphisms hold promise for use as genetic markers in determining which genes contribute to multigenic or quantitative traits, suitable markers and suitable methods for exploiting those markers have not been found and brought to bare on the genes related to disorders of the brain and nervous system. The basis for accomplishment of these goals is to use genetic association analysis to detect markers that predict susceptibility for these traits. Recently, advances in the fields of genetics and molecular biology have allowed identification of forms, or alleles, of human genes that lead to diseases. Most of the genetic variations responsible for human diseases identified so far, belong to the class of single gene disorders. As this name implies, the development of single gene disorders is determined, or largely influenced, by the alleles of a single gene. The alleles that cause these disorders are, in general, highly deleterious (and highly penetrant) to individuals who carry them. Therefore, these alleles and their associated diseases, with some exceptions, tend to be very rare in the human population. In contrast, most common diseases and non-disease traits, such as a physiological response to a pharmaceutical agent, can be viewed as the result of many complex factors. These can include environmental exposures (toxins, allergens, infectious agents, climate, and trauma) as well as multiple genetic factors. Association studies seek to analyze the distributions of chromosomes that have occurred in populations of unrelated (at least not directly related) individuals. An assumption in this type of study is that genetic alleles that result in susceptibility for a common trait arose by ancient mutational events on chromosomes that have been passed down through many generations in the population. These alleles can become common throughout the population in part because the trait they influence, if deleterious, is only expressed in a fraction of those individuals who carry them. Identification of these “ancestral” chromosomes is made difficult by the fact that genetic markers are likely to have become separated from the trait susceptibility allele through the process of recombination, except in regions of DNA which immediately surround the allele. The identities of genetic markers contained within the fragments of DNA surrounding a susceptibility allele will be the same as those from the ancestral chromosome on which the allele arose. Therefore, individuals from the population who express a complex trait might be expected to carry the same set of genetic markers in the vicinity of a susceptibility allele more often than those who do not express the trait; that is these markers will show an association with the trait.

Schizophrenia
Schizophrenia is one of the most severe and debilitating of the major psychiatric diseases. It usually starts in late adolescence or early adult life and often becomes chronic and disabling. Men and women are at equal risk of developing this illness; however, most males become ill between 16 and 25 years old, while females develop symptoms between 25 and 30. People with schizophrenia often experience both “positive” symptoms (e.g., delusions, hallucinations, disorganized thinking, and agitation) and “negative” symptoms (e.g., lack of drive or initiative, social withdrawal, apathy, and emotional unresponsiveness). Schizophrenia affects 1% of the world population. There are an estimated 45 million people with schizophrenia in the world, with more than 33 million of them in the developing countries. This disease places a heavy burden on the patient's family and relatives, both in terms of the direct and indirect costs involved and the social stigma associated with the illness, sometimes over generations. Such stigma often leads to isolation and neglect. Moreover, schizophrenia accounts for one fourth of all mental health costs and takes up one in three psychiatric hospital beds. Most schizophrenia patients are never able to work. The cost of schizophrenia to society is enormous. In the United States, for example, the direct cost of treatment of schizophrenia has been estimated to be close to 0.5% of the gross national product. Standardized mortality ratios (SMRs) for schizophrenic patients are estimated to be two to four times higher than the general population, and their life expectancy overall is 20% shorter than for the general population. The most common cause of death among schizophrenic patients is suicide (in 10% of patients) which represents a 20 times higher risk than for the general population. Deaths from heart disease and from diseases of the respiratory and digestive system are also increased among schizophrenic patients.

Bipolar Disorder

Bipolar disorders are relatively common disorders with severe and potentially disabling effects. In addition to the severe effects on patients' social development, suicide completion rates among bipolar patients are reported to be about 15%. Bipolar disorders are characterized by phases of excitement and often including depression; the excitement phases, referred to as mania or hypomania, and depression can alternate or occur in various admixtures, and can occur to different degrees of severity and over varying time periods. Because bipolar disorders can exist in different forms and display different symptoms, the classification of bipolar disorder has been the subject of extensive studies resulting in the definition of bipolar disorder subtypes and widening of the overall concept to include patients previously thought to be suffering from different disorders. Bipolar disorders often share certain clinical signs, symptoms, treatments and neurobiological features with psychotic illnesses in general and therefore present a challenge to the psychiatrist to make an accurate diagnosis. Furthermore, because the course of bipolar disorders and various mood and psychotic disorders can differ greatly, it is critical to characterize the illness as early as possible in order to offer means to manage the illness over a long term. Bipolar disorders appear in about 1.3% of the population and have been reported to constitute about half of the mood disorders seen in
a psychiatric clinic. Bipolar disorders have been found to vary with gender depending of the type of disorder; for example, bipolar disorder I is found equally among men and women, while bipolar disorder II is reportedly more common in women. The age of onset of bipolar disorders is typically in the teenage years and diagnosis is typically made in the patient's early twenties. Bipolar disorders also occur among the elderly, generally as a result of a medical or neurological disorder. The costs of bipolar disorders to society are enormous. The mania associated with the disease impairs performance and causes psychosis, and often results in hospitalization. This disease places a heavy burden on the patient's family and relatives, both in terms of the direct and indirect costs involved and the social stigma associated with the illness, sometimes over generations. Such stigma often leads to isolation and neglect. Furthermore, the earlier the onset, the more severe are the effects of interrupted education and social development. The DSM-IV classification of bipolar disorder distinguishes among four types of disorders based on the degree and duration of mania or hypomania as well as two types of disorders which are evident typically with medical conditions or their treatments, or to substance abuse. Mania is recognized by elevated, expansive or irritable mood as well as by distractibility, impulsive behavior, increased activity, grandiosity, elation, racing thoughts, and pressured speech. Of the four types of bipolar disorder characterized by the particular degree and duration of mania, DSM-IV includes: - bipolar disorder I, including patients displaying mania for at least one week; - bipolar disorder II, including patients displaying hypomania for at least 4 days, characterized by milder symptoms of excitement than mania, who have not previously displayed mania, and have previously suffered from episodes of major depression; - bipolar disorder not otherwise specified (NOS), including patients otherwise displaying features of bipolar disorder II but not meeting the 4 day duration for the excitement phase, or who display hypomania without an episode of major depression; and - cyclothymia, including patients who show numerous manic and depressive symptoms that do not meet the criteria for hypomania or major depression, but which are displayed for over two years without a symptom-free interval of more than two months. The remaining two types of bipolar disorder as classified in DSM-VI are disorders evident or caused by various medical disorder and their treatments, and disorders involving or related to substance abuse. Medical disorders which can cause bipolar disorders typically include endocrine disorders and cerebrovascular injuries, and medical treatments causing bipolar disorder are known to include glucocorticoids and the abuse of stimulants. The disorder associated with the use or abuse of a substance is referred to as "substance induced mood disorder with manic or mixed features". Diagnosis of bipolar disorder can be very challenging. One particularly troublesome difficulty is that some patients exhibit mixed states, simultaneously manic and dysphoric or depressive, but do not fall into the DSM-IV classification because not all required criteria for mania and major depression are met daily for at least one week. Other difficulties include classification of patients in the DSM-IV groups based on duration of phase since patients often cycle between excited and depressive episodes at different rates. In particular, it is reported that the use of antidepressants
may alter the course of the disease for the worse by causing "rapid-cycling". Also making diagnosis more difficult is the fact that bipolar patients, particularly at what is known as Stage III mania, share symptoms of disorganized thinking and behavior with bipolar disorder patients. Furthermore, psychiatrists must distinguish between agitated depression and mixed mania; it is common that patients with major depression (14 days or more) exhibit agitation, resulting in bipolar-like features. A yet further complicating factor is that bipolar patients have an exceptionally high rate of substance, particularly alcohol abuse. While the prevalence of mania in alcoholic patients is low, it is well known that substance abusers can show excited symptoms. Difficulties therefore result for the diagnosis of bipolar patients with substance abuse.

Treatment

As there are currently no cures for bipolar disorder or schizophrenia, the objective of treatment is to reduce the severity of the symptoms, if possible to the point of remission. Due to the similarities in symptoms, schizophrenia and bipolar disorder are often treated with some of the same medicaments. Both diseases are often treated with antipsychotics and neuroleptics. For schizophrenia, for example, antipsychotic medications are the most common and most valuable treatments. There are four main classes of antipsychotic drugs which are commonly prescribed for schizophrenia. The first, neuroleptics, exemplified by chlorpromazine (Thorazine), has revolutionized the treatment of schizophrenic patients by reducing positive (psychotic) symptoms and preventing their recurrence. Patients receiving chlorpromazine have been able to leave mental hospitals and live in community programs or their own homes. But these drugs are far from ideal. Some 20% to 30% of patients do not respond to them at all, and others eventually relapse. These drugs were named neuroleptics because they produce serious neurological side effects, including rigidity and tremors in the arms and legs, muscle spasms, abnormal body movements, and akathisia (restless pacing and fidgeting). These side effects are so troublesome that many patients simply refuse to take the drugs. Besides, neuroleptics do not improve the so-called negative symptoms of schizophrenia and the side effects may even exacerbate these symptoms. Thus, despite the clear beneficial effects of neuroleptics, even some patients who have a good short-term response will ultimately deteriorate in overall functioning. The well known deficiencies in the standard neuroleptics have stimulated a search for new treatments and have led to a new class of drugs termed atypical neuroleptics. The first atypical neuroleptic, Clozapine, is effective for about one third of patients who do not respond to standard neuroleptics. It seems to reduce negative as well as positive symptoms, or at least exacerbates negative symptoms less than standard neuroleptics do. Moreover, it has beneficial effects on overall functioning and may reduce the chance of suicide in schizophrenic patients. It does not produce the troubling neurological symptoms of the standard neuroleptics, or raise blood levels of the hormone prolactin, excess of which may cause menstrual irregularities and infertility in women, impotence or breast enlargement in men. Many patients who cannot tolerate standard neuroleptics have been able to take clozapine. However, clozapine has serious limitations.
It was originally withdrawn from the market because it can cause agranulocytosis, a potentially lethal inability to produce white blood cells. Agranulocytosis remains a threat that requires careful monitoring and periodic blood tests. Clozapine can also cause seizures and other disturbing side effects (e.g., drowsiness, lowered blood pressure, drooling, bed-wetting, and weight gain). Thus it is usually taken only by patients who do not respond to other drugs. Researchers have developed a third class of antipsychotic drugs that have the virtues of clozapine without its defects. One of these drugs is risperidone (Risperdal). Early studies suggest that it is as effective as standard neuroleptic drugs for positive symptoms and may be somewhat more effective for negative symptoms. It produces more neurological side effects than clozapine but fewer than standard neuroleptics.

However, it raises prolactin levels. Risperidone is now prescribed for a broad range of psychotic patients, and many clinicians seem to use it before clozapine for patients who do not respond to standard drugs, because they regard it as safer. Another new drug is Olanzapine (Zyprexa) which is at least as effective as standard drugs for positive symptoms and more effective for negative symptoms. It has few neurological side effects at ordinary clinical doses, and it does not significantly raise prolactin levels. Although it does not produce most of clozapine's most troubling side effects, including agranulocytosis, some patients taking olanzapine may become sedated or dizzy, develop dry mouth, or gain weight. In rare cases, liver function tests become transiently abnormal. Outcome studies in schizophrenia are usually based on hospital treatment studies and may not be representative of the population of schizophrenia patients. At the extremes of outcome, 20% of patients seem to recover completely after one episode of psychosis, whereas 14-19% of patients develop a chronic unremitting psychosis and never fully recover. In general, clinical outcome at five years seems to follow the rule of thirds: with about 35% of patients in the poor outcome category; 36% in the good outcome category, and the remainder with intermediate outcome. Prognosis in schizophrenia does not seem to worsen after five years. Whatever the reasons, there is increasing evidence that leaving schizophrenia untreated for long periods early in course of the illness may negatively affect the outcome. However, the use of drugs is often delayed for patients experiencing a first episode of the illness. The patients may not realize that they are ill, or they may be afraid to seek help; family members sometimes hope the problem will simply disappear or cannot persuade the patient to seek treatment; clinicians may hesitate to prescribe antipsychotic medications when the diagnosis is uncertain because of potential side effects. Indeed, at the first manifestation of the disease, schizophrenia is difficult to distinguish from bipolar manic-depressive disorders, severe depression, drug-related disorders, and stress-related disorders. Since the optimum treatments differ among these diseases, the long term prognosis of the disorder also differs the beginning of the treatment. For both schizophrenia and bipolar disorder, all the known molecules used for the treatment of schizophrenia have side effects and act only against the symptoms of the disease. There is a strong need for new molecules without associated side effects and directed against targets which are involved in the causal mechanisms of schizophrenia and bipolar disorder.
Therefore, tools facilitating the discovery and characterization of these targets are necessary and useful. Schizophrenia and bipolar disorder are now considered to be brain diseases, and emphasis is placed on biological determinants in researching the conditions. In the case of schizophrenia, neuroimaging and neuropathological studies have shown evidence of brain abnormalities in schizophrenic patients. The timing of these pathological changes is unclear but are likely to be a defect in early brain development. Profound changes have also occurred in hypotheses concerning neurotransmitter abnormalities in schizophrenia. The dopamine hypothesis has been extensively revised and is no longer considered as a primary causative model. The aggregation of schizophrenia and bipolar disorder in families, the evidence from twin and adoption studies, and the lack of variation in incidence worldwide, indicate that schizophrenia and bipolar disorder are primarily genetic conditions, although environmental risk factors are also involved at some level as necessary, sufficient, or interactive causes. For example, schizophrenia occurs in 1% of the general population. But, if there is one grandparent with schizophrenia, the risk of getting the illness increases to about 3%; one parent with Schizophrenia, to about 10%. When both parents have schizophrenia, the risk rises to approximately 40%. Consequently, there is a strong need to identify genes involved in schizophrenia and bipolar disorder. The knowledge of these genes will allow researchers to understand the etiology of schizophrenia and bipolar disorder and could lead to drugs and medications which are directed against the cause of the diseases, not just against their symptoms. There is also a great need for new methods for detecting a susceptibility to schizophrenia and bipolar disorder, as well as for preventing or following up the development of the disease. Diagnostic tools could also prove extremely useful. Indeed, early identification of subjects at risk of developing schizophrenia would enable early and/or prophylactic treatment to be administered. Moreover, accurate assessments of the eventual efficacy of a medicament as well as the patient's eventual tolerance to it may enable clinicians to enhance the benefit/risk ratio of schizophrenia and bipolar disorder treatment regimes.

**SUMMARY OF THE INVENTION**

The present invention stems from the identification of novel polymorphisms including biallelic markers associated with the DAO gene and from the identification of genetic associations between alleles of biallelic markers of the DAO gene and disease, as confirmed and characterized in a panel of human subjects. The present invention is based on the discovery of a set of novel biallelic markers of the D-amino acid oxidase gene. Furthermore, association studies have correlated alleles of these biallelic markers to CNS disorders, specifically schizophrenia. The position of these markers and knowledge of the surrounding sequence has been used to design polynucleotide compositions which are useful in determining the identity of nucleotides at the marker position, as well as more complex association and haplotyping studies which are useful in determining the genetic basis for disease states involving amino acid metabolism. In addition, the markers can be
used in methods of the invention to determine whether an individual is at risk for developing schizophrenia or any trait, to identify targets for the development of pharmaceutical agents and diagnostic methods, as well as the characterization of the differential efficacious responses to and side effects from said pharmaceutical agents.

Furthermore, an object of the invention consists of recombinant vectors comprising any of the nucleic acid sequences described in the present invention, and in particular of recombinant vectors comprising the promoter region of DAO or a sequence encoding the DAO enzyme, as well as cell hosts comprising said nucleic acid sequences or recombinant vectors. The invention is also directed to biallelic markers that are located within the DAO genomic sequence (SEQ ID NO:1), these biallelic markers representing useful tools in order to identify a statistically significant association between specific alleles of the DAO gene and one or several CNS disorders, particularly schizophrenia.

The present invention pertains to nucleic acid molecules comprising the genomic sequences of novel human genes encoding sbg1, g34665, sbg2, g35017 and g35018 proteins, proteins encoded thereby, as well as antibodies thereto. The sbg1, g34665, sbg2, g35017 and g35018 genomic sequences may also comprise regulatory sequence located upstream (5'-end) and downstream (3'-end) of the transcribed portion of said gene, these regulatory sequences being also part of the invention. The invention also deals with the cDNA sequence encoding the sbg1 and g35018 proteins.

Oligonucleotide probes or primers hybridizing specifically with a sbg1, g34665, sbg2, g35017 or g35018 genomic or cDNA sequence are also part of the present invention, as well as DNA amplification and detection methods using said primers and probes.

A further object of the invention consists of recombinant vectors comprising any of the nucleic acid sequences described above, and in particular of recombinant vectors comprising a sbg1, g34665, sbg2, g35017 or g35018 regulatory sequence or a sequence encoding a sbg1, g34665, sbg2, g35017 or g35018 protein, as well as of cell hosts and transgenic non human animals comprising said nucleic acid sequences or recombinant vectors.

The invention also concerns to biallelic markers of the sbg1, g34665, sbg2, g35017 or g35018 gene and the use thereof. Included are probes and primers for use in genotyping biallelic markers of the invention.

An embodiment of the invention encompasses any polynucleotide of the invention attached to a solid support polynucleotide may comprise a sequence disclosed in the present specification; optionally, said polynucleotide may comprise, consist of, or consist essentially of any polynucleotide described in the present specification; optionally, said determining may be performed in a hybridization assay, sequencing assay, microsequencing assay, or an enzyme-based mismatch detection assay; optionally, said polynucleotide may be attached to a solid support, array, or addressable array; optionally, said polynucleotide may be labeled.
A further preferred embodiment of the invention is directed to methods of using the DAO biallelic markers of the invention in forensic analyses, particularly as chromosomal markers or in DNA fingerprinting, in forensic procedures to identify individuals, or in diagnostic procedures to identify individuals having a genetic disease.

Finally, the invention is directed to drug screening assays and methods for the screening of substances for the treatment of schizophrenia, bipolar disorder or a related CNS disorder based on the role of DAO nucleotides and polynucleotides in disease.

As noted above, certain aspects of the present invention stem from the identification of genetic associations between schizophrenia and alleles of biallelic markers located in the DAO gene. The invention provides appropriate tools for establishing further genetic associations between alleles of biallelic markers in the DAO gene and either side effects or benefit resulting from the administration of agents acting on schizophrenia or other CNS disorder, or schizophrenia or other CNS disorder symptoms, including agents like chlorpromazine, clozapine, risperidone, olanzapine, sertindole, quetiapine and ziprasidone.

The invention provides appropriate tools for establishing further genetic associations between alleles of biallelic markers in the DAO gene and a trait. Methods and products are provided for the molecular detection of a genetic susceptibility in humans to schizophrenia, bipolar disorder, or other CNS disorder. They can be used for diagnosis, staging, prognosis and monitoring of this disease, which processes can be further included within treatment approaches. The invention also provides for the efficient design and evaluation of suitable therapeutic solutions including individualized strategies for optimizing drug usage, and screening of potential new medicament candidates.

Additional embodiments are set forth in the Detailed Description of the Invention and in the Examples.

**BRIEF DESCRIPTION OF THE SEQUENCES PROVIDED IN THE SEQUENCE LISTING**

SEQ ID NO:1 genomic sequence of D-amino acid oxidase with locations of biallelic markers of the invention.

SEQ ID NOs:2 and 3 cDNA and polypeptide sequence of human DAO, respectively.

SEQ ID NO:4 polynucleotides comprising biallelic marker 27/1-61 located outside the genomic sequence of SEQ ID NO:1.

SEQ ID NOs:5 and 6 cDNA and protein sequence of human DAO, respectively.

SEQ ID NO:7 and 8 cDNAs of human D-aspartate oxidase (DDO).

SEQ ID NO:9 human DDO polypeptide sequence encoded by polynucleotides of SEQ ID NO:7.

SEQ ID NO:10 human DDO polypeptide sequence encoded by polynucleotides of SEQ ID
NO:8.

SEQ ID NO:11 47-mer polynucleotides comprising biallelic marker 27-81-180.
SEQ ID NO:12 47-mer polynucleotides comprising biallelic marker 27-30-249.
SEQ ID NO:13 47-mer polynucleotides comprising biallelic marker 27-2-106.
SEQ ID NO:14 47-mer polynucleotides comprising biallelic marker 27-29-224.
SEQ ID NO:15 47-mer polynucleotides comprising biallelic marker 27-1-61.

In accordance with the regulations relating to Sequence Listings, the following codes have been used in the Sequence Listing to indicate the locations of biallelic markers within the sequences and to identify each of the alleles present at the polymorphic base. The code “r” in the sequences indicates that one allele of the polymorphic base is a guanine, while the other allele is an adenine. The code “y” in the sequences indicates that one allele of the polymorphic base is a thymine, while the other allele is a cytosine. The code “m” in the sequences indicates that one allele of the polymorphic base is an adenine, while the other allele is a cytosine. The code “k” in the sequences indicates that one allele of the polymorphic base is a guanine, while the other allele is a thymine. The code “s” in the sequences indicates that one allele of the polymorphic base is a guanine, while the other allele is a cytosine. The code “w” in the sequences indicates that one allele of the polymorphic base is an adenine, while the other allele is an thymine.

**DETAILED DESCRIPTION OF THE INVENTION**

The identification of genes involved in a particular trait such as a specific central nervous system disorder, like schizophrenia, can be carried out through two main strategies currently used for genetic mapping: linkage analysis and association studies. Linkage analysis requires the study of families with multiple affected individuals and is now useful in the detection of mono- or oligogenic inherited traits. Conversely, association studies examine the frequency of marker alleles in unrelated trait (T+) individuals compared with trait negative (T-) controls, and are generally employed in the detection of polygenic inheritance. The methodology used to validate genetic markers, such as biallelic markers, and perform association studies to correlate a genotype at one or more markers with a trait or a haplotype of two or more markers with a trait have been previously detailed in a related US Patent Application 09/539,333 and an International Application PCT/IB00/00435, both filed March 30, 2000, which disclosures are hereby incorporated by reference in their entireties.

Genetic link or “linkage” is based on an analysis of which of two neighboring sequences on a chromosome contains the least recombinations by crossing-over during meiosis. Using this technique, it has been possible to localize several genes demonstrating a genetic predisposition to a trait. However, linkage analysis is limited by its reliance on the choice of a genetic model suitable for each studied trait. Furthermore, the resolution attainable using linkage analysis is limited, and complementary studies are required to refine the analysis of the typical 20 Mb regions initially
identified through this method. In addition, linkage analysis have proven difficult when applied to complex genetic traits, such as those due to the combined action of multiple genes and/or environmental factors. In such cases, too great an effort and cost are needed to recruit the adequate number of affected families required for applying linkage analysis to these situations. Finally, linkage analysis cannot be applied to the study of traits for which no large informative families are available.

In the present invention alternative means for conducting association studies rather than linkage analysis between markers of the DAO gene and a trait, preferably schizophrenia or bipolar disorder, are disclosed.

In the present application, novel biallelic markers of the DAO gene are disclosed. Further, biallelic markers of the DAO gene associated with schizophrenia are disclosed. The identification of these biallelic markers in association with schizophrenia can allow for the further definition of the chromosomal region suspected of containing a genetic determinant involved in a predisposition to develop schizophrenia and can result in the identification of novel gene sequences which are associated with a predisposition to develop schizophrenia. Additionally, the sequence information provides a resource for the further identification of new genes in that region. Additionally, the sequences comprising the the schizophrenia-associated genes are useful, for example, for the isolation of other genes in putative gene families, the identification of homologs from other species, treatment of disease and as probes and primers for diagnostic or screening assays as described herein.

These identified polymorphisms are used in the design of assays for the reliable detection of genetic susceptibility to schizophrenia and bipolar disorder. They can also be used in the design of drug screening protocols to provide an accurate and efficient evaluation of the therapeutic and side-effect potential of new or already existing medicament or treatment regime.

**Definitions**

As used interchangeably herein, the term “oligonucleotides”, and “polynucleotides” include RNA, DNA, or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form. The term “nucleotide” as used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. The term “nucleotide” is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. Although the term “nucleotide” is also used herein to encompass “modified nucleotides” which comprise at least one modifications (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar, for examples of analogous linking
groups, purine, pyrimidines, and sugars see for example PCT publication No. WO 95/04064. However, the polynucleotides of the invention are preferably comprised of greater than 50% conventional deoxyribose nucleotides, and most preferably greater than 90% conventional deoxyribose nucleotides. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, ex vivo generation, or a combination thereof, as well as utilizing any purification methods known in the art.

The term “purified” is used herein to describe a polynucleotide or polynucleotide vector of the invention which has been separated from other compounds including, but not limited to other nucleic acids, carbohydrates, lipids and proteins (such as the enzymes used in the synthesis of the polynucleotide), or the separation of covalently closed polynucleotides from linear polynucleotides. A polynucleotide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polynucleotide sequence and conformation (linear versus covalently close). A substantially pure polynucleotide typically comprises about 50%, preferably 60 to 90% weight/weight of a nucleic acid sample, more usually about 95%, and preferably is over 99% pure. Polynucleotide purity or homogeneity may be indicated by a number of means well known in the art, such as agarose or polyacrylamide gel electrophoresis of a sample, followed by visualizing a single polynucleotide band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art.

The term “isolated” requires that the material be removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition, and still be isolated in that the vector or composition is not part of its natural environment.

The term “primer” denotes a specific oligonucleotide sequence which is complementary to a target nucleotide sequence and used to hybridize to the target nucleotide sequence. A primer serves as an initiation point for nucleotide polymerization catalyzed by either DNA polymerase, RNA polymerase or reverse transcriptase.

The term “probe” denotes a defined nucleic acid segment (or nucleotide analog segment, e.g., polynucleotide as defined herein) which can be used to identify a specific polynucleotide sequence present in samples, said nucleic acid segment comprising a nucleotide sequence complementary of the specific polynucleotide sequence to be identified.

The terms “trait” and “phenotype” are used interchangeably herein and refer to any clinically distinguishable, detectable or otherwise measurable property of an organism such as symptoms of, or susceptibility to a disease for example. Typically the terms “trait” or “phenotype” are used herein to refer to symptoms of, or susceptibility to schizophrenia or bipolar disorder; or to
refer to an individual's response to an agent acting on schizophrenia or bipolar disorder; or to refer to symptoms of, or susceptibility to side effects to an agent acting on schizophrenia or bipolar disorder.

The term “allele” is used herein to refer to variants of a nucleotide sequence. A biallelic polymorphism has two forms. Typically the first identified allele is designated as the original allele whereas other alleles are designated as alternative alleles. Diploid organisms may be homozygous or heterozygous for an allelic form.

The term “heterozygosity rate” is used herein to refer to the incidence of individuals in a population, which are heterozygous at a particular allele. In a biallelic system the heterozygosity rate is on average equal to 2P_a(1-P_a), where P_a is the frequency of the least common allele. In order to be useful in genetic studies a genetic marker should have an adequate level of heterozygosity to allow a reasonable probability that a randomly selected person will be heterozygous.

The term “genotype” as used herein refers the identity of the alleles present in an individual or a sample. In the context of the present invention a genotype preferably refers to the description of the biallelic marker alleles present in an individual or a sample. The term “genotyping” a sample or an individual for a biallelic marker involves determining the specific allele or the specific nucleotide(s) carried by an individual at a biallelic marker.

The term “mutation” as used herein refers to a difference in DNA sequence between or among different genomes or individuals which has a frequency below 1%.

The term “haplotype” refers to a combination of alleles present in an individual or a sample on a single chromosome. In the context of the present invention a haplotype preferably refers to a combination of biallelic marker alleles found in a given individual and which may be associated with a phenotype.

The term “polymorphism” as used herein refers to the occurrence of two or more alternative genomic sequences or alleles between or among different genomes or individuals. “Polymorphic” refers to the condition in which two or more variants of a specific genomic sequence can be found in a population. A “polymorphic site” is the locus at which the variation occurs. A polymorphism may comprise a substitution, deletion or insertion of one or more nucleotides. A single nucleotide polymorphism is a single base pair change. Typically a single nucleotide polymorphism is the replacement of one nucleotide by another nucleotide at the polymorphic site. Deletion of a single nucleotide or insertion of a single nucleotide, also give rise to single nucleotide polymorphisms. In the context of the present invention “single nucleotide polymorphism” preferably refers to a single nucleotide substitution. Typically, between different genomes or between different individuals, the polymorphic site may be occupied by two different nucleotides.

The terms “biallelic polymorphism” and “biallelic marker” are used interchangeably herein to refer to a polymorphism having two alleles at a fairly high frequency in the population, preferably a single nucleotide polymorphism. A “biallelic marker allele” refers to the nucleotide variants present at a biallelic marker site. Typically the frequency of the less common allele of the biallelic
markers of the present invention has been validated to be greater than 1%, preferably the frequency is greater than 10%, more preferably the frequency is at least 20% (i.e. heterozygosity rate of at least 0.32), even more preferably the frequency is at least 30% (i.e. heterozygosity rate of at least 0.42). A biallelic marker wherein the frequency of the less common allele is 30% or more is termed a "high quality biallelic marker." All of the genotyping, haplotyping, association, and interaction study methods of the invention may optionally be performed solely with high quality biallelic markers.

