Method of Identifying Disease Risk Factors

Allen D. Roses

Related Applications

This application claims priority to U.S. Provisional Application No. 61/088,203, filed August 12, 2008; U.S. Provisional Application No. 61/186,673, filed June 12, 2009; and U.S. Provisional Application No. 61/224,647, filed July 10, 2009, the disclosures of each of which is incorporated by reference herein in its entirety.

Field of the Invention

The present invention relates to the field of genomics, genetics, pharmacogenetics, and bioinformatics, including genome analysis and the study of DNA sequence variation. The invention also relates to studies of association between variations in DNA sequences and anticipation of an individual's susceptibility to a particular disease, disorder, or condition and/or response to a particular drug or treatment.

Background of the Invention

The search for genetic markers associated with complex diseases is ongoing. Genome-wide scanning studies with SNP arrays continue to highlight the ApoE region as the most important area for investigation in the study of Alzheimer's disease (Coon et al., J. Clin. Psychiatry 68: 613-8 (2007); Li et al., Arch. Neurol. 65: 45-53 (2007)).

The ApoE 4 isoform has previously been strongly associated with increased risk of developing late-onset Alzheimer's disease. (Pericak-Vance et al., Am. J. Hum. Genet. 48, 1034-50 (1991); Martin et al., 2000, US Patent No. 6,027,896 to Roses, et al., US Patent No. 5,716,828 to Roses et al.) The relationship is dose dependent (Yoshizawa et al.,1994; Schellenberg, 1995). That is to say, a carrier of two ApoE 4 alleles is more likely to develop late-onset Alzheimer's disease (LOAD) than a carrier of only one ApoE 4 allele, and at an earlier age (Corder et al., Science 261, 921-3 (1993)).

Nevertheless, E4 alleles only account for roughly 50% of hereditary Alzheimer's disease. One explanation is that ApoE 4 is merely serving as a surrogate marker for something in linkage disequilibrium nearby. Alternatively, considering the recent discovery of a mechanistic role for ApoE 4 in mitochondrial toxicity, the negative effects of ApoE 4 may be abrogated or exacerbated by another gene product encoded nearby (Chang et al., 2005).

As ApoE status is also associated with risk for coronary artery disease and likely also a host of other diseases and disorders, the implications of the study of the ApoE region are not limited to Alzheimer's disease, but are potentially far-reaching (Mahley et al., Proc. Natl. Acad. Sci. USA 103: 5644-51 (2006)). More broadly, the examination of variant sequences for

processes or pathways surrounding genes in linkage disequilibrium with other genetic regions known to be involved in complex disease processes will provide valuable information in deciphering the mechanisms of those diseases.

Summary of the Invention

Provided herein is a method for identifying a genetic variant that is associated with development of a condition of interest (e.g., earlier or later onset of a disease of interest), comprising: *(a)* determining from biological samples containing DNA the nucleotide sequences carried by a plurality of individual human subjects at a genetic locus of interest, wherein subjects include both *(i)* subjects affected with the condition of interest and *(ii)* subjects unaffected with the condition of interest; *(b)* identifying genetic variants at said genetic locus from nucleotide sequences observed in said plurality of subjects (e.g., using a multiple sequence alignment analysis); *(c)* mapping said genetic variants by constructing a phylogenetic tree of said nucleotide sequences of said subjects, said tree comprising branches that identify variant changes between said subjects (e.g., variant changes on the same cistron); *(d)* examining the genetic variants represented as branches in said tree and determining the ratio of affected and unaffected subjects to identify those changes that lead to a changed ratio of affected to unaffected subjects (preferably wherein the starting point is the genetic variant representing the greatest number of subjects); and then *(e)* identifying a genetic variant or group of variants (a haplotype) where the ratio of affected to unaffected subjects is substantially different from one or more adjacent variants on said tree (e.g., at least 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, or 90 % different) to thereby identify a genetic variant associated with the development of said condition of interest.

In some embodiments, all subjects carry a same known polymorphism that is associated with the condition of interest.

In some embodiments, the condition of interest is a neurodegenerative disease, a metabolic disease (e.g., dyslipidemia), a cardiovascular disease, a psychiatric disorder, or cancer. In some embodiments, the disease of interest is a disease in which ApoE and/or TOMM40 are implicated in disease pathogenesis.

In some embodiments, the condition of interest is associated with increased or decreased mitochondrial dysfunction. In some embodiments, the condition of interest is schizophrenia. In some embodiments, the condition of interest is coronary artery disease. In some embodiments, the condition of interest is diabetes mellitus, type II. In some embodiments, the condition of interest is Parkinson's disease. In some embodiments, the condition of interest is Alzheimer's disease.

In some embodiments, the known polymorphism risk factor is the Apolipoprotein E allele (e.g., ApoE 2, ApoE 3 or ApoE 4).

In some embodiments, the genetic locus of interest is in linkage disequilibrium with the known polymorphism. In some embodiments, the genetic locus of interest is on the same chromosome and less than 10, 20, 30, 40, or 50 kilobases away from the known polymorphism. In some embodiments, the genetic locus is TOMM40.

Also provided is a method of determining increased risk for development of a condition of interest, comprising: *(a)* determining from a biological sample containing DNA a genetic variant identified by the method of any of the preceding paragraphs carried by an individual subject; and then *(b)* determining the subject is at increased risk for development of the condition of interest when the genetic variant is present.

Further provided is a method of determining increased risk for development of Alzheimer's disease in a subject (e.g., a subject carrying at least one Apo E3 allele), comprising: *(a)* detecting from a biological sample containing DNA taken from the subject the presence or absence of a genetic variant of the TOMM40 gene associated with increased or decreased risk of Alzheimer's disease; and *(b)* determining the subject is at increased or decreased risk of Alzheimer's disease when the genetic variant is present or absent.

In some embodiments, it is determined whether the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject. In some embodiments, it is determined whether the subject is an Apo E3/E3 or E3/E4 subject.

In some embodiments, the method further includes the step of: *(c)* administering an anti-Alzheimer's disease active agent to the subject in a treatment effective amount when the subject is determined to be at increased risk of Alzheimer's disease.

In some embodiments, the administering step is carried out in the subject at an earlier age when the subject is determined to be at increased risk by the presence or absence of the genetic variant as compared to a subject in which the genetic variant is not present or absent (e.g., for an ApoE 4/4 subject, beginning at age 45, 46, 47, 48, 49, 50, 51, 52, or 53, and continuously through each year thereafter, rather than beginning at age 55 or more; for an ApoE 4/3 subject, at age 50, 51, 52, 53, 54, 55, 56, 57, or 58, and continuously through each year thereafter, rather than beginning at age 60 or more; for an ApoE 3/3 subject, at age 55, 56, 57, 58, 59, 60, 61, 62, or 63, and continuously through each year thereafter, rather than beginning at age 65 or more; and for an ApoE 2/3 subject, at age 60, 61, 62, 63, 64, 65, 66, 67, or 68, and continuously through each year thereafter, rather than beginning at age 70 or more).

In some embodiments, the active agent is selected from the group consisting of acetylcholinesterase inhibitors, NMDA receptor antagonists, PPAR agonists or modulators (e.g.,

drugs in the thiazolidinedione or glitazar classes), antibodies, fusion proteins, therapeutic RNA molecules, and combinations thereof. In some embodiments, the active agent is rosiglitazone or a pharmaceutically acceptable salt thereof.

In some embodiments, the genetic variant of the TOMM40 is a variant listed in **Table 1** as set forth below.

Also provided is a method of treating a subject (e.g., a subject having at least one ApoE 3) allele for Alzheimer's disease by administering an anti-Alzheimer's disease active agent to the subject in a treatment-effective amount; the improvement comprising: administering the active agent to the subject at an earlier age when the subject carries a genetic variant of the TOMM40 gene associated with increased risk of Alzheimer's disease as compared to a corresponding subject who does not carry the genetic variant (e.g., for an ApoE 4/4 subject, beginning at age 45, 46, 47, 48, 49, 50, 51, 52, or 53, and continuously through each year thereafter, rather than beginning at age 55 or more; for an ApoE 4/3 subject, at age 50, 51, 52, 53, 54, 55, 56, 57, or 58, and continuously through each year thereafter, rather than beginning at age 60 or more; for an ApoE 3/3 subject, at age 55, 56, 57, 58, 59, 60, 61, 62, or 63, and continuously through each year thereafter, rather than beginning at age 65 or more; and for an ApoE 2/3 subject, at age 60, 61, 62, 63, 64, 65, 66, 67, or 68, and continuously through each year thereafter, rather than beginning at age 70 or more).

In some embodiments, the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, E4/E4 subject. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

In some embodiments, the active agent is selected from the group consisting of acetylcholinesterase inhibitors, NMDA receptor antagonists, PPAR agonists or modulators (e.g., drugs in the thiazolidinedione or glitazar classes), antibodies, fusion proteins, therapeutic RNA molecules, and combinations thereof. In some embodiments, the active agent is rosiglitazone or a pharmaceutically acceptable salt thereof.

In some embodiments, the genetic variant of the TOMM40 gene is a deletion/insertion polymorphism (DIP). In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 bp poly-T).

In some embodiments, the genetic variant of the TOMM40 is a variant listed in **Table 1** as set forth below. In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

Further provided is a method of treatment for a condition of interest, wherein the condition of interest is associated with ApoE and/or TOMM40, for a patient in need thereof, the method including the steps: *(a)* determining the presence or absence of a genetic variant

identified by the method of paragraph 1-12 carried by an individual subject to generate a genetic profile of the patient; and then, if the profile is indicative of the patient being responsive to an active agent, *(b)* administering the active agent to the subject in a treatment effective amount to treat the condition of interest.

In some embodiments, the active agent is selected from the group consisting of acetylcholinesterase inhibitors, NMDA receptor antagonists, PPAR agonists or modulators (e.g., drugs in the thiazolidinedione or glitazar classes), antibodies, fusion proteins, therapeutic RNA molecules, and combinations thereof. In some embodiments, the active agent is rosiglitazone or a pharmaceutically acceptable salt thereof.

In some embodiments, the genetic variant of the TOMM40 gene is a deletion/insertion polymorphism (DIP). In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 bp poly-T insertion).

In some embodiments, the genetic variant of the TOMM40 is a variant of TOMM40 listed in **Table 1** as set forth below. In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

Also provided is a method of treatment for Alzheimer's disease in a subject, including: *(a)* detecting from a biological sample containing DNA taken from the subject the presence or absence of a genetic variant of the TOMM40 gene associated with responsiveness to an active agent; and, if the genetic variant is present, *(b)* administering the active agent to the subject in a treatment effective amount to treat the Alzheimer's disease.

In some embodiments, the subject carries at least one ApoE 3 allele. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

In some embodiments, the active agent is selected from the group consisting of acetylcholinesterase inhibitors, NMDA receptor antagonists, PPAR agonists or modulators (e.g., drugs in the thiazolidinedione or glitazar classes), antibodies, fusion proteins, therapeutic RNA molecules, and combinations thereof. In some embodiments, the active agent is rosiglitazone or a pharmaceutically acceptable salt thereof.

In some embodiments, the genetic variant of the TOMM40 gene is a deletion/insertion polymorphism (DIP). In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 bp poly-T).

In some embodiments, the genetic variant of the TOMM40 gene is a variant listed in **Table 1** as set forth below. In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

Further provided is the use of an anti-Alzheimer's disease active agent for the preparation of a medicament for carrying out a method of treatment for Alzheimer's disease in accordance with the paragraphs set forth above. Also provided is the use of an anti-Alzheimer's disease active agent for carrying out a method of treatment for Alzheimer's disease.

A method of determining a prognosis for a patient at risk for developing Alzheimer's disease is provided, including obtaining a patient profile, wherein the obtaining a patient profile includes: detecting the presence or absence of at least one ApoE allele in a biological sample of the patient, and detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene, and then, converting the patient profile into the prognosis, wherein the presence of the ApoE allele and the presence of the at least one TOMM40 DIP polymorphism identifies the patient as a patient at risk for developing Alzheimer's disease.

In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 poly-T).

In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

In some embodiments, the method further includes detecting whether the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, E4/E4 subject. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

Also provided is a method for stratifying a subject into a subgroup of a clinical trial of a therapy for the treatment of Alzheimer's disease, the method including: detecting the presence or absence of at least one ApoE allele in a biological sample of the patient, and detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene, wherein the subject is stratified into the subgroup for the clinical trial of the therapy based upon the presence or absence of the at least one ApoE and/or TOMM40 DIP allele.

In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 poly-T insertion).

In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

In some embodiments, the method further includes detecting whether the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, E4/E4 subject. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

Further provided is a method for identifying a patient in a clinical trial of a treatment for Alzheimer's disease including: *a)* identifying a patient diagnosed with Alzheimer's disease; and *b)* determining a prognosis for the patient diagnosed with Alzheimer's disease comprising obtaining a patient profile, wherein the patient profile comprises *i)* detecting the presence or absence of at least one ApoE allele in a biological sample of the patient, *ii)* detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene, and *iii)* converting the patient profile into the prognosis, the prognosis including a prediction of whether the patient is a candidate for the clinical trial for the treatment of Alzheimer's disease.

In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 poly-T).

In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

In some embodiments, the method further includes detecting whether the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, E4/E4 subject. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

A kit for determining if a subject is at increased risk of developing late onset Alzheimer's disease is provided, including: (A) at least one reagent that specifically detects ApoE 3, ApoE 4, or ApoE 2, wherein the reagent is selected from the group consisting of antibodies that selectively bind ApoE 3, ApoE 4, or ApoE 2, and oligonucleotide probes that selectively bind to DNA encoding the same; (B) at least one reagent that specifically detects the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene; and (C) instructions for determining that the subject is at increased risk of developing late onset Alzheimer's disease by: *(i)* detecting the presence or absence of an ApoE isoform in the subject with the at least one reagent; *(ii)* detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene; and *(iii)* observing whether or not the subject is at increased risk of developing late onset Alzheimer's disease by observing if the presence of ApoE isoform and the TOMM40 DIP is or is not detected with the at least one reagent, wherein the presence of the ApoE isoform and the TOMM40 DIP indicates the subject is at increased risk of developing late onset Alzheimer's disease.

In some embodiments, the at least one reagent and the instructions are packaged in a single container.

In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 poly-T).

In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

In some embodiments, the determining step further includes detecting whether the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

A kit is provided for determining if a subject is responsive to treatment for a condition of interest, wherein the condition of interest is associated with ApoE and/or TOMM40, with an active agent, the kit including: (A) at least one reagent that specifically detects ApoE 3, ApoE 4, or ApoE 2, wherein the reagent is selected from the group consisting of antibodies that selectively bind ApoE 3, ApoE 4, or ApoE 2, and oligonucleotide probes that selectively bind to DNA encoding the same; (B) at least one reagent that specifically detects the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene; and (C) instructions for determining that the subject is responsive to treatment for the condition of interest with the active agent of interest by: *(i)* detecting the presence or absence of an ApoE isoform in the subject with the at least one reagent; *(ii)* detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) located in intron 6 or intron 9 of the TOMM40 gene; and *(iii)* determining whether or not the subject is responsive to treatment by observing if the presence of the ApoE isoform and the TOMM40 DIP is or is not detected with the at least one reagent, wherein the presence of ApoE 3 and the TOMM40 DIP indicates that the subject is responsive to the treatment with the active agent.

In some embodiments, the at least one reagent and the instructions are packaged in a single container.

In some embodiments, the DIP is an insertion polymorphism. In some embodiments, the DIP is poly-T deletion/insertion polymorphism (e.g., between 5 and 100, or 10 and 80, or 20 and 50 bp poly-T).

In some embodiments, the DIP is rs10524523, rs10602329 or DIP3. In some embodiments, the DIP is rs10524523.

In some embodiments, the determining step further includes detecting whether the subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject. In some embodiments, the subject is an Apo E3/E3 or E3/E4 subject.

It will be understood that all of the foregoing embodiments can be combined in any way and/or combination. The foregoing and other objects and aspects of the present invention are explained in greater detail in the drawings provided herewith and in the specification set forth below.

Brief Description of the Drawings

Figure 1 shows a general flowchart for identifying a genetic variant in a predetermined region of genomic sequence in a genetic locus of interest, which may be associated with a condition of interest, according to some embodiments.

Figure 2 shows a graph of the mean age of onset of Alzheimer's disease as a function of the inheritance of the five common ApoE genotypes, and representing ApoE 4 as a risk factor for Alzheimer's disease (1993).

Figure 3 shows Regions A, B, and C on Chromosome 19, which are exemplary genetic loci of interest. The TOMM40 gene is in close proximity to the ApoE gene and encodes a 40 kD protein directed to the outer mitochondrial membrane. TOMM40 interacts with ApoE directly in regulation of mitochondrial protein import, and a present hypothesis is that the presence of a particular TOMM40 variant(s) exacerbates the increased risk for Alzheimer's disease associated with the dose-dependent presence of the ApoE 3 allele.

Figure 4 shows, in schematic form, an evolutionary map of Region B as denoted in Figure 3 created according to some embodiments detailed herein. Each horizontal oval represents an observed sequence variant (e.g., a nucleotide change), and the size of the oval represents its frequency. Each variant is called a node in the tree, and the connection of one node to another node by a branch represents an observation of the two variants in *cis* carried by an individual subject.

Figure 5 is a schematic diagram of the phylogenetic tree based on Region B constructed for TOMM40, showing the percentages of ApoE phenotypes in two major groupings, or clades, of the TOMM40 variants in this region.

Figure 6 is a schematic overview of the *TOMM40-ApoE* locus including an LD plot showing haplotype blocks and regions subject to primary sequencing in the exploratory (R1) (23 Kb) and confirmatory (R2) (10 Kb) studies (NCBI Build 36.3). The LD plot is shown for Hapmap data (CEU analysis panel), solid spine haplotype block definition, r^2 values with *D*'/LOD color scheme represented by different line characteristics.

Figure 7 shows representations of the phylogenetic trees with separation of variants. **7A**: SNP variants, clade A vs. B, E6 – E10 represent *TOMM40* exons and vertical lines indicate the approximate locations of the SNPs. Separation of the two main branches has strong bootstrap support (973/1000). **7B**: rs10524523 length polymorphisms. Descriptive statistics are provided

for each group of length polymorphisms. Several long haplotypes that formed individual outgroups, in the tree or very small clades, are in the group identified as 'Remainder.'

Figure 8 presents histograms of the length of the rs10524523 length polymorphism stratified by ApoE genotypes 3/3 (**8A**), 3/4 (**8B**), and 4/4 (**8C**). N=210 haplotypes (AS cohort).

Figure 9 shows the association between AD age of onset and length of the rs10524523 polymorphism for AD patients with onset between 60 and 86 years. Box plots indicate the 95% range (vertical lines), median (horizontal line in box) and interquartile range (box).

Detailed Description

The present invention is explained in greater detail below. This description is not intended to be a detailed catalog of all the different ways in which the invention may be implemented, or all of the features that may be added to the instant invention. For example, features illustrated with respect to one embodiment may be incorporated into other embodiments, and features illustrated with respect to a particular embodiment may be deleted from that embodiment. In addition, numerous variations and additions to the various embodiments suggested herein will be apparent to those skilled in the art in light of the instant disclosure which do not depart from the instant invention. Hence, the following specification is intended to illustrate some particular embodiments of the invention, and not to exhaustively specify all permutations, combinations and variations thereof.

As used in the description of the invention and the appended claims, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. Also, as used herein, "and/or" refers to and encompasses any and all possible combinations of one or more of the listed items, as well as the lack of combinations when interpreted in the alternative ("or").

The present invention is directed to methods for revealing genetic variation in regions of particular interest for complex diseases and disorders. It also relates to the discovery of the most informative genetic markers on the basis of associations with phenotype information. In one embodiment, the invention may be used to locate genetic markers associated with susceptibility to a particular disease, disorder, or condition. In another embodiment, data regarding subject response to a candidate treatment or drug may be included in a phylogenetic analysis for the location of genetic markers associated with a beneficial response to that treatment or drug (i.e., pharmacogenetics). The methods can be applied on any data set of genetic variation from a particular locus. See **Figure 1** for a flowchart of the approach to finding genetic risk factors according to the present invention.

In one aspect, the analysis of the genetic variation is based on variant sequence data. In a second aspect, the structure is uncovered using diploid genotype data, thereby avoiding the need

to either experimentally or computationally infer the component haplotypes (see, e.g., U.S. Patent No. 6,027,896 to Roses et al.). In another aspect, the present method can be applied onto uncharacterized allelic variation that results from the interrogation of a target nucleic acid with an experimental procedure that provides a record of the sequence variation present but does not actually provide the entire sequence. The underlying structure of genetic variation is also useful for the deduction of the constituent haplotypes from diploid genotype data.

It is preferred and contemplated that the methods described herein be used in conjunction with other clinical diagnostic information known or described in the art which are used in evaluation of subjects with diseases or disorders (e.g., those believed to involve mitochondrial dysfunction (e.g. Alzheimer's disease or other neurodegenerative diseases)) or for evaluation of subjects suspected to be at risk for developing such disease. The invention is also applicable for discovery of genetic risk factors for other complex diseases, disorders, or conditions.

The disclosures of all United States patent references cited herein are hereby incorporated by reference herein in their entirety.

1. Definitions. The following definitions are used herein:

"Condition of interest" refers to a specific condition, disease, or disorder designated for phylogenetic study and/or subsequent diagnosis or prognosis. "Condition" as used herein includes, but is not limited to, conditions associated with ApoE and/or TOMM40 and/or mitochondrial dysfunction, e.g., neurodegenerative diseases, metabolic diseases, psychiatric disorders, and cancer.

Examples of conditions in which ApoE and/or TOMM40 have been implicated include, but are not limited to, cardiovascular disease; metabolic disease; neurodegenerative disease; neurological trauma or disease; autoimmune disease (e.g., multiple sclerosis (Pinholt M, et al. Apo E in multiple sclerosis and optic neuritis: the apo E-epsilon4 allele is associated with progression of multiple sclerosis. *Mult Scler.* **11**:511-5 (2005); Masterman, T. & Hillert, J. The telltale scan: APOE ε4 in multiple sclerosis. Lancet Neurol. 3: 331 (2004), neuropsychiatric systemic lupus erythematosus (Pullmann Jr. R, et al. Apolipoprotein E polymorphism in patients with neuropsychiatric SLE. *Clin Rheumatol.* **23**: 97-101 (2004)), etc.)); viral infection (e.g., liver disease associated with hepatitis C infection (Wozniak MA, et al. Apolipoprotein E-∈4 protects against severe liver disease caused by hepatitis C virus. *Hepatol.* **36**: 456-463 (2004)), HIV disease (Burt TD, et al. Apolipoprotein (apo) E4 enhances HIV-1 cell entry in vitro, and the APOE epsilon4/epsilon4 genotype accelerates HIV disease progression. *Proc Natl Acad Sci U S A.* **105**:8718-23 (2008)), etc.)); hip fracture/osteoporosis (Pluijm SM, et al. Effects of gender and age on the association of apolipoprotein E epsilon4 with bone mineral density, bone turnover and the risk of fractures in older people. *Osteoporos Int.* **13:** 701-9 (2002)); mitochondrial diseases

(Chang S, et al. Lipid- and receptor-binding regions of apolipoprotein E4 fragments act in concert to cause mitochondrial dysfunction and neurotoxicity. *Proc Natl Acad Sci U S A.* **102**:18694-9 (2005)); aging (Schächter F, et al. Genetic associations with human longevity at the APOE and ACE loci. *Nat Genet.* **6**:29-32 (1994); Rea IM, et al., Apolipoprotein E alleles in nonagenarian subjects in the Belfast Elderly Longitudinal Free-living Ageing Study (BELFAST). *Mech. Aging and Develop.* **122**: 1367-1372 (2001)); inflammation (Li L, et al., Infection induces a positive acute phase apolipoprotein E response from a negative acute phase gene: role of hepatic LDL receptors. *J Lipid Res.* **49**:1782-93 (2008)); and memory dysfunction (Caselli RJ, et al. Longitudinal modeling of age-related memory decline and the APOE epsilon4 effect. *N Engl J Med.* **361**:255-63 (2009)).

"Cardiovascular disease" as used herein refers to a disease involving the heart and/or blood vessels, including, but not limited to, coronary artery disease (Song Y, et al. Meta-analysis: apolipoprotein E genotypes and risk for coronary heart disease. *Ann Intern Med.* **141**:137-47 (2004); Bennet AM, et al., Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA* **298**:1300-11 (2007)), atherosclerosis (Norata GD, et al. Effects of PCSK9 variants on common carotid artery intima media thickness and relation to ApoE alleles. *Atherosclerosis* (2009) Jun 27. [Epub ahead of print], doi:10.1016/j.atherosclerosis 2009.06.023; Paternoster L, et al. Association Between Apolipoprotein E Genotype and Carotid Intima-Media Thickness May Suggest a Specific Effect on Large Artery Atherothrombotic Stroke. *Stroke* 39:48–54 (2008)), ischemic heart disease (Schmitz F, et al., Robust association of the APOE 4 allele with premature myocardial infarction especially in patients without hypercholesterolaemia: the Aachen study. *Eur. J. Clin. Investigation* **37**: 106-108 (2007)), vascular disease such as ischemic stroke (Peck G, et al. The genetics of primary haemorrhagic stroke, subarachnoid haemorrhage and ruptured intracranial aneurysms in adults. *PLoS One.* **3**:e3691 (2008); Paternoster L, et al. Association Between Apolipoprotein E Genotype and Carotid Intima-Media Thickness May Suggest a Specific Effect on Large Artery Atherothrombotic Stroke. *Stroke* 39:48–54 (2008)), vascular dementia (Bang OY, et al. Important link between dementia subtype and apolipoprotein E: a meta-analysis. *Yonsei Med J.* **44**:401-13 (2003); Baum L, et al. Apolipoprotein E epsilon4 allele is associated with vascular dementia. *Dement Geriatr Cogn Disord.* **22**:301-5 (2006)), etc.