The location of nucleotides in a polynucleotide with respect to the center of the polynucleotide are described herein in the following manner. When a polynucleotide has an odd number of nucleotides, the nucleotide at an equal distance from the 3' and 5' ends of the polynucleotide is considered to be "at the center" of the polynucleotide, and any nucleotide immediately adjacent to the nucleotide at the center, or the nucleotide at the center itself is considered to be "within 1 nucleotide of the center." With an odd number of nucleotides in a polynucleotide any of the five nucleotides positions in the middle of the polynucleotide would be considered to be within 2 nucleotides of the center, and so on. When a polynucleotide has an even number of nucleotides, there would be a bond and not a nucleotide at the center of the polynucleotide. Thus, either of the two central nucleotides would be considered to be "within 1 nucleotide of the center" and any of the four nucleotides in the middle of the polynucleotide would be considered to be "within 2 nucleotides of the center", and so on. For polymorphisms which involve the substitution, insertion or deletion of 1 or more nucleotides, the polymorphism, allele or biallelic marker is "at the center" of a polynucleotide if the difference between the distance from the substituted, inserted, or deleted polynucleotides of the polymorphism and the 3' end of the polynucleotide, and the distance from the substituted, inserted, or deleted polynucleotides of the polymorphism and the 5' end of the polynucleotide is zero or one nucleotide. If this difference is 0 to 3, then the polymorphism is considered to be "within 1 nucleotide of the center." If the difference is 0 to 5, the polymorphism is considered to be "within 2 nucleotides of the center." If the difference is 0 to 7, the polymorphism is considered to be "within 3 nucleotides of the center," and so on. For polymorphisms which involve the substitution, insertion or deletion of 1 or more nucleotides, the polymorphism, allele or biallelic marker is "at the center" of a polynucleotide if the difference between the distance from the substituted, inserted, or deleted polynucleotides of the polymorphism and the 3' end of the polynucleotide, and the distance from the substituted, inserted, or deleted polynucleotides of the polymorphism and the 5' end of the polynucleotide is zero or one nucleotide. If this difference is 0 to 3, then the polymorphism is considered to be "within 1 nucleotide of the center." If the difference is 0 to 5, the polymorphism is considered to be "within 2 nucleotides of the center." If the difference is 0 to 7, the polymorphism is considered to be "within 3 nucleotides of the center," and so on.

The term "upstream" is used herein to refer to a location which, is toward the 5' end of the polynucleotide from a specific reference point.
The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another by virtue of their sequence identities in a manner like that found in double-helical DNA with thymine or uracil residues linked to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, L., *Biochemistry*, 4th edition, 1995).

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with another specified polynucleotide throughout the entirety of the complementary region. This term is applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind.

As used herein the term "DAO related biallelic marker" relates to a set of biallelic markers in linkage disequilibrium with the DAO gene or a DAO nucleotide sequence. The term DAO related biallelic marker encompasses the biallelic markers disclosed herein, particularly 27-2/106, 27-1/61, 27-81/180, 27-29/224, and 27-30/249 of which nucleotide sequence and polymorphic alleles are described in the sequence listing (SEQ ID NOs:1, 4, and 11-15).

The term "polypeptide" refers to a polymer of amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude prokaryotic modifications of polypeptides, for example, polypeptides which include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems etc.), polypeptides with substituted linkages, as well as other modifications known in the art, both naturally occurring and non-naturally occurring.

The term "purified" is used herein to describe a polypeptide of the invention which has been separated from other compounds including, but not limited to nucleic acids, lipids, carbohydrates and other proteins. A polypeptide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polypeptide sequence. A substantially pure polypeptide typically comprises about 50%, preferably 60 to 90% weight/weight of a protein sample, more usually about 95%, and preferably is over about 99% pure. Polypeptide purity or homogeneity is indicated by a number of means well known in the art, such as agarose or polyacrylamide gel electrophoresis of a sample, followed by visualizing a single polypeptide band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art.

As used herein, the term "non-human animal" refers to any non-human vertebrate, birds and more usually mammals, preferably primates, farm animals such as swine, goats, sheep, donkeys, and horses, rabbits or rodents, more preferably rats or mice. As used herein, the term "animal" is used to
refer to any vertebrate, preferable a mammal. Both the terms “animal” and “mammal” expressly embrace human subjects unless preceded with the term “non-human”.

As used herein, the term “antibody” refers to a polypeptide or group of polypeptides which are comprised of at least one binding domain, where an antibody binding domain is formed from the folding of variable domains of an antibody molecule to form three-dimensional binding spaces with an internal surface shape and charge distribution complementary to the features of an antigenic determinant of an antigen, which allows an immunological reaction with the antigen. Antibodies include recombinant proteins comprising the binding domains, as well as fragments, including Fab, Fab', F(ab)₂, and F(ab')₂ fragments.

As used herein, an “antigenic determinant” is the portion of an antigen molecule, in this case an sbg1 polypeptide, that determines the specificity of the antigen-antibody reaction. An “epitope” refers to an antigenic determinant of a polypeptide. An epitope can comprise as few as 3 amino acids in a spatial conformation which is unique to the epitope. Generally an epitope comprises at least 6 such amino acids, and more usually at least 8-10 such amino acids. Methods for determining the amino acids which make up an epitope include x-ray crystallography, 2-dimensional nuclear magnetic resonance, and epitope mapping e.g. the Pepsan method described by Geysen et al. 1984; PCT Publication No. WO 84/03564; and PCT Publication No. WO 84/03506.

**Stringent Hybridization Conditions**

By way of example and not limitation, procedures using conditions of high stringency are as follows: Prehybridization of filters containing DNA is carried out for 8 h to overnight at 65°C in buffer composed of 6X SSC, 50 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA, and 500 µg/ml denatured salmon sperm DNA. Filters are hybridized for 48 h at 65°C, the preferred hybridization temperature, in prehybridization mixture containing 100 µg/ml denatured salmon sperm DNA and 5-20 X 10⁶ cpm of ³²P-labeled probe. Subsequently, filter washes can be done at 37°C for 1 h in a solution containing 2 X SSC, 0.01% PVP, 0.01% Ficoll, and 0.01% BSA, followed by a wash in 0.1 X SSC at 50°C for 45 min. Following the wash steps, the hybridized probes are detectable by autoradiography. Other conditions of high stringency which may be used are well known in the art and as cited in Sambrook et al., 1989; and Ausubel et al., 1989. These hybridization conditions are suitable for a nucleic acid molecule of about 20 nucleotides in length. There is no need to say that the hybridization conditions described above are to be adapted according to the length of the desired nucleic acid, following techniques well known to the one skilled in the art. The suitable hybridization conditions may for example be adapted according to the teachings disclosed in the book of Hames and Higgins (1985) or in Sambrook et al.(1989).

**Oligonucleotide Probes And Primers**

The polynucleotides of the invention are useful in order to detect the presence of at least a copy of a nucleotide sequence of SEQ ID No. 1 or of the biallelic markers, complement, or variant
thereof in a test sample.

 Particularly preferred probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides, to the extent that said span is consistent with the length of the nucleotide position range, of SEQ ID No 1.

 Probes and primers of the invention also include isolated, purified, or recombinant polynucleotides having at least 70, 75, 80, 85, 90, or 95% nucleotide identity with a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides of nucleotide positions 40939 to 78463 of SEQ ID No. 1. Preferred probes and primers of the invention also include isolated, purified, or recombinant polynucleotides comprising DAO nucleotide sequence having at least 70, 75, 80, 85, 90, or 95% nucleotide identity with at least one sequence selected SEQ ID NO:2 or SEQ ID NO:5. Preferred probes and primers of the invention also include isolated, purified, or recombinant polynucleotides comprising DDO nucleotide sequence having at least 70, 75, 80, 85, 90, or 95% nucleotide identity with at least one sequence selected SEQ ID NO:7 or SEQ ID NO:8.

 Another set of probes and primers of the invention include isolated, purified, or recombinant polynucleotides comprising a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides of SEQ ID No. 1 or the complements thereof, wherein said contiguous span comprises at least 1, 2, 3, 5, or 10 nucleotide positions of any one of the ranges of nucleotide position 41118 to 78451, of SEQ ID No. 1.

 The invention also relates to nucleic acid probes characterized in that they hybridize specifically, under the stringent hybridization conditions defined above, with a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides of nucleotide positions 40939 to 78463 of SEQ ID No. 1, or a variant thereof or a sequence complementary thereto. Particularly preferred are nucleic acid probes characterized in that they hybridize specifically, under the stringent hybridization conditions defined above.

 The formation of stable hybrids depends on the melting temperature (Tm) of the DNA. The Tm depends on the length of the primer or probe, the ionic strength of the solution and the G+C content. The higher the G+C content of the primer or probe, the higher is the melting temperature because G:C pairs are held by three H bonds whereas A:T pairs have only two. The GC content in the probes of the invention usually ranges between 10 and 75 %, preferably between 35 and 60 %, and more preferably between 40 and 55 %.

 A probe or a primer according to the invention may be between 8 and 2000 nucleotides in length, or is specified to be at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 100, 250, 500, 1000 nucleotides in length. More particularly, the length of these probes can range from 8, 10, 15, 20, or 30 to 100 nucleotides, preferably from 10 to 50, more preferably from 15 to 30 nucleotides. Shorter probes tend to lack specificity for a target nucleic acid sequence and generally require cooler
temperatures to form sufficiently stable hybrid complexes with the template. Longer probes are expensive to produce and can sometimes self-hybridize to form hairpin structures. The appropriate length for primers and probes under a particular set of assay conditions may be empirically determined by one of skill in the art.

The primers and probes can be prepared by any suitable method, including, for example, cloning and restriction of appropriate sequences and direct chemical synthesis by a method such as the phosphodiester method of Narang et al. (1979), the phosphodiester method of Brown et al. (1979), the diethylphosphoramidite method of Beaucage et al. (1981) and the solid support method described in EP 0 707 592.

Detection probes are generally nucleic acid sequences or uncharged nucleic acid analogs such as, for example peptide nucleic acids which are disclosed in International Patent Application WO 92/20702, morpholino analogs which are described in U.S. Patents Numbered 5,185,444; 5,034,506 and 5,142,047. The probe may have to be rendered "non-extendable" in that additional dNTPs cannot be added to the probe. In and of themselves analogs usually are non-extendable and nucleic acid probes can be rendered non-extendable by modifying the 3' end of the probe such that the hydroxyl group is no longer capable of participating in elongation. For example, the 3' end of the probe can be functionalized with the capture or detection label to thereby consume or otherwise block the hydroxyl group. Alternatively, the 3' hydroxyl group simply can be cleaved, replaced or modified; U.S. Patent Application Serial No. 07/049,061 filed April 19, 1993, describes modifications which can be used to render a probe non-extendable.

Any of the polynucleotides of the present invention can be labeled, if desired, by incorporating a label detectable by spectroscopic, photochemical, biochemical, immunochemical, or chemical means. For example, useful labels include radioactive substances (\(^{32}\)P, \(^{35}\)S, \(^{3}\)H, \(^{125}\)I), fluorescent dyes (5-bromodesoxyuridin, fluorescein, acetylaminofluorene, digoxigenin) or biotin. Preferably, polynucleotides are labeled at their 3' and 5' ends. Examples of non-radioactive labeling of nucleic acid fragments are described in the French patent No. FR-7810975 or by Urdea et al. (1988) or Sanchez-Pescador et al. (1988). In addition, the probes according to the present invention may have structural characteristics such that they allow the signal amplification, such structural characteristics being, for example, branched DNA probes as those described by Urdea et al. in 1991 or in the European patent No. EP 0 225 807 (Chiron).

A label can also be used to capture the primer, so as to facilitate the immobilization of either the primer or a primer extension product, such as amplified DNA, on a solid support. A capture label is attached to the primers or probes and can be a specific binding member which forms a binding pair with the solid's phase reagent's specific binding member (e.g. biotin and streptavidin). Therefore depending upon the type of label carried by a polynucleotide or a probe, it may be employed to capture or to detect the target DNA. Further, it will be understood that the polynucleotides, primers or probes provided herein, may, themselves, serve as the capture label. For
example, in the case where a solid phase reagent's binding member is a nucleic acid sequence, it may be selected such that it binds a complementary portion of a primer or probe to thereby immobilize the primer or probe to the solid phase. In cases where a polynucleotide probe itself serves as the binding member, those skilled in the art will recognize that the probe will contain a sequence or "tail" that is not complementary to the target. In the case where a polynucleotide primer itself serves as the capture label, at least a portion of the primer will be free to hybridize with a nucleic acid on a solid phase. DNA Labeling techniques are well known to the skilled technician.

The probes of the present invention are useful for a number of purposes. They can be notably used in Southern hybridization to genomic DNA. The probes can also be used to detect PCR amplification products. They may also be used to detect mismatches in a sequence comprising a polynucleotide of SEQ ID Nos 1, 2, 4, 5, 7, 8, and 11-15, or an sbg1, g34665, sbg2, g35017 or g35018 polynucleotide or gene or mRNA using other techniques.

Any of the polynucleotides, primers and probes of the present invention can be conveniently immobilized on a solid support. Solid supports are known to those skilled in the art and include the walls of wells of a reaction tray, test tubes, polystyrene beads, magnetic beads, nitrocellulose strips, membranes, microparticles such as latex particles, sheep (or other animal) red blood cells, duracytes and others. The solid support is not critical and can be selected by one skilled in the art. Thus, latex particles, microparticles, magnetic or non-magnetic beads, membranes, plastic tubes, walls of microtiter wells, glass or silicon chips, sheep (or other suitable animal's) red blood cells and duracytes are all suitable examples. Suitable methods for immobilizing nucleic acids on solid phases include ionic, hydrophobic, covalent interactions and the like. A solid support, as used herein, refers to any material which is insoluble, or can be made insoluble by a subsequent reaction. The solid support can be chosen for its intrinsic ability to attract and immobilize the capture reagent. Alternatively, the solid phase can retain an additional receptor which has the ability to attract and immobilize the capture reagent. The additional receptor can include a charged substance that is oppositely charged with respect to the capture reagent itself or to a charged substance conjugated to the capture reagent. As yet another alternative, the receptor molecule can be any specific binding member which is immobilized upon (attached to) the solid support and which has the ability to immobilize the capture reagent through a specific binding reaction. The receptor molecule enables the indirect binding of the capture reagent to a solid support material before the performance of the assay or during the performance of the assay. The solid phase thus can be a plastic, derivatized plastic, magnetic or non-magnetic metal, glass or silicon surface of a test tube, microtiter well, sheet, bead, microparticle, chip, sheep (or other suitable animal's) red blood cells, duracytes and other configurations known to those of ordinary skill in the art. The polynucleotides of the invention can be attached to or immobilized on a solid support individually or in groups of at least 2, 5, 8, 10, 12, 15, 20, or 25 distinct polynucleotides of the invention to a single solid support. In addition, polynucleotides other than those of the invention may be attached to the same solid support as one or
more polynucleotides of the invention.

Consequently, the invention also comprises a method for detecting the presence of a nucleic acid comprising a nucleotide sequence selected from a group consisting of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11, 12, 13, 14, and 15, a fragment or a variant thereof or a complementary sequence thereto in a sample, said method comprising the following steps of:

a) bringing into contact a nucleic acid probe or a plurality of nucleic acid probes which can hybridize with a nucleotide sequence included in a nucleic acid selected form the group consisting of the nucleotide sequences of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11, 12, 13, 14, and 15, a fragment or a variant thereof or a complementary sequence thereto and the sample to be assayed; and

b) detecting the hybrid complex formed between the probe and a nucleic acid in the sample.

The invention further concerns a kit for detecting the presence of a nucleic acid comprising a nucleotide sequence selected from a group consisting of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11, 12, 13, 14, and 15, a fragment or a variant thereof or a complementary sequence thereto in a sample, said kit comprising:

a) a nucleic acid probe or a plurality of nucleic acid probes which can hybridize with a nucleotide sequence included in a nucleic acid selected form the group consisting of the nucleotide sequences of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11, 12, 13, 14, and 15, a fragment or a variant thereof or a complementary sequence thereto; and

b) optionally, the reagents necessary for performing the hybridization reaction.

In a first preferred embodiment of this detection method and kit, said nucleic acid probe or the plurality of nucleic acid probes are labeled with a detectable molecule. In a second preferred embodiment of said method and kit, said nucleic acid probe or the plurality of nucleic acid probes has been immobilized on a substrate. In a third preferred embodiment, the nucleic acid probe or the plurality of nucleic acid probes comprise either a sequence which is selected from the group consisting of the nucleotide sequences identified in SEQ ID NO: 1 as 27-81.rp, 27-81.pu, 27-29.rp, 27-29.pu, 27-2.rp, 27-2.pu, 27-30.rp, 27-30.pu, 27-81-180.mis, 27-81-180.mis complement, 27-29-224.mis, 27-29-224.mis complement, 27-2-106.mis, 27-2-106.mis complement, 27-30-249.mis, 27-30-249.mis complement, 27-81-180.probe, 27-29-224.probe, 27-2-106.probe, and 27-30-249.probe, or a sequence which is selected from the group consisting of the nucleotide sequences identified in SEQ ID NO: 4 as 27-1-61.probe, 27-1-61.mis, 27-1-61.mis complement, 27-1.pu, and 27-1.rp complement, and the complementary sequences thereto, or a nucleotide sequence comprising a biallelic marker selected from the group consisting of 27-2-106, 27-81-180, 27-29-224, and 27-30-249 as identified in SEQ ID NO: 1 and 27-1-61 as identified in SEQ ID NO: 4 or the complements thereto.

**Biallelic markers of the inventions**

**Advantages of the biallelic markers of the present invention**

The biallelic marker of the inventions of the present invention offer a number of important
advantages over other genetic markers such as RFLP (Restriction fragment length polymorphism) and VNTR (Variable Number of Tandem Repeats) markers.

The first generation of markers, were RFLPs, which are variations that modify the length of a restriction fragment. But methods used to identify and to type RFLPs are relatively wasteful of materials, effort, and time. The second generation of genetic markers were VNTRs, which can be categorized as either minisatellites or microsatellites. Minisatellites are tandemly repeated DNA sequences present in units of 5-50 repeats which are distributed along regions of the human chromosomes ranging from 0.1 to 20 kilobases in length. Since they present many possible alleles, their informative content is very high. Minisatellites are scored by performing Southern blots to identify the number of tandem repeats present in a nucleic acid sample from the individual being tested. However, there are only \(10^4\) potential VNTRs that can be typed by Southern blotting. Moreover, both RFLP and VNTR markers are costly and time-consuming to develop and assay in large numbers.

Single nucleotide polymorphism or biallelic markers can be used in the same manner as RFLPs and VNTRs but offer several advantages. Single nucleotide polymorphisms are densely spaced in the human genome and represent the most frequent type of variation. An estimated number of more than \(10^7\) sites are scattered along the \(3 \times 10^9\) base pairs of the human genome. Therefore, single nucleotide polymorphism occur at a greater frequency and with greater uniformity than RFLP or VNTR markers which means that there is a greater probability that such a marker will be found in close proximity to a genetic locus of interest. Single nucleotide polymorphisms are less variable than VNTR markers but are mutationally more stable.

Also, the different forms of a characterized single nucleotide polymorphism, such as the biallelic markers of the present invention, are often easier to distinguish and can therefore be typed easily on a routine basis. Biallelic markers have single nucleotide based alleles and they have only two common alleles, which allows highly parallel detection and automated scoring. The biallelic markers of the present invention offer the possibility of rapid, high-throughput genotyping of a large number of individuals.

Biallelic markers are densely spaced in the genome, sufficiently informative and can be assayed in large numbers. The combined effects of these advantages make biallelic markers extremely valuable in genetic studies. Biallelic markers can be used in linkage studies in families, in allele sharing methods, in linkage disequilibrium studies in populations, in association studies of case-control populations. An important aspect of the present invention is that biallelic markers allow association studies to be performed to identify genes involved in complex traits. Association studies examine the frequency of marker alleles in unrelated case- and control-populations and are generally employed in the detection of polygenic or sporadic traits. Association studies may be conducted within the general population and are not limited to studies performed on related individuals in affected families (linkage studies). Biallelic markers in different genes can be screened in parallel
for direct association with disease or response to a treatment. This multiple gene approach is a powerful tool for a variety of human genetic studies as it provides the necessary statistical power to examine the synergistic effect of multiple genetic factors on a particular phenotype, drug response, sporadic trait, or disease state with a complex genetic etiology.

Preferred biallelic markers of the present invention are listed in the Sequence listing, specifically 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4, and in Table 1 below. Primer pairs used to amplify the region of the marker in a sample of genomic DNA, amplicons, are indicated in SEQ ID NO:1 and SEQ ID NO:4 by the prefix “.rp” and “.pu complement”. Microsequencing primer pairs used in the methods of genotyping of an allele specifically to determine the base at a particular biallelic marker are indicated in SEQ ID NO:1 and SEQ ID NO:4 by the prefix “.mis” and “.mis complement”.

Table 1:

<table>
<thead>
<tr>
<th>BIALLELIC MARKER</th>
<th>ALLELE1</th>
<th>ALLELE2</th>
<th>POSITION IN SEQ ID NO:1</th>
<th>POSITION IN SEQ ID NO:4</th>
<th>47 MER SEQ ID NO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>27-2-106</td>
<td>C</td>
<td>A</td>
<td>74320</td>
<td>N/A</td>
<td>13</td>
</tr>
<tr>
<td>27-1-61</td>
<td>A</td>
<td>G</td>
<td>N/A</td>
<td>61</td>
<td>15</td>
</tr>
<tr>
<td>27-81-180</td>
<td>G</td>
<td>A</td>
<td>41118</td>
<td>N/A</td>
<td>11</td>
</tr>
<tr>
<td>27-29-224</td>
<td>T</td>
<td>G</td>
<td>69461</td>
<td>N/A</td>
<td>14</td>
</tr>
<tr>
<td>27-30-249</td>
<td>C</td>
<td>T</td>
<td>78451</td>
<td>N/A</td>
<td>12</td>
</tr>
</tbody>
</table>

Polymorphisms, Biallelic Markers And Polynucleotides Comprising Them

In one aspect, the invention concerns biallelic markers associated with schizophrenia. Also included are biallelic markers in linkage disequilibrium with the biallelic markers of the invention.

The polynucleotides of the invention may consist of, consist essentially of, or comprise a contiguous span of nucleotides of a sequence from any of SEQ ID NOs:1, 2, 4, 5, 7, 8, and 11-15 as well as sequences which are complementary thereto ("complements thereof"). The “contiguous span” may be at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 nucleotides in length, to the extent that a contiguous span of these lengths is consistent with the lengths of the particular Sequence ID.

The present invention encompasses polynucleotides for use as primers and probes in the methods of the invention. These polynucleotides may consist of, consist essentially of, or comprise a contiguous span of nucleotides of a sequence from either of SEQ ID Nos. 1 or 4 as well as sequences which are complementary thereto ("complements thereof"). The “contiguous span” may be at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 nucleotides in length, to the extent that a contiguous span of these lengths is consistent with the lengths of the particular Sequence ID. It should be noted that the polynucleotides of the present invention are not
limited to having the exact flanking sequences surrounding the polymorphic bases which, are
enumerated in the Sequence Listing. Rather, it will be appreciated that the flanking sequences
surrounding the biallelic markers and other polymorphisms of the invention, or any of the primers of
probes of the invention which, are more distant from the markers, may be lengthened or shortened to
any extent compatible with their intended use and the present invention specifically contemplates
such sequences. It will be appreciated that the polynucleotides of SEQ ID NOs: 1, 2, 4, 5, 7, 8, and
11-15 may be of any length compatible with their intended use. Also the flanking regions outside of
the contiguous span need not be homologous to native flanking sequences which actually occur in
human subjects. The addition of any nucleotide sequence, which is compatible with the nucleotides
intended use is specifically contemplated. The contiguous span may optionally include the biallelic
markers of the invention in said sequence. Biallelic markers generally comprise a polymorphism at
one single base position. Each biallelic marker therefore corresponds to two forms of a
polynucleotide sequence which, when compared with one another, present a nucleotide modification
at one position. Usually, the nucleotide modification involves the substitution of one nucleotide for
another. Optionally allele 1 or allele 2 of the biallelic markers disclosed in Table 1 or SEQ ID NO:1
or 4 may be specified as being present at the biallelic marker of the invention. The contiguous span
may optionally include a nucleotide at a polymorphism position described in Table 1 or SEQ ID
NO:1 or 4, including single nucleotide substitutions, deletions as well as multiple nucleotide
deletions. The polymorphisms of Table 1 or SEQ ID NO:1 or 4 have been validated as biallelic
markers. Preferred polynucleotides may consist of, consist essentially of, or comprise a contiguous
span of nucleotides of a sequence from SEQ ID NO:1 or 4 as well as sequences which are
complementary thereto. The “contiguous span” may be at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50,
70, 80, 100, 250, 500, 1000 or 2000 nucleotides in length, to the extent that a contiguous span of
these lengths is consistent with the lengths of the particular Sequence ID.

A preferred probe or primer comprises a nucleic acid comprising a polynucleotide selected
from the group of the nucleotide sequences indicated in SEQ ID NO:1 or 4.

The invention also relates to polynucleotides that hybridize, under conditions of high or
intermediate stringency, to a polynucleotide of any of SEQ ID NOs: 1, 2, 4, 5, 7, 8, and 11-15 as well
as sequences, which are complementary thereto. Preferably such polynucleotides are at least 20, 25,
35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 nucleotides in length, to the extent that a
polynucleotide of these lengths is consistent with the lengths of the particular Sequence ID.
Preferred polynucleotides comprise a polymorphism of the invention. Optionally either allele 1 or
allele 2 of the polymorphism disclosed in Table 1 of in the sequence listing may be specified as
being present at the polymorphism of the invention. Particularly preferred polynucleotides comprise
a biallelic marker of the invention. Optionally either allele 1 or allele 2 of the biallelic markers
disclosed in Table 1 or the sequence listing may be specified as being present at the biallelic marker
of the invention. Conditions of high stringency are further described herein.
The primers of the present invention may be designed from the disclosed sequences for any method known in the art. A preferred set of primers is fashioned such that the 3' end of the contiguous span of identity with the sequences of any of SEQ ID NOs. 1 and 4 is present at the 3' end of the primer. Such a configuration allows the 3' end of the primer to hybridize to a selected nucleic acid sequence and dramatically increases the efficiency of the primer for amplification or sequencing reactions. In a preferred set of primers the contiguous span is found in one of the sequences described in the sequence listing (SEQ ID NO:1 and SEQ ID NO:4). Allele specific primers may be designed such that a biallelic marker or other polymorphism of the invention is at the 3' end of the contiguous span and the contiguous span is present at the 3' end of the primer. Such allele specific primers tend to selectively prime an amplification or sequencing reaction so long as they are used with a nucleic acid sample that contains one of the two alleles present at said marker. The 3' end of primer of the invention may be located within or at least 2, 4, 6, 8, 10, 12, 15, 18, 20, 25, 50, 100, 250, 500, or 1000 nucleotides upstream of a biallelic marker of the invention in said sequence or at any other location which is appropriate for their intended use in sequencing, amplification or the location of novel sequences or markers. Primers with their 3' ends located 1 nucleotide upstream of an biallelic marker of the invention have a special utility as microsequencing assays. Preferred microsequencing primers are described in the sequence listing (SEQ ID NO:1 and SEQ ID NO:4).

The probes of the present invention may be designed from the disclosed sequences for any method known in the art, particularly methods which allow for testing if a particular sequence or marker disclosed herein is present. A preferred set of probes may be designed for use in the hybridization assays of the invention in any manner known in the art such that they selectively bind to one allele of a biallelic marker or other polymorphism, but not the other under any particular set of assay conditions. Preferred hybridization probes may consists of, consist essentially of, or comprise a contiguous span which ranges in length from 8, 10, 12, 15, 18 or 20 to 25, 35, 40, 50, 60, 70, or 80 nucleotides, or be specified as being 12, 15, 18, 20, 25, 35, 40, or 50 nucleotides in length and including an biallelic marker or other polymorphism of the invention in said sequence. In a preferred embodiment, either of allele 1 or 2 disclosed in the sequence listing (SEQ ID NO:1 and 4) may be specified as being present at the biallelic marker site. In another preferred embodiment, said biallelic marker may be within 6, 5, 4, 3, 2, or 1 nucleotides of the center of the hybridization probe or at the center of said probe.

In one embodiment the invention encompasses isolated, purified, and recombinant polynucleotides comprising, consisting of, or consisting essentially of a contiguous span of 8 to 50 nucleotides of any one of SEQ ID NOs:1, 2, 4, 5, 7, 8, and 11-15 and the complement thereof, wherein said span includes a polymorphism of the invention; optionally, wherein said polymorphism is a biallelic marker, and the complements thereof, or optionally the biallelic markers in linkage disequilibrium therewith.
In another embodiment the invention encompasses isolated, purified and recombinant polynucleotides comprising, consisting of, or consisting essentially of a contiguous span of 8 to 50 nucleotides of any one of SEQ ID NOs: 1, 2, 4, 5, 7, 8, and 11-15, or the complement thereof, wherein the 3' end of said contiguous span is located at the 3' end of said polynucleotide, and wherein the 3' end of said polynucleotide is located within 20 nucleotides upstream of a biallelic marker of the invention and the complements thereof, or optionally the biallelic markers in linkage disequilibrium therewith. In a further embodiment, the invention encompasses isolated, purified, or recombinant polynucleotides comprising, consisting of, or consisting essentially of a sequence selected from the following sequences: SEQ ID NO: 11-15.

In an additional embodiment, the invention encompasses polynucleotides for use in hybridization assays, sequencing assays, and enzyme-based mismatch detection assays for determining the identity of the nucleotide at a biallelic marker in SEQ ID Nos. 1 or 4 or the complement thereof, as well as polynucleotides for use in amplifying segments of nucleotides comprising a biallelic marker of the invention.