"Neurodegenerative disease" as used herein refers to Alzheimer's disease (Corder EH, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**:921-3 (1993).; Corder EH, et al. There is a pathologic relationship between ApoE-epsilon 4 and Alzheimer's disease. *Arch Neurol.* **52**:650-1 (1995)), Parkinson's disease (Huang X, et al. Apolipoprotein E and dementia in Parkinson disease: a meta-analysis. *Arch*

Neurol. **63**:189-93 (2006); Huang X et al. APOE-[epsilon]2 allele associated with higher prevalence of sporadic Parkinson disease. *Neurology* **62**:2198-202 (2004); Martinez, M. *et al.* Apolipoprotein E4 is probably responsible for the chromosome 19 linkage peak for Parkinson's disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **136B**, 172–174 (2005)), Huntington's disease, and a plurality of less common diseases and disorders which cause neurons to decline, e.g., age-related macular degeneration (Thakkinstian A, et al. Association between apolipoprotein E polymorphisms and age-related macular degeneration: A HuGE review and meta-analysis. *Am J Epidemiol.* **164**:813-22 (2006); Bojanowski CM, et al. An apolipoprotein E variant may protect against age-related macular degeneration through cytokine regulation. *Environ Mol Mutagen.* **47**:594-602 (2006)).

"Neurological trauma or disease" includes, but is not limited to, outcome after head injury (Zhou W, et al. Meta-analysis of APOE4 allele and outcome after traumatic brain injury. *J Neurotrauma.* **25**:279-90 (2008); Lo TY, et al. Modulating effect of apolipoprotein E polymorphisms on secondary brain insult and outcome after childhood brain trauma. *Childs Nerv Syst.* **25**:47-54 (2009)), migraine (Gupta R, et al. Polymorphism in apolipoprotein E among migraineurs and tension-type headache subjects. *J Headache Pain.* **10**:115-20 (2009)), vasogenic edema (James ML, et al. Apolipoprotein E modifies neurological outcome by affecting cerebral edema but not hematoma size after intracerebral hemorrhage in humans. *J Stroke Cerebrovasc Dis.* **18**:144-9 (2009); James ML, et al. Pharmacogenomic effects of apolipoprotein e on intracerebral hemorrhage. *Stroke* **40**:632-9 (2009)), etc.

"Metabolic disease" as used herein includes, but is not limited to, dyslipidemia (Willer CJ, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet.* **40**:161-9 (2008); Bennet AM, et al., Association of apolipoprotein E genotypes with lipid levels and coronary risk. *JAMA* **298**:1300-11 (2007)), end stage renal disease (Oda H, et al. Apolipoprotein E polymorphism and renal disease. *Kidney Int Suppl.* **71**:S25-7 (1999); Hubacek JA, et al. Apolipoprotein E Polymorphism in Hemodialyzed Patients and Healthy Controls. *Biochem Genet.* (2009) Jun 30. [Epub ahead of print] DOI 10.1007/s10528-009-9266-y.), chronic kidney disease (Yoshida T, et al. Association of a polymorphism of the apolipoprotein E gene with chronic kidney disease in Japanese individuals with metabolic syndrome. *Genomics* **93**:221-6 (2009); Leiva E, et al. Relationship between Apolipoprotein E polymorphism and nephropathy in type-2 diabetic patients. *Diabetes Res Clin Pract.* **78**:196-201 (2007)), gallbladder disease (Boland LL,et al. Apolipoprotein E genotype and gallbladder disease risk in a large population-based cohort. *Ann Epidemiol.* **16**:763-9 (2006); Andreotti G, et al. Polymorphisms of genes in the lipid metabolism pathway and risk of biliary tract cancers and stones: a population-based case-control study in Shanghai, China. *Cancer*

Epidemiol Biomarkers Prev. **17**:525-34 (2008)), diabetes mellitus (type II) (Elosua R, et al. Obesity Modulates the Association among *APOE* Genotype, Insulin, and Glucose in Men. *Obes Res.* **11**:1502–1508 (2003); Moreno JA, et al. The Apolipoprotein E Gene Promoter (–219G/T) Polymorphism Determines Insulin Sensitivity in Response to Dietary Fat in Healthy Young Adults. *J. Nutr.* **135**:2535-2540 (2005)), metabolic syndrome, cholelithiasis (Abu Abeid S, et al. Apolipoprotein-E genotype and the risk of developing cholelithiasis following bariatric surgery: a clue to prevention of routine prophylactic cholecystectomy. *Obes Surg.* **12**:354-7 (2002)), etc.

"Psychiatric Disorder" as used herein refers to schizophrenia (Kampman O, et al. Apolipoprotein E polymorphism is associated with age of onset in schizophrenia. *J Hum Genet.* **49**:355-9 (2004); Dean B. et al., Plasma apolipoprotein E is decreased in schizophrenia spectrum and bipolar disorder. *Psychiatry Res.* **158**:75-78 (2008)), obsessive compulsive disorder (OCD), addictive behavior (smoking addiction, alcohol addiction, etc.), bipolar disorder (Dean B. et al., Plasma apolipoprotein E is decreased in schizophrenia spectrum and bipolar disorder. *Psychiatry Res.* **158**:75-78 (2008)), and other diseases, disorders, or conditions of a psychiatric nature.

"Development of a condition" as used herein refers to either an initial diagnosis of a disease, disorder, or other medical condition, or exacerbation of an existing disease, disorder, or medical condition for which the subject has already been diagnosed.

"Diagnosis" or "prognosis" as used herein refers to the use of information (e.g., genetic information or data from other molecular tests on biological samples, signs and symptoms, physical exam findings, cognitive performance results, etc.) to anticipate the most likely outcomes, timeframes, and/or response to a particular treatment for a given disease, disorder, or condition, based on comparisons with a plurality of individuals sharing common nucleotide sequences, symptoms, signs, family histories, or other data relevant to consideration of a patient's health status.

"Biological sample" as used herein refers to a material suspected of containing a nucleic acid of interest. Biological samples containing DNA include hair, skin, cheek swab, and biological fluids such as blood, serum, plasma, sputum, lymphatic fluid, semen, vaginal mucus, feces, urine, spinal fluid, and the like. Isolation of DNA from such samples is well known to those skilled in the art.

A "subject" according to some embodiments is an individual whose genotype(s) or haplotype(s) are to be determined and recorded in conjunction with the individual's condition (i.e., disease or disorder status) and/or response to a candidate drug or treatment. Nucleotide sequences from a plurality of subjects are used to construct a phylogenetic tree, and then analogous nucleotide sequences from an individual subject may be compared to those on the phylogenetic tree for diagnostic or prognostic purposes.

"Gene" as used herein means a segment of DNA that contains all the information for the regulated biosynthesis of an RNA product, including promoters, exons, introns, and other untranslated regions that control expression.

"Genetic locus" or "locus" as used herein means a location on a chromosome or DNA molecule, often corresponding to a gene or a physical or phenotypic feature or to a particular nucleotide or stretch of nucleotides. Loci is the plural form of locus.

"Amplification," as applied to nucleic acids herein refers to any method that results in the formation of one or more copies of a nucleic acid, where preferably the amplification is exponential. One such method for enzymatic amplification of specific sequences of DNA is known as the polymerase chain reaction (PCR), as described by Saiki et al., 1986, Science 230:1350-1354. Primers used in PCR normally vary in length from about 10 to 50 or more nucleotides, and are typically selected to be at least about 15 nucleotides to ensure sufficient specificity. The double stranded fragment that is produced is called an "amplicon," and may vary in length from as few as about 30 nucleotides, to 20,000 or more.

A "marker" or "genetic marker" as used herein is a known variation of a DNA sequence at a particular locus. The variation may be present in an individual due to mutation or inheritance. A genetic marker may be a short DNA sequence, such as a sequence surrounding a single base-pair change (single nucleotide polymorphism, SNP), or a long one, like minisatellites. Markers can be used to study the relationship between an inherited disease and its genetic cause (for example, a particular mutation of a gene that results in a defective or otherwise undesirable form of protein).

A "genetic risk factor" as used herein means a genetic marker that is associated with increased susceptibility to a condition, disease, or disorder. It may also refer to a genetic marker that is associated with a particular response to a selected drug or treatment of interest.

"Associated with" as used herein means the occurrence together of two or more characteristics more often than would be expected by chance alone. An example of association involves a feature on the surface of white blood cells called HLA (HLA stands for human leukocyte antigen). A particular HLA type, HLA type B-27, is associated with an increased risk for a number of diseases including ankylosing spondylitis. Ankylosing spondylitis is 87 times more likely to occur in people with HLA B-27 than in the general population.

A subject "at increased risk of developing a condition" due to a genetic risk factor is one who is predisposed to the condition, has genetic susceptibility for the condition, and/or is more likely to develop the condition than subjects in which the genetic risk factor is absent. For example, a subject who is "at increased risk of developing Alzheimer's disease" due to the

presence of one or two ApoE 4 alleles is more likely to develop Alzheimer's disease than a subject who does not carry an ApoE 4 allele.

"Polymorphism" as used herein refers to the existence of two or more different nucleotide sequences at a particular locus in the DNA of the genome. Polymorphisms can serve as genetic markers and may also be referred to as genetic variants. Polymorphisms include nucleotide substitutions, insertions, deletions and microsatellites, and may, but need not, result in detectable differences in gene expression or protein function. A polymorphic site is a nucleotide position within a locus at which the nucleotide sequence varies from a reference sequence in at least one individual in a population.

A "deletion/insertion polymorphism" or "DIP" as used herein is an insertion of one or more nucleotides in one version of a sequence relative to another. If it is known which of the alleles represent minor alleles, the term "deletion" is used when the minor allele is a deletion of a nucleotide, and the term "insertion" is used when the minor allele is an addition of a nucleotide. The term "deletion/insertion polymorphism" is also used when there are multiple forms or lengths and the minor allele is not apparent. For example, for the poly-T polymorphisms described herein, multiple lengths of polymorphisms are observed.

"Polymorphism data" as used herein means information concerning one or more of the following for a specific gene: location of polymorphic sites; sequence variation at those sites; frequency of polymorphisms in one or more populations; the different genotypes and/or haplotypes determined for the gene; frequency of one or more of these genotypes and/or haplotypes in one or more populations; and any known association(s) between a trait and a genotype or a haplotype for the gene.

"Haplotype" as used herein refers to a genetic variant or combination of variants carried on at least one chromosome in an individual. A haplotype often includes multiple contiguous polymorphic loci. All parts of a haplotype as used herein occur on the same copy of a chromosome or haploid DNA molecule. Absent evidence to the contrary, a haplotype is presumed to represent a combination of multiple loci that are likely to be transmitted together during meiosis. Each human carries a pair of haplotypes for any given genetic locus, consisting of sequences inherited on the homologous chromosomes from two parents. These haplotypes may be identical or may represent two different genetic variants for the given locus. Haplotyping is a process for determining one or more haplotypes in an individual. Haplotyping may include use of family pedigrees, molecular techniques and/or statistical inference.

A "variant," "variance," or "genetic variant" as used herein, refers to a specific isoform of a haplotype in a population, the specific form differing from other forms of the same haplotype in the sequence of at least one, and frequently more than one, variant sites or

nucleotides within the sequence of the gene. The sequences at these variant sites that differ between different alleles of a gene are termed "gene sequence variants," "alleles," "variances" or "variants." The term "alternative form" refers to an allele that can be distinguished from other alleles by having at least one, and frequently more than one, variant sites within the gene sequence. Other terms known in the art to be equivalent to "variances" or "variants" include mutations and single nucleotide polymorphisms (SNPs). Reference to the presence of a variance or variances means particular variances, i.e., particular nucleotides at particular polymorphic sites, rather than just the presence of any variance in the gene.

"Isoform" as used herein means a particular form of a gene, mRNA, cDNA or the protein encoded thereby, distinguished from other forms by its particular sequence and/or structure. For example, the ApoE 4 isoform of apolipoprotein E as opposed to the ApoE2 or ApoE 3 isoforms.

"Cistron" as used herein means a section of DNA found on a single chromosome that contains the genetic code for a single polypeptide and functions as a hereditary unit. A cistron includes exons, introns, and regulatory elements related to a single functional unit (i.e., a gene). The term derives from the classic *cis-trans* test for determining whether genetic elements were able to functionally interact regardless of whether they were located on the same DNA molecule ("trans" complementation) or only when they were located on the same DNA molecule ("cis" acting elements).

The term "genotype" in the context of this invention refers to the particular allelic form of a gene, which can be defined by the particular nucleotide(s) present in a nucleic acid sequence at a particular site(s). Genotype may also indicate the pair of alleles present at one or more polymorphic loci. For diploid organisms, such as humans, two haplotypes make up a genotype. Genotyping is any process for determining a genotype of an individual, e.g., by nucleic acid amplification, antibody binding, or other chemical analysis. The resulting genotype may be unphased, meaning that the sequences found are not known to be derived from one parental chromosome or the other.

"Linkage disequilibrium" as used herein means the non-random association of alleles at two or more loci. Linkage disequilibrium describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected from a random formation of haplotypes from alleles based on their frequencies. Non-random associations between polymorphisms at different loci are measured by the degree of linkage disequilibrium.

"Multiple sequence alignment" or "MSA" as used herein means alignment of three or more nucleotide sequences from genomic DNA derived from a plurality of individuals to determine homology and heterology between the sequences. In general, the input set of query

sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. Computer algorithms are most often used to perform the analysis of aligned sequences.

Some embodiments of the present invention are described with reference to block diagrams illustrating methods (e.g., **Figure 1**), which may include steps implemented by a computer and/or computer program products. It will be understood that each block of the block diagrams and/or operational illustrations, and combinations of blocks in the block diagrams and/or operational illustrations, can be implemented by analog and/or digital hardware, and/or computer program instructions. These computer program instructions may be provided to a processor of a general purpose computer, special purpose computer, ASIC, and/or other programmable data processing apparatus, such that the instructions, which execute via the processor of the computer and/or other programmable data processing apparatus, create means for implementing the functions/acts specified in the block diagrams and/or operational illustrations. Accordingly, it will be appreciated that the block diagrams and operational illustrations support apparatus, methods and computer program products.