These arrays may generally be produced using mechanical synthesis methods or light directed synthesis methods, which incorporate a combination of photolithographic methods and solid phase oligonucleotide synthesis (Fodor et al., Science, 251:767-777, 1991). The immobilization of arrays of oligonucleotides on solid supports has been rendered possible by the development of a technology generally identified as “Very Large Scale Immobilized Polymer Synthesis” (VLSIPS™) in which, typically, probes are immobilized in a high density array on a solid surface of a chip. Examples of VLSIPS™ technologies are provided in US Patents 5,143,854 and 5,412,087 and in PCT Publications WO 90/15070, WO 92/10092 and WO 95/11995, which describe methods for forming oligonucleotide arrays through techniques such as light-directed synthesis technique. In designing strategies aimed at providing arrays of nucleotides immobilized on solid supports, further presentation strategies were developed to order and display the oligonucleotide arrays on the chips in an attempt to maximize hybridization patterns and sequence information. Examples of such presentation strategies are disclosed in PCT Publications WO 94/12305, WO 94/11530, WO 97/29212 and WO 97/31256.

Oligonucleotide arrays may comprise at least one of the sequences selected from the group consisting of SEQ ID NOs: 1, 2, 4, 5, 7, 8, and 11-15; and the sequences complementary thereto or a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 consecutive nucleotides, to the extent that fragments of these lengths is consistent with the lengths of the particular Sequence ID, for determining whether a sample contains one or more alleles of the biallelic markers of the present invention. Oligonucleotide arrays may also comprise at least one of the sequences selected from the group consisting of SEQ ID NOs: 1, 2, 4, 5, 7, 8, and 11-15; and the sequences complementary thereto or a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 consecutive nucleotides, to the extent that fragments of
these lengths is consistent with the lengths of the particular Sequence ID, for amplifying one or more alleles of the biallelic markers of Table 1 in the sequence listing. In other embodiments, arrays may also comprise at least one of the sequences selected from the group consisting of SEQ ID NOs:1, 2, 4, 5, 7, 8, and 11-15; and the sequences complementary thereto or a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 consecutive nucleotides, to the extent that fragments of these lengths is consistent with the lengths of the particular Sequence ID, for conducting microsequencing analyses to determine whether a sample contains one or more alleles of the biallelic markers of the invention. In still further embodiments, the oligonucleotide array may comprise at least one of the sequences selecting from the group consisting of SEQ ID NOs:1, 2, 4, 5, 7, 8, and 11-15; and the sequences complementary thereto or a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 35, 40, 50, 70, 80, 100, 250, 500, 1000 or 2000 nucleotides in length, to the extent that fragments of these lengths is consistent with the lengths of the particular Sequence ID, for determining whether a sample contains one or more alleles of the polymorphisms and biallelic markers of the present invention.

A further object of the invention relates to an array of nucleic acid sequences comprising either at least one of the sequences selected from the group consisting of amplicons or microsequencing primers as defined above, or the sequences complementary thereto or a fragment thereof of at least 8, 10, 12, 15, 18, 20, 25, 30, or 40 consecutive nucleotides thereof, or at least one sequence comprising at least 1, 2, 3, 4, 5, 10, 20 biallelic markers selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4, or the complements thereof. The invention also pertains to an array of nucleic acid sequences comprising either at least 1, 2, 3, 4, 5, 10, 20 of the sequences selected from the group consisting of amplicons or microsequencing primers as defined above or the sequences complementary thereto or a fragment thereof of at least 8 consecutive nucleotides thereof, or at least two sequences comprising a biallelic marker selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4 or the complements thereto.

The present invention also encompasses diagnostic kits comprising one or more polynucleotides of the invention, optionally with a portion or all of the necessary reagents and instructions for genotyping a test subject by determining the identity of a nucleotide at an biallelic marker of the invention. The polynucleotides of a kit may optionally be attached to a solid support, or be part of an array or addressable array of polynucleotides. The kit may provide for the determination of the identity of the nucleotide at a marker position by any method known in the art including, but not limited to, a sequencing assay method, a microsequencing assay method, a hybridization assay method, or enzyme-based mismatch detection assay. Optionally such a kit may include instructions for scoring the results of the determination with respect to the test subjects' predisposition to schizophrenia, or likely response to an agent acting on schizophrenia, or chances of suffering from side effects to an agent acting on schizophrenia.
Finally, in any embodiments of the present invention, a biallelic marker may optionally comprise:

(a) a biallelic marker selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4, or more preferably a biallelic marker selected from the group consisting of 27-2-106 and 27-29-224 of SEQ ID NO:1;

(b) a biallelic marker selected from the group consisting of 27-2-106 and 27-29-224 of SEQ ID NO:1;

(c) a biallelic marker 27-2-106 of SEQ ID NO:1; or

(d) a biallelic marker 27-29-224 of SEQ ID NO:1.

Optionally, in any of the embodiments described herein, a DAO related biallelic marker may be selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4. Optionally, in any of the embodiments described herein, a DAO related biallelic marker may be selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4. A set of said DAO related biallelic markers may comprise at least 1, 2, 3, 4, 5, 10, 20, 40, 50, 100 or 200 of said biallelic markers, respectively.

Optionally, any of the compositions of methods described herein may specifically exclude at least 1, 2, or 3 biallelic markers.

Furthermore, in any of the embodiments of the present invention, a set of DAO related biallelic markers may comprise at least 1, 2, 3, 4, or 5 of said biallelic markers.

**Methods For De Novo Identification Of Biallelic Markers**

Any of a variety of methods can be used to screen a genomic fragment for single nucleotide polymorphisms such as differential hybridization with oligonucleotide probes, detection of changes in the mobility measured by gel electrophoresis or direct sequencing of the amplified nucleic acid. A preferred method for identifying biallelic markers involves comparative sequencing of genomic DNA fragments from an appropriate number of unrelated individuals.

In a first embodiment, DNA samples from unrelated individuals are pooled together, following which the genomic DNA of interest is amplified and sequenced. The nucleotide sequences thus obtained are then analyzed to identify significant polymorphisms. One of the major advantages of this method resides in the fact that the pooling of the DNA samples substantially reduces the number of DNA amplification reactions and sequencing reactions, which must be carried out. Moreover, this method is sufficiently sensitive so that a biallelic marker obtained thereby usually demonstrates a sufficient frequency of its less common allele to be useful in conducting association studies. Usually, the frequency of the least common allele of a biallelic marker identified by this method is at least 10%.
In a second embodiment, the DNA samples are not pooled and are therefore amplified and sequenced individually. This method is usually preferred when biallelic markers need to be identified in order to perform association studies within candidate genes. Preferably, highly relevant gene regions such as promoter regions or exon regions may be screened for biallelic markers. A biallelic marker obtained using this method may show a lower degree of informativeness for conducting association studies, e.g. if the frequency of its less frequent allele may be less than about 10%. Such a biallelic marker will however be sufficiently informative to conduct association studies and it will further be appreciated that including less informative biallelic markers in the genetic analysis studies of the present invention, may allow in some cases the direct identification of causal mutations, which may, depending on their penetrance, be rare mutations.

The following is a description of the various parameters of a preferred method used by the inventors for the identification of the biallelic markers of the present invention.

**Genomic DNA samples**

The genomic DNA samples from which the biallelic markers of the present invention are generated are preferably obtained from unrelated individuals corresponding to a heterogeneous population of known ethnic background. The number of individuals from whom DNA samples are obtained can vary substantially, preferably from about 10 to about 1000, more preferably from about 50 to about 200 individuals. Usually, DNA samples are collected from at least about 100 individuals in order to have sufficient polymorphic diversity in a given population to identify as many markers as possible and to generate statistically significant results.

As for the source of the genomic DNA to be subjected to analysis, any test sample can be foreseen without any particular limitation. These test samples include biological samples, which can be tested by the methods of the present invention described herein, and include human and animal body fluids such as whole blood, serum, plasma, cerebrospinal fluid, urine, lymph fluids, and various external secretions of the respiratory, intestinal and genitourinary tracts, tears, saliva, milk, white blood cells, myelomas and the like; biological fluids such as cell culture supernatants; fixed tissue specimens including tumor and non-tumor tissue and lymph node tissues; bone marrow aspirates and fixed cell specimens. The preferred source of genomic DNA used in the present invention is from peripheral venous blood of each donor. Techniques to prepare genomic DNA from biological samples are well known to the skilled technician. Details of a preferred embodiment are provided in Example 1. The person skilled in the art can choose to amplify pooled or unpooled DNA samples.

**DNA Amplification**

The identification of biallelic markers in a sample of genomic DNA may be facilitated through the use of DNA amplification methods. DNA samples can be pooled or unpooled for the amplification step. DNA amplification techniques are well known to those skilled in the art. Various methods to amplify DNA fragments carrying biallelic markers are further described.
hereinafter herein. The PCR technology is the preferred amplification technique used to identify new biallelic markers.

In a first embodiment, biallelic markers are identified using genomic sequence information generated by the inventors. Genomic DNA fragments, such as the inserts of the BAC clones described above, are sequenced and used to design primers for the amplification of 500 bp fragments. These 500 bp fragments are amplified from genomic DNA and are scanned for biallelic markers. Primers may be designed using the OSP software (Hillier L. and Green P., 1991). All primers may contain, upstream of the specific target bases, a common oligonucleotide tail that serves as a sequencing primer. Those skilled in the art are familiar with primer extensions, which can be used for these purposes.

In another embodiment of the invention, genomic sequences of candidate genes are available in public databases allowing direct screening for biallelic markers. Preferred primers, useful for the amplification of genomic sequences encoding the candidate genes, focus on promoters, exons and splice sites of the genes. A biallelic marker present in these functional regions of the gene have a higher probability to be a causal mutation.

**Sequencing Of Amplified Genomic DNA And Identification Of Single Nucleotide Polymorphisms**

The amplification products generated as described above, are then sequenced using any method known and available to the skilled technician. Methods for sequencing DNA using either the dideoxy-mediated method (Sanger method) or the Maxam-Gilbert method are widely known to those of ordinary skill in the art. Such methods are for example disclosed in Maniatis et al. (Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Second Edition, 1989). Alternative approaches include hybridization to high-density DNA probe arrays as described in Chee et al. (Science 274, 610, 1996).

Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. The products of the sequencing reactions are run on sequencing gels and the sequences are determined using gel image analysis. The polymorphism search is based on the presence of superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position. Because each dideoxy terminator is labeled with a different fluorescent molecule, the two peaks corresponding to a biallelic site present distinct colors corresponding to two different nucleotides at the same position on the sequence. However, the presence of two peaks can be an artifact due to background noise. To exclude such an artifact, the two DNA strands are sequenced and a comparison between the peaks is carried out. In order to be registered as a polymorphic sequence, the polymorphism has to be detected on both strands.

The above procedure permits those amplification products, which contain biallelic markers to be identified. The detection limit for the frequency of biallelic polymorphisms detected by
sequencing pools of 100 individuals is approximately 0.1 for the minor allele, as verified by sequencing pools of known allelic frequencies. However, more than 90% of the biallelic polymorphisms detected by the pooling method have a frequency for the minor allele higher than 0.25. Therefore, the biallelic markers selected by this method have a frequency of at least 0.1 for the minor allele and less than 0.9 for the major allele. Preferably at least 0.2 for the minor allele and less than 0.8 for the major allele, more preferably at least 0.3 for the minor allele and less than 0.7 for the major allele, thus a heterozygosity rate higher than 0.18, preferably higher than 0.32, more preferably higher than 0.42.

In another embodiment, biallelic markers are detected by sequencing individual DNA samples, the frequency of the minor allele of such a biallelic marker may be less than 0.1.

**Validation of the biallelic markers of the present invention**

The polymorphisms are evaluated for their usefulness as genetic markers by validating that both alleles are present in a population. Validation of the biallelic markers is accomplished by genotyping a group of individuals by a method of the invention and demonstrating that both alleles are present. Microsequencing is a preferred method of genotyping alleles. The validation by genotyping step may be performed on individual samples derived from each individual in the group or by genotyping a pooled sample derived from more than one individual. The group can be as small as one individual if that individual is heterozygous for the allele in question. Preferably the group contains at least three individuals, more preferably the group contains five or six individuals, so that a single validation test will be more likely to result in the validation of more of the biallelic markers that are being tested. It should be noted, however, that when the validation test is performed on a small group it may result in a false negative result if as a result of sampling error none of the individuals tested carries one of the two alleles. Thus, the validation process is less useful in demonstrating that a particular initial result is an artifact, than it is at demonstrating that there is a *bona fide* biallelic marker at a particular position in a sequence. All of the genotyping, haplotyping, association, and interaction study methods of the invention may optionally be performed solely with validated biallelic markers.

**Evaluation of the frequency of the biallelic markers of the present invention**

The validated biallelic markers are further evaluated for their usefulness as genetic markers by determining the frequency of the least common allele at the biallelic marker site. The determination of the least common allele is accomplished by genotyping a group of individuals by a method of the invention and demonstrating that both alleles are present. This determination of frequency by genotyping step may be performed on individual samples derived from each individual in the group or by genotyping a pooled sample derived from more than one individual. The group must be large enough to be representative of the population as a whole. Preferably the group
contains at least 20 individuals, more preferably the group contains at least 50 individuals, most preferably the group contains at least 100 individuals. Of course the larger the group the greater the accuracy of the frequency determination because of reduced sampling error. A biallelic marker wherein the frequency of the less common allele is 30% or more is termed a "high quality biallelic marker." All of the genotyping, haplotyping, association, and interaction study methods of the invention may optionally be performed solely with high quality biallelic markers.

Another embodiment of the invention comprises methods of estimating the frequency of an allele in a population comprising genotyping individuals from said population for a DAO related biallelic marker and determining the proportional representation of said biallelic marker in said population. In addition, the methods of estimating the frequency of an allele in a population encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: Optionally, said DAO related biallelic marker may be in a sequence selected individually or in any combination from the group consisting of SEQ NOs:1, 4, and 11-15; and the complements thereof; optionally, said DAO related biallelic marker may be selected from the biallelic markers described in Table 1; optionally, determining the frequency of a biallelic marker allele in a population may be accomplished by determining the identity of the nucleotides for both copies of said biallelic marker present in the genome of each individual in said population and calculating the proportional representation of said nucleotide at said DAO related biallelic marker for the population; optionally, determining the frequency of a biallelic marker allele in a population may be accomplished by performing a genotyping method on a pooled biological sample derived from a representative number of individuals, or each individual, in said population, and calculating the proportional amount of said nucleotide compared with the total.

**Methods Of Genotyping An Individual For Biallelic Markers**

Methods are provided to genotype a biological sample for one or more biallelic markers of the present invention, all of which may be performed *in vitro*. Such methods of genotyping comprise determining the identity of a nucleotide at a biallelic marker of the invention by any method known in the art. These methods find use in genotyping case-control populations in association studies as well as individuals in the context of detection of alleles of biallelic markers which, are known to be associated with a given trait, in which case both copies of the biallelic marker present in individual's genome are determined so that an individual may be classified as homozygous or heterozygous for a particular allele.

These genotyping methods can be performed nucleic acid samples derived from a single individual or pooled DNA samples.

Genotyping can be performed using similar methods as those described above for the identification of the biallelic markers, or using other genotyping methods such as those further described below. In preferred embodiments, the comparison of sequences of amplified genomic fragments from different individuals is used to identify new biallelic markers whereas
microsequecning is used for genotyping known biallelic markers in diagnostic and association study applications.

Another embodiment of the invention encompasses methods of genotyping a biological sample comprising determining the identity of a nucleotide at a DAO related biallelic marker. In addition, the genotyping methods of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: Optionally, said DAO related biallelic marker may be in a sequence selected individually or in any combination from the group consisting of marker 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4, and the complements thereof; optionally, said DAO related biallelic marker may be selected individually or in any combination from the biallelic markers described in Table 1, SEQ ID NOs:1, 4, or SEQ ID NOs:11-15, optionally, said method further comprises determining the identity of a second nucleotide at said biallelic marker, wherein said first nucleotide and second nucleotide are not base paired (by Watson & Crick base pairing) to one another; optionally, said biological sample is derived from a single individual or subject; optionally, said method is performed in vitro; optionally, said biallelic marker is determined for both copies of said biallelic marker present in said individual's genome; optionally, said biological sample is derived from multiple subjects or individuals; optionally, said method further comprises amplifying a portion of said sequence comprising the biallelic marker prior to said determining step; optionally, wherein said amplifying is performed by PCR, LCR, or replication of a recombinant vector comprising an origin of replication and said portion in a host cell; optionally, wherein said determining is performed by a hybridization assay, sequencing assay, microsequecning assay, or an enzyme-based mismatch detection assay.

**Source of DNA for genotyping**

Any source of nucleic acids, in purified or non-purified form, can be utilized as the starting nucleic acid, provided it contains or is suspected of containing the specific nucleic acid sequence desired. DNA or RNA may be extracted from cells, tissues, body fluids and the like as described herein. While nucleic acids for use in the genotyping methods of the invention can be derived from any mammalian source, the test subjects and individuals from which nucleic acid samples are taken are generally understood to be human.

**Amplification Of DNA Fragments Comprising Biallelic Markers**

Methods and polynucleotides are provided to amplify a segment of nucleotides comprising one or more biallelic marker of the present invention. It will be appreciated that amplification of DNA fragments comprising biallelic markers may be used in various methods and for various purposes and is not restricted to genotyping. Nevertheless, many genotyping methods, although not all, require the previous amplification of the DNA region carrying the biallelic marker of interest. Such methods specifically increase the concentration or total number of sequences that span the biallelic marker or include that site and sequences located either distal or proximal to it. Diagnostic
assays may also rely on amplification of DNA segments carrying a biallelic marker of the present invention.

Amplification of DNA may be achieved by any method known in the art. The established PCR (polymerase chain reaction) method or by developments thereof or alternatives. Amplification methods which can be utilized herein include but are not limited to Ligase Chain Reaction (LCR) as described in EP A 320 308 and EP A 439 182, Gap LCR (Wolcott, M.J.), the so-called "NASBA" or "3SR" technique described in Guatelli J.C. et al. (1990) and in Compton J. (1991), Q-beta amplification as described in EP A 4544 610, strand displacement amplification as described in Walker et al. (1996) and EP A 684 315 and, target mediated amplification as described in PCT Publication WO 9322461.

LCR and Gap LCR are exponential amplification techniques, both depend on DNA ligase to join adjacent primers annealed to a DNA molecule. In Ligase Chain Reaction (LCR), probe pairs are used which include two primary (first and second) and two secondary (third and fourth) probes, all of which are employed in molar excess to target. The first probe hybridizes to a first segment of the target strand and the second probe hybridizes to a second segment of the target strand, the first and second segments being contiguous so that the primary probes abut one another in 5' phosphate-3'hydroxyl relationship, and so that a ligase can covalently fuse or ligate the two probes into a fused product. In addition, a third (secondary) probe can hybridize to a portion of the first probe and a fourth (secondary) probe can hybridize to a portion of the second probe in a similar abutting fashion. Of course, if the target is initially double stranded, the secondary probes also will hybridize to the target complement in the first instance. Once the ligated strand of primary probes is separated from the target strand, it will hybridize with the third and fourth probes which can be ligated to form a complementary, secondary ligated product. It is important to realize that the ligated products are functionally equivalent to either the target or its complement. By repeated cycles of hybridization and ligation, amplification of the target sequence is achieved. A method for multiplex LCR has also been described (WO 9320227). Gap LCR (GLCR) is a version of LCR where the probes are not adjacent but are separated by 2 to 3 bases.

For amplification of mRNAs, it is within the scope of the present invention to reverse transcribe mRNA into cDNA followed by polymerase chain reaction (RT-PCR); or, to use a single enzyme for both steps as described in U.S. Patent No. 5,322,770 or, to use Asymmetric Gap LCR (RT-AGLCR) as described by Marshall R.L. et al. (1994). AGLCR is a modification of GLCR that allows the amplification of RNA.

Some of these amplification methods are particularly suited for the detection of single nucleotide polymorphisms and allow the simultaneous amplification of a target sequence and the identification of the polymorphic nucleotide as it is further described herein.

The PCR technology is the preferred amplification technique used in the present invention. A variety of PCR techniques are familiar to those skilled in the art. For a review of PCR technology,
see Molecular Cloning to Genetic Engineering White, B.A. Ed. (1997) and the publication entitled “PCR Methods and Applications” (1991, Cold Spring Harbor Laboratory Press). In each of these PCR procedures, PCR primers on either side of the nucleic acid sequences to be amplified are added to a suitably prepared nucleic acid sample along with dNTPs and a thermostable polymerase such as Taq polymerase, Pfu polymerase, or Vent polymerase. The nucleic acid in the sample is denatured and the PCR primers are specifically hybridized to complementary nucleic acid sequences in the sample. The hybridized primers are extended. Thereafter, another cycle of denaturation, hybridization, and extension is initiated. The cycles are repeated multiple times to produce an amplified fragment containing the nucleic acid sequence between the primer sites. PCR has further been described in several patents including US Patents 4,683,195, 4,683,202 and 4,965,188.

Primers can be prepared by any suitable method. As for example, direct chemical synthesis by a method such as the phosphodiester method of Narang S.A. et al. (1979), the phosphodiester method of Brown E.L. et al. (1979), the diethylphosphoramidite method of Beaucage et al. (1981) and the solid support method described in EP 0 707 592.

In some embodiments the present invention provides primers for amplifying a DNA fragment containing one or more biallelic markers of the present invention. It will be appreciated that the primers listed are merely exemplary and that any other set of primers which produce amplification products containing one or more biallelic markers of the present invention.

The spacing of the primers determines the length of the segment to be amplified. In the context of the present invention amplified segments carrying biallelic markers can range in size from at least about 25 bp to 35 kbp. Amplification fragments from 25-3000 bp are typical, fragments from 50-1000 bp are preferred and fragments from 100-600 bp are highly preferred. It will be appreciated that amplification primers for the biallelic markers may be any sequence which allow the specific amplification of any DNA fragment carrying the markers. Amplification primers may be labeled or immobilized on a solid support as described in the section titled “Oligonucleotide Probes and Primers”.

**Methods of Genotyping DNA samples for Biallelic Markers**

Any method known in the art can be used to identify the nucleotide present at a biallelic marker site. Since the biallelic marker allele to be detected has been identified and specified in the present invention, detection will prove simple for one of ordinary skill in the art by employing any of a number of techniques. Many genotyping methods require the previous amplification of the DNA region carrying the biallelic marker of interest. While the amplification of target or signal is often preferred at present, ultrasensitive detection methods which do not require amplification are also encompassed by the present genotyping methods. Methods well-known to those skilled in the art that can be used to detect biallelic polymorphisms include methods such as, conventional dot blot analyzes, single strand conformational polymorphism analysis (SSCP) described by Orita et al. (1989), denaturing gradient gel electrophoresis (DGGE), heteroduplex analysis, mismatch cleavage

Preferred methods involve directly determining the identity of the nucleotide present at a biallelic marker site by sequencing assay, enzyme-based mismatch detection assay, or hybridization assay. The following is a description of some preferred methods. A highly preferred method is the microsequencing technique. The term “sequencing assay” is used herein to refer to polymerase extension of duplex primer/template complexes and includes both traditional sequencing and microsequencing.

1) Sequencing assays

The nucleotide present at a polymorphic site can be determined by sequencing methods. In a preferred embodiment, DNA samples are subjected to PCR amplification before sequencing as described above. DNA sequencing methods are described in herein. Preferably, the amplified DNA is subjected to automated dideoxy terminator sequencing reactions using a dye-primer cycle sequencing protocol. Sequence analysis allows the identification of the base present at the biallelic marker site.

2) Microsequencing assays

In microsequencing methods, a nucleotide at the polymorphic site that is unique to one of the alleles in a target DNA is detected by a single nucleotide primer extension reaction. This method involves appropriate microsequencing primers which, hybridize just upstream of a polymorphic base of interest in the target nucleic acid. A polymerase is used to specifically extend the 3' end of the primer with one single ddNTP (chain terminator) complementary to the selected nucleotide at the polymorphic site. Next the identity of the incorporated nucleotide is determined in any suitable way.

Typically, microsequencing reactions are carried out using fluorescent ddNTPs and the extended microsequencing primers are analyzed by electrophoresis on ABI 377 sequencing machines to determine the identity of the incorporated nucleotide as described in EP 412 883. Alternatively capillary electrophoresis can be used in order to process a higher number of assays simultaneously. An example of a typical microsequencing procedure that can be used in the context of the present invention is provided in example 4.

Different approaches can be used to detect the nucleotide added to the microsequencing primer. A homogeneous phase detection method based on fluorescence resonance energy transfer has been described by Chen and Kwok (1997) and Chen et al. (1997). In this method amplified genomic DNA fragments containing polymorphic sites are incubated with a 5'-fluorescein-labeled primer in the presence of allelic dye-labeled dideoxyribonucleoside triphosphates and a modified Taq polymerase. The dye-labeled primer is extended one base by the dye-terminator specific for the allele present on the template. At the end of the genotyping reaction, the fluorescence intensities of
the two dyes in the reaction mixture are analyzed directly without separation or purification. All these steps can be performed in the same tube and the fluorescence changes can be monitored in real time. Alternatively, the extended primer may be analyzed by MALDI-TOF Mass Spectrometry. The base at the polymorphic site is identified by the mass added onto the microsequencing primer (see Haff L.A. and Smirnov L.P., 1997).

Microsequencing may be achieved by the established microsequencing method or by developments or derivatives thereof. Alternative methods include several solid-phase microsequencing techniques. The basic microsequencing protocol is the same as described previously, except that the method is conducted as a heterogenous phase assay, in which the primer or the target molecule is immobilized or captured onto a solid support. To simplify the primer separation and the terminal nucleotide addition analysis, oligonucleotides are attached to solid supports or are modified in such ways that permit affinity separation as well as polymerase extension. The 5' ends and internal nucleotides of synthetic oligonucleotides can be modified in a number of different ways to permit different affinity separation approaches, e.g., biotinylation. If a single affinity group is used on the oligonucleotides, the oligonucleotides can be separated from the incorporated terminator reagent. This eliminates the need of physical or size separation. More than one oligonucleotide can be separated from the terminator reagent and analyzed simultaneously if more than one affinity group is used. This permits the analysis of several nucleic acid species or more nucleic acid sequence information per extension reaction. The affinity group need not be on the priming oligonucleotide but could alternatively be present on the template. For example, immobilization can be carried out via an interaction between biotinylated DNA and streptavidin-coated microtitration wells or avidin-coated polystyrene particles. In the same manner oligonucleotides or templates may be attached to a solid support in a high-density format. In such solid phase microsequencing reactions, incorporated ddNTPs can be radiolabeled (Svännen, 1994) or linked to fluorescein (Livak and Hainer, 1994). The detection of radiolabeled ddNTPs can be achieved through scintillation-based techniques. The detection of fluorescein-linked ddNTPs can be based on the binding of antifluorescein antibody conjugated with alkaline phosphatase, followed by incubation with a chromogenic substrate (such as p-nitrophenyl phosphate). Other possible reporter-detection pairs include: ddNTP linked to dinitrophenyl (DNP) and anti-DNP alkaline phosphatase conjugate (Harju et al., 1993) or biotinylated ddNTP and horseradish peroxidase-conjugated streptavidin with o-phenylenediamine as a substrate (WO 92/15712). As yet another alternative solid-phase microsequencing procedure, Nyren et al. (1993) described a method relying on the detection of DNA polymerase activity by an enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA).

Pastinen et al. (1997), describe a method for multiplex detection of single nucleotide polymorphism in which the solid phase minisequencing principle is applied to an oligonucleotide array format. High-density arrays of DNA probes attached to a solid support (DNA chips) are
further described in herein.

In one aspect the present invention provides polynucleotides and methods to genotype one or more biallelic markers of the present invention by performing a microsequencing assay. Preferred microsequencing primers include those listed in the SEQ ID NO:1 and SEQ ID NO:4 (as described previously as prefix "mis" and "mis complement"). It will be appreciated that the microsequencing primers listed in the sequence listing are merely exemplary and that, any primer having a 3' end immediately adjacent to a polymorphic nucleotide may be used. Similarly, it will be appreciated that microsequencing analysis may be performed for any biallelic marker or any combination of biallelic markers of the present invention. One aspect of the present invention is a solid support which includes one or more microsequencing primers listed in SEQ ID NO:1 and 4, or fragments comprising at least 8, at least 12, at least 15, or at least 20 consecutive nucleotides thereof and having a 3' terminus immediately upstream of the corresponding biallelic marker, for determining the identity of a nucleotide at biallelic marker site.

3) Mismatch detection assays based on polymerases and ligases

In one aspect the present invention provides polynucleotides and methods to determine the allele of one or more biallelic markers of the present invention in a biological sample, by mismatch detection assays based on polymerases and/or ligases. These assays are based on the specificity of polymerases and ligases. Polymerization reactions places particularly stringent requirements on correct base pairing of the 3' end of the amplification primer and the joining of two oligonucleotides hybridized to a target DNA sequence is quite sensitive to mismatches close to the ligation site, especially at the 3' end. The terms "enzyme based mismatch detection assay" are used herein to refer to any method of determining the allele of a biallelic marker based on the specificity of ligases and polymerases. Preferred methods are described below. Methods, primers and various parameters to amplify DNA fragments comprising biallelic markers of the present invention are further described herein.

Allele specific amplification

Discrimination between the two alleles of a biallelic marker can also be achieved by allele specific amplification, a selective strategy, whereby one of the alleles is amplified without amplification of the other allele. This is accomplished by placing a polymorphic base at the 3' end of one of the amplification primers. Because the extension forms from the 3'end of the primer, a mismatch at or near this position has an inhibitory effect on amplification. Therefore, under appropriate amplification conditions, these primers only direct amplification on their complementary allele. Designing the appropriate allele-specific primer and the corresponding assay conditions are well within the ordinary skill in the art.

Ligation/amplification based methods

The “Oligonucleotide Ligation Assay” (OLA) uses two oligonucleotides which are designed to be capable of hybridizing to abutting sequences of a single strand of a target molecule. One of
the oligonucleotides is biotinylated, and the other is detectably labeled. If the precise complementary sequence is found in a target molecule, the oligonucleotides will hybridize such that their termini abut, and create a ligation substrate that can be captured and detected. OLA is capable of detecting biallelic markers and may be advantageously combined with PCR as described by Nickerson D.A. et al. (1990). In this method, PCR is used to achieve the exponential amplification of target DNA, which is then detected using OLA.