Other software, such as an operating system, also may be included. It will be further appreciated that the functionality of the multiple sequence alignment module, mapping module and/or other modules described herein may be embodied, at least in part, using discrete hardware components, one or more Application Specific Integrated Circuits (ASIC) and/or one or more special purpose digital processors and/or computers.

"Mapping" as used herein means creating a phylogenetic tree by assigning a node to each new nucleotide sequence variant observed, connecting that node to another node representing a known sequence carried by the same individual on the same chromosome or cistron, and counting the numbers of each type of subject represented at each node. See **Figure 4** for an example of a phylogenetic tree developed in this manner.

"Phylogenetic" means related to the study of evolutionary connections among various groups of organisms or individuals within a species. Before genetic information was readily available, phylogeny was based mostly on phenotypic observation. "Phylogenetic mapping" as used herein means using DNA sequence data to connect related sequence variants carried by a plurality of individuals in order to determine evolutionary connections and the chronology of divergence. A "phylogenetic tree" is the result of mapping the connections between variants.

"Node" as used herein means a polymorphism data point on a phylogenetic tree representing an actual variant sequence carried by at least one subject. A node is connected by a branch to another node representing a variant sequence carried by the same individual on the same chromosome and in the same cistron but at a different genetic locus within the cistron. The

presence of a node indicates that at least one subject carried both the sequence indicated by the node as well as the sequence represented by the neighboring node to which it is connected by a branch.

"Branch" as used herein means a connection between two nodes representing two distinct variant sequences or haplotypes, wherein the two variants are located on the same chromosome and in the same cistron from an individual subject. "Branching point" means any node from which more than two branches extend, but it is especially used herein to refer to a root node from which three or more nodes extend. A "root node" represents the genetic sequence of a common evolutionary ancestor from which genetic divergence has generated the variety of nearby sequence variants represented by the connected nodes.

"Iteratively" as used herein refers to repetitive calculation of values for each character in a series. For example, each node on a phylogenetic tree is analyzed to calculate the ratio of the number of subjects affected with a condition of interest (such as Alzheimer's disease) to control unaffected subjects; this ratio is compared with the connected nodes to locate correlations with increased or decreased risk for developing a disease, disorder, or condition of interest. A substantial change in this ratio between one node and the next indicates the presence of a variant that either increases or decreases the risk of earlier disease onset. "Iteratively examining the genetic variants" means beginning the analysis with nodes representing the sequences shared by the greatest numbers of individual subjects and successively analyzing the nodes connected by branches extending from that node, followed by the second level of nodes, and so on. The analysis then moves overall from the roots of the tree toward the outer branches and nodes of the tree.

"Treatment" as used herein includes any drug, procedure, lifestyle change, or other adjustment introduced in attempt to effect a change in a particular aspect of a subject's health (i.e. directed to a particular disease, disorder, or condition).

"Drug" as used herein refers to a chemical entity or biological product, or combination of chemical entities or biological products, administered to a person to treat or prevent or control a disease or condition. The term "drug" as used herein is synonymous with the terms "medicine," "medicament," "therapeutic intervention," or "pharmaceutical product." Most preferably the drug is approved by a government agency for treatment of at least one specific disease or condition.

"Disease," "disorder," and "condition" are commonly recognized in the art and designate the presence of signs and/or symptoms in an individual or patient that are generally recognized as abnormal and/or undesirable. Diseases or conditions may be diagnosed and categorized based on pathological changes. The disease or condition may be selected from the types of diseases

listed in standard texts such as Harrison's Principles of Internal Medicine, 1997, or Robbins Pathologic Basis of Disease, 1998.

"Mitochondrial dysfunction" as used herein means any detrimental abnormalities of the mitochondria within a cell or cells. Some diseases, disorders, or conditions presently known in the art to be associated with mitochondrial dysfunction include Alzheimer's disease, Parkinson's disease, and other neurodegenerative diseases, ischemia-reperfusion injury in stroke and heart attack, epilepsy, diabetes, and aging. Many other diseases, disorders, and conditions have been associated with mitochondrial dysfunction in the art. Indeed, the mitochondrion is critical for proper functioning of most cell types, and mitochondrial decline often leads to cell death. This mitochondrial dysfunction causes cell damage and death by compromising ATP production, disrupting calcium homeostasis and increasing oxidative stress. Furthermore, mitochondrial damage can lead to apoptotic cell death by causing the release of cytochrome c and other pro-apoptotic factors into the cytoplasm (for review, see Wallace, 1999; Schapira, 2006). Regarding a specific example found herein, the ApoE 3 and ApoE 4 isoforms are hypothesized to cause mitochondrial dysfunction through interactions with TOMM40. Some TOMM40 variants may act synergistically with ApoE 3 isoform to accelerate mitochondrial decline. This mitochondrial mechanism is believed to contribute to many complex genetic diseases, disorders, and conditions.

"Subjects" as used herein are preferably, but not limited to, human subjects. The subjects may be male or female and may be of any race or ethnicity, including, but not limited to, Caucasian, African-American, African, Asian, Hispanic, Indian, etc. The subjects may be of any age, including newborn, neonate, infant, child, adolescent, adult, and geriatric. Subjects may also include animal subjects, particularly mammalian subjects such as canines, felines, bovines, caprines, equines, ovines, porcines, rodents (e.g., rats and mice), lagomorphs, primates (including non-human primates), etc., screened for veterinary medicine or pharmaceutical drug development purposes.

"Treat," "treating," or "treatment" as used herein refers to any type of measure that imparts a benefit to a patient afflicted with a disease, including improvement in the condition of the patient (e.g., in one or more symptoms), delay in the onset or progression of the disease, etc.

"Late-onset Alzheimer's disease" or "LOAD" as used herein is known in the art, and is the classification used if the Alzheimer's disease has an onset or is diagnosed after the age of 65. It is the most common form of Alzheimer's disease.

2. Methods for Identifying Genetic Variants

While lists of associations derived from genome-wide scans are useful, they are generally inadequate to explain disease complexity. Families, pathways, and interactions of genes can

provide specificities. High-resolution variance mapping may reveal answers to complex genetic interactions. This is particularly applicable where one known genetic risk factor which does not itself entirely explain an association to the disease, disorder, or condition of interest may present an excellent candidate genetic locus for more detailed investigations. Furthermore, pharmacogenetics, while useful for drug development, can also extend biological relevance. The analysis of sequence data from large numbers of individuals to discover variances in the gene sequence between individuals in a population will result in detection of a greater fraction of all the variants in the population.

The initial sequence information to be analyzed by the method of the present invention is derived from the genomic DNA of a plurality of subjects. The organism can be any organism for which multiple sequences are available, but is preferably from human. In identifying new variances it is often useful to screen different population groups based on race, ethnicity, gender, and/or geographic origin because particular variances may differ in frequency between such groups. Most preferably, for diseases or disorders believed to be multigenic (genetically complex diseases/disorders), the phenotypes represented by the subject population are from the extremes of a spectrum. Biological samples containing DNA may be blood, semen, cheek swab, etc. Isolation of DNA from such samples is well known in the art.

In some embodiments, the invention relates to the analysis of nucleotide sequence data from a plurality of subjects having at least one known risk factor for a given disease, disorder, or condition (genetic or otherwise). The nucleotide sequences are analyzed to generate haplotype data, and the haplotypes or genetic variants are then mapped onto a phylogenetic tree to demonstrate the evolution of the sequences represented. By comparing this tree to phenotype data about the plurality of subjects, a prognosis or diagnosis is possible for an individual subject carrying haplotypes observed on the phylogenetic tree.

In other embodiments, the invention relates to the fields of pharmacogenetics and pharmacogenomics and the use of genetic haplotype information to predict an individual's susceptibility to disease and/or their response to a particular drug or drugs, so that drugs tailored to genetic differences of population groups may be developed and/or administered to individuals with the appropriate genetic profile.

Nucleotide sequence information is derived from genomic DNA. Genomic sequence data used may be obtained from clinical or non-human animals or from cultured cells or isolated tissue studies. The organism can be any organism for which multiple sequences are available, but is preferably from human. In identifying new variances it is often useful to screen different population groups based on race, ethnicity, gender, and/or geographic origin because particular variances may differ in frequency between such groups. Most preferably, for diseases or

disorders believed to be multigenic (genetically complex diseases/disorders), the phenotypes represented by the subject population are extreme opposites.

Biological samples containing DNA may be blood, semen, cheek swab, etc. Isolation of DNA from such samples is well known in the art. Methods for determining DNA sequence at a particular genetic locus of interest are also known in the art. Automated sequencing is now widely available and requires only an isolated DNA sample and at least one primer that is specifically designed to recognize a highly conserved sequence within or in close proximity to the genetic locus of interest.

According to some embodiments, a defined genetic region or locus of interest (e.g., defined by a set of forward and reverse PCR primers) is carefully sequenced from a cohort of people inclusive of patients who are well characterized for a particular disorder.

A consensus sequence is determined, and all observed sequence variants for a given genetic locus are compiled into a list. Loci having the greatest number of observed variants represent evolutionary divergence from a common ancestor. As such, these loci are connected in *cis* to loci having only one or very few observed variants. During initial phases of investigation at least, it is preferred that populations be parsed into groups of subjects sharing a common general phenotype representing similar ancestry. Otherwise, analysis of these data through construction of phylogenetic trees will require a prohibitively large number of subjects.

3. Multiple sequence alignments.

Determining the presence of a particular variance or plurality of variances in a gene or gene region in a population can be performed in a variety of ways, all of which involve locating a particular genetic locus by targeting sequences within the region of interest that are known to be highly conserved. From the highly conserved locus, the contiguous sequences are easily obtained through one of many techniques well-known in the art.

The first step in analyzing parallel DNA sequences from a plurality of subjects is multiple sequence alignment ("MSA"). MSA is typically used to display sequence alignment from homologous samples with polymorphic differences within genes or gene regions to show conserved areas and variant sequences. MSAs of the sequence information obtained at the locus of interest may be constructed using one or more various known techniques and publicly available software, and are publicly available from many sources including the Internet. Methods for analyzing multiple sequence alignments known in the art include, e.g., those described in U.S. Patent 6,128,587 to Sjolander; U.S. Patent 6,291,182 to Schork et al.; and U.S. Patent 6,401,043 to Stanton et al.

4. Phylogenetic trees and analysis.

Various methods for construction of "phylogenetic trees" are known in the art (See, e.g., Sanderson, 2008). Sun et al. used "haplotype block" analyses to study associations between toll-like receptor (TLR) variants and prostate cancer (2005) and Bardel et al. (2005) used a cladistic analysis approach to investigate associations between CARD15 gene variants and Crohn's disease. However, neither utilized genetic loci previously associated with the disease to investigate linkages.

Phylogenetic trees according to some embodiments may be constructed with a topology in which haplotype sequence variants observed in individual human subjects studied form nodes (representing each sequence observed in the data) on a tree. Nodes may be joined to other nodes, and the common ancestor is found at the branching site, common root or root node of the tree. A phylogenetic tree reflects the evolutionary relationship between genetic loci for which data are analyzed (see Sanderson, 2008; Tzeng, 2005; Seltman, 2003). **Figure 4** shows a detailed phylogenetic tree constructed for Region B of the genetic locus shown in **Figure 3**.

The starting point for phylogenetic tree estimation is generally an MSA (see above). Multiple software applications are available for constructing phylogenetic trees based on sequence data. See, e.g., U.S. Patent 7,127,466 and U.S. Patent 6,532,467 to Brocklebank, et al. The basic premise is that a genetic locus exhibiting many variants is represented by these variants connected in *cis*. Polymorphisms create branching points (nodes) in the tree that define groups of related sequences or haplotypes.

The phylogenetic tree is utilized for information by iteratively examining ratios of subjects affected with a condition to unaffected control subjects; the calculations begin with nodes observed in the greatest numbers of subjects and move toward the periphery of the tree to nodes observed in fewer subjects. The goal is to locate a branching point, branch, or node where there is substantial change in the ratio of subjects affected with the condition of interest to unaffected control subjects. Such a branching point represents the evolutionary divergence of higher risk subjects from lower risk subjects or vice versa.

Statistical analysis of the phylogenetic tree generated may be performed in accordance with the methods known in the art. One art-recognized method is the calculation of bootstrap confidence levels (see Efron et al., Proc. Natl. Acad. Sci. USA 93, 13429-13434 (1996)).

5. Patient evaluation.

Once a phylogenetic tree has been generated for a particular genetic locus, an individual subject may be evaluated by comparing their DNA sequence to the sequences that comprise the phylogenetic tree. The presence of haplotypes or sequence variants corresponding with regions of the tree representing subjects with higher incidence of the condition of interest (i.e., higher

ratios of subjects affected with the disease or disorder to unaffected control subjects) would mean that the individual subject is also at increased risk. Conversely, substantially lower ratios correspond to reduced risk of developing the condition of interest.

Phylogenetic trees may also be analyzed based upon responsiveness of the condition of interest to treatment with an active agent or treatment method of interest according to some embodiments.

6. APOE and TOMM40.

ApoE phenotypes and genotypes are well known in the art. The established nomenclature system as well as the phenotypes and genotypes for ApoE are described in, for example, Zannis et al., 1982, which is incorporated by reference herein.