Other methods which are particularly suited for the detection of biallelic markers include LCR (ligase chain reaction), Gap LCR (GLCR) which are described herein. As mentioned above, LCR uses two pairs of probes to exponentially amplify a specific target. The sequences of each pair of oligonucleotides, is selected to permit the pair to hybridize to abutting sequences of the same strand of the target. Such hybridization forms a substrate for a template-dependant ligase. In accordance with the present invention, LCR can be performed with oligonucleotides having the proximal and distal sequences of the same strand of a biallelic marker site. In one embodiment, either oligonucleotide will be designed to include the biallelic marker site. In such an embodiment, the reaction conditions are selected such that the oligonucleotides can be ligated together only if the target molecule either contains or lacks the specific nucleotide(s) that is complementary to the biallelic marker on the oligonucleotide. In an alternative embodiment, the oligonucleotides will not include the biallelic marker, such that when they hybridize to the target molecule, a “gap” is created as described in WO 90/01069. his gap is then “filled” with complementary dNTPs (as mediated by DNA polymerase), or by an additional pair of oligonucleotides. Thus at the end of each cycle, each single strand has a complement capable of serving as a target during the next cycle and exponential allele-specific amplification of the desired sequence is obtained.

Ligase/Polymerase-mediated Genetic Bit Analysis™ is another method for determining the identity of a nucleotide at a preselected site in a nucleic acid molecule (WO 95/21271). This method involves the incorporation of a nucleoside triphosphate that is complementary to the nucleotide present at the preselected site onto the terminus of a primer molecule, and their subsequent ligation to a second oligonucleotide. The reaction is monitored by detecting a specific label attached to the reaction’s solid phase or by detection in solution.

4) Hybridization assay methods

A preferred method of determining the identity of the nucleotide present at a biallelic marker site involves nucleic acid hybridization. The hybridization probes, which can be conveniently used in such reactions, preferably include the probes defined herein. Any hybridization assay may be used including Southern hybridization, Northern hybridization, dot blot hybridization and solid-phase hybridization (see Sambrook et al., Molecular Cloning – A Laboratory Manual, Second Edition, Cold Spring Harbor Press, N.Y., 1989).

Hybridization refers to the formation of a duplex structure by two single stranded nucleic acids due to complementary base pairing. Hybridization can occur between exactly complementary
nucleic acid strands or between nucleic acid strands that contain minor regions of mismatch. Specific probes can be designed that hybridize to one form of a biallelic marker and not to the other and therefore are able to discriminate between different allelic forms. Allele-specific probes are often used in pairs, one member of a pair showing perfect match to a target sequence containing the original allele and the other showing a perfect match to the target sequence containing the alternative allele. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Stringent, sequence specific hybridization conditions, under which a probe will hybridize only to the exactly complementary target sequence are well known in the art (Sambrook et al., Molecular Cloning – A Laboratory Manual, Second Edition, Cold Spring Harbor Press, N.Y., 1989). Stringent conditions are sequence dependent and will be different in different circumstances. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. By way of example and not limitation, procedures using conditions of high stringency are as follows: Prehybridization of filters containing DNA is carried out for 8 h to overnight at 65°C in buffer composed of 6X SSC, 50 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.02% PVP, 0.02% Ficoll, 0.02% BSA, and 500 µg/ml denatured salmon sperm DNA. Filters are hybridized for 48 h at 65°C, the preferred hybridization temperature, in prehybridization mixture containing 100 µg/ml denatured salmon sperm DNA and 5-20 X 10^6 cpm of ^32P-labeled probe. Alternatively, the hybridization step can be performed at 65°C in the presence of SSC buffer, 1 x SSC corresponding to 0.15M NaCl and 0.05 M Na citrate. Subsequently, filter washes can be done at 37°C for 1 h in a solution containing 2X SSC, 0.01% PVP, 0.01% Ficoll, and 0.01% BSA, followed by a wash in 0.1X SSC at 50°C for 45 min. Alternatively, filter washes can be performed in a solution containing 2 x SSC and 0.1% SDS, or 0.5 x SSC and 0.1% SDS, or 0.1 x SSC and 0.1% SDS at 68°C for 15 minute intervals. Following the wash steps, the hybridized probes are detectable by autoradiography. By way of example and not limitation, procedures using conditions of intermediate stringency are as follows: Filters containing DNA are prehybridized, and then hybridized at a temperature of 60°C in the presence of a 5 x SSC buffer and labeled probe. Subsequently, filters washes are performed in a solution containing 2x SSC at 50°C and the hybridized probes are detectable by autoradiography. Other conditions of high and intermediate stringency which may be used are well known in the art and as cited in Sambrook et al. (Molecular Cloning - A Laboratory Manual, Second Edition, Cold Spring Harbor Press, N.Y., 1989) and Ausubel et al. (Current Protocols in Molecular Biology, Green Publishing Associates and Wiley Interscience, N.Y., 1989).

Although such hybridizations can be performed in solution, it is preferred to employ a solid-phase hybridization assay. The target DNA comprising a biallelic marker of the present invention may be amplified prior to the hybridization reaction. The presence of a specific allele in the sample is determined by detecting the presence or the absence of stable hybrid duplexes formed between the
probe and the target DNA. The detection of hybrid duplexes can be carried out by a number of methods. Various detection assay formats are well known which utilize detectable labels bound to either the target or the probe to enable detection of the hybrid duplexes. Typically, hybridization duplexes are separated from unhybridized nucleic acids and the labels bound to the duplexes are then detected. Those skilled in the art will recognize that wash steps may be employed to wash away excess target DNA or probe. Standard heterogeneous assay formats are suitable for detecting the hybrids using the labels present on the primers and probes.

Two recently developed assays allow hybridization-based allele discrimination with no need for separations or washes (see Landegren U. et al., 1998). The TaqMan assay takes advantage of the 5' nuclease activity of Taq DNA polymerase to digest a DNA probe annealed specifically to the accumulating amplification product. TaqMan probes are labeled with a donor-acceptor dye pair that interacts via fluorescence energy transfer. Cleavage of the TaqMan probe by the advancing polymerase during amplification dissociates the donor dye from the quenching acceptor dye, greatly increasing the donor fluorescence. All reagents necessary to detect two allelic variants can be assembled at the beginning of the reaction and the results are monitored in real time (see Livak et al., 1995). In an alternative homogeneous hybridization-based procedure, molecular beacons are used for allele discriminations. Molecular beacons are hairpin-shaped oligonucleotide probes that report the presence of specific nucleic acids in homogeneous solutions. When they bind to their targets they undergo a conformational reorganization that restores the fluorescence of an internally quenched fluorophore (Tyagi et al., 1998).

By assaying the hybridization to an allele specific probe, one can detect the presence or absence of a biallelic marker allele in a given sample.

High-Throughput parallel hybridizations in array format are specifically encompassed within “hybridization assays” and are described below.

**Hybridization to addressable arrays of oligonucleotides**

Hybridization assays based on oligonucleotide arrays rely on the differences in hybridization stability of short oligonucleotides to perfectly matched and mismatched target sequence variants. Efficient access to polymorphism information is obtained through a basic structure comprising high-density arrays of oligonucleotide probes attached to a solid support (the chip) at selected positions. Each DNA chip can contain thousands to millions of individual synthetic DNA probes arranged in a grid-like pattern and miniaturized to the size of a dime.

The chip technology has already been applied with success in numerous cases. For example, the screening of mutations has been undertaken in the BRCA1 gene, in *S. cerevisiae* mutant strains, and in the protease gene of HIV-1 virus (Hacia et al., 1996; Shoemaker et al., 1996; Kozal et al., 1996). Chips of various formats for use in detecting biallelic polymorphisms can be produced on a customized basis by Affymetrix (GeneChip™), Hyseq (HyChip and HyGnostics), and Protogene Laboratories.
In general, these methods employ arrays of oligonucleotide probes that are complementary to target nucleic acid sequence segments from an individual which, target sequences include a polymorphic marker. EP785280, describes a tiling strategy for the detection of single nucleotide polymorphisms. Briefly, arrays may generally be “tiled” for a large number of specific polymorphisms. By “tiling” is generally meant the synthesis of a defined set of oligonucleotide probes which is made up of a sequence complementary to the target sequence of interest, as well as preselected variations of that sequence, e.g., substitution of one or more given positions with one or more members of the basis set of monomers, i.e. nucleotides. Tiling strategies are further described in PCT application No. WO 95/11995. In a particular aspect, arrays are tiled for a number of specific, identified biallelic marker sequences. In particular the array is tiled to include a number of detection blocks, each detection block being specific for a specific biallelic marker or a set of biallelic markers. For example, a detection block may be tiled to include a number of probes, which span the sequence segment that includes a specific polymorphism. To ensure probes that are complementary to each allele, the probes are synthesized in pairs differing at the biallelic marker. In addition to the probes differing at the polymorphic base, monosubstituted probes are also generally tiled within the detection block. These monosubstituted probes have bases at and up to a certain number of bases in either direction from the polymorphism, substituted with the remaining nucleotides (selected from A, T, G, C and U). Typically the probes in a tiled detection block will include substitutions of the sequence positions up to and including those that are 5 bases away from the biallelic marker. The monosubstituted probes provide internal controls for the tiled array, to distinguish actual hybridization from artefactual cross-hybridization. Upon completion of hybridization with the target sequence and washing of the array, the array is scanned to determine the position on the array to which the target sequence hybridizes. The hybridization data from the scanned array is then analyzed to identify which allele or alleles of the biallelic marker are present in the sample. Hybridization and scanning may be carried out as described in PCT application No. WO 92/10092 and WO 95/11995 and US patent No. 5,424,186.

Thus, in some embodiments, the chips may comprise an array of nucleic acid sequences of fragments of about 15 nucleotides in length. In further embodiments, the chip may comprise an array including at least one of the sequences selected from the group consisting of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 and the sequences complementary thereto, or a fragment thereof at least about 8 consecutive nucleotides, preferably 10, 15, 20, more preferably 25, 30, 40, 47, or 50 consecutive nucleotides. In some embodiments, the chip may comprise an array of at least 2, 3, 4, 5, 6, 7, 8 or more of these polynucleotides of the invention. Solid supports and polynucleotides of the present invention attached to solid supports are further described in the section titled “Oligonucleotide probes and Primers”.

5) Integrated Systems
Another technique, which may be used to analyze polymorphisms, includes multicomponent integrated systems, which miniaturize and compartmentalize processes such as PCR and capillary electrophoresis reactions in a single functional device. An example of such technique is disclosed in US patent 5,589,136, which describes the integration of PCR amplification and capillary electrophoresis in chips.

Integrated systems can be envisaged mainly when microfluidic systems are used. These systems comprise a pattern of microchannels designed onto a glass, silicon, quartz, or plastic wafer included on a microchip. The movements of the samples are controlled by electric, electroosmotic or hydrostatic forces applied across different areas of the microchip. For genotyping biallelic markers, the microfluidic system may integrate nucleic acid amplification, microsequencing, capillary electrophoresis and a detection method such as laser-induced fluorescence detection.

**Methods Of Genetic Analysis Using The Biallelic Markers Of The Present Invention**

Different methods are available for the genetic analysis of complex traits (see Lander and Schork, 1994). The search for disease-susceptibility genes is conducted using two main methods: the linkage approach in which evidence is sought for cosegregation between a locus and a putative trait locus using family studies, and the association approach in which evidence is sought for a statistically significant association between an allele and a trait or a trait causing allele (Khoury J. et al, 1993). In general, the biallelic markers of the present invention find use in any method known in the art to demonstrate a statistically significant correlation between a genotype and a phenotype. The biallelic markers may be used in parametric and non-parametric linkage analysis methods. Preferably, the biallelic markers of the present invention are used to identify genes associated with detectable traits using association studies, an approach which does not require the use of affected families and which permits the identification of genes associated with complex and sporadic traits.

The genetic analysis using the biallelic markers of the present invention may be conducted on any scale. The whole set of biallelic markers of the present invention or any subset of biallelic markers of the present invention may be used. Further, any set of genetic markers including a biallelic marker of the present invention may be used. As mentioned above, it should be noted that the biallelic markers of the present invention may be included in any complete or partial genetic map of the human genome. These different uses are specifically contemplated in the present invention and claims.

**Linkage analysis**

Linkage analysis is based upon establishing a correlation between the transmission of genetic markers and that of a specific trait throughout generations within a family. Thus, the aim of linkage analysis is to detect marker loci that show cosegregation with a trait of interest in pedigrees.

**Parametric methods**

When data are available from successive generations there is the opportunity to study the
degree of linkage between pairs of loci. Estimates of the recombination fraction enable loci to be ordered and placed onto a genetic map. With loci that are genetic markers, a genetic map can be established, and then the strength of linkage between markers and traits can be calculated and used to indicate the relative positions of markers and genes affecting those traits (Weir, B.S., 1996). The classical method for linkage analysis is the logarithm of odds (lod) score method (see Morton N.E., 1955; Ott J, 1991). Calculation of lod scores requires specification of the mode of inheritance for the disease (parametric method). Generally, the length of the candidate region identified using linkage analysis is between 2 and 20Mb. Once a candidate region is identified as described above, analysis of recombinant individuals using additional markers allows further delineation of the candidate region. Linkage analysis studies have generally relied on the use of a maximum of 5,000 microsatellite markers, thus limiting the maximum theoretical attainable resolution of linkage analysis to about 600 kb on average.

Linkage analysis has been successfully applied to map simple genetic traits that show clear Mendelian inheritance patterns and which have a high penetrance (i.e., the ratio between the number of trait positive carriers of allele $a$ and the total number of $a$ carriers in the population). However, parametric linkage analysis suffers from a variety of drawbacks. First, it is limited by its reliance on the choice of a genetic model suitable for each studied trait. Furthermore, as already mentioned, the resolution attainable using linkage analysis is limited, and complementary studies are required to refine the analysis of the typical 2Mb to 20Mb regions initially identified through linkage analysis. In addition, parametric linkage analysis approaches have proven difficult when applied to complex genetic traits, such as those due to the combined action of multiple genes and/or environmental factors. It is very difficult to model these factors adequately in a lod score analysis. In such cases, too large an effort and cost are needed to recruit the adequate number of affected families required for applying linkage analysis to these situations, as recently discussed by Risch, N. and Merikangas, K. (1996).

**Non-parametric methods**

The advantage of the so-called non-parametric methods for linkage analysis is that they do not require specification of the mode of inheritance for the disease, they tend to be more useful for the analysis of complex traits. In non-parametric methods, one tries to prove that the inheritance pattern of a chromosomal region is not consistent with random Mendelian segregation by showing that affected relatives inherit identical copies of the region more often than expected by chance. Affected relatives should show excess “allele sharing” even in the presence of incomplete penetrance and polygenic inheritance. In non-parametric linkage analysis the degree of agreement at a marker locus in two individuals can be measured either by the number of alleles identical by state (IBS) or by the number of alleles identical by descent (IBD). Affected sib pair analysis is a well-known special case and is the simplest form of these methods.

The biallelic markers of the present invention may be used in both parametric and non-
parametric linkage analysis. Preferably biallelic markers may be used in non-parametric methods which allow the mapping of genes involved in complex traits. The biallelic markers of the present invention may be used in both IBD- and IBS- methods to map genes affecting a complex trait. In such studies, taking advantage of the high density of biallelic markers, several adjacent biallelic marker loci may be pooled to achieve the efficiency attained by multi-allelic markers (Zhao et al., 1998).

However, both parametric and non-parametric linkage analysis methods analyse affected relatives, they tend to be of limited value in the genetic analysis of drug responses or in the analysis of side effects to treatments. This type of analysis is impractical in such cases due to the lack of availability of familial cases. In fact, the likelihood of having more than one individual in a family being exposed to the same drug at the same time is extremely low.

**Population Association Studies**

The present invention comprises methods for identifying one or several genes among a set of candidate genes that are associated with a detectable trait using the biallelic markers of the present invention. In one embodiment the present invention comprises methods to detect an association between a biallelic marker allele or a biallelic marker haplotype and a trait. Further, the invention comprises methods to identify a trait causing allele in linkage disequilibrium with any biallelic marker allele of the present invention.

As described above, alternative approaches can be employed to perform association studies: genome-wide association studies, candidate region association studies and candidate gene association studies. The candidate region analysis clearly provides a short-cut approach to the identification of genes and gene polymorphisms related to a particular trait when some information concerning the biology of the trait is available. Further, the biallelic markers of the present invention may be incorporated in any map of genetic markers of the human genome in order to perform genome-wide association studies. Methods to generate a high-density map of biallelic markers has been described in US Provisional Patent application serial number 60/082,614. The biallelic markers of the present invention may further be incorporated in any map of a specific candidate region of the genome (a specific chromosome or a specific chromosomal segment for example).

As mentioned above, association studies may be conducted within the general population and are not limited to studies performed on related individuals in affected families. Association studies are extremely valuable as they permit the analysis of sporadic or multifactor traits. Moreover, association studies represent a powerful method for fine-scale mapping enabling much finer mapping of trait causing alleles than linkage studies. Studies based on pedigrees often only narrow the location of the trait causing allele. Association studies using the biallelic markers of the present invention can therefore be used to refine the location of a trait causing allele in a candidate region identified by Linkage Analysis methods. Biallelic markers of the present invention can be
used to identify the involved gene; such uses are specifically contemplated in the present invention and claims.

1) Determining the frequency of a biallelic marker allele or of a biallelic marker haplotype in a population

Another embodiment of the present invention encompasses methods of estimating the frequency of a haplotype for a set of biallelic markers in a population, comprising the steps of: a) genotyping each individual in said population for at least one DAO related biallelic marker, b) genotyping each individual in said population for a second biallelic marker by determining the identity of the nucleotides at said second biallelic marker for both copies of said second biallelic marker present in the genome; and c) applying a haplotype determination method to the identities of the nucleotides determined in steps a) and b) to obtain an estimate of said frequency. In addition, the methods of estimating the frequency of a haplotype of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: optionally said haplotype determination method is selected from the group consisting of asymmetric PCR amplification, double PCR amplification of specific alleles, the Clark method, or an expectation maximization algorithm; optionally, said second biallelic marker is a DAO related biallelic marker in a sequence selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4, or SEQ ID NOs:11-15, and the complements thereof; optionally, said DAO related biallelic marker may be selected individually or in any combination from the biallelic markers described in Table 1; optionally, the identity of the nucleotides at the biallelic markers in everyone of the sequences of SEQ ID NOs:1, 4, or 11-15 is determined in steps a) and b).

Association studies explore the relationships among frequencies for sets of alleles between loci.

Determining the frequency of an allele in a population

Allelic frequencies of the biallelic markers in a population can be determined using one of the methods described above under the heading “Methods for genotyping an individual for biallelic markers”, or any genotyping procedure suitable for this intended purpose. Genotyping pooled samples or individual samples can determine the frequency of a biallelic marker allele in a population. One way to reduce the number of genotypings required is to use pooled samples. A major obstacle in using pooled samples is in terms of accuracy and reproducibility for determining accurate DNA concentrations in setting up the pools. Genotyping individual samples provides higher sensitivity, reproducibility and accuracy and; is the preferred method used in the present invention. Preferably, each individual is genotyped separately and simple gene counting is applied to determine the frequency of an allele of a biallelic marker or of a genotype in a given population.

Determining the frequency of a haplotype in a population

The gametic phase of haplotypes is unknown when diploid individuals are heterozygous at
more than one locus. Using genealogical information in families gametic phase can sometimes be inferred (Perlin et al., 1994). When no genealogical information is available different strategies may be used. One possibility is that the multiple-site heterozygous diploids can be eliminated from the analysis, keeping only the homozygotes and the single-site heterozygote individuals, but this approach might lead to a possible bias in the sample composition and the underestimation of low-frequency haplotypes. Another possibility is that single chromosomes can be studied independently, for example, by asymmetric PCR amplification (see Newton et al., 1989; Wu et al., 1989) or by isolation of single chromosome by limit dilution followed by PCR amplification (see Ruano et al., 1990). Further, a sample may be haplotyped for sufficiently close biallelic markers by double PCR amplification of specific alleles (Sarkar, G. and Sommer S.S., 1991). These approaches are not entirely satisfying either because of their technical complexity, the additional cost they entail, their lack of generalisation at a large scale, or the possible biases they introduce. To overcome these difficulties, an algorithm to infer the phase of PCR-amplified DNA genotypes introduced by Clark A.G. (1990) may be used. Briefly, the principle is to start filling a preliminary list of haplotypes present in the sample by examining unambiguous individuals, that is, the complete homozygotes and the single-site heterozygotes. Then other individuals in the same sample are screened for the possible occurrence of previously recognised haplotypes. For each positive identification, the complementary haplotype is added to the list of recognised haplotypes, until the phase information for all individuals is either resolved or identified as unresolved. This method assigns a single haplotype to each multiheterozygous individual, whereas several haplotypes are possible when there are more than one heterozygous site. Alternatively, one can use methods estimating haplotype frequencies in a population without assigning haplotypes to each individual. Preferably, a method based on an expectation-maximization (EM) algorithm (Dempster et al., J. R. 1977) leading to maximum-likelihood estimates of haplotype frequencies under the assumption of Hardy-Weinberg proportions (random mating) is used (see Excoffier L. and Slatkin M., 1995). The EM algorithm is a generalised iterative maximum-likelihood approach to estimation that is useful when data are ambiguous and/or incomplete. The EM algorithm is used to resolve heterozygotes into haplotypes. Haplotype estimations are further described below under the heading “Statistical methods». Any other method known in the art to determine or to estimate the frequency of a haplotype in a population may also be used.

2) Linkage Disequilibrium analysis

Linkage disequilibrium is the non-random association of alleles at two or more loci and represents a powerful tool for mapping genes involved in disease traits (see Ajioka R.S. et al., 1997). Biallelic markers, because they are densely spaced in the human genome and can be genotyped in more numerous numbers than other types of genetic markers (such as RFLP or VNTR markers), are particularly useful in genetic analysis based on linkage disequilibrium. The biallelic markers of the present invention may be used in any linkage disequilibrium analysis method known
in the art.

Briefly, when a disease mutation is first introduced into a population (by a new mutation or the immigration of a mutation carrier), it necessarily resides on a single chromosome and thus on a single “background” or “ancestral” haplotype of linked markers. Consequently, there is complete disequilibrium between these markers and the disease mutation: one finds the disease mutation only in the presence of a specific set of marker alleles. Through subsequent generations recombinations occur between the disease mutation and these marker polymorphisms, and the disequilibrium gradually dissipates. The pace of this dissipation is a function of the recombination frequency, so the markers closest to the disease gene will manifest higher levels of disequilibrium than those that are further away. When not broken up by recombination, “ancestral” haplotypes and linkage disequilibrium between marker alleles at different loci can be tracked not only through pedigrees but also through populations. Linkage disequilibrium is usually seen as an association between one specific allele at one locus and another specific allele at a second locus.

The pattern or curve of disequilibrium between disease and marker loci is expected to exhibit a maximum that occurs at the disease locus. Consequently, the amount of linkage disequilibrium between a disease allele and closely linked genetic markers may yield valuable information regarding the location of the disease gene. For fine-scale mapping of a disease locus, it is useful to have some knowledge of the patterns of linkage disequilibrium that exist between markers in the studied region. As mentioned above the mapping resolution achieved through the analysis of linkage disequilibrium is much higher than that of linkage studies. The high density of biallelic markers combined with linkage disequilibrium analysis provides powerful tools for fine-scale mapping. Different methods to calculate linkage disequilibrium are described below under the heading “Statistical Methods”.

3) Population-based case-control studies of trait-marker associations

As mentioned above, the occurrence of pairs of specific alleles at different loci on the same chromosome is not random and the deviation from random is called linkage disequilibrium. Association studies focus on population frequencies and rely on the phenomenon of linkage disequilibrium. If a specific allele in a given gene is directly involved in causing a particular trait, its frequency will be statistically increased in an affected (trait positive) population, when compared to the frequency in a trait negative population or in a random control population. As a consequence of the existence of linkage disequilibrium, the frequency of all other alleles present in the haplotype carrying the trait-causing allele will also be increased in trait positive individuals compared to trait negative individuals or random controls. Therefore, association between the trait and any allele (specifically a biallelic marker allele) in linkage disequilibrium with the trait-causing allele will suffice to suggest the presence of a trait-related gene in that particular region. Case-control populations can be genotyped for biallelic markers to identify associations that narrowly locate a trait causing allele. As any marker in linkage disequilibrium with one given marker associated with
a trait will be associated with the trait. Linkage disequilibrium allows the relative frequencies in case-control populations of a limited number of genetic polymorphisms (specifically biallelic markers) to be analysed as an alternative to screening all possible functional polymorphisms in order to find trait-causing alleles. Association studies compare the frequency of marker alleles in unrelated case-control populations, and represent powerful tools for the dissection of complex traits.

**Case-control populations (inclusion criteria)**

Population-based association studies do not concern familial inheritance but compare the prevalence of a particular genetic marker, or a set of markers, in case-control populations. They are case-control studies based on comparison of unrelated case (affected or trait positive) individuals and unrelated control (unaffected or trait negative or random) individuals. Preferably the control group is composed of unaffected or trait negative individuals. Further, the control group is ethnically matched to the case population. Moreover, the control group is preferably matched to the case-population for the main known confusion factor for the trait under study (for example age-matched for an age-dependent trait). Ideally, individuals in the two samples are paired in such a way that they are expected to differ only in their disease status. In the following "trait positive population," "case population" and "affected population" are used interchangeably.

An important step in the dissection of complex traits using association studies is the choice of case-control populations (see Lander and Schork, 1994). A major step in the choice of case-control populations is the clinical definition of a given trait or phenotype. Any genetic trait may be analysed by the association method proposed here by carefully selecting the individuals to be included in the trait positive and trait negative phenotypic groups. Four criteria are often useful: clinical phenotype, age at onset, family history and severity. The selection procedure for continuous or quantitative traits (such as blood pressure for example) involves selecting individuals at opposite ends of the phenotype distribution of the trait under study, so as to include in these trait positive and trait negative populations individuals with non-overlapping phenotypes. Preferably, case-control populations comprise phenotypically homogeneous populations. Trait positive and trait negative populations comprise phenotypically uniform populations of individuals representing each between 1 and 98%, preferably between 1 and 80%, more preferably between 1 and 50%, and more preferably between 1 and 30%, most preferably between 1 and 20% of the total population under study, and selected among individuals exhibiting non-overlapping phenotypes. The clearer the difference between the two trait phenotypes, the greater the probability of detecting an association with biallelic markers. The selection of those drastically different but relatively uniform phenotypes enables efficient comparisons in association studies and the possible detection of marked differences at the genetic level, provided that the sample sizes of the populations under study are significant enough.

In preferred embodiments, a first group of between 50 and 300 trait positive individuals, preferably about 100 individuals, are recruited according to their phenotypes. A similar number of
trait negative individuals are included in such studies.

In the present invention, typical examples of inclusion criteria include affection by schizophrenia.

**Association analysis**

The general strategy to perform association studies using biallelic markers derived from a region carrying a candidate gene is to scan two groups of individuals (case-control populations) in order to measure and statistically compare the allele frequencies of the biallelic markers of the present invention in both groups.

If a statistically significant association with a trait is identified for at least one or more of the analysed biallelic markers, one can assume that: either the associated allele is directly responsible for causing the trait (the associated allele is the trait causing allele), or more likely the associated allele is in linkage disequilibrium with the trait causing allele. The specific characteristics of the associated allele with respect to the gene function usually gives further insight into the relationship between the associated allele and the trait (causal or in linkage disequilibrium). If the evidence indicates that the associated allele within the gene is most probably not the trait causing allele but is in linkage disequilibrium with the real trait causing allele, then the trait causing allele can be found by sequencing the vicinity of the associated marker.

Another embodiment of the present invention encompasses methods of detecting an association between a haplotype and a phenotype, comprising the steps of: a) estimating the frequency of at least one haplotype in a trait positive population according to a method of estimating the frequency of a haplotype of the invention; b) estimating the frequency of said haplotype in a control population according to the method of estimating the frequency of a haplotype of the invention; and c) determining whether a statistically significant association exists between said haplotype and said phenotype. In addition, the methods of detecting an association between a haplotype and a phenotype of the invention encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination: Optionally, said DAO related biallelic marker may be in a sequence selected individually or in any combination from the group consisting of SEQ ID Nos 1, 2, 4, 5, 7, 8, and 11-15, and the complements thereof; optionally, said DAO related biallelic marker may be selected individually or in any combination from the biallelic markers described in Tables 6b and 6c; optionally, said control population may be a trait negative population, or a random population; optionally, said phenotype is a disease involving schizophrenia, a response to an agent acting on schizophrenia, or a side effects to an agent acting on schizophrenia.

**Haplotype analysis**

As described above, when a chromosome carrying a disease allele first appears in a population as a result of either mutation or migration, the mutant allele necessarily resides on a chromosome having a set of linked markers: the ancestral haplotype. This haplotype can be tracked
through populations and its statistical association with a given trait can be analysed. Complementing single point (allelic) association studies with multi-point association studies also called haplotype studies increases the statistical power of association studies. Thus, a haplotype association study allows one to define the frequency and the type of the ancestral carrier haplotype. A haplotype analysis is important in that it increases the statistical power of an analysis involving individual markers.

In a first stage of a haplotype frequency analysis, the frequency of the possible haplotypes based on various combinations of the identified biallelic markers of the invention is determined. The haplotype frequency is then compared for distinct populations of trait positive and control individuals. The number of trait positive individuals, which should be, subjected to this analysis to obtain statistically significant results usually ranges between 30 and 300, with a preferred number of individuals ranging between 50 and 150. The same considerations apply to the number of unaffected individuals (or random control) used in the study. The results of this first analysis provide haplotype frequencies in case-control populations, for each evaluated haplotype frequency a p-value and an odd ratio are calculated. If a statistically significant association is found the relative risk for an individual carrying the given haplotype of being affected with the trait under study can be approximated.

**Interaction Analysis**

The biallelic markers of the present invention may also be used to identify patterns of biallelic markers associated with detectable traits resulting from polygenic interactions. The analysis of genetic interaction between alleles at unlinked loci requires individual genotyping using the techniques described herein. The analysis of allelic interaction among a selected set of biallelic markers with appropriate level of statistical significance can be considered as a haplotype analysis. Interaction analysis comprises stratifying the case-control populations with respect to a given haplotype for the first loci and performing a haplotype analysis with the second loci with each subpopulation.

Statistical methods used in association studies are further described herein.

4) **Testing for linkage in the presence of association**

The biallelic markers of the present invention may further be used in TDT (transmission/disequilibrium test). TDT tests for both linkage and association and is not affected by population stratification. TDT requires data for affected individuals and their parents or data from unaffected sibs instead of from parents (see Spielmann S. et al., 1993; Schaid D.J. et al., 1996, Spielmann S. and Ewens W.J, 1998). Such combined tests generally reduce the false - positive errors produced by separate analyses.

**Statistical methods**

In general, any method known in the art to test whether a trait and a genotype show a statistically significant correlation may be used.
1) Methods in linkage analysis

Statistical methods and computer programs useful for linkage analysis are well-known to those skilled in the art (see Terwilliger J.D. and Ott J., 1994; Ott J., 1991).

2) Methods to estimate haplotype frequencies in a population

As described above, when genotypes are scored, it is often not possible to distinguish heterozygotes so that haplotype frequencies cannot be easily inferred. When the gametic phase is not known, haplotype frequencies can be estimated from the multilocus genotypic data. Any method known to person skilled in the art can be used to estimate haplotype frequencies (see Lange K., 1997; Weir, B.S., 1996) Preferably, maximum-likelihood haplotype frequencies are computed using an Expectation-Maximization (EM) algorithm (see Dempster et al., 1977; Excoffier L. and Slatkin M., 1995). This procedure is an iterative process aiming at obtaining maximum-likelihood estimates of haplotype frequencies from multi-locus genotype data when the gametic phase is unknown.

Haplotype estimations are usually performed by applying the EM algorithm using for example the EM-HAPLO program (Hawley M.E. et al., 1994) or the Arlequin program (Schneider et al., 1997). The EM algorithm is a generalised iterative maximum likelihood approach to estimation and is briefly described below.

In the following part of this text, phenotypes will refer to multi-locus genotypes with unknown phase. Genotypes will refer to known-phase multi-locus genotypes. Suppose a sample of N unrelated individuals typed for K markers. The data observed are the unknown-phase K-locus phenotypes that can categorised in F different phenotypes. Suppose that we have H underlying possible haplotypes (in case of K biallelic markers, H=2^K).

For phenotype j, suppose that c_j genotypes are possible. We thus have the following equation

\[ P_j = \sum_{i=1}^{c_j} pr(genotype_i) = \sum_{i=1}^{c_j} pr(h_k, h_l) \]  \hspace{1cm} \text{Equation 1}

where \( P_j \) is the probability of the phenotype j, \( h_k \) and \( h_l \) are the two haplotypes constituent the genotype i. Under the Hardy-Weinberg equilibrium, \( pr(h_k, h_l) \) becomes:

\[ pr(h_k, h_l) = pr(h_k)^2 \text{ if } h_k = h_l, \text{ pr}(h_k, h_l) = 2 \text{ pr}(h_k) \cdot \text{pr}(h_l) \text{ if } h_k \neq h_l. \]  \hspace{1cm} \text{Equation 2}

The successive steps of the E-M algorithm can be described as follows:

Starting with initial values of the of haplotypes frequencies, noted \( P_1^{(0)}, P_2^{(0)}, \ldots, P_H^{(0)} \), these initial values serve to estimate the genotype frequencies (Expectation step) and then estimate another set of haplotype frequencies (Maximisation step), noted \( P_1^{(1)}, P_2^{(1)}, \ldots, P_H^{(1)} \), these two steps are iterated until changes in the sets of haplotypes frequency are very small.

A stop criterion can be that the maximum difference between haplotype frequencies between two iterations is less than 10^{-7}. These values can be adjusted according to the desired precision of
estimations. In details, at a given iteration s, the Expectation step comprises calculating the genotypes frequencies by the following equation:

\[
pr(\text{genotype}_i)_{(s)} = pr(\text{phenotype}_j).pr(\text{genotype}_i|\text{phenotype}_j)_{(s)}
\]

\[
= \frac{n_j}{N} \cdot \frac{pr(h_k, h_l)}{P_j^{(s)}}
\]

Equation 3

where genotype i occurs in phenotype j, and where \(h_k\) and \(h_l\) constitute genotype i. Each probability is derived according to eq.1, and eq.2 described above.

Then the Maximisation step simply estimates another set of haplotype frequencies given the genotypes frequencies. This approach is also known as gene-counting method (Smith, 1957).

\[
p_i^{(s+1)} = \frac{1}{2} \sum_{j=1}^{F} \sum_{i=1}^{C_i} \delta_{it} \cdot pr(\text{genotype}_i)_{(s)}
\]

Equation 4

Where \(\delta_{it}\) is an indicator variable which count the number of time haplotype t in genotype i. It takes the values of 0, 1 or 2.

To ensure that the estimation finally obtained is the maximum-likelihood estimation several values of departures are required. The estimations obtained are compared and if they are different the estimations leading to the best likelihood are kept.

3) Methods to calculate linkage disequilibrium between markers

A number of methods can be used to calculate linkage disequilibrium between any two genetic positions, in practice linkage disequilibrium is measured by applying a statistical association test to haplotype data taken from a population. Linkage disequilibrium between any pair of biallelic markers comprising at least one of the biallelic markers of the present invention (M_i, M_j) having alleles (a_i/b_i) at marker M_i and alleles (a_j/b_j) at marker M_j can be calculated for every allele combination (a_i,a_j; a_i,b_j; b_i,a_j and b_i,b_j), according to the Piazza formula:

\[
\Delta_{ij} = \sqrt{\theta_4 - \sqrt{(\theta_4 + \theta_3) (\theta_4 + \theta_2)}}, \text{ where }:
\]

\[
\theta_4 = - - = \text{frequency of genotypes not having allele } a_i \text{ at } M_i \text{ and not having allele } a_j \text{ at } M_j
\]

\[
\theta_3 = - + = \text{frequency of genotypes not having allele } a_i \text{ at } M_i \text{ and having allele } a_j \text{ at } M_j
\]

\[
\theta_2 = + - = \text{frequency of genotypes having allele } a_i \text{ at } M_i \text{ and not having allele } a_j \text{ at } M_j
\]

Linkage disequilibrium (LD) between pairs of biallelic markers (M_i, M_j) can also be calculated for every allele combination (a_i,a_j; a_i,b_j; b_i,a_j and b_i,b_j), according to the maximum-likelihood estimate (MLE) for delta (the composite genotypic disequilibrium coefficient), as described by Weir (B.S., 1996). The MLE for the composite linkage disequilibrium is:

\[
D_{ij} = (2n_1 + n_2 + n_3 + n_4)/N - 2(pr(a_i).pr(a_j))
\]

where \(n_1 = \Sigma \text{ phenotype (a_i/a_i, a_j/a_j)}, n_2 = \Sigma \text{ phenotype (a_i/a_i, a_j/b_j)}, n_3 = \Sigma \text{ phenotype (a_i/b_i, a_j/a_j)}, n_4 = \Sigma \text{ phenotype (a_i/b_i, a_j/b_j)}\) and N is the number of individuals in the sample. This formula allows
linkage disequilibrium between alleles to be estimated when only genotype, and not haplotype, data are available.

Another means of calculating the linkage disequilibrium between markers is as follows. For a couple of biallelic markers, \( M_i (a/b) \) and \( M_j (a/b) \), fitting the Hardy-Weinberg equilibrium, one can estimate the four possible haplotype frequencies in a given population according to the approach described above.

The estimation of gametic disequilibrium between \( ai \) and \( aj \) is simply:

\[
D_{ai|aj} = pr(haplotype(a_i, a_j)) - pr(a_i).pr(a_j).
\]

Where \( pr(ai) \) is the probability of allele \( a_i \) and \( pr(aj) \) is the probability of allele \( a_j \) and where \( pr(haplotype \ (a_i, a_j)) \) is estimated as in Equation 3 above.

For a couple of biallelic marker only one measure of disequilibrium is necessary to describe the association between \( M_i \) and \( M_j \).

Then a normalised value of the above is calculated as follows:

\[
\begin{align*}
D'_{ai|aj} & = \frac{D_{ai|aj}}{max (-pr(ai).pr(aj), -pr(bi).pr(bj))} \text{ with } D_{ai|aj} < 0 \\
D'_{ai|aj} & = \frac{D_{ai|aj}}{max (pr(bi).pr(ai), pr(a).pr(bi))} \text{ with } D_{ai|aj} \geq 0
\end{align*}
\]

The skilled person will readily appreciate that other LD calculation methods can be used without undue experimentation.

Linkage disequilibrium among a set of biallelic markers having an adequate heterozygosity rate can be determined by genotyping between 50 and 1000 unrelated individuals, preferably between 75 and 200, more preferably around 100.

4) Testing for association

Methods for determining the statistical significance of a correlation between a phenotype and a genotype, in this case an allele at a biallelic marker or a haplotype made up of such alleles, may be determined by any statistical test known in the art and with any accepted threshold of statistical significance being required. The application of particular methods and thresholds of significance are well within the skill of the ordinary practitioner of the art.

Testing for association is performed by determining the frequency of a biallelic marker allele in case and control populations and comparing these frequencies with a statistical test to determine if their is a statistically significant difference in frequency which would indicate a correlation between the trait and the biallelic marker allele under study. Similarly, a haplotype analysis is performed by estimating the frequencies of all possible haplotypes for a given set of biallelic markers in case and control populations, and comparing these frequencies with a statistical test to determine if their is a statistically significant correlation between the haplotype and the phenotype (trait) under study. Any statistical tool useful to test for a statistically significant association between a genotype and a phenotype may be used. Preferably the statistical test employed is a chi-square test with one degree of freedom. A P-value is calculated (the P-value is the probability that a statistic as large or larger than the observed one would occur by chance).
Statistical significance

In preferred embodiments, significance for diagnosis purposes, either as a positive basis for further diagnostic tests or as a preliminary starting point for early preventive therapy, the p value related to a biallelic marker association is preferably about $1 \times 10^{-2}$ or less, more preferably about $1 \times 10^{-4}$ or less, for a single biallelic marker analysis and about $1 \times 10^{-3}$ or less, still more preferably $1 \times 10^{-6}$ or less and most preferably of about $1 \times 10^{-8}$ or less, for a haplotype analysis involving several markers. These values are believed to be applicable to any association studies involving single or multiple marker combinations.

The skilled person can use the range of values set forth above as a starting point in order to carry out association studies with biallelic markers of the present invention. In doing so, significant associations between the biallelic markers of the present invention and diseases involving schizophrenia can be revealed and used for diagnosis and drug screening purposes.

Phenotypic permutation

In order to confirm the statistical significance of the first stage haplotype analysis described above, it might be suitable to perform further analyses in which genotyping data from case-control individuals are pooled and randomised with respect to the trait phenotype. Each individual genotyping data is randomly allocated to two groups, which contain the same number of individuals as the case-control populations used to compile the data obtained in the first stage. A second stage haplotype analysis is preferably run on these artificial groups, preferably for the markers included in the haplotype of the first stage analysis showing the highest relative risk coefficient. This experiment is reiterated preferably at least between 100 and 10000 times. The repeated iterations allow the determination of the percentage of obtained haplotypes with a significant p-value level.

Assessment of statistical association

To address the problem of false positives similar analysis may be performed with the same case-control populations in random genomic regions. Results in random regions and the candidate region are compared as described in US Provisional Patent Application entitled “Methods, software and apparatus for identifying genomic regions harbouring a gene associated with a detectable trait”.

5) Evaluation of risk factors

The association between a risk factor (in genetic epidemiology the risk factor is the presence or the absence of a certain allele or haplotype at marker loci) and a disease is measured by the odds ratio (OR) and by the relative risk (RR). If $P(R^+)$ is the probability of developing the disease for individuals with R and $P(R^-)$ is the probability for individuals without the risk factor, then the relative risk is simply the ratio of the two probabilities, that is:

$$RR = \frac{P(R^+)/P(R^-)}$$

In case-control studies, direct measures of the relative risk cannot be obtained because of the sampling design. However, the odds ratio allows a good approximation of the relative risk for low-incidence diseases and can be calculated:

$$OR = \frac{F^+}{F^-}$$
F^c is the frequency of the exposure to the risk factor in cases and F^r is the frequency of the exposure to the risk factor in controls. F^c and F^r are calculated using the allelic or haplotype frequencies of the study and further depend on the underlying genetic model (dominant, recessive, additive...). One can further estimate the attributable risk (AR) which describes the proportion of individuals in a population exhibiting a trait due to a given risk factor. This measure is important in quantitating the role of a specific factor in disease etiology and in terms of the public health impact of a risk factor. The public health relevance of this measure lies in estimating the proportion of cases of disease in the population that could be prevented if the exposure of interest were absent. AR is determined as follows:

$$AR = \frac{P_E (RR-1)}{(P_E (RR-1)+1)}$$

AR is the risk attributable to a biallelic marker allele or a biallelic marker haplotype. P_E is the frequency of exposure to an allele or a haplotype within the population at large; and RR is the relative risk which, is approximated with the odds ratio when the trait under study has a relatively low incidence in the general population.

AR is the risk attributable to a biallelic marker allele or a biallelic marker haplotype. P_E is the frequency of exposure to an allele or a haplotype within the population at large; and RR is the relative risk which, is approximated with the odds ratio when the trait under study has a relatively low incidence in the general population.

**Association of Biallelic Markers of the Invention with Schizophrenia**

In the context of the present invention, an association between DAO related biallelic markers and schizophrenia were established. Several association studies using different populations and screening samples thereof, and with different sets of biallelic markers distributed in or near the DAO gene were carried out. Further details concerning these association studies and the results are provided herein in Table 3.

This information is extremely valuable. The knowledge of a potential genetic predisposition to schizophrenia, even if this predisposition is not absolute, might contribute in a very significant manner to treatment efficacy of schizophrenia and to the development of new therapeutic and diagnostic tools.

**Identification Of Biallelic Markers In Linkage Disequilibrium With The Biallelic Markers of the Invention**

Once a first biallelic marker has been identified in a genomic region of interest, the practitioner of ordinary skill in the art, using the teachings of the present invention, can easily identify additional biallelic markers in linkage disequilibrium with this first marker. As mentioned before, any marker in linkage disequilibrium with a first marker associated with a trait will be
associated with the trait. Therefore, once an association has been demonstrated between a given biallelic marker and a trait, the discovery of additional biallelic markers associated with this trait is of great interest in order to increase the density of biallelic markers in this particular region. The causal gene or mutation will be found in the vicinity of the marker or set of markers showing the highest correlation with the trait.

Identification of additional markers in linkage disequilibrium with a given marker involves: (a) amplifying a genomic fragment comprising a first biallelic marker from a plurality of individuals; (b) identifying of second biallelic markers in the genomic region harboring said first biallelic marker; (c) conducting a linkage disequilibrium analysis between said first biallelic marker and second biallelic markers; and (d) selecting said second biallelic markers as being in linkage disequilibrium with said first marker. Subcombinations comprising steps (b) and (c) are also contemplated.

Methods to identify biallelic markers and to conduct linkage disequilibrium analysis are described herein and can be carried out by the skilled person without undue experimentation. The present invention then also concerns biallelic markers and other polymorphisms which are in linkage disequilibrium with the specific biallelic markers of the invention and which are expected to present similar characteristics in terms of their respective association with a given trait. In a preferred embodiment, the invention concerns biallelic markers which are in linkage disequilibrium with the specific biallelic markers.

**Identification Of Functional Mutations**

Once a positive association is confirmed with a biallelic marker of the present invention, the associated candidate gene sequence can be scanned for mutations by comparing the sequences of a selected number of trait positive and trait negative individuals. In a preferred embodiment, functional regions such as exons and splice sites, promoters and other regulatory regions of the gene are scanned for mutations. Preferably, trait positive individuals carry the haplotype shown to be associated with the trait and trait negative individuals do not carry the haplotype or allele associated with the trait. The mutation detection procedure is essentially similar to that used for biallelic site identification.

The method used to detect such mutations generally comprises the following steps: (a) amplification of a region of the candidate DNA sequence comprising a biallelic marker or a group of biallelic markers associated with the trait from DNA samples of trait positive patients and trait negative controls; (b) sequencing of the amplified region; (c) comparison of DNA sequences from trait-positive patients and trait-negative controls; and (d) determination of mutations specific to trait-positive patients. Subcombinations which comprise steps (b) and (c) are specifically contemplated.

It is preferred that candidate polymorphisms be then verified by screening a larger population of cases and controls by means of any genotyping procedure such as those described herein, preferably using a microsequencing technique in an individual test format. Polymorphisms
are considered as candidate mutations when present in cases and controls at frequencies compatible with the expected association results.

Candidate polymorphisms and mutations of the sbg1 nucleic acid sequences suspected of being involved in a predisposition to schizophrenia can be confirmed by screening a larger population of affected and unaffected individuals using any of the genotyping procedures described herein. Preferably the microsequencing technique is used. Such polymorphisms are considered as candidate “trait-causing” mutations when they exhibit a statistically significant correlation with the detectable phenotype.

**Biallelic Markers Of The Invention In Methods Of Genetic Diagnostics**

The biallelic markers and other polymorphisms of the present invention can also be used to develop diagnostics tests capable of identifying individuals who express a detectable trait as the result of a specific genotype or individuals whose genotype places them at risk of developing a detectable trait at a subsequent time. The trait analyzed using the present diagnostics may be any detectable trait, including predisposition to schizophrenia, age of onset of detectable symptoms, a beneficial response to or side effects related to treatment against schizophrenia. Such a diagnosis can be useful in the monitoring, prognosis and/or prophylactic or curative therapy for schizophrenia.

The diagnostic techniques of the present invention may employ a variety of methodologies to determine whether a test subject has a genotype associated with an increased risk of developing a detectable trait or whether the individual suffers from a detectable trait as a result of a particular mutation, including methods which enable the analysis of individual chromosomes for haplotyping, such as family studies, single sperm DNA analysis or somatic hybrids.

The diagnostic techniques concern the detection of specific alleles present within or near the DAO gene. More particularly, the invention concerns the detection of a nucleic acid comprising at least one of the nucleotide sequences of SEQ ID NOs:1, 4, 11-15, or a fragment thereof or a complementary sequence thereto including the polymorphic base.

These methods involve obtaining a nucleic acid sample from the individual and, determining, whether the nucleic acid sample contains at least one allele or at least one biallelic marker haplotype, indicative of a risk of developing the trait or indicative that the individual expresses the trait as a result of possessing a particular DAO related biallelic marker (polymorphism or mutation (trait-causing allele)).

Preferably, in such diagnostic methods, a nucleic acid sample is obtained from the individual and this sample is genotyped using methods described above in “Methods Of Genotyping DNA Samples For Biallelic markers.” The diagnostics may be based on a single biallelic marker or a on group of biallelic markers.

In each of these methods, a nucleic acid sample is obtained from the test subject and the biallelic marker pattern of one or more of the biallelic markers of the invention is determined.

In one embodiment, a PCR amplification is conducted on the nucleic acid sample to
amplify regions in which polymorphisms associated with a detectable phenotype have been identified. The amplification products are sequenced to determine whether the individual possesses one or more DAO related biallelic markers with a detectable phenotype. The primers used to generate amplification products may comprise the primers listed in SEQ ID NO:1 or 4 (as defined previously as having the prefix “.rp” and “.pu complement”). Alternatively, the nucleic acid sample is subjected to microsequencing reactions as described above to determine whether the individual possesses one or more DAO related biallelic markers (polymorphisms) associated with a detectable phenotype. The primers used in the microsequencing reactions may include the primers listed in SEQ ID NO:1 or 4 (as previously defined as having the prefix “.mis” and “.mis complement”). In another embodiment, the nucleic acid sample is contacted with one or more allele specific oligonucleotide probes which, specifically hybridize to one or more DAO related alleles associated with a detectable phenotype. The probes used in the hybridization assay may include the probes listed in SEQ ID NO:1 or 4 (defined as having the prefix “.probe”). In another embodiment, the nucleic acid sample is contacted with a second oligonucleotide capable of producing an amplification product when used with the allele specific oligonucleotide in an amplification reaction. The presence of an amplification product in the amplification reaction indicates that the individual possesses one or more DAO related alleles associated with a detectable phenotype.

In a preferred embodiment the identity of the nucleotide present at at least one biallelic marker selected from the group consisting of 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4 and the complements thereof, is determined and the detectable trait is schizophrenia. Diagnostic kits comprise any of the polynucleotides of the present invention.

These diagnostic methods are extremely valuable as they can, in certain circumstances, be used to initiate preventive treatments or to allow an individual carrying a significant haplotype to foresee warning signs such as minor symptoms.

Diagnostics, which analyze and predict response to a drug or side effects to a drug, may be used to determine whether an individual should be treated with a particular drug. For example, if the diagnostic indicates a likelihood that an individual will respond positively to treatment with a particular drug, the drug may be administered to the individual. Conversely, if the diagnostic indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects.

Clinical drug trials represent another application for the markers of the present invention. One or more markers indicative of response to an agent acting against schizophrenia or to side effects to an agent acting against schizophrenia may be identified using the methods described above. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to
experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and without risking undesirable safety problems.

PREVENTION, DIAGNOSIS AND TREATMENT OF PSYCHIATRIC DISEASE

Sbg1 in Methods of Diagnosis or Detecting Predisposition

Individuals affected by or predisposed to schizophrenia and bipolar disorder may possess a particular allele of the DAO gene. In one aspect of the present invention is a method for determining whether an individual is at risk of suffering from or is currently suffering from schizophrenia, bipolar disorder or other psychotic disorders, mood disorders, autism, substance dependence or alcoholism, mental retardation, or other psychiatric diseases including cognitive, anxiety, eating, impulse-control, and personality disorders, as defined with the Diagnosis and Statistical Manual of Mental Disorders fourth edition (DSM-IV) classification, comprising determining whether the individual has a particular allele of the DAO gene as determined by the association studies described herein.

Biallelic Markers Of The Invention In Methods Of Genetic Diagnostics

The biallelic markers and other polymorphisms of the present invention can also be used to develop diagnostics tests capable of identifying individuals who express a detectable trait as the result of a specific genotype or individuals whose genotype places them at risk of developing a detectable trait at a subsequent time. The trait analyzed using the present diagnostics may be used to diagnose any detectable trait, including predisposition to schizophrenia or bipolar disorder, age of onset of detectable symptoms, a beneficial response to or side effects related to treatment against schizophrenia or bipolar disorder. Such a diagnosis can be useful in the monitoring, prognosis and/or prophylactic or curative therapy for schizophrenia or bipolar disorder.

The diagnostic techniques of the present invention may employ a variety of methodologies to determine whether a test subject has a genotype associated with an increased risk of developing a detectable trait or whether the individual suffers from a detectable trait as a result of a particular mutation, including methods which enable the analysis of individual chromosomes for haplotyping, such as family studies, single sperm DNA analysis or somatic hybrids.

The diagnostic techniques concern the detection of specific alleles present within or near the DAO gene. More particularly, the invention concerns the detection of a nucleic acid comprising at least one of the nucleotide sequences of SEQ ID NOs:1, 4, 11-15 or a fragment thereof or a complementary sequence thereto including the polymorphic base.

These methods involve obtaining a nucleic acid sample from the individual and, determining, whether the nucleic acid sample contains at least one allele or at least one biallelic marker haplotype, indicative of a risk of developing the trait or indicative that the individual expresses the trait as a result of possessing a particular DAO related biallelic marker (polymorphism
or mutation (trait-causing allele)).

Preferably, in such diagnostic methods, a nucleic acid sample is obtained from the individual and this sample is genotyped using methods described above in “Methods Of Genotyping DNA Samples For Biallelic markers.” The diagnostics may be based on a single biallelic marker or a on group of biallelic markers.

In each of these methods, a nucleic acid sample is obtained from the test subject and the biallelic marker pattern of one or more of a biallelic marker of the invention is determined.

In one embodiment, a PCR amplification is conducted on the nucleic acid sample to amplify regions in which polymorphisms associated with a detectable phenotype have been identified. The amplification products are sequenced to determine whether the individual possesses one or more DAO related biallelic markers (polymorphisms) associated with a detectable phenotype. The primers used to generate amplification products may comprise the primers listed in SEQ ID NO:1 or 4. Alternatively, the nucleic acid sample is subjected to microsequencing reactions as described above to determine whether the individual possesses one or more DAO related biallelic markers (polymorphisms) associated with a detectable phenotype resulting from a mutation or a polymorphism in or near the DAO gene. The primers used in the microsequencing reactions may include the primers listed in SEQ ID NO:1 or 4. In another embodiment, the nucleic acid sample is contacted with one or more allele specific oligonucleotide probes which, specifically hybridize to one or more DAO related alleles associated with a detectable phenotype. The probes used in the hybridization assay may include the probes listed in SEQ ID NO:1 or 4. In another embodiment, the nucleic acid sample is contacted with a second oligonucleotide capable of producing an amplification product when used with the allele specific oligonucleotide in an amplification reaction. The presence of an amplification product in the amplification reaction indicates that the individual possesses one or more DAO related alleles associated with a detectable phenotype. In a preferred embodiment, the detectable trait is schizophrenia or bipolar disorder. Diagnostic kits comprise any of the polynucleotides of the present invention.

These diagnostic methods are extremely valuable as they can, in certain circumstances, be used to initiate preventive treatments or to allow an individual carrying a significant haplotype to foresee warning signs such as minor symptoms.

Diagnostics, which analyze and predict response to a drug or side effects to a drug, may be used to determine whether an individual should be treated with a particular drug. For example, if the diagnostic indicates a likelihood that an individual will respond positively to treatment with a particular drug, the drug may be administered to the individual. Conversely, if the diagnostic indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects.

Clinical drug trials represent another application for the markers of the present invention.
One or more markers indicative of response to an agent acting against schizophrenia or to side effects to an agent acting against schizophrenia may be identified using the methods described above. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and without risking undesirable safety problems.

**Prevention And Treatment Of Disease Using Biallelic Markers**

In large part because of the risk of suicide, the detection of susceptibility to schizophrenia, bipolar disorder as well as other psychiatric disease in individuals is very important. Consequently, the invention concerns a method for the treatment of schizophrenia or bipolar disorder, or a related disorder comprising the following steps:

- selecting an individual whose DNA comprises alleles of a DAO related biallelic marker, or of a group of biallelic markers of DAO related markers, and more preferably DAO related markers associated with schizophrenia or bipolar disorder;

- following up said individual for the appearance (and optionally the development) of the symptoms related to schizophrenia or bipolar disorder; and

- administering a treatment acting against schizophrenia or bipolar disorder or against symptoms thereof to said individual at an appropriate stage of the disease.

Another embodiment of the present invention comprises a method for the treatment of schizophrenia or bipolar disorder comprising the following steps:

- selecting an individual whose DNA comprises alleles of a DAO related biallelic marker, or of a group of biallelic markers of DAO related markers, and more preferably DAO related markers associated with schizophrenia or bipolar disorder;

- administering a preventive treatment of schizophrenia or bipolar disorder to said individual.

In a further embodiment, the present invention concerns a method for the treatment of schizophrenia or bipolar disorder comprising the following steps:

- selecting an individual whose DNA comprises alleles of a DAO related biallelic marker, or of a group of biallelic markers of DAO related markers, and more preferably DAO related markers associated with schizophrenia or bipolar disorder;

- administering a preventive treatment of schizophrenia or bipolar disorder to said individual;

- following up said individual for the appearance and the development of schizophrenia or bipolar disorder symptoms; and optionally

- administering a treatment acting against schizophrenia or bipolar disorder or against symptoms thereof to said individual at the appropriate stage of the disease.
For use in the determination of the course of treatment of an individual suffering from disease, the present invention also concerns a method for the treatment of schizophrenia or bipolar disorder comprising the following steps:

- selecting an individual suffering from schizophrenia or bipolar disorder whose DNA comprises alleles of a DAO related biallelic marker or of a group of DAO related biallelic markers, preferably markers associated with the gravity of schizophrenia or bipolar disorder or of the symptoms thereof; and

- administering a treatment acting against schizophrenia or bipolar disorder or symptoms thereof to said individual.

The invention also concerns a method for the treatment of schizophrenia or bipolar disorder in a selected population of individuals. The method comprises:

- selecting an individual suffering from schizophrenia or bipolar disorder and whose DNA comprises alleles of a DAO related biallelic marker or of a group of DAO related biallelic markers, preferably markers associated with a positive response to treatment with an effective amount of a medicament acting against schizophrenia or bipolar disorder or symptoms thereof;

- and/or whose DNA does not comprise alleles of a biallelic marker or of a group of DAO related biallelic markers, preferably DAO related markers associated with a negative response to treatment with said medicament; and

- administering at suitable intervals an effective amount of said medicament to said selected individual.

In the context of the present invention, a “positive response” to a medicament can be defined as comprising a reduction of the symptoms related to the disease. In the context of the present invention, a “negative response” to a medicament can be defined as comprising either a lack of positive response to the medicament which does not lead to a symptom reduction or which leads to a side-effect observed following administration of the medicament.