TOMM40 (**T**he **O**uter **M**itochondrial **M**embrane channel subunit, 40kDa) phenotypes and genotypes are also known. TOMM40 functions as a channel-forming subunit of the translocase found in mitochondrial membrane that is essential for protein import into mitochondria.

Genome-wide association scanning data from studies of Alzheimer's disease patients have unequivocally identified the linkage disequilibrium region that contains the apolipoprotein E (ApoE) gene. The ApoE 4 variant has been widely replicated as a confirmed susceptibility gene since the initial publications in 1993 (see, e.g., Corder et al.). However, the genome-wide association scanning data resulted in a remarkable "coincidence" observed in cell biology studies involving the co-localization ApoE and TOMM40 to the outer mitochondrial membrane. This other gene, TOMM40, was first encountered during studies modeling linkage disequilibrium around ApoE in 1998. The polymorphisms were located adjacent to ApoE within a small linkage disequilibrium region.

ApoE co-localizes to the outer mitochondrial membrane, suggesting isoform-specific interactions leading to a potential role for ApoE-induced mitochondrial apoptosis as an early step in Alzheimer's disease expression. Biological data have demonstrated that the proportion of mobile mitochondria in neuronal cell culture, as well as the speed at which they move and the distance that they traverse, are factors affecting increased mitochondrial apoptosis. Phylogenetic data suggest an independent genetic effect on the development of Alzheimer's disease for TOMM40.

ApoE binds specifically to mitochondria in human neuronal cultures (Chang, 2005), and sequencing of this linkage disequilibrium region in hundreds of Alzheimer's disease patients and matched controls, combined with mapping the genetic variant evolution of TOMM40, defines a region of particular interest for ApoE-TOMM40 interactions, as shown in **Figure 3**. These evolutionary data further support the genetic association between ApoE and TOMM40, and

suggest that mitochondrial dysfunction could be responsible for neuronal death occurring slowly over many years. The age of onset distribution for Alzheimer's disease (see, e.g., U.S. Patent 6,027,896 to Roses et al.) might reflect the inheritance of tightly linked variants of two biochemically interacting proteins that lead to the clinical expression of disease.

As detailed herein, the interaction between multiple haplotypes of TOMM40 variants and ApoE alleles contribute to Alzheimer's disease pathogenesis; in particular, haplotypes of TOMM40 in linkage to the E 3 allele of ApoE contribute to disease pathogenesis. Several of the TOMM40 gene variants evolved only *cis*-linked to ApoE 3. (Similarly, specific TOMM40 variants may have evolved *cis*-linked to ApoE 4 or ApoE 2.) Thus, any added genetic effect of the TOMM40 variants segregates independently of ApoE 4 but the two variant protein products may functionally interact, in *trans*, to produce a given observable phenotype or trait. Thus, any added genetic effect of the TOMM40 variants segregates independently from ApoE 4. This "coincidence" of adjacent interacting genes may account for the extraordinarily significant statistical association data found in all Alzheimer's disease genome-wide association scanning studies. It is of interest to note that the initial commercially available genome-wide association scanning platforms did not contain any ApoE polymorphisms, but were identified with TOMM40 and ApoC1 SNPs – but the region is virtually always referred to as the "ApoE region."

These data, which combine disease genetics and putative molecular mechanisms of pathogenesis, can also be viewed within a pharmacogenetics context. Because of the strong genetic effect of inheriting an ApoE 4 allele, ApoE 4 has been referred to as a complex susceptibility gene for more than a decade. Consistent replications of the age of onset distributions as a function of ApoE genotype confirm that the role of ApoE 3 inheritance is not totally benign, but is a lower risk factor observed at a slower rate of disease onset. There are genetic variants of TOMM40 that are located only on DNA strands containing ApoE 3 in the linkage disequilibrium regions (Roses et al., unpublished data), and thus not in Hardy-Weinberg equilibrium as was required for SNPs in genome-wide association panels. Evolutionary changes in TOMM40 sequences that are *cis*-linked only to ApoE 3 act to increase the risk of Alzheimer's disease associated with ApoE 3, while other variants of TOMM40 *cis*-linked to ApoE 3 decrease the risk associated with ApoE 3. An independent genetic test would be to determine whether those TOMM40 polymorphisms associated with less Alzheimer's disease segregate at a later age in age of onset distribution plots for ApoE 3 containing genotypes [ApoE 3/3 or ApoE 4/3].

Detecting the presence or absence of ApoE 2, 3 or 4, and/or TOMM40 haplotypes or of DNA encoding the same (including, in some embodiments, the number of alleles for each) in a subject may be carried out either directly or indirectly by any suitable means. A variety of

techniques are known to those skilled in the art. All generally involve the step of collecting a sample of biological material containing either DNA or protein from the subject, and then detecting whether or not the subject possesses the haplotype of interest. For example, the detecting step with respect to ApoE may be carried out by collecting an ApoE sample from the subject (for example, from cerebrospinal fluid, or any other fluid or tissue containing ApoE), and then determining the presence or absence of an ApoE 2, 3, or 4 isoform in the ApoE sample (e.g., by isoelectric focusing or immunoassay).

Determining the presence or absence of DNA encoding an ApoE and/or TOMM40 isoform may be carried out by direct sequencing of the genomic DNA region of interest, with an oligonucleotide probe labeled with a suitable detectable group, and/or by means of an amplification reaction such as a polymerase chain reaction or ligase chain reaction (the product of which amplification reaction may then be detected with a labeled oligonucleotide probe or a number of other techniques). Further, the detecting step may include the step of detecting whether the subject is heterozygous or homozygous for the gene encoding an ApoE and/or TOMM40 haplotype. Numerous different oligonucleotide probe assay formats are known which may be employed to carry out the present invention. See, e.g., U.S. Pat. No. 4,302,204 to Wahl et al.; U.S. Pat. No. 4,358,535 to Falkow et al.; U.S. Pat. No. 4,563,419 to Ranki et al.; and U.S. Pat. No. 4,994,373 to Stavrianopoulos et al. (applicants specifically intend that the disclosures of all U.S. Patent references cited herein be incorporated herein by reference).

In some embodiments, detection may include multiplex amplification of the DNA (e.g., allele-specific fluorescent PCR). In some embodiments, detection may include hybridization to a microarray (a chip, beads, etc.). In some embodiments, detection may include sequencing appropriate portions of the gene containing the haplotypes sought to be detected. In some embodiments, haplotypes that change susceptibility to digestion by one or more endonuclease restriction enzymes may be used for detection. For example, restriction fragment length polymorphism (RFLP), which refers to the digestion pattern when various restriction enzymes are applied to DNA, may be used. In some embodiments, the presence of one or more haplotypes can be determined by allele specific amplification. In some embodiments, the presence of haplotypes can be determined by primer extension. In some embodiments, the presence of haplotypes can be determined by oligonucleotide ligation. In some embodiments, the presence of haplotypes can be determined by hybridization with a detectably labeled probe. See, e.g., U.S. Patent Application Publication No. 2008/0153088 to Sun et al.; Kobler et al., Identification of an 11T allele in the polypyrimidine tract of intron 8 of the CFTR gene, Genetics in Medicine 8(2):125-8 (2006); Costa et al., Multiplex Allele-Specific Fluorescent PCR for Haplotyping the IVS8 (TG)m(T)n Locus in the CFTR Gene, Clin. Chem., 54:1564-1567 (2008); Johnson et al., A

Comparative Study of Five Technologically Diverse CFTR Testing Platforms, J. Mol. Diagnostics, 9(3) (2007); Pratt et al., Development of Genomic Reference Materials for Cystic Fibrosis Genetic Testing, J. Mol. Diagnostics, 11:186-193 (2009).

Amplification of a selected, or target, nucleic acid sequence may be carried out by any suitable means on DNA isolated from biological samples. See generally D. Kwoh and T. Kwoh, 1990. Examples of suitable amplification techniques include, but are not limited to, polymerase chain reaction, ligase chain reaction, strand displacement amplification (see generally Walker et al., 1992a; Walker et al., 1992b), transcription-based amplification (see Kwoh et al., 1989), self-sustained sequence replication (or "3SR") (see Guatelli et al., 1990), the Qβ replicase system (see Lizardi et al., 1988), nucleic acid sequence-based amplification (or "NASBA") (see Lewis, 1992), the repair chain reaction (or "RCR") (see Lewis, supra), and boomerang DNA amplification (or "BDA") (see Lewis, supra). Polymerase chain reaction is currently preferred.

DNA amplification techniques such as the foregoing can involve the use of a probe, a pair of probes, or two pairs of probes which specifically bind to DNA encoding ApoE 4, but do not bind to DNA encoding ApoE 2 or ApoE 3 under the same hybridization conditions, and which serve as the primer or primers for the amplification of the ApoE 4 DNA or a portion thereof in the amplification reaction. Likewise, one may use a probe, a pair of probes, or two pairs of probes which specifically bind to DNA encoding ApoE 2, but do not bind to DNA encoding ApoE 3 or ApoE 4 under the same hybridization conditions, and which serve as the primer or primers for the amplification of the ApoE 2 DNA or a portion thereof in the amplification reaction; and one may use a probe, a pair of probes, or two pairs of probes which specifically bind to DNA encoding ApoE 3, but do not bind to DNA encoding ApoE 2 or ApoE 4 under the same hybridization conditions, and which serve as the primer or primers for the amplification of the ApoE 3 DNA or a portion thereof in the amplification reaction.

Similarly, one may use a probe, a pair of probes, or two pairs of probes which specifically bind to DNA encoding a TOMM40 haplotype of interest, but do not bind to other TOMM40 haplotypes under the same hybridization conditions, and which serve as the primer or primers for the amplification of the TOMM40 DNA or a portion thereof in the amplification reaction.

In general, an oligonucleotide probe which is used to detect DNA encoding ApoE and/or TOMM40 haplotypes is an oligonucleotide probe which binds to DNA encoding the haplotype of interest, but does not bind to DNA encoding other haplotypes under the same hybridization conditions. The oligonucleotide probe is labeled with a suitable detectable group, such as those set forth below in connection with antibodies.

Polymerase chain reaction (PCR) may be carried out in accordance with known techniques. See, e.g., U.S. Pat. Nos. 4,683,195; 4,683,202; 4,800,159; and 4,965,188. In general, PCR involves, first, treating a nucleic acid sample (e.g., in the presence of a heat stable DNA polymerase) with one oligonucleotide primer for each strand of the specific sequence to be detected under hybridizing conditions so that an extension product of each primer is synthesized which is complementary to each nucleic acid strand, with the primers sufficiently complementary to each strand of the specific sequence to hybridize therewith so that the extension product synthesized from each primer, when it is separated from its complement, can serve as a template for synthesis of the extension product of the other primer, and then treating the sample under denaturing conditions to separate the primer extension products from their templates if the sequence or sequences to be detected are present. These steps are cyclically repeated until the desired degree of amplification is obtained. Detection of the amplified sequence may be carried out by adding to the reaction product an oligonucleotide probe capable of hybridizing to the reaction product (e.g., an oligonucleotide probe of the present invention), the probe carrying a detectable label, and then detecting the label in accordance with known techniques, or by direct visualization on a gel.

When PCR conditions allow for amplification of all ApoE allelic types, the types can be distinguished by hybridization with allelic specific probe, by restriction endonuclease digestion, by electrophoresis on denaturing gradient gels, or other techniques. A PCR protocol for determining the ApoE genotype is described in Wenham et al. (1991), incorporated by reference herein. Examples of primers effective for amplification and identification of the ApoE isoforms are described therein. Primers specific for the ApoE polymorphic region (whether ApoE 4, E3 or E2) can be employed. In Wenham, for example, PCR primers are employed which amplify a 227 bp region of DNA that spans the ApoE polymorphic sites (codons 112 and 158, which contain nucleotides 3745 and 3883). The amplified fragments are then subjected to restriction endonuclease CfoI which provides different restriction fragments from the six possible ApoE genotypes which may be recognizable on an electrophoresis gel. See also, Hixon et al. (1990); Houlston et al. (1989) Wenham et al. (1991); and Konrula et al. (1990) for additional methods, all of which are incorporated by reference herein.

In addition to Alzheimer's disease, there are several other genetically complex diseases and disorders for which the methods of the present invention provide advantages over existing analyses. For example, data from multiple type 2 diabetes mellitus genetic studies support the view that very large clinical case/control series will be necessary to provide statistical significance for loci defined by genome-wide association studies.

7. Active agents, compositions and treatment.

As noted above, phylogenetic trees created using the methods detailed herein may also be analyzed based upon responsiveness of the condition of interest to treatment with an active agent or treatment method of interest according to some embodiments, and treatment decisions for a subject or patient may be based upon specific genetic variants identified.

Active agents. Active agents include those known for treatment of a condition of interest, and are inclusive of anti-Alzheimer's disease active agents, including, but are not limited to, acetylcholinesterase inhibitors, NMDA receptor antagonists, and peroxisome proliferator-activated receptor (PPAR) agonists or modulators, including but not limited to those drugs in the thiazolidinedione or glitazar classes. The active agent could also be a biopharmaceutical product, for example an antibody (e.g., monoclonal, polyclonal, derivatives of or modified antibodies such as Domain Antibodies™, Bapineuzumab, etc.), fusion proteins or therapeutic RNA molecules. The active agent could also be a combination of any of these products.