The invention also relates to a method of determining whether a subject is likely to respond positively to treatment with a medicament. The method comprises identifying a first population of individuals who respond positively to said medicament and a second population of individuals who respond negatively to said medicament. One or more biallelic markers is identified in the first population which is associated with a positive response to said medicament or one or more biallelic markers is identified in the second population which is associated with a negative response to said medicament. The biallelic markers may be identified using the techniques described herein.

A DNA sample is then obtained from the subject to be tested. The DNA sample is analyzed to determine whether it comprises alleles of one or more biallelic markers associated with a positive response to treatment with the medicament and/or alleles of one or more biallelic markers associated with a negative response to treatment with the medicament.

In some embodiments, the medicament may be administered to the subject in a clinical trial.
if the DNA sample contains alleles of one or more biallelic markers associated with a positive response to treatment with the medicament and/or if the DNA sample lacks alleles of one or more biallelic markers associated with a negative response to treatment with the medicament. In preferred embodiments, the medicament is a drug acting against schizophrenia or bipolar disorder.

Using the method of the present invention, the evaluation of drug efficacy may be conducted in a population of individuals likely to respond favorably to the medicament.

Another aspect of the invention is a method of using a medicament comprising obtaining a DNA sample from a subject, determining whether the DNA sample contains alleles of one or more biallelic markers associated with a positive response to the medicament and/or whether the DNA sample contains alleles of one or more biallelic markers associated with a negative response to the medicament, and administering the medicament to the subject if the DNA sample contains alleles of one or more biallelic markers associated with a positive response to the medicament and/or if the DNA sample lacks alleles of one or more biallelic markers associated with a negative response to the medicament.

The invention also concerns a method for the clinical testing of a medicament, preferably a medicament acting against schizophrenia or or bipolar disorder or symptoms thereof. The method comprises the following steps:

- administering a medicament, preferably a medicament susceptible of acting against schizophrenia or or bipolar disorder or symptoms thereof to a heterogeneous population of individuals,
- identifying a first population of individuals who respond positively to said medicament and a second population of individuals who respond negatively to said medicament,
- identifying biallelic markers in said first population which are associated with a positive response to said medicament,
- selecting individuals whose DNA comprises biallelic markers associated with a positive response to said medicament, and
- administering said medicament to said individuals.

In any of the methods for the prevention, diagnosis and treatment of schizophrenia and bipolar disorder, including methods of using a medicament, clinical testing of a medicament, determining whether a subject is likely to respond positively to treatment with a medicament, said biallelic marker may optionally comprise:

(a) a biallelic marker selected from the group consisting of biallelic markers 27-81-180, 27-29-224, 27-2-106, and 27-30-249 of SEQ ID NO:1, and 27-1-61 of SEQ ID NO:4;
(b) a biallelic marker selected from the group consisting of biallelic markers 27-29-224, and 27-2-106 of SEQ ID NO:1;
(c) a biallelic marker 27-2-106 of SEQ ID NO:1; or
(d) a biallelic marker 27-29-224 of SEQ ID NO:1;
Such methods are deemed to be extremely useful to increase the benefit/risk ratio resulting from the administration of medicaments which may cause undesirable side effects and/or be inefficacious to a portion of the patient population to which it is normally administered.

Once an individual has been diagnosed as suffering from schizophrenia or bipolar disorder, selection tests are carried out to determine whether the DNA of this individual comprises alleles of a biallelic marker or of a group of biallelic markers associated with a positive response to treatment or with a negative response to treatment which may include either side effects or unresponsiveness.

The selection of the patient to be treated using the method of the present invention can be carried out through the detection methods described above. The individuals which are to be selected are preferably those whose DNA does not comprise alleles of a biallelic marker or of a group of biallelic markers associated with a negative response to treatment. The knowledge of an individual's genetic predisposition to unresponsiveness or side effects to particular medicaments allows the clinician to direct treatment toward appropriate drugs against schizophrenia or bipolar disorder or symptoms thereof.

Once the patient's genetic predispositions have been determined, the clinician can select appropriate treatment for which negative response, particularly side effects, has not been reported or has been reported only marginally for the patient.

The biallelic markers of the invention have demonstrated an association with schizophrenia and bipolar disorders. However, the present invention also comprises any of the prevention, diagnostic, prognosis and treatment methods described herein using the biallelic markers of the invention in methods of preventing, diagnosing, managing and treating related disorders, particularly related CNS disorders. By way of example, related disorders may comprise psychotic disorders, mood disorders, autism, substance dependence and alcoholism, mental retardation, and other psychiatric diseases including cognitive, anxiety, eating, impulse-control, and personality disorders, as defined with the Diagnosis and Statistical Manual of Mental Disorders fourth edition (DSM-IV) classification made using electroporation, such as described by Thomas et al. (1987). The cells subjected to electroporation are screened (e.g. by selection via selectable markers, by PCR or by Southern blot analysis) to find positive cells which have integrated the exogenous recombinant polynucleotide into their genome, preferably via an homologous recombination event. An illustrative positive-negative selection procedure that may be used according to the invention is described by Mansour et al. (1988).

Then, the positive cells are isolated, cloned and injected into 3.5 days old blastocysts from mice, such as described by Bradley (1987). The blastocysts are then inserted into a female host animal and allowed to grow to term.

Alternatively, the positive ES cells are brought into contact with embryos at the 2.5 days old 8-16 cell stage (morulae) such as described by Wood et al. (1993) or by Nagy et al. (1993), the ES cells being internalized to colonize extensively the blastocyst including the cells which will give rise
to the germ line.

The offspring of the female host are tested to determine which animals are transgenic e.g. include the inserted exogenous DNA sequence and which are wild-type.

Thus, the present invention also concerns a transgenic animal containing a nucleic acid, a recombinant expression vector or a recombinant host cell according to the invention.

**Recombinant Cell Lines Derived From The Transgenic Animals Of The Invention.**

A further object of the invention comprises recombinant host cells obtained from a transgenic animal described herein. In one embodiment the invention encompasses cells derived from non-human host mammals and animals comprising a recombinant vector of the invention or a gene comprising an sbg1, g34665, sbg2, g35017 or g35018 nucleic acid sequence disrupted by homologous recombination with a knock out vector.

Recombinant cell lines may be established *in vitro* from cells obtained from any tissue of a transgenic animal according to the invention, for example by transfection of primary cell cultures with vectors expressing one-genes such as SV40 large T antigen, as described by Chou (1989) and Shay et al. (1991).

**Computer-Related Embodiments**

As used herein the term “nucleic acid codes of the invention” encompass the nucleotide sequences comprising, consisting essentially of, or consisting of any one of the following:

a) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides of SEQ ID No. 1, and the complements thereof, wherein said contiguous span comprises at least one of the following nucleotide positions of SEQ ID No. 1: 40939 to 78463; or

b) a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides of any of SEQ ID NOs: 1, 2, 4, 5, 7, 8, or 11-15, and the complements thereof, to the extent that such a length is consistent with the particular sequence ID.

The “nucleic acid codes of the invention” further encompass nucleotide sequences homologous to a contiguous span of at least 30, 35, 40, 50, 60, 70, 80, 90, 100, 150, 200, 500, 1000 or 2000 nucleotides, to the extent that such a length is consistent with the particular sequence of SEQ ID NOs: 1, 2, 4, 5, 7, 8, or 11-15, and the complements thereof. The “nucleic acid codes of the invention” also encompass nucleotide sequences homologous to a contiguous span of at least 12, 15, 18, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 90 or 100 nucleotides of SEQ ID No. 1 or the complements thereof, wherein said contiguous span comprises at least one of the following nucleotide positions of SEQ ID No. 1:

(i) 40939 to 78463; or
(ii) 41118, 69461, 74320, or 78451;

Homologous sequences refer to a sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, or 75% homology to these contiguous spans. Homology may be determined using
any method described herein, including BLAST2N with the default parameters or with any modified parameters. Homologous sequences also may include RNA sequences in which uridines replace the thymines in the nucleic acid codes of the invention. It will be appreciated that the nucleic acid codes of the invention can be represented in the traditional single character format (See the inside back cover of Stryer, Lubert. *Biochemistry*, 3rd edition. W. H Freeman & Co., New York.) or in any other format or code which records the identity of the nucleotides in a sequence.

As used herein the term “polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10” encompasses the polypeptide sequence of SEQ ID Nos 3, 6, 9, and 10, polypeptide sequences homologous to the polypeptides of SEQ ID Nos. 3, 6, 9, and 10, or fragments of any of the preceding sequences. Homologous polypeptide sequences refer to a polypeptide sequence having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75% homology to one of the polypeptide sequences of SEQ ID Nos. 3, 6, 9, and 10. Homology may be determined using any of the computer programs and parameters described herein, including FASTA with the default parameters or with any modified parameters. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error as described above. The polypeptide fragments comprise at least 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 50, 75, 100, or 150 consecutive amino acids of the polypeptides of SEQ ID Nos. 3, 6, 9, and 10. Preferably, the fragments are novel fragments. It will be appreciated that the polypeptide codes of the SEQ ID Nos. 3, 6, 9, and 10 can be represented in the traditional single character format or three letter format (See the inside back cover of Starrier, Lubert. *Biochemistry*, 3rd edition. W. H Freeman & Co., New York.) or in any other format which relates the identity of the polypeptides in a sequence.

It will be appreciated by those skilled in the art that the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11-15 and polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 can be stored, recorded, and manipulated on any medium which can be read and accessed by a computer. As used herein, the words “recorded” and “stored” refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any of the presently known methods for recording information on a computer readable medium to generate embodiment comprising one or more of nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11-15, or one or more of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, 20, 25, 30, or 50 nucleic acid codes of SEQ ID Nos 1, 2, 4, 5, 7, 8, 11-15. Another aspect of the present invention is a computer readable medium having recorded thereon at least 2, 5, 10, 15, 20, 25, 30, or 50 polypeptide codes of SEQ ID Nos 3, 6, 9, and 10.

Computer readable media include magnetically readable media, optically readable media, electronically readable media and magnetic/optical media. For example, the computer readable media may be a hard disk, a floppy disk, a magnetic tape, CD-ROM, Digital Versatile Disk (DVD), Random Access Memory (RAM), or Read Only Memory (ROM) as well as other types of other media known to those skilled in the art.
Embellishments of the present invention include systems, particularly computer systems which store and manipulate the sequence information described herein. One example of a computer system 100 is illustrated in block diagram form in Figure 19. As used herein, “a computer system” refers to the hardware components, software components, and data storage components used to analyze the nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11-15, or the amino acid sequences of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10. In one embodiment, the computer system 100 is a Sun Enterprise 1000 server (Sun Microsystems, Palo Alto, CA). The computer system 100 preferably includes a processor for processing, accessing and manipulating the sequence data. The processor 105 can be any well-known type of central processing unit, such as the Pentium III from Intel Corporation, or similar processor from Sun, Motorola, Compaq or International Business Machines.

Preferably, the computer system 100 is a general purpose system that comprises the processor 105 and one or more internal data storage components 110 for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

In one particular embodiment, the computer system 100 includes a processor 105 connected to a bus which is connected to a main memory 115 (preferably implemented as RAM) and one or more internal data storage devices 110, such as a hard drive and/or other computer readable media having data recorded thereon. In some embodiments, the computer system 100 further includes one or more data retrieving device 118 for reading the data stored on the internal data storage devices 110.

The data retrieving device 118 may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, etc. In some embodiments, the internal data storage device 110 is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system 100 may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device.

The computer system 100 includes a display 120 which is used to display output to a computer user. It should also be noted that the computer system 100 can be linked to other computer systems 125a-c in a network or wide area network to provide centralized access to the computer system 100.

Software for accessing and processing the nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11-15, or the amino acid sequences of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 (such as search tools, compare tools, and modeling tools etc.) may reside in main memory 115 during execution.

In some embodiments, the computer system 100 may further comprise a sequence comparer for comparing the above-described nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11-15 or polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 stored on a computer readable medium to reference nucleotide or polypeptide sequences stored on a computer readable medium. A “sequence comparer” refers to one or more programs which are implemented on the computer system 100 to compare a
nucleotide or polypeptide sequence with other nucleotide or polypeptide sequences and/or compounds including but not limited to peptides, peptidomimetics, and chemicals stored within the data storage means. For example, the sequence comparer may compare the nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, 11-15, or the amino acid sequences of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 stored on a computer readable medium to reference sequences stored on a computer readable medium to identify homologies, motifs implicated in biological function, or structural motifs. The various sequence comparer programs identified elsewhere in this patent specification are particularly contemplated for use in this aspect of the invention.

A process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database. The database of sequences can be a private database stored within the computer system 100, or a public database such as GENBANK, PIR OR SWISSPROT that is available through the Internet. The methodology for such a process has been previously described in a related US Patent Application 09/539,333 and international application PCT/IB00/00435.

The process 200 begins at a start state 201 and then moves to a state 202 wherein the new sequence to be compared is stored to a memory in a computer system 100. The memory could be any type of memory, including RAM or an internal storage device.

The process 200 then moves to a state 204 wherein a database of sequences is opened for analysis and comparison. The process 200 then moves to a state 206 wherein the first sequence stored in the database is read into a memory on the computer. A comparison is then performed at a state 210 to determine if the first sequence is the same as the second sequence. It is important to note that this step is not limited to performing an exact comparison between the new sequence and the first sequence in the database. Well-known methods are known to those of skill in the art for comparing two nucleotide or protein sequences, even if they are not identical. For example, gaps can be introduced into one sequence in order to raise the homology level between the two tested sequences. The parameters that control whether gaps or other features are introduced into a sequence during comparison are normally entered by the user of the computer system.

Once a comparison of the two sequences has been performed at the state 210, a determination is made at a decision state 210 whether the two sequences are the same. Of course, the term "same" is not limited to sequences that are absolutely identical. Sequences that are within the homology parameters entered by the user will be marked as "same" in the process 200.

If a determination is made that the two sequences are the same, the process 200 moves to a state 214 wherein the name of the sequence from the database is displayed to the user. This state notifies the user that the sequence with the displayed name fulfills the homology constraints that were entered. Once the name of the stored sequence is displayed to the user, the process 200 moves to a decision state 218 wherein a determination is made whether more sequences exist in the database. If no more sequences exist in the database, then the process 200 terminates at an end state 220. However, if more
sequences do exist in the database, then the process 200 moves to a state 224 wherein a pointer is moved to the next sequence in the database so that it can be compared to the new sequence. In this manner, the new sequence is aligned and compared with every sequence in the database.

It should be noted that if a determination had been made at the decision state 212 that the sequences were not homologous, then the process 200 would move immediately to the decision state 218 in order to determine if any other sequences were available in the database for comparison.

Accordingly, one aspect of the present invention is a computer system comprising a processor, a data storage device having stored thereon a nucleic acid code of SEQ ID NOs.1, 2, 4, 5, 7, 8, and 11-15 or a polypeptide code of SEQ ID Nos 3, 6, 9, and 10, a data storage device having retrievably stored thereon reference nucleotide sequences or polypeptide sequences to be compared to the nucleic acid code of SEQ ID NOs.1, 2, 4, 5, 7, 8, and 11-15 or a polypeptide code of SEQ ID Nos 3, 6, 9, and 10 and a sequence comparer for conducting the comparison. The sequence comparer may indicate a homology level between the sequences compared or identify structural motifs in the above described nucleic acid code of SEQ ID NOs.1, 2, 4, 5, 7, 8, and 11-15 or a polypeptide code of SEQ ID Nos 3, 6, 9, and 10 or it may identify structural motifs in sequences which are compared to these nucleic acid codes and polypeptide codes. In some embodiments, the data storage device may have stored thereon the sequences of at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of SEQ ID NOs.1, 2, 4, 5, 7, 8, and 11-15 or a polypeptide code of SEQ ID Nos 3, 6, 9, and 10.

Another aspect of the present invention is a method for determining the level of homology between a nucleic acid code of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 and a reference nucleotide sequence, comprising the steps of reading the nucleic acid code and the reference nucleotide sequence through the use of a computer program which determines homology levels and determining homology between the nucleic acid code and the reference nucleotide sequence with the computer program. The computer program may be any of a number of computer programs for determining homology levels, including those specifically enumerated herein, including BLAST2N with the default parameters or with any modified parameters. The method may be implemented using the computer systems described above. The method may also be performed by reading 2, 5, 10, 15, 20, 25, 30, or 50 of the above described nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 through use of the computer program and determining homology between the nucleic acid codes and reference nucleotide sequences.

Another embodiment is directed to a process 250 in a computer for determining whether two sequences are homologous. The process 250 begins at a start state 252 and then moves to a state 254 wherein a first sequence to be compared is stored to a memory. The second sequence to be compared is then stored to a memory at a state 256. The process 250 then moves to a state 260 wherein the first character in the first sequence is read and then to a state 262 wherein the first character of the second sequence is read. It should be understood that if the sequence is a nucleotide
sequence, then the character would normally be either A, T, C, G or U. If the sequence is a protein sequence, then it should be in the single letter amino acid code so that the first and sequence sequences can be easily compared.

A determination is then made at a decision state 264 whether the two characters are the same. If they are the same, then the process 250 moves to a state 268 wherein the next characters in the first and second sequences are read. A determination is then made whether the next characters are the same. If they are, then the process 250 continues this loop until two characters are not the same. If a determination is made that the next two characters are not the same, the process 250 moves to a decision state 274 to determine whether there are any more characters either sequence to read.

If there aren’t any more characters to read, then the process 250 moves to a state 276 wherein the level of homology between the first and second sequences is displayed to the user. The level of homology is determined by calculating the proportion of characters between the sequences that were the same out of the total number of sequences in the first sequence. Thus, if every character in a first 100 nucleotide sequence aligned with a every character in a second sequence, the homology level would be 100%.

Alternatively, the computer program may be a computer program which compares the nucleotide sequences of the nucleic acid codes of the present invention, to reference nucleotide sequences in order to determine whether the nucleic acid code of SEQ ID NOs:1, 2, 4, 5, 7, 8, and 11-15 differs from a reference nucleic acid sequence at one or more positions. Optionally such a program records the length and identity of inserted, deleted or substituted nucleotides with respect to the sequence of either the reference polynucleotide or the nucleic acid code of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15. In one embodiment, the computer program may be a program which determines whether the nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 contain a biallelic marker or single nucleotide polymorphism (SNP) with respect to a reference nucleotide sequence. This single nucleotide polymorphism may comprise a single base substitution, insertion, or deletion, while this biallelic marker may comprise about one to ten consecutive bases substituted, inserted or deleted.

Another aspect of the present invention is a method for determining the level of homology between a polypeptide code of SEQ ID Nos. 3, 6, 9, and 10 and a reference polypeptide sequence, comprising the steps of reading the polypeptide code of SEQ ID Nos. 3, 6, 9, and 10 and the reference polypeptide sequence through use of a computer program which determines homology levels and determining homology between the polypeptide code and the reference polypeptide sequence using the computer program. Accordingly, another aspect of the present invention is a method for determining whether a nucleic acid code of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 differs at one or more nucleotides from a reference nucleotide sequence comprising the steps of reading the nucleic acid code and the reference
nucleotide sequence through use of a computer program which identifies differences between nucleic acid sequences and identifying differences between the nucleic acid code and the reference nucleotide sequence with the computer program. In some embodiments, the computer program is a program which identifies single nucleotide polymorphisms. The method may be implemented by the computer systems described above and the method illustrated in Figure 21. The method may also be performed by reading at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 and the reference nucleotide sequences through the use of the computer program and identifying differences between the nucleic acid codes and the reference nucleotide sequences with the computer program.

In other embodiments the computer based system may further comprise an identifier for identifying features within the nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the amino acid sequences of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10.

An “identifier” refers to one or more programs which identifies certain features within the above-described nucleotide sequences of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the amino acid sequences of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10. In one embodiment, the identifier may comprise a program which identifies an open reading frame in the cDNAs codes of SEQ ID Nos 2, 5, 7, and 8.

Another embodiment is an identifier process 300 for detecting the presence of a feature in a sequence. The process 300 begins at a start state 302 and then moves to a state 304 wherein a first sequence that is to be checked for features is stored to a memory 115 in the computer system 100. The process 300 then moves to a state 306 wherein a database of sequence features is opened. Such a database would include a list of each feature’s attributes along with the name of the feature. For example, a feature name could be “Initiation Codon” and the attribute would be “ATG”. Another example would be the feature name “TAATAA Box” and the feature attribute would be “TAATAA”. An example of such a database is produced by the University of Wisconsin Genetics Computer Group (www.gcg.com).

Once the database of features is opened at the state 306, the process 300 moves to a state 308 wherein the first feature is read from the database. A comparison of the attribute of the first feature with the first sequence is then made at a state 310. A determination is then made at a decision state 316 whether the attribute of the feature was found in the first sequence. If the attribute was found, then the process 300 moves to a state 318 wherein the name of the found feature is displayed to the user.

The process 300 then moves to a decision state 320 wherein a determination is made whether move features exist in the database. If no more features do exist, then the process 300 terminates at an end state 324. However, if more features do exist in the database, then the process 300 reads the next sequence feature at a state 326 and loops back to the state 310 wherein the attribute of the next feature is compared against the first sequence.
It should be noted, that if the feature attribute is not found in the first sequence at the
decision state 316, the process 300 moves directly to the decision state 320 in order to determine if
any more features exist in the database.

In another embodiment, the identifier may comprise a molecular modeling program which
determines the 3-dimensional structure of the polypeptides codes of SEQ ID Nos. 3, 6, 9, and 10. In
some embodiments, the molecular modeling program identifies target sequences that are most
compatible with profiles representing the structural environments of the residues in known three-
dimensional protein structures. (See, e.g., Eisenberg et al., U.S. Patent No. 5,436,850 issued July 25,
1995). In another technique, the known three-dimensional structures of proteins in a given family
are superimposed to define the structurally conserved regions in that family. This protein modeling
technique also uses the known three-dimensional structure of a homologous protein to approximate
the structure of the polypeptide codes of SEQ ID Nos. 4 to 8. (See e.g., Srinivasan, et al., U.S.
have been used routinely to build models of proteases and antibodies. (Sowdhambini et al., Protein
Engineering 10:207, 215 (1997)). Comparative approaches can also be used to develop three-
dimensional protein models when the protein of interest has poor sequence identity to template
proteins. In some cases, proteins fold into similar three-dimensional structures despite having very
weak sequence identities. For example, the three-dimensional structures of a number of helical
cytokines fold in similar three-dimensional topology in spite of weak sequence homology.

The recent development of threading methods now enables the identification of likely
folding patterns in a number of situations where the structural relatedness between target and
template(s) is not detectable at the sequence level. Hybrid methods, in which fold recognition is
performed using Multiple Sequence Threading (MST), structural equivalencies are deduced from the
threading output using a distance geometry program DRAGON to construct a low resolution model,
and a full-atom representation is constructed using a molecular modeling package such as
QUANTA.

According to this 3-step approach, candidate templates are first identified by using the novel
fold recognition algorithm MST, which is capable of performing simultaneous threading of multiplealigned sequences onto one or more 3-D structures. In a second step, the structural equivalencies
obtained from the MST output are converted into interresidue distance restraints and fed into the
distance geometry program DRAGON, together with auxiliary information obtained from secondary
structure predictions. The program combines the restraints in an unbiased manner and rapidly
generates a large number of low resolution model confirmations. In a third step, these low
resolution model confirmations are converted into full-atom models and subjected to energy
minimization using the molecular modeling package QUANTA. (See e.g., Aszódi et al.,
Proteins:Structure, Function, and Genetics, Supplement 1:38-42 (1997)).

The results of the molecular modeling analysis may then be used in rational drug design
techniques to identify agents which modulate the activity of the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10.

Accordingly, another aspect of the present invention is a method of identifying a feature within the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 comprising reading the nucleic acid code(s) or the polypeptide code(s) through the use of a computer program which identifies features therein and identifying features within the nucleic acid code(s) or polypeptide code(s) with the computer program. In one embodiment, computer program comprises a computer program which identifies open reading frames. In a further embodiment, the computer program identifies structural motifs in a polypeptide sequence. In another embodiment, the computer program comprises a molecular modeling program. The method may be performed by reading a single sequence or at least 2, 5, 10, 15, 20, 25, 30, or 50 of the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 through the use of the computer program and identifying features within the nucleic acid codes or polypeptide codes with the computer program.

The nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 may be stored and manipulated in a variety of data processor programs in a variety of formats. For example, the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10 may be stored as text in a word processing file, such as Microsoft WORD or WORDPERFECT or as an ASCII file in a variety of database programs familiar to those of skill in the art, such as DB2, SYBASE, or ORACLE. In addition, many computer programs and databases may be used as sequence comparers, identifiers, or sources of reference nucleotide or polypeptide sequences to be compared to the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10. The following list is intended not to limit the invention but to provide guidance to programs and databases which are useful with the nucleic acid codes of SEQ ID Nos. 1, 2, 4, 5, 7, 8, and 11-15 or the polypeptide codes of SEQ ID Nos. 3, 6, 9, and 10. The programs and databases which may be used include, but are not limited to: MacPattern (EMBL), DiscoveryBase (Molecular Applications Group), GeneMine (Molecular Applications Group), Look (Molecular Applications Group), MacLook (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul et al, J. Mol. Biol. 215: 403 (1990)), FASTA (Pearson and Lipman, Proc. Natl. Acad. Sci. USA, 85: 2444 (1988)), FASTDB (Brutlag et al. Comp. App. Biosci. 6:237-245, 1990), Catalyst (Molecular Simulations Inc.), Catalyst/SHAPE (Molecular Simulations Inc.), Cerius2 DBAccess (Molecular Simulations Inc.), HypoGen (Molecular Simulations Inc.), Insight II, (Molecular Simulations Inc.), Discover (Molecular Simulations Inc.), CHARMm (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), DelPhi, (Molecular Simulations Inc.), QuanteMM, (Molecular Simulations Inc.), Homology (Molecular Simulations Inc.), Modeler (Molecular Simulations Inc.), ISIS (Molecular Simulations Inc.), Quanta/Protein Design (Molecular Simulations Inc.), WebLab (Molecular Simulations Inc.), WebLab
Diversity Explorer (Molecular Simulations Inc.), Gene Explorer (Molecular Simulations Inc.), SeqFold (Molecular Simulations Inc.), the EMBL/Swissprotein database, the MDL Available Chemicals Directory database, the MDL Drug Data Report database, the Comprehensive Medicinal Chemistry database, Derwents's World Drug Index database, the BioByteMasterFile database, the Genbank database, and the Genseq database. Many other programs and data bases would be apparent to one of skill in the art given the present disclosure.

Motifs which may be detected using the above programs include sequences encoding leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences encoding signal peptides which direct the secretion of the encoded proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

Throughout this application, various publications, patents, and published patent applications are cited. The disclosures of the publications, patents, and published patent specifications referenced in this application are all hereby incorporated by reference in their entireties into the present disclosure to more fully describe the state of the art to which this invention pertains.

**EXAMPLES**

Several of the methods of the present invention are described in the following examples, which are offered by way of illustration and not by way of limitation. Many other modifications and variations of the invention as herein set forth can be made without departing from the spirit and scope thereof and therefore only such limitations should be imposed as are indicated by the appended claims.

**Example 1**

**Identification Of Biallelic Markers: DNA Extraction**

Donors were unrelated and healthy. They presented a sufficient diversity for being representative of a heterogeneous population. The DNA from 100 individuals was extracted and tested for the detection of the biallelic markers.

30 ml of peripheral venous blood were taken from each donor in the presence of EDTA. Cells (pellet) were collected after centrifugation for 10 minutes at 2000 rpm. Red cells were lysed by a lysis solution (50 ml final volume: 10 mM Tris pH 7.6; 5 mM MgCl2; 10 mM NaCl). The solution was centrifuged (10 minutes, 2000 rpm) as many times as necessary to eliminate the residual red cells present in the supernatant, after resuspension of the pellet in the lysis solution.

The pellet of white cells was lysed overnight at 42°C with 3.7 ml of lysis solution composed of:

- 3 ml TE 10-2 (Tris-HCl 10 mM, EDTA 2 mM) / NaCl 0.4 M
- 200 μl SDS 10%
- 500 μl K-proteinase (2 mg K-proteinase in TE 10-2 / NaCl 0.4 M).

For the extraction of proteins, 1 ml saturated NaCl (6M) (1/3.5 v/v) was added. After vigorous agitation, the solution was centrifuged for 20 minutes at 10000 rpm.

For the precipitation of DNA, 2 to 3 volumes of 100% ethanol were added to the previous supernatant, and the solution was centrifuged for 30 minutes at 2000 rpm. The DNA solution was rinsed three times with 70% ethanol to eliminate salts, and centrifuged for 20 minutes at 2000 rpm. The pellet was dried at 37°C, and resuspended in 1 ml TE 10-1 or 1 ml water. The DNA concentration was evaluated by measuring the OD at 260 nm (1 unit OD = 50 μg/ml DNA). To determine the presence of proteins in the DNA solution, the OD 260 / OD 280 ratio was determined. Only DNA preparations having a OD 260 / OD 280 ratio between 1.8 and 2 were used in the subsequent examples described below.

The pool was constituted by mixing equivalent quantities of DNA from each individual.

Example 2

Identification Of Biallelic Markers: Amplification Of Genomic DNA By PCR

The amplification of specific genomic sequences of the DNA samples of Example 1 was carried out on the pool of DNA obtained previously. In addition, 50 individual samples were similarly amplified.

PCR assays were performed using the following protocol:

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final volume</td>
<td>25 μl</td>
</tr>
<tr>
<td>DNA</td>
<td>2 ng/μl</td>
</tr>
<tr>
<td>MgCl₂</td>
<td>2 mM</td>
</tr>
<tr>
<td>dNTP (each)</td>
<td>200 μM</td>
</tr>
<tr>
<td>primer (each)</td>
<td>2.9 ng/μl</td>
</tr>
<tr>
<td>Ampli Taq Gold DNA polymerase</td>
<td>0.05 unit/μl</td>
</tr>
<tr>
<td>PCR buffer (10x = 0.1 M TrisHCl pH8.3 0.5M KCl)</td>
<td>1x</td>
</tr>
</tbody>
</table>

Each pair of first primers was designed using the sequence information of genomic DNA sequences of SEQ ID Nos 1 and 4 disclosed herein and the OSP software (Hillier & Green, 1991). This first pair of primers was about 20 nucleotides in length and had the sequences disclosed in SEQ ID NO:1, indicated by 27-81_rp and 27-81_pu complement. This primer pair will amplify the region of marker 27-81-180. Primer pairs for the other biallelic markers of the invention are listed in SEQ ID NO:1 and 4 in an analogous manner.

Preferably, the primers contained a common oligonucleotide tail upstream of the specific bases targeted for amplification which was useful for sequencing.

The synthesis of these primers was performed following the phosphoramidite method, on a
GENSET UFPS 24.1 synthesizer.

DNA amplification was performed on a Genius II thermocycler. After heating at 95°C for 10 min, 40 cycles were performed. Each cycle comprised: 30 sec at 95°C, 54°C for 1 min, and 30 sec at 72°C. For final elongation, 10 min at 72°C ended the amplification. The quantities of the amplification products obtained were determined on 96-well microtiter plates, using a fluorometer and Picogreen as intercalant agent (Molecular Probes).

Example 3

Identification of Polymorphisms

a) Identification of Biallelic Markers from Amplified Genomic DNA of Example 2

The sequencing of the amplified DNA obtained in Example 2 was carried out on ABI 377 sequencers. The sequences of the amplification products were determined using automated dideoxy terminator sequencing reactions with a dye terminator cycle sequencing protocol. The products of the sequencing reactions were run on sequencing gels and the sequences were determined using gel image analysis (ABI Prism DNA Sequencing Analysis software (2.1.2 version)).

The sequence data were further evaluated to detect the presence of biallelic markers within the amplified fragments. The polymorphism search was based on the presence of superimposed peaks in the electrophoresis pattern resulting from different bases occurring at the same position as described previously.

The localization of the biallelic markers detected in the fragments of amplification are as shown below in Table 2.

<table>
<thead>
<tr>
<th>Amplicon</th>
<th>Marker Name</th>
<th>Polymorphism</th>
<th>SEQ ID No.</th>
<th>BM position in SEQID</th>
<th>Position of probes in SEQ ID No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>27-81</td>
<td>27-81-180</td>
<td>G A</td>
<td>1</td>
<td>41118</td>
<td>41106 41130</td>
</tr>
<tr>
<td>27-29</td>
<td>27-29-224</td>
<td>T g</td>
<td>1</td>
<td>69461</td>
<td>69449 69473</td>
</tr>
<tr>
<td>27-2</td>
<td>27-2-106</td>
<td>C A</td>
<td>1</td>
<td>74320</td>
<td>74308 74332</td>
</tr>
<tr>
<td>27-30</td>
<td>27-30-249</td>
<td>C T</td>
<td>1</td>
<td>78451</td>
<td>78439 78463</td>
</tr>
<tr>
<td>27-1</td>
<td>27-1-61</td>
<td>A G</td>
<td>4</td>
<td>61</td>
<td>49 73</td>
</tr>
</tbody>
</table>

BM refers to “biallelic marker”. All1 and all2 refer respectively to allele 1 and allele 2 of the biallelic marker.

b) Identification of Polymorphisms by Comparison of Genomic DNA from Overlapping BACs

Genomic DNA from multiple BACs derived from the same DNA donor sample and
overlapping in regions of genomic DNA of SEQ ID No. 1 was sequenced. Sequencing was carried out on ABI 377 sequencers. The sequences of the amplification products were determined using automated dideoxy terminator sequencing reactions with a dye terminator cycle sequencing protocol. The products of the sequencing reactions were run on sequencing gels and the sequences were determined using gel image analysis (ABI Prism DNA Sequencing Analysis software (2.1.2 version)).

Example 4

Validation Of The Polymorphisms Through Microsequencing

The biallelic markers identified in Example 3 were further confirmed and their respective frequencies were determined through microsequencing. Microsequencing was carried out for each individual DNA sample described in Example 1.

Amplification from genomic DNA of individuals was performed by PCR as described above for the detection of the biallelic markers with the same set of PCR primers described in SEQ ID NO:1 and 4 (prefixed “.rp” and “.pu complement”).

The preferred primers used in microsequencing were about 19 nucleotides in length and hybridized just upstream of the considered polymorphic base. According to the invention, the primers used in microsequencing are detailed in SEQ ID NO:1 and 4 (prefixed “.mis” and “.mis complement”).

As example, for biallelic marker 27-2-106, amplification primers 27-2.rp and 27-2. pu complement are used to amplify the DNA (as Example 1) and microsequencing primers 27-2-106.mix and 27-2-106.mis complement are used according to the microsequencing reaction performed as follows:

After purification of the amplification products, the microsequencing reaction mixture was prepared by adding, in a 20 µl final volume: 10 pmol microsequencing oligonucleotide, 1 U Thermosequenase (Amersham E79000G), 1.25 µl Thermosequenase buffer (260 mM Tris HCl pH 9.5, 65 mM MgCl₂), and the two appropriate fluorescent ddNTPs (Perkin Elmer, Dye Terminator Set 401095) complementary to the nucleotides at the polymorphic site of each biallelic marker tested, following the manufacturer’s recommendations. After 4 minutes at 94°C, 20 PCR cycles of 15 sec at 55°C, 5 sec at 72°C, and 10 sec at 94°C were carried out in a Tetrad PTC-225 thermocycler (MJ Research). The unincorporated dye terminators were then removed by ethanol precipitation. Samples were finally resuspended in formamide-EDTA loading buffer and heated for 2 min at 95°C before being loaded on a polyacrylamide sequencing gel. The data were collected by an ABI PRISM 377 DNA sequencer and processed using the GENESCAN software (Perkin Elmer).

Following gel analysis, data were automatically processed with software that allows the determination of the alleles of biallelic markers present in each amplified fragment.

The software evaluates such factors as whether the intensities of the signals resulting from the above microsequencing procedures are weak, normal, or saturated, or whether the signals are
ambiguous. In addition, the software identifies significant peaks (according to shape and height criteria). Among the significant peaks, peaks corresponding to the targeted site are identified based on their position. When two significant peaks are detected for the same position, each sample is categorized classification as homozygous or heterozygous type based on the height ratio.

Example 5

Association Study Between Schizophrenia And The Biallelic Markers Of The Invention:

Collection Of DNA Samples From Affected And Non-Affected Individuals

A) Affected population

All the samples were collected from a large epidemiological study of schizophrenia undertaken in hospital centers of Quebec from October 1995 to April 1997. The population was composed of French Caucasian individuals. The study design consisted in the ascertainment of cases and two of their first degree relatives (parents or siblings).

As a whole, 956 schizophrenic cases were ascertained according to the following inclusion criteria:

- the diagnosis had been done by a psychiatrist;
- the diagnosis had been done at least 3 years before recruitment time, in order to exclude individuals suffering from transient manic-depressive psychosis or depressive disorders;
- the patient ancestors had been living in Quebec for at least 6 generations;
- it was possible to get a blood sample from 2 close relatives.

Among the 956 schizophrenic ascertained cases, 834 individuals were included in the study for the following reasons:

- for the included individual cases, the diagnosis of schizophrenia was established according to the DSM-IV (Diagnostic and Statistical Manual, Fourth edition, Revised 1994, American Psychiatric Press);
- samples from individuals suffering from schizoaffective disorder were discarded;
- individuals suffering from catatonic schizophrenia were also excluded from the population of schizophrenic cases;
- were also excluded the individuals having a first degree relative or 2 or more second degree relatives suffering from depression or mood disorder;
- individuals having had severe head trauma, severe obstetrical complications, encephalitis, or meningitis before onset of symptoms were also excluded;
- has also been excluded from the population of schizophrenic cases a patient suffering from epilepsy and treated with anticonvulsants.

The age at onset was not added as an inclusion criteria.

B) Unaffected population
Control cases were respectively ascertained based on the following cumulative criteria:
- the individual must not be affected by schizophrenia or any other psychiatric disorder;
- the individual must have 35 years old or more;
- the individual must belong to the French-Canadian population;
- the individual must have one or two first degree relative available for blood sampling.
Controls were matched with cases sex when possible.

C) Cases and Control Populations Selected for the Association Study

The unaffected population retained for the study was composed of 241 individuals. The initial sample of the clinical study was composed of 215 cases and 214 controls. The controls were composed of 116 males and 98 females while the cases were composed of 154 males and 64 females. For each control, two first degree relatives (father, mother, sisters and brothers) were available. In order to match the sex of cases and controls, the parents of female controls were substituted for the female controls where possible and where the parents were known to be unaffected by schizophrenia or other psychosis. The parents of 27 female controls were thus substituted for the respective females, resulting in a total control sample size of 241 individuals.

The association data that are presented below in Table 3 wherein the individuals have been randomly selected from the populations described above.

### Table 3

**ASSOCIATION RESULTS DAAO - PJ 27 Algene sample (213 cases, 241 controls)**

<table>
<thead>
<tr>
<th>Mks</th>
<th>Location</th>
<th>Chosen Allele</th>
<th>Allelic freq Diff.</th>
<th>Chi Sq.</th>
<th>p.value</th>
<th>Chi sq.</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>27-8/180</td>
<td>intron 2</td>
<td>C</td>
<td>0.021</td>
<td>0.3768</td>
<td>5.39E-01</td>
<td>0.4763</td>
<td>7.88E-01</td>
</tr>
<tr>
<td>27-29/224</td>
<td>intron V</td>
<td>C</td>
<td>0.073</td>
<td>5.6175</td>
<td>1.78E-02</td>
<td>5.7353</td>
<td>5.68E-02</td>
</tr>
<tr>
<td>27-7/2/106</td>
<td>intron 4</td>
<td>T</td>
<td>0.104</td>
<td>9.8824</td>
<td>1.67E-03</td>
<td>11.7828</td>
<td>2.76E-03</td>
</tr>
<tr>
<td>27-30/249</td>
<td>intron 6</td>
<td>A</td>
<td>0.013</td>
<td>4.97E-01</td>
<td>(*)</td>
<td>8.57E-02</td>
<td>(*)</td>
</tr>
<tr>
<td>27-1/61</td>
<td>3' of gene</td>
<td>G</td>
<td>0.022</td>
<td>0.4412</td>
<td>5.07E-01</td>
<td>0.5981</td>
<td>7.42E-01</td>
</tr>
</tbody>
</table>

(*) : exact test

Both case and control populations form two groups, each group consisting of unrelated individuals that do not share a known common ancestor. Additionally, the individuals of the control population were selected among those having no family history of schizophrenia or schizophrenic disorder.

**Genotyping of affected and control individuals**

A) Results from the genotyping

The general strategy to perform the association studies was to individually scan the DNA samples from all individuals in each of the populations described above in order to establish the
allele frequencies of biallelic markers, and among them the biallelic markers of the invention, in the diploid genome of the tested individuals belonging to each of these populations.

Allelic frequencies of every biallelic marker in each population (cases and controls) were determined by performing microsequencing reactions on amplified fragments obtained by genomic PCR performed on the DNA samples from each individual. Genomic PCR and microsequencing were performed as detailed above in Examples 1 to 3 using the described PCR and microsequencing primers.

**Single biallelic marker frequency analysis**

For each allele of the biallelic markers included in this study, the difference between the allelic frequency in the unaffected population and in the population affected by schizophrenia was calculated and the absolute value of the difference was determined. The more the difference in allelic frequency for a particular biallelic marker or a particular set of biallelic markers, the more probable an association between the genomic region harboring this particular biallelic marker or set of biallelic markers and schizophrenia. Allelic frequencies were also useful to check that the markers used in the haplotype studies meet the Hardy-Weinberg proportions (random mating).

In the association study described herein, several individual biallelic markers were shown to be significantly associated with schizophrenia. In particular, 27-2-106 and 27-29-224 showed significant association with schizophrenia.

**Haplotype frequency analysis**

Analysis of markers Haplotype analysis for association of chromosome DAO related biallelic markers and schizophrenia was performed by estimating the frequencies of all possible 2, 3 and 4 marker haplotypes in the affected and control populations described above. Haplotype estimations were performed by applying the Expectation-Maximization (EM) algorithm (Excoffier and Slatkin, 1995), using the EM-HAPLO program (Hawley et al., 1994) as described above. Estimated haplotype frequencies in the affected and control population were compared by means of a chi-square statistical test (one degree of freedom).

**Example 6**

**Forensic Matching by DNA Sequencing**

In one exemplary method, DNA samples are isolated from forensic specimens of, for example, hair, semen, blood or skin cells by conventional methods. A panel of PCR primers based on a number of the 5' ESTs, or cDNAs or genomic DNAs isolated therefrom as described above, is then utilized in accordance with Example 41 to amplify DNA of approximately 100-200 bases in length from the forensic specimen. Corresponding sequences are obtained from a test subject. Each of these identification DNAs is then sequenced using standard techniques, and a simple database comparison determines the differences, if any, between the sequences from the subject and those from the sample. Statistically significant differences between the suspect's DNA sequences and those from the sample conclusively prove a lack of identity. This lack of identity can be proven, for example, with only one
sequence. Identity, on the other hand, should be demonstrated with a large number of sequences, all matching. Preferably, a minimum of 50 statistically identical sequences of 100 bases in length are used to prove identity between the suspect and the sample.

**Positive Identification by DNA Sequencing**

The technique outlined in the previous example may also be used on a larger scale to provide a unique fingerprint-type identification of any individual. In this technique, primers are prepared from those described in SEQ ID NO:1 and 4, or cDNA or genomic DNA sequences obtainable therefrom. These primers are used to obtain a corresponding number of PCR-generated DNA segments from the individual in question in accordance with Example 1. The database of sequences generated through this procedure uniquely identifies the individual from whom the sequences were obtained. The same panel of primers may then be used at any later time to absolutely correlate tissue or other biological specimen with that individual.

**Southern Blot Forensic Identification**

The procedure above is repeated to obtain a panel of at least 5 amplified sequences from an individual and a specimen. This PCR-generated DNA is then digested with one or a combination of, preferably, four base specific restriction enzymes. Such enzymes are commercially available and known to those of skill in the art. After digestion, the resultant gene fragments are size separated in multiple duplicate wells on an agarose gel and transferred to nitrocellulose using Southern blotting techniques well known to those with skill in the art.

A panel of probes based on the sequences of the 5’ ESTs (or cDNAs or genomic DNAs obtainable therefrom), or fragments thereof of at least 8, 10, 12, 15, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300, 500, or 1000 bases, are radioactively or colorimetrically labeled using methods known in the art, such as nick translation or end labeling, and hybridized to the Southern blot using techniques known in the art.

Preferably, at least 5 of these labeled probes are used. The resultant bands appearing from the hybridization of a large sample of 5’ ESTs (or cDNAs or genomic DNAs obtainable therefrom) will be a unique identifier. Since the restriction enzyme cleavage will be different for every individual, the band pattern on the Southern blot will also be unique. Increasing the number of probes derived from 5’ ESTs (or cDNAs or genomic DNAs obtainable therefrom) will provide a statistically higher level of confidence in the identification since there will be an increased number of sets of bands used for identification.

**Alternative "Fingerprint" Identification Technique**

20-mer oligonucleotides are prepared from primers directed at the biallelic markers of the invention. Cell samples from the test subject are processed for DNA using techniques well known to those with skill in the art. The nucleic acid is digested with restriction enzymes such as EcoRI and XbaI. Following digestion, samples are applied to wells for electrophoresis. The procedure, as known in the art, may be modified to accommodate polyacrylamide electrophoresis, however in this example,
samples containing 5 µg of DNA are loaded into wells and separated on 0.8% agarose gels. The gels are transferred onto nitrocellulose using standard Southern blotting techniques.

10 ng of each of the oligonucleotides are pooled and end-labeled with $^{32}$P. The nitrocellulose is prehybridized with blocking solution and hybridized with the labeled probes. Following hybridization and washing, the nitrocellulose filter is exposed to X-Omat AR X-ray film. The resulting hybridization pattern will be unique for each individual. It is additionally contemplated within this example that the number of probe sequences used can be varied for additional accuracy or clarity.
The disclosures of all issued patents, published PCT applications, scientific references or other publications cited herein are incorporated herein by reference in their entireties.

Although this invention has been described in terms of certain preferred embodiments, other embodiments which will be apparent to those of ordinary skill in the art of view of the disclosure herein are also within the scope of this invention. Accordingly, the scope of the invention is intended to be defined only by reference to the appended claims.
CLAIMS

1. A method of determining a genotype of an individual comprising the steps of:
   a) obtaining a biological sample containing a polymucleotide from said individual; and
   b) determining the identity of a nucleotide at a biallelic marker of the DAO gene of SEQ ID
      NO:1 or 4, in said polymucleotide, wherein said nucleotide at said biallelic marker is
      selected from the group consisting of: nucleotide G at biallelic marker 27-81-180,
      nucleotide A at biallelic marker 27-81-180, nucleotide T at biallelic marker 27-29-224,
      nucleotide G at biallelic marker 27-29-224, nucleotide C at biallelic marker 27-2-106,
      nucleotide A at biallelic marker 27-2-106, nucleotide C at biallelic marker 27-30-249,
      nucleotide T at biallelic marker 27-30-249, nucleotide A at biallelic marker 27-1-61,
      nucleotide G at biallelic marker 27-1-61;
   and wherein said nucleotide determines said genotype of said individual.

2. A method of determining whether an association between an allele of a biallelic marker of the
   DAO gene of SEQ ID NO:1 or 4 and schizophrenia exists, wherein said allele is defined by the
   identity of a nucleotide at said biallelic marker, and wherein said nucleotide and said biallelic
   marker is selected from the group consisting of: nucleotide G at biallelic marker 27-81-180,
   nucleotide A at biallelic marker 27-81-180, nucleotide T at biallelic marker 27-29-224, nucleotide
   G at biallelic marker 27-29-224, nucleotide C at biallelic marker 27-2-106, nucleotide A at
   biallelic marker 27-2-106, nucleotide C at biallelic marker 27-30-249, nucleotide T at biallelic
   marker 27-30-249, nucleotide A at biallelic marker 27-1-61, nucleotide G at biallelic marker 27-
   1-61, comprising the steps of:
   a) determining a schizophrenia-positive frequency of said allele in a schizophrenia-positive
      population of at least 50 individuals;
   b) determining a schizophrenia-negative frequency of said allele in a schizophrenia-negative
      population of at least 50 individuals; and
   c) using said schizophrenia-positive frequency of step a) and said schizophrenia-negative
      frequency of step b) to determine statistically whether said association between said allele and
      schizophrenia exists.
GENSET

BIALLELIC MARKERS OF D-AMINO ACID OXIDASE AND USES THEREOF

160.WO1

60/340,400
2001-12-12
15
Patent.pm

1
86592
DNA
Homo sapiens
misc_feature
40665...64577
5' regulatory region
exon
64578...64711
exon U
exon
69488...69690
exon 2
exon
71942...72056
exon 3
exon
73962...74038
exon 4
exon
74701...74766
exon 5
exon
77478...77532
exon 6
exon
78762...78866
exon 7
exon
The text contains various annotations related to detected features and alleles. Here is a structured representation of the content:

- **Exon 8**: 81507..81589
- **Exon 9**: 83181..83298
- **Exon 10**: 83879..83977
- **Exon 11**: 84906..85430

- **Misc Feature**: 85431..86592

- **3' Regulatory Region**:

  - **Allele**: 41118
  - **Polymorphic Base**: G or A

- **Allele**: 69461
  - **Polymorphic Base**: T or G

- **Allele**: 74320
  - **Polymorphic Base**: C or A

- **Allele**: 78451
  - **Polymorphic Base**: C or T

- **Primer Bind**
  - **27-81**: 40939..40956
    - **RP**: 27-81
  - **27-81 Complement**: 41224..41242
  - **Polymorphic Base**: C or T
  - **27-29**: 69408..69428
    - **RP**: 27-29
  - **27-29 Complement**: 69666..69684
<220>
<221> primer_bind
<222> 74252..74272
<223> 27-2.rp

<220>
<221> primer_bind
<222> 74408..74425
<223> 27-2.pu complement

<220>
<221> primer_bind
<222> 78385..78404
<223> 27-30.rp

<220>
<221> primer_bind
<222> 78682..78699
<223> 27-30.pu complement

<220>
<221> primer_bind
<222> 41099..41117
<223> 27-81-180.mis

<220>
<221> primer_bind
<222> 41119..41137
<223> 27-81-180.mis complement

<220>
<221> primer_bind
<222> 69442..69460
<223> 27-29-224.mis

<220>
<221> primer_bind
<222> 69462..69480
<223> 27-29-224.mis complement

<220>
<221> primer_bind
<222> 74301..74319
<223> 27-2-106.mis

<220>
<221> primer_bind
<222> 74321..74339
<223> 27-2-106.mis complement

<220>
<221> primer_bind
<222> 78432..78450
<223> 27-30-249.mis

<220>
<221> primer_bind
<222> 78452..78470
<223> 27-30-249.mis complement

<220>
atttattgga atttattgga ttg 60
tagacacatg ctagg a 120
gctcctgctc atctag 300
agctcctcag caggc 360
atagttgtat tattattgt ggg 420
cattttttat cctcctcctc 480
agaagctcct ctgag 540
gacttttgct tctgctg 600
ctccaccctt gactggcctct acq 840
ctttccacag acq 1020
aacacgca cctcctgctg 1080
gctacagct caagc 1140
cacgcttgg ctgctg 1200
ctctcctcag ggtgacgc 1260
tgggtgttt gctctcctcag 1320
ctgggttaaa cacccccc 1380
tcgctgactc ctgggg 1440
cactgtgatc cctgctgctg 1500
gagaggga 1560
gatcagctgc cag 1620
gtggccctc aag 1740
ccttcttgg agtgctctt gg 1800
taccatct taccatctt g 1860
caggctcca aac 1920
ttgaggcat ctgggtgctg 1980
cattttgct g 2040
agacctgc aagtgagtt g 2100
gacgac aatctgctgctg 2220
tcgaggaat c 2280
tatttgc gctccttgcctgctg 2400
aatcgg gtttgccttctgcct 2460
atacgtgatc cctgctgctg 2520
tg 2580
<table>
<thead>
<tr>
<th>Line</th>
<th>Sequence</th>
<th>Description</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gtgtttctgg</td>
<td>tcctttccca</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>tcttgtaaccgc</td>
<td>tcgtggtaaggtctttgcattgc</td>
<td>29100</td>
</tr>
</tbody>
</table>
| 3     | ccctagacaacacagattcactcttggtgcctttctctccttcctttttggcccgggcatccatcactccccctgttgccgcttggcattttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
gtagggacct gaagagggcg caccaggaa ctggagaaa aagaagccag tggctcggaa 32880
gaccggacat atgtgacacct gcaatattatt ggtatgatt gaatacagc tgtacaccattt 32940
gatgacaaaa agatgaagaacct tattataatta gctgacagg agggaagaag cagagagaga 33040
atagccccctg acctcctggt tttcctctct cagccacctct tagaatgtag tgtgtagcag 33060
ctacctctctc gacatccagtt cagagacgtt cagacatcct ctagacct ctgagcagatg 33120
acgagaagct accaaaccatt ccaaaaatatt tctctcagct cagacatcctct ttagagcttt 33180
cagcagagaag agcagagagg cagagacat caggttggag ggtttaaagt cagagacacta 33240
caccagggac caccacccac ccccaaggct tgggttaaat ttagagcttt ggtttaaagt 33300
tcctcttctcct cacatcctt cagacatcct ttagagcattt cagagagactg cagagagactg 33360
tgtgtcaacct tggccacact ccccaaggct tgggttaaat ttagagcattt cagagagactg 33420
tggtctgacag ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct 33480
tcctgccctgcc ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct 33540
tgtggccctg ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct 33600
tagtctaaact acatctgtt ggtgagcagct ggtgagcagct ggtgagcagct ggtgagcagct 33660
tagctgactct gtagctctct gtagctctct gtagctctct gtagctctct gtagctctct 33720
ggctgacact cacccaccac gaaacgctct ggtggtagtg cttgctgcag tctcctctct 33780
tgctgactct gtagctctct gtagctctct gtagctctct gtagctctct gtagctctct 33840
caagacccag aagagagact aagagagact aagagagact aagagagact aagagagact 33900
taacatcctct gtagctctct gtagctctct gtagctctct gtagctctct gtagctctct 34020
tcctctctctct ctagctctct ctagctctct ctagctctct ctagctctct ctagctctct 34100
aacatcctct gtagctctct gtagctctct gtagctctct gtagctctct gtagctctct 34160
tcctctctctct ctagctctct ctagctctct ctagctctct ctagctctct ctagctctct 34200
taatagaaaga aaagaaacact aagaaacact aagaaacact aagaaacact aagaaacact 34260
taagacttact atctaatcact atctaatcact atctaatcact atctaatcact atctaatcact 34320
cacagagtct ctaagctact taaatatat taaatatat taaatatat taaatatat 34380
cacagagtct ctaagctact taaatatat taaatatat taaatatat taaatatat 34440
atggttggag cttggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt 34500
gggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt 34560
gggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt cttggtgaggt 34620

PCT/IB02/04811
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>acctacccctt</td>
<td>gctctccaaa</td>
<td>gtgtggttgt</td>
<td>tacagccggtg</td>
<td>agccacggtctg</td>
<td>ctggtcccccc</td>
</tr>
<tr>
<td>tttcaatgggtt</td>
<td>tttatggtagt</td>
<td>ttttgtgact</td>
<td>gaatccccctg</td>
<td>gaagggacagc</td>
<td>tcagctctga</td>
</tr>
<tr>
<td>ctggagtctta</td>
<td>ccgtgaacact</td>
<td>tctttggatgt</td>
<td>gcaaatagtg</td>
<td>gtgtgggttg</td>
<td>agacaagaac</td>
</tr>
<tr>
<td>aaggttagtga</td>
<td>aatgtggagtct</td>
<td>acataagctttag</td>
<td>gttacagagtc</td>
<td>gttctgagcc</td>
<td>gtgtgcaact</td>
</tr>
<tr>
<td>gcacctttccg</td>
<td>agggcaatgtg</td>
<td>gggagattt</td>
<td>ctgtagccca</td>
<td>ggtgtcaag</td>
<td>acacagcagg</td>
</tr>
<tr>
<td>gcacactgtgc</td>
<td>aaggcctccct</td>
<td>ctcctcaaaa</td>
<td>atacacaaat</td>
<td>taccggregc</td>
<td>gtgtggtgta</td>
</tr>
<tr>
<td>ttgcctatgtt</td>
<td>cccagcatatt</td>
<td>cggagacgt</td>
<td>agatgggagg</td>
<td>attcgctaaa</td>
<td>cctggagaggt</td>
</tr>
<tr>
<td>ttggagtatct</td>
<td>ttatggttggag</td>
<td>ttttgtgggtt</td>
<td>ttgatgtaga</td>
<td>gatacaatgt</td>
<td>ttttgcttggg</td>
</tr>
<tr>
<td>tgtgtggagt</td>
<td>ctctgtgctc</td>
<td>ctgtatcga</td>
<td>gataatagg</td>
<td>acttgcttgg</td>
<td>agatgtgtag</td>
</tr>
<tr>
<td>atatctgtttg</td>
<td>gattggagag</td>
<td>acacgccttc</td>
<td>tgaggggtggt</td>
<td>ggtagagta</td>
<td>tattagggag</td>
</tr>
<tr>
<td>aatctcattgg</td>
<td>gataaggtc</td>
<td>tgttgact</td>
<td>gagaattgg</td>
<td>ttgtaagttg</td>
<td>tctccttctt</td>
</tr>
<tr>
<td>cttcctcatct</td>
<td>tatttattaa</td>
<td>ctcagctttct</td>
<td>ctctctctct</td>
<td>cctctccccat</td>
<td>ccccctcccc</td>
</tr>
<tr>
<td>acacagtgggt</td>
<td>acacactgct</td>
<td>ctggtgcttt</td>
<td>ctgcctcgtt</td>
<td>ctcgctcctc</td>
<td>ccgccgctcga</td>
</tr>
<tr>
<td>ctggggagtga</td>
<td>atgtgacagt</td>
<td>atgtaggttag</td>
<td>atgagggatag</td>
<td>gtgtctactc</td>
<td>atgtgtgagga</td>
</tr>
<tr>
<td>aaatgtggagtct</td>
<td>aatgtgtagt</td>
<td>taatggagtc</td>
<td>atggtaggtg</td>
<td>atgtggagta</td>
<td>atgcctgagct</td>
</tr>
<tr>
<td>gttggttaggt</td>
<td>gtttaggttg</td>
<td>agaattggtg</td>
<td>gagaagactgt</td>
<td>ttgtaaggg</td>
<td>cagacgtact</td>
</tr>
<tr>
<td>gttgctgctg</td>
<td>gttgtaggttag</td>
<td>gtttagttgg</td>
<td>ggaggagagt</td>
<td>ggatggttca</td>
<td>gttctggctag</td>
</tr>
<tr>
<td>gttggctgcttt</td>
<td>gttgtgtgtag</td>
<td>ttgtagtagtc</td>
<td>gttatggtg</td>
<td>ttgtaatgag</td>
<td>gttggaagg</td>
</tr>
<tr>
<td>gttggggtgtg</td>
<td>gtttgtagttg</td>
<td>gttatatagtc</td>
<td>gttgtaggat</td>
<td>ttttgctcttc</td>
<td>cccgggctcag</td>
</tr>
<tr>
<td>gcagccagag</td>
<td>gcagttgcgtg</td>
<td>ccagttggga</td>
<td>cttgtgctct</td>
<td>atggagttcg</td>
<td>ggatgagagt</td>
</tr>
<tr>
<td>gacatcagttt</td>
<td>cctgtccactc</td>
<td>ctcctggcag</td>
<td>ctggagctc</td>
<td>gtttccccct</td>
<td>gggagatcct</td>
</tr>
<tr>
<td>aagacacgtt</td>
<td>gaggagttc</td>
<td>ttcctgatct</td>
<td>ctttgagag</td>
<td>aacgagctaac</td>
<td>ggcggcggag</td>
</tr>
<tr>
<td>gatcctgggag</td>
<td>cagagaggt</td>
<td>ccacagcagc</td>
<td>aggaggtgct</td>
<td>agggtggagt</td>
<td>ggcggcagag</td>
</tr>
<tr>
<td>gatccctgcc</td>
<td>ggtcgggtg</td>
<td>gcacacgctg</td>
<td>gcaggttgat</td>
<td>gtcgggtgtg</td>
<td>gacagacgct</td>
</tr>
<tr>
<td>gactcctggtt</td>
<td>cttcctgact</td>
<td>cgtctcagct</td>
<td>cttgagaggg</td>
<td>cttgagaggg</td>
<td>gcagcagtg</td>
</tr>
<tr>
<td>gattcctcat</td>
<td>ttcctggtct</td>
<td>atacacatg</td>
<td>ctttgagag</td>
<td>ctttgagag</td>
<td>gcagcagctt</td>
</tr>
<tr>
<td>aacagctgct</td>
<td>ggaggaggt</td>
<td>cttcctgta</td>
<td>cttgagaggg</td>
<td>cttgagaggg</td>
<td>gcagcagctt</td>
</tr>
<tr>
<td>gatccctggt</td>
<td>gctgagccag</td>
<td>ctcgagag</td>
<td>ctttgagag</td>
<td>ctttgagag</td>
<td>gcagcagctt</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
<tr>
<td>gactgaggtt</td>
<td>acttacat</td>
<td>ttttaagttt</td>
<td>gaaagtctg</td>
<td>ttcgcttc</td>
<td>ggttgatc</td>
</tr>
</tbody>
</table>
tcccaacaggg aaccccttttc a tagagaacctg ttttagtcca ttttccttgtt gttataacagag 85800
aataactgga aactgaggtat ttttttttttt ttttttgaga caggtcata ctcggtcacc 85860
ccaggctgcttg tgtcgatgagcc tcactgtcca ctcgctcctc caggtc aa 85920
tatacctcctg ccctccttcc cggactgac ccgtgtgttc acctgc 85980
gctataatttt tggggttggtg tgtgggttggtg tgtgttttgtg agaggtgttg tgtgtgagtc 86040
tggggttgtgc cagttgctgct tgaactcttg ggatcaagtg atctcttcgtcg 86100
ttcagctccttc aaggtgggttg gattataggg ataagctcctcc acgcctgtg cggaaactgcgg 86160
aataaatagaaa aagaagataa ttttattatttatcgcttataa gctctgagag cccaaagtttg 86220
aggccaccaactcggaggtacgcttcctc tgctttgtggtgctgagttggg aggcttgcgtg 86280
gtgcctc agggccccacagctagttaagtggctgtgtgcc tggattgtgctaatgagttgag 86340
tctcgccttgcc tacgctggcc tagtaagctataactccttcc tccaagaggtaccattatatg 86400
cctattatatcaatataaccattttgtagtatataatccatttagaagcagcggctg 86460
cctagattccatctcctttctcagccgcc acctctctagc accaccagctgacttgggcc 86520
cctactaggacatgtctccagtcctgttggc cagagcgttaacagacgagcagtaagcttaa 86580
acccctcacaag gc 86592