Examples of acetylcholinesterase inhibitors include, but are not limited to, donepezil (commercially available as ARICEPT), galantamine (commercially available as RAZADYNE), and rivastigmine (commercially available as EXELON) and the pharmaceutically acceptable salts thereof. Additional examples include, but are not limited to, those described in U.S. Patent Nos. 6,303,633; 5,965,569; 5,595,883; 5,574,046; and 5,171,750 (the disclosures of all U.S. Patent references cited herein are to be incorporated by reference herein in their entirety).

Examples of NMDA receptor antagonists include, but are not limited to, memantine (commercially available as AKATINOL, AXURA, EBIXIA/ABIXIA, MEMOX and NAMENDA) and the pharmaceutically acceptable salts thereof. Additional examples include, but are not limited to, those described in US Patent Nos. 6,956,055; 6,828,462; 6,642,267; 6,432,985; and 5,990126.

Examples of thiazolidinediones include, but are not limited to, rosiglitazone (commercially available as AVANDIA) and the pharmaceutically acceptable salts thereof. Additional examples include, but are not limited to: 5-(4-[2-(N-methyl-N-(2-benzothiazolyl)amino)ethoxy]benzyl)-2,4-thiazolidine dione; 5-(4-[2-(N-methyl-N-(2-benzothiazolyl)amino)ethoxy]benzylidene)-2,4-thiazol idinedione; 5-(4-[2-(N-methyl-N-(2-benzoxazolyl)amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-benzoxazolyl)amino)ethoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-pyrimidinyl)amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-pyrimidinyl)amino)ethoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-(2-(N-methyl-N-[2-(4,5-dimethylthiazolyl)]amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4,5-dimethylthiazolyl)]amino)ethoxy]benzylidene)-2, 4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-

thiazolyl)amino)ethoxy]benzyl) -2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-thiazolyl)amino)ethoxy]benzylidene)-2,4-thiazolidinedione; 5-[4-(2-(N-methyl-N-(2-(4-phenylthiazolyl))amino) ethoxy)benzyl]-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-(4-phenylthiazolyl))amino) ethoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4-phenyl-5-methylthiazolyl)]amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4-phenyl-5-methylthiazolyl)]amino)ethoxy] benzylidene)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4-methyl-5-phenylthiazolyl)]amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4-methyl-5-phenylthiazolyl)]amino)ethoxy]benzyliden e)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4-methylthiazolyl)]amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4-methylthiazolyl)]amino) ethoxy]benzylidene)-2,4-thiazolidinedione; 5-[4-(2-(N-methyl-N-[2-(5-phenyloxazolyl)]amino) ethoxy)benzyl]-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(5-phenyloxazolyl)]amino) ethoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-[2-(4,5-dimethyloxazolyl)]amino) ethoxy]benzyl)-2,4-thiazolidinedione;

5-(4-[2-(N-methyl-N-[2-(4,5-dimethyloxazolyl)]amino)ethoxy]benzylidene)-2,4-thiazolidinedione; 5-[4-(2-(2-pyrimidinylamino)ethoxy)benzyl]-2,4-thiazolidinedione; 5-[4-(2-(2-pyrimidinylamino)ethoxy)benzylidene]-2,4-thiazolidinedione; 5-(4-[2-(N-acetyl-N-(2-pyrimidinyl)amino)ethoxy]benzyl) -2,4-thiazolidinedione; 5-(4-(2-(N-(2-benzothiazolyl)-N-benzylamino)ethoxy) benzylidene)-2,4-thiazolidinedione; 5-(4-(2-(N-(2-benzothiazolyl)-N-benzylamino)ethoxy) benzyl)-2,4-thiazolidinedione; 5-(4-[3-(N-methyl-N-(2-benzoxazolyl)amino)propoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[3-(N-methyl-N-(2-benzoxazolyl)amino)propoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-[2-(N-methyl-N-(2-pyridyl)amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4 [2-(N-methyl-N-(2-pyridyl)amino)ethoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-[4-(N-methyl-N-(2-benzoxazolyl)amino)butoxy]benzylidene)-2,4-thiazolidinedione; 5-(4-[4-(N-methyl-N-(2-benzoxazolyl)amino)butoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-(2-benzoxazolyl)amino)ethoxy]benzylidene)2,4-thiazolidinedione; 5-(4-[2-(N-(2-benzoxazolyl)amino)ethoxy]benzyl)-2,4-thiazolidinedione; 5-(4-[2-(N-isopropyl-N-(2-benzoxazolyl)amino)ethoxy]benzyl)-2,4-thiazolidin edione, and pharmaceutically acceptable salts thereof. See, e.g., U.S. Patent No. 5,002953.

The active agents disclosed herein can, as noted above, be prepared in the form of their pharmaceutically acceptable salts. Pharmaceutically acceptable salts are salts that retain the desired biological activity of the parent compound and do not impart undesired toxicological effects. Examples of such salts are (a) acid addition salts formed with inorganic acids, for example hydrochloric acid, hydrobromic acid, sulfuric acid, phosphoric acid, nitric acid and the

like; and salts formed with organic acids such as, for example, acetic acid, oxalic acid, tartaric acid, succinic acid, maleic acid, fumaric acid, gluconic acid, citric acid, malic acid, ascorbic acid, benzoic acid, tannic acid, palmitic acid, alginic acid, polyglutamic acid, naphthalenesulfonic acid, methanesulfonic acid, p-toluenesulfonic acid, naphthalenedisulfonic acid, polygalacturonic acid, and the like; (b) salts formed from elemental anions such as chlorine, bromine, and iodine, and (c) salts derived from bases, such as ammonium salts, alkali metal salts such as those of sodium and potassium, alkaline earth metal salts such as those of calcium and magnesium, and salts with organic bases such as dicyclohexylamine and N-methyl-D-glucamine.

Active agents can be administered as prodrugs. "Prodrugs" as used herein refers to those prodrugs of the compounds of the present invention which are, within the scope of sound medical judgment, suitable for use in contact with the tissues of humans and lower animals without undue toxicity, irritation, allergic response and the like, commensurate with a reasonable risk/benefit ratio, and effective for their intended use, as well as the zwitterionic forms, where possible, of the compounds of the invention. The term "prodrug" refers to compounds that are rapidly transformed in vivo to yield the parent compound of the above formulae, for example, by hydrolysis in blood. A thorough discussion is provided in T. Higuchi and V. Stella, Prodrugs as Novel delivery Systems, Vol. 14 of the A.C.S. Symposium Series and in Edward B. Roche, ed., Bioreversible Carriers in Drug Design, American Pharmaceutical Association and Pergamon Press, 1987, both of which are incorporated by reference herein. See also U.S. Patent No. 6,680,299 Examples include a prodrug that is metabolized in vivo by a subject to an active drug having an activity of active compounds as described herein, wherein the prodrug is an ester of an alcohol or carboxylic acid group, if such a group is present in the compound; an acetal or ketal of an alcohol group, if such a group is present in the compound; an N-Mannich base or an imine of an amine group, if such a group is present in the compound; or a Schiff base, oxime, acetal, enol ester, oxazolidine, or thiazolidine of a carbonyl group, if such a group is present in the compound, such as described in U.S. Patent No. 6,680,324 and U.S. Patent No. 6,680,322.

Compositions. The active agents described above may be formulated for administration in a pharmaceutical carrier in accordance with known techniques. See, e.g., Remington, *The Science And Practice of Pharmacy* (9th Ed. 1995). In the manufacture of a pharmaceutical formulation according to the invention, the active compound (including the physiologically acceptable salts thereof) is typically admixed with, *inter alia*, an acceptable carrier. The carrier must, of course, be acceptable in the sense of being compatible with any other ingredients in the formulation and must not be deleterious to the patient. The carrier may be a solid or a liquid, or both, and is preferably formulated with the compound as a unit-dose formulation, for example, a tablet, which may contain from 0.01 or 0.5% to 95% or 99% by weight of the active compound.

One or more active compounds may be incorporated in the formulations of the invention, which may be prepared by any of the well known techniques of pharmacy comprising admixing the components, optionally including one or more accessory ingredients.

The formulations of the invention include those suitable for oral, rectal, topical, buccal (e.g., sub-lingual), vaginal, parenteral (e.g., subcutaneous, intramuscular, intradermal, or intravenous), topical (i.e., both skin and mucosal surfaces, including airway surfaces) and transdermal administration, although the most suitable route in any given case will depend on the nature and severity of the condition being treated and on the nature of the particular active compound which is being used.

Formulations suitable for oral administration may be presented in discrete units, such as capsules, cachets, lozenges, or tablets, each containing a predetermined amount of the active compound; as a powder or granules; as a solution or a suspension in an aqueous or non-aqueous liquid; or as an oil-in-water or water-in-oil emulsion. Such formulations may be prepared by any suitable method of pharmacy which includes the step of bringing into association the active compound and a suitable carrier (which may contain one or more accessory ingredients as noted above). In general, the formulations of the invention are prepared by uniformly and intimately admixing the active compound with a liquid or finely divided solid carrier, or both, and then, if necessary, shaping the resulting mixture. For example, a tablet may be prepared by compressing or molding a powder or granules containing the active compound, optionally with one or more accessory ingredients. Compressed tablets may be prepared by compressing, in a suitable machine, the compound in a free-flowing form, such as a powder or granules optionally mixed with a binder, lubricant, inert diluent, and/or surface active/dispersing agent(s). Molded tablets may be made by molding, in a suitable machine, the powdered compound moistened with an inert liquid binder.

Formulations suitable for buccal (sub-lingual) administration include lozenges comprising the active compound in a flavored base, usually sucrose and acacia or tragacanth; and pastilles comprising the compound in an inert base such as gelatin and glycerin or sucrose and acacia.

Formulations of the present invention suitable for parenteral administration comprise sterile aqueous and non-aqueous injection solutions of the active compound(s), which preparations are preferably isotonic with the blood of the intended recipient. These preparations may contain anti-oxidants, buffers, bacteriostats and solutes which render the formulation isotonic with the blood of the intended recipient. Aqueous and non-aqueous sterile suspensions may include suspending agents and thickening agents. The formulations may be presented in unit\dose or multi-dose containers, for example sealed ampoules and vials, and may be stored in

a freeze-dried (lyophilized) condition requiring only the addition of the sterile liquid carrier, for example, saline or water-for-injection immediately prior to use. Extemporaneous injection solutions and suspensions may be prepared from sterile powders, granules and tablets of the kind previously described. For example, in one aspect of the present invention, there is provided an injectable, stable, sterile composition comprising an active agent(s), or a salt thereof, in a unit dosage form in a sealed container. The compound or salt is provided in the form of a lyophilizate which is capable of being reconstituted with a suitable pharmaceutically acceptable carrier to form a liquid composition suitable for injection thereof into a subject. The unit dosage form typically comprises from about 10 mg to about 10 grams of the compound or salt. When the compound or salt is substantially water-insoluble, a sufficient amount of emulsifying agent which is physiologically acceptable may be employed in sufficient quantity to emulsify the compound or salt in an aqueous carrier. One such useful emulsifying agent is phosphatidyl choline.

Formulations suitable for topical application to the skin preferably take the form of an ointment, cream, lotion, paste, gel, spray, aerosol, or oil. Carriers which may be used include petroleum jelly, lanoline, polyethylene glycols, alcohols, transdermal enhancers, and combinations of two or more thereof.

Formulations suitable for transdermal administration may be presented as discrete patches adapted to remain in intimate contact with the epidermis of the recipient for a prolonged period of time. Formulations suitable for transdermal administration may also be delivered by iontophoresis (see, for example, *Pharmaceutical Research* 3 (6):318 (1986)) and typically take the form of an optionally buffered aqueous solution of the active compound. Suitable formulations comprise citrate or bis\tris buffer (pH 6) or ethanol/water and contain from 0.1 to 0.2M active ingredient.

In addition to active compound(s), the pharmaceutical compositions may contain other additives, such as pH-adjusting additives. In particular, useful pH-adjusting agents include acids, such as hydrochloric acid, bases or buffers, such as sodium lactate, sodium acetate, sodium phosphate, sodium citrate, sodium borate, or sodium gluconate. Further, the compositions may contain microbial preservatives. Useful microbial preservatives include methylparaben, propylparaben, and benzyl alcohol. The microbial preservative is typically employed when the formulation is placed in a vial designed for multidose use. Of course, as indicated, the pharmaceutical compositions of the present invention may be lyophilized using techniques well known in the art.

Dosage. The therapeutically effective dosage of any specific active agent, the use of which is in the scope of present invention, will vary somewhat from compound to compound,

and patient to patient, and will depend upon the condition of the patient and the route of delivery. For oral administration, a total daily dosage of from 1, 2 or 3 mg, up to 30, 40 or 50 mg, may be used, given as a single daily dose or divided into two or three daily doses.

Treatment. Genetic variants as described herein or discovered using the methods as taught herein may be used to determine the course of treatment of a patient afflicted with a condition (e.g., a condition associated with ApoE and/or TOMM40), by, e.g., determining which active agent and/or course of treatment to administer based upon the presence or absence of the genetic variant or variants. The presence or absence of the genetic variants may indicate efficacy of an active agent and/or course of treatment for the patient, predict age of onset for a condition, indicate preferred dose regimens, etc. A genetic profile may be generated for a patient, and the profile consulted to determine whether the patient is among a group of patients that are likely to be responsive to a particular active agent.

Instructions for use may be packaged with or otherwise associated with an active agent indicating recommendations for treatment, time to treatment, dose regimens, etc., based upon the presence or absence of the genetic variants.

8. Methods of determining a prediction of disease risk or a prognosis.

To determine a prediction of disease risk for a non-symptomatic individual or a prognosis (the prospect of affliction or disease course as anticipated from the usual course of disease or peculiarities of the case) according to some embodiments of the present invention, diagnostic data, including the patient's diagnosis or medical hisotry and genetic data, such as the patient's genotype (e.g., ApoE and/or TOMM40 genotype), may be processed to provide therapeutic options and outcome predictions. Processing may include obtaining a "patient profile" such as the collection of a patient's medical history including age and gender, genotyping of the loci of interest (e.g., using appropriately designed primers and using an RT-PCR or PCR amplification step and/or phenotyping, e.g., using an antibody-mediated method or enzymatic test), and statistical or other analyses that converts this raw data into a prognosis. The prognosis may include a prediction of a patient's age of disease onset, response to drug therapy, time to treatment, treatment efficacy, etc. In some embodiments, the prognosis may include the use of a computer software program to analyze patient data and run statistical cross-checks against relational databases in order to convert the patient data or profile to a prognosis.

A "patient profile" includes data and/or materials pertaining to the patient for whom the predictive and/or prognostic analysis is being performed. Data may include information on the patient's diagnosis, age, gender, and/or genotype. The patient profile may also include materials from the patient such as blood, serum protein samples, cerebrospinal fluid, or purified RNA or DNA.

9. Genotype stratification in clinical trials.

Detection of a genotype taught herein or as determined with the methods herein can be used in conducting a clinical trial in like manner as other genotype information is used to conduct a clinical trial, such as described in, e.g., U.S. Patent Nos. 6,573,049 6,368,797 and 6,291,175.

In some embodiments, such methods advantageously stratify or permit the refinement of the patient population (e.g., by division of the population into one or more subgroups) so that advantages of particular treatment regimens can be more accurately detected, particularly with respect to particular sub-populations of patients with particular genotypes. In some embodiments, such methods comprise administering a test active agent or therapy to a plurality of subjects (a control or placebo therapy typically being administered to a separate but similarly characterized plurality of subjects) and detecting the presence or absence of a genotype (e.g., ApoE and/or TOMM40) as described above in the plurality of subjects. The genotype may be detected before, after, or concurrently with the step of administering the test therapy. The influence of one or more detected alleles on the test therapy can then be determined on any suitable parameter or potential treatment outcome or consequence, including, but not limited to, the efficacy of said therapy, lack of side effects of the therapy, etc.

A clinical trial can be set up to test the efficacy of test compounds to treat any number of diseases for which a particular genotype has been determined to be associated, for subjects who are diagnosed with the disease or are at risk for developing the disease. If subjects are genotyped after the completion of a clinical trial, the analyses may still be aimed at determining a relationship between a treatment for a disease and the allele to be assessed for efficacy. Alternatively, if a symptomatic or asymptomatic subject has not yet been diagnosed with the disease but has been determined to be at risk of developing the disease, a similar clinical trial to the clinical trial described above may be carried out.

The underlying biological mechanisms may also be considered when designing the treatment groups. For example, the ApoE 4 (1-272) fragment binds to mitochondria, decreases mitochondrial cellular dynamics and decreases synaptogenesis more than ApoE 3 (1-272). Rosiglitazone, a drug candidate for the treatment of Alzheimer's disease, increases mitogenesis and increases synaptogenesis – opposing the effects of ApoE fragment binding – for ApoE 3 greater than with ApoE 4. Therefore, the drug or treatment candidate (e.g., rosiglitazone) may be selected based upon an underlying mechanism of action as it relates to the genetic markers used for the stratifications (e.g., ApoE 2, E 3, E 4 and/or TOMM40 variants).

Assessment of the efficacy of a drug chosen for the trial may include monitoring the subject over a period of time and analyzing the delay of onset of the disease and the intensity of

the disease at the time of onset, as well as measuring the onset of symptoms which are associated with the disease. A drug that, in a clinical trial, eliminates or delays the onset of the disease, or reduces the symptoms of the disease may be a beneficial drug to use in patients diagnosed with the disease or at risk of developing the disease. Test compounds which may be used in such trials include the agents as described above, including those previously approved for clinical use and new compounds not yet approved for use, or approved for treating a particular disease. Thus, in some embodiments the clinical trial may include the optimization of drug administration, including dosage, timing of administration, toxicities or side effects, route of administration, and efficacy of the treatment.

10. Kits useful for the detection of genotype variants at loci of interest.

Kits for determining if a subject is at increased risk of developing a disease, developing a disease at an earlier age of onset, and/or a candidate for a particular treatment, where the disease is associated with ApoE and/or TOMM40 (e.g., late onset Alzheimer's disease), are provided herein. The kits include at least one reagent specific for detecting for the presence or absence of an ApoE and/or TOMM40 variant as described herein, and may include instructions to aid in determining whether the subject is at increased risk of developing the disease. The kit may optionally include a nucleic acid for detection of an ApoE gene (e.g., ApoE 2, ApoE 3 and/or ApoE 4) or instructions for isoelectric focusing methods for detecting the ApoE genotype; and/or a nucleic acid for detection of a TOMM40 variant as described herein. In some embodiments, the kit may optionally include one or more antibodies which binds to ApoE 2, ApoE 3, ApoE 4, or to isoforms of TOMM40. The test kit may be packaged in any suitable manner, typically with all elements in a single container along with a sheet of printed instructions for carrying out the test.

In some embodiments, the kit may optionally contain buffers, enzymes, and reagents for amplifying the genomic nucleic acids via primer-directed amplification. The kit also may include one or more devices for detecting the presence or absence of particular haplotypes in the amplified nucleic acid. Such devices may include one or more probes that hybridize to a haplotype nucleic acid, which may be attached to a bio-chip or microarray device, such as any of those described in U.S. Pat. No. 6,355,429. The bio-chip or microarray device optionally has at least one capture probe attached to a surface that can hybridize to a haplotype sequence. In preferred embodiments, the bio-chip or microarray contains multiple probes, and most preferably contains at least one probe for a haplotype sequence which, if present, would be amplified by a set of flanking primers. For example, if five pairs of flanking primers are used for amplification, the device would contain at least one haplotype probe for each amplified product, or at least five probes. The kit also preferably includes instructions for using the components of the kit.

The present invention is explained in greater detail in the following non-limiting Examples.

EXAMPLE 1: Construction of Phylogenetic Trees

All of the known genome-wide scanning studies demonstrate extremely significant *p* values around the apolipoproteinC1 [ApoC1] locus. (Mahley et al., Proc. Natl. Acad. Sci. U S A 103: 5644-51 (2006), Coon et al., J. Clin. Psychiatry 68: 613-8 (2007); Li et al., Arch. Neurol. 65: 45-53 (2007)). Of equal importance is that each series identified a "favored" borderline significant candidate gene outside of the ApoE linkage disequilibrium area, but these favored candidate genes were different in each study. TOMM40 is near ApoC1 and in linkage disequilibrium with ApoE. Interactions between ApoE 3 or ApoE 4 and different TOMM40 isoforms are believed to be associated with increased or decreased risk of developing Alzheimer's disease within an earlier age range. Age of onset curves for Apo 4/4, 3/4, 3/3, 2/4, and 2/3 genotypes is shown in **Figure 2**, indicating a range of risk for earlier development of the disease, depending upon the ApoE profile. ApoE alone does not appear to explain all of the data in these age of onset curves, however.

Various methods for polymorphic profiling of Alzheimer's disease risk associated with the different ApoE alleles have been proposed (see, e.g., U.S. Application of Cox et al., No. 20060228728; U.S. Application of Li and Grupe, No. 20080051318). A phylogenetic approach to the ApoE 4 puzzle is demonstrated herein.

Biological samples, DNA isolation, amplification of loci of interest. A total of 340 subjects included 135 Alzheimer's disease cases and 99 age-matched controls in Group A as well as 57 cases and 49 controls in Group B. All subjects carried the ApoE genotypes previously associated with higher risk for earlier disease onset (i.e. 3/3, 3/4, or 4/4). Biological samples containing DNA were collected from all subjects. Genomic DNA was then isolated according to conventional methods for sequencing of genetic loci on Chromosome 19.

Figure 3 shows the genetic regions on Chromosome 19 targeted for study using genome-wide scanning data from multiple reports. The region is encompassed within GenBank reference sequence AF050154. Software was used to generate multiple sequence alignments for variant loci (e.g., ClustalW2, European Bioinformatics Institute). Subsequently, the multiple sequence alignments were analyzed using software for developing phylogenetic trees (e.g., MEGA version 2.1, Center for Evolutionary Functional Genomics, TREEVOLVE, Department of Zoology, University of Oxford, or parsimony-based construction software such as PAUP, Sinauer Associates). Statistical analyses may be performed with, e.g., Genetic Data Analysis (GDA: Software for the Analysis of Discrete Genetic Data, The Bioinformatics Research Center of

North Carolina State University). The results of Region B analysis are demonstrated in the phylogenetic tree of **Figure 4**.

Each piece of data in **Figure 4** represents an observed sequence variant. These variants may be nucleotide substitutions, insertions, deletions, or microsatellites and may or may not result in detectable differences in gene expression or protein function. Each node represents a variant (or a number of variants) that occurs on more than one chromosome. Adjacent nodes define the boundaries of sequences that are in cis, and therefore more likely to be inherited as a unit, in the region of interest on a subject's chromosome. Nodes that precede the greatest number of subsequent nodes represent evolutionarily ancestral variants from which genetic divergence has occurred over time.

The presence of haplotypes or sequence variants corresponding with regions of the tree representing subjects with substantially higher incidence of Alzheimer's disease (i.e., higher ratios of subjects affected with the disease to unaffected control subjects) would mean that the individual subject is also at increased risk. Conversely, substantially lower ratios correspond to reduced risk of developing Alzheimer's disease.

TOMM40 interacts with ApoE directly in regulation of mitochondrial protein import, and a present hypothesis is that the expression of a particular TOMM40 variant(s) exacerbates the relatively moderate risk for Alzheimer's disease associated with the dose-dependent presence of the ApoE 3 allele. Such a TOMM40 variant is discovered within Region B using the methods of the present invention.

Testing new drugs on human subjects carries immense risk (see Kenter and Cohen, Lancet, 368: 1387-91 (2006)). The use of phylogenetic trees to anticipate individual response to a drug or treatment of interest has potential to alleviate that risk significantly. Preliminary studies indicated that rosiglitazone (Avandia) may have genetic-profile specific efficacy in the treatment of Alzheimer's disease (see Risner et al., The Pharmacogenomics Journal 6, 246-254 (2006); Brodbeck et al., Proc. Nat. Acad. Sci. 105, 1343-6 (2008)). Phase II clinical trial data indicate that Alzheimer's disease patients without an ApoE 4 allele responded better to rosiglitazone than patients who carry either 1 or 2 ApoE 4 alleles (data not shown). This supports the hypothesis that variants identified with the methods taught herein may be used to anticipate individual response to treatment based upon genotype.

EXAMPLE 2: Identification of TOMM40 Variants of Interest

174 sequences (2 from each of 87 subjects) were aligned using the CLUSTAL X program (version 2.0.10, Larkin et al., Clustal W and Clustal X version 2.0. Bioinformatics, 23:2947-2948 (2007)). The multiple sequence alignment was used to construct a phylogenetic tree using a neighbor joining algorithm (Saitou and Nei, The neighbor-joining method: a new method for

reconstructing phylogenetic trees. Mol. Evol. Biol., 4:406-425 (1987)) as implemented on the European Bioinformatics Institute (EBI) website.

The resulting phylogenetic tree has a structure of two major groups (A, B) at the first divergence. The ApoE genotype frequencies for these groups are tabulated and shown in **Figure 5**. It is clear that group B contains subject-haplotypes of primarily ε3/ε3 and ε3/ε4 ApoE genotypes and almost no ε4/ε4. Group A contains almost all of the subject haplotypes with the ε4/ε4 genotype.

The list of polymorphisms generated by the SNP discovery platform (Polymorphic) were used to identify specific variants in the TOMM40 gene that separated the data into the two groups. A likelihood ratio test was used to identify significant variants with a p value less than 0.005.

The list of variants is summarized in **Table 1**. In the table, the term "deletion" is used when the minor allele is a deletion of a nucleotide, and the term "insertion" is used when the minor allele is an addition of a nucleotide. The term "deletion/insertion polymorphism" is used when there are more than two possible forms and the minor allele is not apparent. For example, for the poly-T polymorphisms, there are multiple length polymorphisms observed. The second column of the table provides information on the identities of the specific alleles associated with the variant that divide the sequences into the two groups. For example, T>A indicates that the T allele segregates sequences into group "A" on the phylogenetic tree. When two alleles are listed, e.g. G>B; A>A, each allele uniquely segregates the sequence data into the two groups, while when a single allele is listed it is associated with the predominate separation of the data, and the remaining allele does not uniquely separate the data into a homogenous group, but instead a mixture of both groups.

Table 1. TOMM40 variants associated with groups on phylogenetic tree that distribute by ApoE genotype.