<210>  2
<211>  1580
<212> DNA
<213> Homo sapiens

<220>
<221>  5'UTR
<222>  1..143

<220>
<221> CDS
<222>  144..1187

<220>
<221>  3'UTR
<222>  1188..1580

<220>
<221> polyA_signal
<222>  1549..1554

<400>  2
tgcctctcaag tcggggttgag gcga cagggaggtcctgaa cagacgcgtgacctactcgccttg 60
ctttctctc aatgcaccagag cgcagcggcc caggtcctca cgaatcagtag ctctctcgca 120
gttgtgtcagt tcggggttgac gcc ata tgt ggt tgt tgt att gga gca gga gtc 173

Met Arg Val Val Val Ile Gly Ala Gly Val
1  5  10

atc ggg tgt ctc acc gcc ctc tgt g ac cat gag cgc tac cac tca gtc 221
Ile Gly Leu Ser Thr Ala Leu Cys Ile His Glu Arg Tyr His Ser Val
15 20 25

tgt cag gca ac aata aag gtc tac ggc gac cgc ttc acc cca ctc 269
Leu Gln Pro Leu Asp Ile Lys Tyr Ala Asp Arg Phe Thr Pro Leu
30 35 40

acc acc acc gac tgt gct gcc gcc ctc tgt cag ccc tac tct tct gcc 317
Thr Thr Thr Asp Val Ala Ala Gly Leu Trp Gin Pro Tyr Leu Ser Asp
45 50 55

ccc acc acc cca cag gag gcc cag gtc acc caa cag acc ttt tac tat 365
Pro Asn Asn Pro Gln Glu Cys Ala Asp Trp Ser Gin Thr Phe Asp Tyr
60 65 70

cct cag aac aca cat gtc cat ctc ccc aac gct gaa aac ctg gcc ctg gtc 413
Leu Leu Ser His Val His Ser Pro Asn Ala Glu Asn Leu Gly Phe
75 80 85 90

cct  atc tgg gcc tac aac ctc ttc cat gaa gcc att ccc gcc gcc ctc cct 461
Leu Ile Ser Gly Tyr Asn Leu Phe His Glu Ala Ala Pro Asp Pro Ser
95 100 105

tgg aag gac acc gtt ctc gaa ttt cgg aag ctg acc ccc ccc aag gag ctc 509
Trp Lys Asp Thr Val Leu Gly Phe Arg Lys Thr Pro Arg Gly Leu
110  gat atg  ttc cca  gat tac  ggc tat  ggc tgg  ttc cac  cac  aca  agc  cta  att
Asp Met Phe Pro Asp Tyr Gly Tyr Gly Trp Phe His Thr Ser Leu Ile
125  130  135

tct gaa gga aag  aac  tat  cta  cag  tgg  ctc  act  gaa  agg  tta  act  gag
Leu Gly Lys Asn Tyr Leu Gln Trp Leu Thr Glu Arg Leu Thr Glu
140  145  150

155  aag  gga  gtg  aac  ttc  ctc  cag  gga  aag  aag  att  tct  cca  att  gag  gtc
Arg Gly Val Lys Phe Phe Gln Arg Lys Val Ser Glu Phe Val
170  175  180  185

gca  gaa  ggc  gca  gac  gtc  att  gta  aac  tgc  act  ggg  gta  tgg  gct
Ala Arg Gly Ala Asp Val Ile Val Asn Tyr Val Gln Thr Val Trp Ala
190  195  200

195  ggg  ggc  ctc  cca  cga  gac  ccc  ctc  cag  cca  cca  ggc  cgg  ggg  cag  atc
Gly Ala Leu Gln Arg Asp Pro Leu Leu Gln Pro Gly Arg Gly Lys Ile
210  215

210  atg  aag  gtg  gac  gcc  cct  tgg  atg  aag  cac  ttc  att  ctc  acc  cat  gac
Met Lys Val Asp Ala Pro Met Lys His Phe Ile Leu Thr His Asp
225  230

220  cca  gag  aga  ggc  atc  tac  aat  tcc  ccc  tac  atc  cca  ggc  acc  cag
Pro Glu Arg Gly Ile Tyr Asn Ser Pro Tyr Ile Pro Gly Thr Glu
235  240  245  250

235  acu  gtt  act  ctt  gga  ggc  atc  ttc  cag  tgg  gga  aac  tgg  aat  gaa cta
Thr Val Thr Leu Gly Lys Gln Phe His Tyr Leu Ser Glu Leu
250  255

245  aac  aat  atc  cag  gac  cac  aac  acc  att  tgg  gaa  ggc  tgc  tgc  aat  ctg
Asn Asn Ile Gln Asp His Asn Thr Ile Trp Glu Gly Cys Cys Arg Leu
260  265

260  gag  ccc  aca  ctc  atg  aat  gca  aga  att  att  ggt  gaa  gca  act  ggc  ttc
Glu Pro Thr Leu Lys Gly Lys Gln Phe His Asn Tyr Gly Gly Thr Phe
275  280

280  cgg  cca  gta  cgc  ccc  cag  att  cgg  cta  gaa  aga  gaa  cag  ctt  cgc  act
Arg Pro Val Arg Pro Gln Ile Arg Leu Glu Arg Glu Leu Arg Thr
295  300  305  310

300  gga  cct  tca  aac  aca  gag  gtc  atc  cac  aac  tat  ggc  cat  gga  ggc  tac
Gly Pro Ser Asn Thr Val Thr Tyr Val Asn Tyr Gly Gly Tyr
315  320  325  330

320  ggg  ctc  acc  atc  cac  tgg  gga  tgt  gcc  ctc  gag  gca  gcc  aag  ctc  ttt
Gly Leu Thr Ile His Trp Gly Cys Ala Leu Gln Ala Ala Lys Leu Phe
335  340  345

335  cgg  neg  atc  gtc  aag  aag  aa  aga  aag  ttc  ttg  ctc  aga  atg  cca  cca  tcc  cac
Gly Arg Ala Leu Gly Lys Gly Leu Ser Met Cys Tyr Ser Val Leu Val Ala

<210>  3
<211>  347
<212>  PRT
<213>  Homo sapiens

<400>  3
Met Arg Val Val Val Ile Gly Ala Gly Val Ile Gly Leu Ser Thr Ala
1   5   10   15
Leu Cys Ile His Glu Arg Tyr His Ser Val Leu Gln Pro Leu Asp Ile
20  25  30
Lys Val Tyr Ala Asp Arg Phe Thr Pro Leu Thr Thr Thr Asp Val Ala
 Ala Gly Leu Trp Gln Pro Tyr Leu Ser Asp Pro Asn Pro Gln Glu  
     35  40  45  
 Ala Asp Trp Ser Gln Gln Thr Phe Asp Tyr Leu Leu Ser His Val His  
     50  55  60  
 Ser Pro Asn Ala Glu Asn Leu Gly Leu Phe Leu Ile Ser Gly Tyr Asn  
     65  70  75  80  
 Leu Phe His Glu Ala Ile Pro Asp Pro Ser Trp Lys Asp Thr Val Leu  
     85  90  95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 310 315 320 325 330 335 340 345  
 Gly Tyr Gly Trp Phe His Thr Ser Leu Ile Leu Glu Gly Lys Asn Tyr  
     220  
 Gly Lys Trp Leu Thr Glu Arg Leu Thr Glu Arg Gly Val Lys Phe Phe  
     230 235 240  
 Gln Arg Lys Val Gly Ser Phe Gly Val Ala Arg Glu Gly Ala Asp  
     245 250 255 260 265 270 275 280 285 290 295 300 310 315 320 325 330 335 340 345  
 Val Ile Val Asn Cys Thr Gly Val Trp Ala Gly Ala Leu Gln Arg Asp  
     180 185 190  
 Pro Leu Leu Gln Pro Gly Arg Gly Gln Ile Met Lys Val Asp Ala Pro  
 Trp Met Lys His Phe Ile Leu Thr His Asp Pro Glu Arg Gly Ile Tyr  
     230  
 Asn Ser Pro Tyr Ile Pro Gly Thr Gln Thr Val Thr Leu Gly Gly  
     260 265 270 275 280 285 290 295 300 310 315 320 325 330 335 340 345  
 Asn Thr Ile Trp Glu Gly Cys Cys Arg Leu Glu Pro Thr Leu Lys Asn  
     250  
 Ala Arg Ile Glu Gly Arg Thr Gly Phe Arg Pro Val Arg Pro Gln  
     245 250 255 260 265 270 275 280 285 290 295 300 310 315 320 325 330 335 340 345  
 Ile Arg Leu Glu Arg Glu Gln Leu Arg Thr Gly Pro Ser Asn Thr Glu  
     235  
 Val Ile His Asn Tyr Gly His Gly Gly Tyr Gly Leu Thr Ile His Trp  
     290 295 300 310 315 320 330 335 340 345  
 Gly Cys Ala Leu Glu Ala Ala Lys Leu Phe Gly Arg Ile Leu Glu Glu  
     315 320 325 330 335 340 345  
 Lys Lys Leu Ser Arg Met Pro Pro Ser His Leu  
     310 315 320 325 330 335 340 345  

<210> 4  
<211> 467  
<212> DNA  
<213> Homo sapiens  

<220> 
<221> allele  
<222> 61  
<223> 27-1-61 : polymorphic base A or G  

<220> 
<221> misc_binding  
<222> 49..73  
<223> 27-1-61.probe  

<220> 
<221> primer_bind  
<222> 42..60  
<223> 27-1-61.mis  

<220> 
<221> primer_bind  
<222> 62..80
<223> 27-1-61.mis complement

<220>
<221> primer_bind
<222> 1..18
<223> 27-1.pu

<220>
<221> primer_bind
<222> 134..152
<223> 27-1.rp complement

<400> 4
gacatccact atcatctgta tgrgtaacat tggcatcagag gttatgggca aagcagtag  60
tgggaaatag aaataaagag agaggtgataa aaatattaaggt ttcgagatga  120
tgtgtagtca ttttaaagaagat acatctaaacat taattacaca aacataggg  180
atttggagag ggttttataat ggcacactag tggtaaacag tggcaggtgt gttcttcatgc  240
tcagctaggc tcagctgcca tggcctctgtg gccagcgggt gtcaggacc aagctagct  300
ccagcaacacc ggcctaggggt aatccctcatc caaatgggtct cttggaagtt  360
cagagactc cttcctacac aacgtggtct ctggcaggtta cccctctcag aagtttgct  420
aaaaatcatc cttccctagtc ggcactcatgtag atagttctttt cccctca  467

<410> 5
<411> 1633
<412> DNA
<413> Homo sapiens

<400> 5
tggtggtcaca ttggcaacccgg agggagacagt agagttccca aagcagaaag ggagagcag  60
tgggactcag tggagagcgg acagaggtc gtaaagacag gctccagagat tccaggttt  120
ccctcagaggt aataccagctcc tggtcagcgct ctcagctttgct tccagctgtt  180
ttgattccag gccagttgctg atg cttgctgtgtgtag att gga gca gta atc  240
Met Arg Val Val Val Ile Gly Ala Gly Val Ile

175

5

10

15

20

25

30

35

40

45

50

55

60

65

70

75

80

85

90

95

100

105

110

115

120

125

130

135

140

145

150

155

160

165

170

175

180

185

190

195

200

205

210

215

220

225

230

235

240

245

250

255

260
aga gaa ggc gca gac gtc att gtc aac act ggg gta tgg gcc 761
Arg Glu Gly Ala Asp Val Ile Val Asn Thr Gly Val Thr Ala Gly
160 165 170

gcg cta caa cga gac ccc ctg ctg cag cca ggc cgg ggg cag atc atg 809
Ala Leu Gln Arg Asp Pro Leu Leu Gln Pro Gly Arg Gly Gln Ile Met
175 180 185

aag gtt gac gcc cct tgg atg aag cag ttc att ttc acc cat gac cca 857
Lys Val Asp Ala Pro Trp Met Lys His Phe Ile Leu Thr His Asp Pro
190 195 200

 gag aga ggc atc tac aat tcc ccc tac atc cca ggg acc cag cca 905
Glu Arg Gly Ile Tyr Asn Pro Ser Pro Ile Gly Thr Gln Thr
205 210 215

gtt act ctt gga ggc atc ttc cag tgg gaa aac tgg aat gaa cta aac 953
Val Thr Leu Gly Ile Phe Gln Leu Gly Asn Trp Ser Glu Leu Asn
220 225 230 235

 aat atc cag cac acc acc att tgg gaa ggc tgc aga ctg gag 1001
Asn Ile Gln Asp His Thr Ile Pro Trp Gly Cys Arg Leu Glu
240 245 250

 ccc aca ctg aag aat gca aga att att ggt gaa gca act ggc ttc cgg 1049
Pro Thr Leu Lys Asn Ala Arg Ile Ile Gly Glu Ala Thr Gly Phe Arg
255 260 265

 cca gta cgc ccc cag att cgg cta gaa aga cag ctt cgc act gga 1097
Pro Val Arg Pro Glu Arg Thr Glu Glu Leu Thr Arg Thr Gly
270 275 280

cct tca aac aca gag gtc atc cac aac tat ggc cat gga ggc tac ggc 1145
Pro Ser Asn Thr Glu Val Ile His Asn Tyr Gly His Gly Gly Tyr Gly
285 290 295

 ctc acc atc cac tgg gga tgt gcc ctg gag gca ggc aag ctc ttt ggg 1193
Leu Thr Ile His Thr Gln Tyr Ala Gly Ala Ala Leu Leu Phe Gly
300 305 310 315

 aga atc ctg gaa gaa aag aad tgt tcc aga atg cca cca tcc cac ctc 1241
Arg Ile Leu Glu Gly Lys Leu Ser Arg Met Pro Pro Ser His Leu
320 325 330

 tga agacccacct gactgctgcc tccccccaca agaactccct tctccccctca 1294
*

* gcgaatgca tcctcataagc ccatgctcct cccctacttc ttctcctcaa 1354
gaagcaatg cgtcgagaa gcccracagtg ctaagccctg gccaggccgt agcccracac 1414
ggggccctcc tctactagtc aactcctctca ctctccctcg gcctgcatc ataaagcaca 1474
gctcgagctg tctactcatg agtctctcaga aagcaggaat cctcagaaat ctaaagccg 1534
cactgcctca gacccacaga aatgtagggc taatgaccct gattacaagt 1594
tgtactacacataaaaggtctctgaaagctggcc 1633

<210> 6
<211> 347
<212> PRT
<213> Homo sapiens

<400> 6
Met Arg Val Val Ile Gly Ala Gly Val Ile Gly Leu Ser Thr Ala 1 5 10 15
Leu Cys Ile His Glu Arg Tyr His Ser Val Leu Gln Pro Leu Asp Ile
20 25 30
Lys Val Tyr Ala Asp Arg Phe Thr Pro Leu Thr Thr Thr Asp Val Ala
35 40 45
Ala Gly Leu Trp Gln Pro Tyr Leu Ser Asp Pro Asn Asn Asn Pro Glu
50 55 60
Ala Asp Trp Ser Gln Gln Thr Phe Asp Tyr Leu Leu Ser His Val His
65 70 75 80
Ser Pro Asn Ala Glu Asn Leu Gly Leu Phe Leu Ile Ser Gly Tyr Asn
85 90 95
Leu Phe His Glu Ala Ile Pro Asp Pro Ser Trp Lys Thr Val Leu
100 105 110
Gly Phe Arg Lys Leu Thr Pro Arg Glu Leu Asp Met Phe Pro Asp Tyr
115 120 125
Gly Tyr Gly Trp Phe His Thr Ser Leu Ile Leu Glu Gly Lys Asn Tyr
130 135 140
Leu Gln Trp Leu Thr Glu Arg Leu Thr Glu Gly Val Lys Phe Phe
145 150 155 160
Gln Arg Lys Val Glu Ser Phe Glu Glu Val Ala Arg Glu Gly Ala Asp
165 170 175
Val Ile Val Asn Cys Thr Gly Val Trp Ala Gly Ala Leu Glu Arg Asp
180 185 190
Pro Leu Leu Gln Pro Gly Arg Gly Glu Ile Met Lys Val Asp Ala Pro
195 200 205
Trp Met Lys His Phe Ile Leu Thr His Asp Pro Glu Arg Gly Ile Tyr
210 215 220
Asn Ser Pro Tyr Ile Ile Pro Gly Thr Glu Thr Val Thr Leu Gly Gly
225 230 235 240
Ile Phe Gln Leu Gly Asp Trp Ser Glu Leu Asn Asn Ile Gln Asp His
245 250 255
Asn Thr Ile Trp Glu Gly Cys Cys Arg Leu Glu Pro Thr Leu Lys Asn
260 265 270
Ala Arg Ile Ile Gly Glu Ala Thr Gly Phe Arg Pro Val Arg Pro Gln
275 280 285
Ile Arg Leu Glu Arg Glu Glu Leu Arg Thr Gly Pro Ser Asn Thr Glu
290 295 300
Val Ile His Asn Tyr Gly His Gly Gly Tyr Gly Leu Thr Ile His Trp
305 310 315 320
Gly Cys Ala Leu Ala Ala Lys Leu Phe Gly Arg Ile Leu Glu Glu
325 330 335
Lys Lys Leu Ser Arg Met Pro Pro Ser His Leu
340 345

<210> 7
<211> 1200
<212> DNA
<213> Homo sapiens

<400> 7
atg gac aca gca cgg att gca gtt gtc ggg gca ggt gtg ggg ctc
1 5 10 15
Met Asp Thr Ala Arg Ile Ala Val Val Gly Ala Gly Val Gly Leu
20
TCC acg gct gtg tgc atc tcc aaa cgt gtt ccc cga tgc tcc gtt acc
25 30
Ser Thr Ala Val Cys Ile Ser Lys Leu Val Pro Arg Cys Ser Thr Val
30
ATC att tca gac aag ttt act cca gat acc acc gat gtt gca gcc
40 45
Ile Ile Ser Asp Lys Phe Thr Pro Asp Thr Thr Ser Asp Val Ala Ala
50 55 60
GGG aat gtt taa cgg att cca gat aca ccc att ccc acg cag
70 75 80
Gly Met Leu Ile Pro His Thr Tyr Pro Asp Thr Pro Ile His Thr Gln
85 90 95
AAG cag tgg tcc aga gaa acc ttt aat ccc ctc att gcc aat
100 105
Lys Gln Trp Phe Arg Glu Thr Phe Asn His Leu Phe Ala Ile Ala Asn
110
TCT gca gaa gct gga gat gct gtt gat tgc gta cca gtt ggg cag
120 125 130
Ser Ala Glu Ala Gly Ala Gly Val His Leu Val Ser Gly Trp Gln
140 145
ATA ttt cag agc act ccc act gaa gta gtc cca ttc tgg gat gac gtt
150 155 160
Val Leu Gln Ser Thr Pro Pro Thr Glu Val Pro Phe Thr Pro Asp Val
170 175 180
GGT gga taa cgg aag atg act gag gct gct aag ccc cct ccc cac
190 195 200
Val Leu Gln Met Thr Gly Ala Glu Leu Lys Phe Pro
210 215 220
CAG gag cgg att cgg ggc att cct gaa pro agt ccg aag
220 225 230
Gln Tyr Val Phe Gly Glu Ala Phe Thr Thr Leu Lys Cys Glu Cys Pro
235 240 245

432
Homo sapiens

aaatgacag caaagagact gagagactgt tgaattcaagc acagaacagg ttcaaatattc 1941
tttccactg catgaaagtt taattagaca ttctcttggt ttcaacattca gaagtggtgt 2001
aacatgtaag tctagcaccgg tagcatgctc atagtcccgct ctaaccttg 2048

<210> 9
<211> 341
<212> PRT
<213> Homo sapiens

<400> 9
Met Asp Thr Ala Arg Ile Ala Val Val Gly Ala Gly Val Val Gly Leu
1    5    10   15
Ser Thr Ala Val Cys Ile Ser Lys Leu Val Pro Arg Cys Ser Val Thr
20   25   30
Ile Ile Ser Asp Lys Phe Thr Pro Asp Thr Thr Ser Asp Val Ala Ala
35   40   45
Gly Met Leu Ile Pro His Thr Tyr Pro Asp Thr Pro Ile His Thr Gln
50   55   60
Lys Gln Trp Phe Arg Glu Thr Phe Asn His Leu Phe Ala Ile Ala Asn
65   70   75   80
Ser Ala Glu Ala Gly Asp Ala Gly Val His Leu Val Ser Gly Trp Gln
85   90   95
Ile Phe Gln Ser Thr Pro Thr Glu Glu Val Pro Phe Trp Ala Asp Val
100 105  110
Val Leu Gly Phe Arg Lys Met Thr Glu Ala Glu Leu Lys Lys Phe Pro
115 120  125
Gln Tyr Val Phe Gly Glu Ala Phe Thr Leu Lys Cys Glu Cys Pro
130 135  140
Ala Tyr Leu Pro Trp Leu Glu Lys Arg Ile Lys Gly Ser Gly Gly Trp
145 150  155  160
Thr Leu Thr Arg Arg Ile Glu Asp Leu Trp Glu Leu His Pro Ser Phe
165 170  175
Asp Ile Val Val Asn Cys Ser Gly Leu Gly Ser Arg Glu Gln Leu Ala Gly
180 185  190
Asp Ser Lys Ile Phe Pro Val Arg Gly Glu Val Leu Gln Val Glu Ala
195 200  205
Pro Trp Val Glu His Phe Ile Arg Asp Gly Ser Gly Leu Thr Tyr Ile
210 215  220
Tyr Pro Gly Thr Ser His Val Thr Leu Gly Gly Thr Arg Glu Lys Gly
225 230  235  240
Asp Trp Asn Leu Ser Pro Asp Ala Glu Asn Ser Arg Glu Ile Leu Ser
245 250  255
Arg Cys Cys Ala Leu Glu Pro Ser Leu His Gly Ala Cys Asn Ile Arg
260 265  270
Glu Lys Val Gly Leu Arg Pro Tyr Arg Pro Gly Val Arg Leu Gln Thr
275 280  285
Glu Leu Leu Ala Arg Asp Gly Glu Arg Leu Pro Val Val His His Tyr
290 295  300
Gly His Gly Ser Gly Gly Ile Ser Val His Thr Gly Thr Ala Leu Glu
305 310  315  320
Ala Ala Arg Leu Val Ser Glu Cys Val His Ala Leu Arg Thr Pro Ile
325 330  335
Pro Lys Ser Asn Leu
340

<210> 10
<211> 282
<212> PRT
<213> Homo sapiens

<400> 10
Met Asp Thr Ala Arg Ile Ala Val Val Gly Ala Gly Val Val Gly Leu
1    5    10   15
Ser Thr Ala Val Cys Ile Ser Lys Leu Val Pro Arg Cys Ser Val Thr  
   20 25 30  
Ile Ile Ser Asp Lys Phe Thr Pro Asp Thr Ser Asp Val Ala Ala  
   35 40 45  
Gly Met Leu Ile Pro His Thr Tyr Pro Asp Thr Pro Ile His Thr Gln  
   50 55 60  
Lys Gln Trp Phe Arg Glu Thr Phe Asn His Leu Phe Ala Ile Ala Asn  
   65 70 75 80  
Ser Ala Glu Ala Gly Asp Ala Gly Val His Leu Val Ser Gly Ile Lys  
   85 90 95  
Gly Ser Gly Gly Trp Thr Leu Thr Thr Arg Arg Ile Glu Asp Leu Trp Glu  
   100 105 110  
Leu His Pro Ser Phe Asp Ile Val Val Asn Cys Ser Gly Leu Gly Ser  
   115 120 125  
Arg Gln Leu Ala Gly Asp Ser Lys Ile Phe Pro Val Arg Gly Gln Val  
   130 135 140  
Leu Gln Val Gln Ala Pro Trp Val Glu His Phe Ile Arg Asp Gly Ser  
   145 150 155 160  
Gly Leu Thr Tyr Ile Tyr Pro Gly Thr Ser His Val Thr Leu Gly Gly  
   165 170 175  
Thr Arg Gln Lys Gly Asp Trp Asn Leu Ser Pro Asp Ala Glu Asn Ser  
   180 185 190  
Arg Glu Ile Leu Ser Arg Cys Cys Ala Leu Glu Pro Ser Leu His Gly  
   195 200 205  
Ala Cys Asn Ile Arg Glu Lys Val Gly Arg Pro Tyr Arg Pro Gly  
   210 215 220  
Val Arg Leu Gln Thr Glu Leu Leu Ala Arg Asp Gly Gln Arg Leu Pro  
   225 230 235 240  
Val Val His His Tyr Gly His Gly Ser Gly Gly Ile Ser Val His Trp  
   245 250 255  
Gly Thr Ala Leu Glu Ala Ala Arg Leu Val Ser Glu Cys Val His Ala  
   260 265 270  
Leu Arg Thr Pro Ile Pro Lys Ser Asn Leu  
   275 280  

<210> 11
<211> 47
<212> DNA
<213> Artificial Sequence

<220>
<223> oligonucleotide 27-81-180

<220>
<221> allele
<222> 24
<223> polymorphic base A or G

<400> 11
taaggtgca cgcacagac tgacacag ccacagag cccacag 47

<210> 12
<211> 47
<212> DNA
<213> Artificial Sequence

<220>
<223> oligonucleotide 27-30-249

<220>
<221> allele
<222> 24
<223> polymorphic base C or T
<400> 12
aagatttccc ccaccaag aacygttta caaccacttt tatattc

<210> 13
<211> 47
<212> DNA
<213> Artificial Sequence

<220>
<223> oligonucleotide 27-2-106

<220>
<221> allele
<222> 24
<223> polymorphic base A or C

<400> 13
cttggctgtg tgaccttagg aagkcaactta gcctctctga aactcag

<210> 14
<211> 47
<212> DNA
<213> Artificial Sequence

<220>
<223> oligonucleotide 27-29-224

<220>
<221> allele
<222> 24
<223> polymorphic base T or G

<400> 14
tgatggagat gatggtagt atgktgtgcc tcaacccttc cttccca

<210> 15
<211> 47
<212> DNA
<213> Artificial Sequence

<220>
<223> oligonucleotide 27-1-61

<220>
<221> allele
<222> 24
<223> polymorphic base A or G

<400> 15
ttcaggcac cagccccatct ccctgcacg ctgtgggag agaataa