Variant	Allele>tree group	Genomic Location (NCBI Build 36.3)	Function	UCSC Classification
50,092,565	T>A	50,092,565	Intron 6	single
50,092,587	T>A	50,092,587	Intron 6	single
rs8106922	G>B; A>A	50,093,506	Intron 6	single
rs34896370, rs55821237, rs56290633	T12_C_T15, T12_C_T16, T13_C_T14, T13_C_T15, T13_C_T16>A; T14_C_T14, T14_C_T15>B	50,093,609	Intron 6	complex
rs34878901	T>B; C>A	50,094,317	Intron 6	single
rs35568738	C>B	50,094,558	Intron 6	single
rs10602329	T16,17,18>A T14,15>B	50,094,716	Intron 6	insertion/deletion
50,094,733	->A	50,094,733	Intron 7	insertion
rs10524523	T12,14,15,16,17>B T21,22,26,27,28,29, 30,31,32,33,34 35,36>A	50,094,889	Intron 6	insertion/deletion
rs1160985	T>B; C>A	50,095,252	Intron 6	single
50,095,506	T>A	50,095,506	Intron 6	single
rs760136	A>A; G>B	50,095,698	Intron 6	single
rs1160984	T>B	50,095,764	Intron 6	single
rs741780	C>B; T>A	50,096,271	Intron 8	single
rs405697	A>A	50,096,531	Intron 9	single
50,096,647 (DIP3)	->A	50,096,647	Intron 9	deletion
50,096,697	C>A	50,096,697	Intron 9	single
rs1038025	C>B; T>A	50,096,812	Intron 9	single
rs1038026	G>B; A>A	50,096,902	Intron 9	single
rs1305062	C>B; G>A	50,097,361	Intron 9	single
rs34215622	G>B; ->A	50,098,378	Exon 10	insertion
rs10119	A>A	50,098,513	Exon 10	single
rs7259620	G>A; A>B	50,099,628	unknown	single

EXAMPLE 3: Two distinct forms of ApoE 3 : those linked to TOMM40 haplotypes that increase risk and decrease age of onset, and those that decrease risk

The association of apolipoprotein E (ApoE) genotypes, particularly ApoE ε4 (ApoE 4), with the risk and age of onset of Alzheimer's disease (AD) remains the most confirmed genetic association for any complex disease. Estimates of the heritability of ApoE 4 for late onset AD

range from 58% to 79%, and the population attributable risk due to the ApoE 4 allele is between 20% and 70%. These estimates suggest that other genetic variants and/or interactions between variants incur additional disease risk and modify age of onset distributions.

Genome wide scan association results for AD have consistently reproduced the extraordinary association of the LD region containing ApoE. TOMM40, the protein translocase of the outer mitochondrial membrane, is in high LD with ApoE, and codes for the membrane channel through which cytoplasmic peptides and proteins traverse in order to synthesize new mitochondria. Our objectives were to identify additional haplotypes within the LD region that increase the estimates of heritability.

Methods: We examined the LD region containing both ApoE and TOMM40 using deep (10X) primary sequencing in AD patients and controls. We performed phylogenetic analyses of the LD region covering TOMM40 and ApoE in 66 patients and 66 age-matched controls with respect to risk and age of onset distribution.

Conclusion: We found that unique and distinct inherited families of different TOMM40 variants are located on the same genomic interval as ApoE 3, but not on the ApoE 4-containing genomic interval, and can either increase or decrease the age of risk distribution of AD. Therefore, the genetic inheritance of these TOMM40 variants are independent of the inheritance of ApoE 4, effectively providing a differentiation of two distinct forms of ApoE 3 : those linked to TOMM40 haplotypes that increase risk and decrease age of onset, and those that decrease risk. These data increase the accuracy of genetic age of onset risk, dependent on age, ApoE and TOMM40 genotypes and provide the opportunity to define high risk of AD over the next 5-7 years, versus lower risk of AD.

EXAMPLE 4: Analysis of three identified TOMM40 DIP variants

Three of the TOMM40 variants identified in this application are deletion/insertion polymorphisms (DIPs) located in intron 6 or intron 9. These DIPs are identified as rs10524523 and rs10602329 in the National Center for Biotechnology Information dbSNP database, and a previously undescribed polymorphism, designated as DIP3. These polymorphisms are located at chr19:50,094,889, chr19:50,094,731, and chr19:50,096,647, respectively, according to NCBI build 36. This invention describes the identification of these DIPs using phylogenetic analysis of the TOMM40 gene, specifically of a 10 Kb fragment of the gene, and that the DIPs are associated with different evolutionary groups determined by phylogenetic analysis. This invention further discloses the utility of these DIPs for (1) determining risk of a healthy person for developing Alzheimer's disease in the future, and (2) for predicting age of onset of AD within an approximately 8 year time-frame.

The three DIP polymorphisms characterized herein correspond to different lengths of DIP poly-T repeats in the TOMM40 gene. The association of DIP poly-T variants with disease risk has precedence. For example a poly-T variant in intron 8 of the cystic fibrosis transmembrane conductance regulator (CFTR) gene is associated with skipping of exon 9 and the development of cystic fibrosis (Groman et al., Am J Hum Genet 74(1):176-9 (2004)). Herein is disclosed: (1) use of the novel method - phylogenetic association analysis (described above) - to identify DIPs that are predictive of disease risk and/or differences in age of disease onset, (2) the identity of three specific DIPs associated with differences in AD age of onset and AD risk, (3) the use of these SNPs individually, together, or with other sequence variants in TOMM40 or ApoE to diagnose disease or predict or determine disease characteristics such as age of disease onset, disease prognosis, disease sub-types, disease severity, and also to analyze or determine the response to drugs.

Phylogenetic analysis reveals the distribution of rs10524523 and rs10602329 DIPs into two different clades. This analysis reveals that shorter poly-T lengths at these loci map to the phylogenetically-identified clades in group B, the group that also comprises higher percentages of ApoE e3/e3 genotype subjects, effectively few (0%) ApoE e4/e4 subjects and lower case/control ratios (i.e., AD disease risk) (**Figure 5**). The association between DIP length and phylogenetic group is statistically significant ($p < 0.0001$) by the likelihood ratio test or Pearson Chi-square test.

Due to the genomic architecture, the high linkage disequilibrium and the evolutionary relationships as indicated the phylogenetic analysis, between the two genes, and the putative physical interaction between the two gene products, the influence of TOMM40 genotype is likely to extend to other diseases that are influenced by ApoE genotype. These diseases include, but are not limited to, Parkinson's disease, Multiple sclerosis, cardiovascular disease, dyslipidemia, recovery from traumatic brain injury, recovery from brain ischemic events, response to anaesthetics, and response to drugs used to treat AD and the diseases listed here.

These polymorphisms could also be used in drug discovery efforts for the screening of compounds useful for treating diseases influenced by variations in TOMM40 or ApoE protein or gene variants.

In addition, the variants may influence or determine therapies based on specifically targeted biopharmaceuticals as exemplified by monoclonal antibodies and siRNA molecules.

The DIP polymorphisms in TOMM40 that are disclosed herein can be identified from an individual's DNA sample using many different molecular nucleotide analysis methodologies, including, but not limited to, DNA sequencing with the primers denoted in **Table 4** listed below.

EXAMPLE 5: Longer poly-T tracts at rs10524523 are significantly correlated with earlier age of onset of LOAD

Phylogenetic analysis has been used to identify genomic relationships between low frequency genetic variants and to cluster evolutionarily related haplotypes (Hahn et al. Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* **167**, 867-77 (2004)). This methodology was employed to explore the ApoE-TOMM40 LD block for the existence of novel risk determinants for LOAD. In an exploratory study, 23 Kb of DNA containing the TOMM40 and ApoE genes were amplified and sequenced, and phase-resolved haplotypes were determined, for 72 LOAD cases and 60 age-matched controls (Li et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* **65**, 45-53 (2008)). It was possible to construct a distinct phylogenetic tree for 10 Kb, encoding exons 2-10, of this region. Two clades (A and B) were distinguished with strong bootstrap support (98%, 1000 replicates). There was a significant difference in the distribution of the ApoE genotypes between the two clades of TOMM40 haplotypes on this phylogenetic tree, suggesting that this region could be functionally significant. Both clades contained subjects with the ε3/ε3 genotype, but 98% of all clade B haplotypes occurred in *cis* with the ApoE ε3 allele ($P = 1.2 \times 10^{-18}$, Fisher's exact test, two-tailed).

The phylogenetic structure of this 10 Kb region of TOMM40, the ApoE ε3-specific inheritance of particular haplotypes, and the identify of the clade-specific polymorphisms were subsequently confirmed in two independent LOAD case/control cohorts, including one cohort with autopsy-confirmed AD status and age of disease onset data. The association between the two clades and disease risk and age of disease onset, where the data was available, was also explored for these two cohorts. The first cohort (AS) comprised AD cases ($n = 74$) and controls ($n = 31$) ascertained at the Arizona Alzheimer's Disease Research Center (ADRC). The second cohort (DS) was assembled at the Duke Bryan ADRC and comprised ApoE ε3/ε4 subjects only (40 autopsy-confirmed cases with known age of disease onset and 33 controls) (**Table 2**). Although DNA sequencing was successful for a subset of the DS cohort who had disease onset from 50 to 68 years of age, association analyses were limited to a subset of patients who developed AD after the age of 60.

Table 2. Cohort compositions. The number of cases and controls, mean age, and percentage that are female are shown for each series. Mean age is given as age-at-diagnosis of AD for cases and age-at-examination for controls. The standard deviation from the mean is given in parenthesis.

Series	*n*		Mean Age (SD)				% Females	
	Cases	Controls	Cases		Controls		Cases	Controls
AS	74	31	81.7	(8.01)	77	(8.93)	56.3	46.7
DS	40	33	69.3	(8.3)	71.9	(7.5)	70	66.7

A phylogenetic tree of similar structure to that generated in the exploratory study was developed with strong bootstrap support (97%, 1000 replicates) for the AS cohort. ApoE ε4/ε4 subjects occurred only in clade A (98% separation between groups, $P = 2.0 \times 10^{-4}$ Fisher's exact test, two-tailed), while the remaining ApoE genotypes were distributed between clades A and B (**Figure 6**). That is, ApoE ε4 was always in LD with clade A variants whereas ApoE ε3 occurred in both clade A and clade B haplotypes. Examination of the distribution of the few ApoE ε2/ε4 subjects on the phylogenetic tree suggests that ApoE ε2-TOMM40 haplotypes share a similar evolutionary history with ApoEε3-TOMM40 haplotypes (data not shown). To verify the phylogenetic structure using a separate method, and to ensure that recombination within the genetic interval did not confound the phylogenetic tree structure developed for the AS cohort, haplotype networks were also constructed using statistical parsimony (TCS version 1.21 (Clement et al. TCS: a computer program to estimate gene genealogies. *Mol Ecol* **9**, 1657-9 (2000))). The major subject-haplotype clusters derived from the two methods (maximum parsimony and TCS) were congruent.

Clade A was more frequently associated with AD cases than was clade B (OR = 1.44, 95% CI = 0.76 - 2.70). ApoE ε3/ε4 heterozygotes ($n = 36$) were analyzed to estimate disease risk associated with clade A haplotypes while controlling for the effect of ApoE ε4. There was a trend to higher incidence of LOAD for the subset that was homozygous for TOMM40 clade A relative to the subset that was heterozygous for clade A and clade B (OR = 1.36, 95% CI = 0.40 – 4.61) and thus it was postulated that at least some of the TOMM40 variants which define clade A confer ApoE ε4-independent risk of LOAD.

Analysis of the AS cohort sequence data identified 39 polymorphic sites in the TOMM40 10 Kb region, of which there were 30 parsimony-informative sites (at least two different nucleotides, each represented in at least two sequences). Of the 30 parsimony-informative sites, 18 had a minor allele frequency (MAF) > 0.10 and six SNPs were outside the boundary of the

TOMM40 gene. 10 SNPs occurred exclusively in the context of ApoE ε3 (P = 6.07 x 10^{-50}, Fisher's exact test, two-tailed, n = 210) and were never observed in ApoE ε4/ε4 homozygous subjects (n = 16). The majority of the ε3-specific TOMM40 variants were located in intronic regions.

Figure 7 illustrates the 10 SNPs and 6 insertion/deletion polymorphisms that distinguish TOMM40 clades A and B (at $P < 0.001$) for the ApoE ε3/ε3 subjects from the AS cohort. These polymorphisms were tested individually and as haplotypes for association with LOAD risk (**Table 3**). The odds ratios for disease risk for each clade B allele, in all cases the minor allele, suggest that the clade B alleles are protective of AD risk in the AS cohort, however, in each case the association narrowly missed significance. To account for the effect of ApoE ε4 on the odds ratios reported in **Table 3**, a balanced set of 48 AD cases and 48 AD controls was constructed by selecting sequences at random from ApoE ε3/ε4 subjects from the pooled AS and DS cohorts. Single SNPs again were not significantly associated with LOAD in this balanced data set. However, the minor alleles of four of the SNPs (rs8106922, rs1160985, rs760136, rs741780) that distinguish TOMM40 clade B were assayed previously in three LOAD case/control genome-wide association studies and were found to be protective of disease risk (OR < 1 in each case), which is consistent with the trend observed in our study (Abraham et al. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics* **1**, 44 (2008); Carrasquillo et al. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet* **41**, 192-198 (2009); Takei et al. Genetic association study on in and around the ApoE in late-onset Alzheimer disease in Japanese. *Genomics* **93**, 441-448 (2009)).

Table 3. Descriptive statistics and allelic and genotypic association results for the individual SNPs.

All *ApoE* genotypes																		
SNP ID	Position	Allele	clade B allele	MAF (all)	MAF (cases)	MAF (controls)	LOAD (M)	LOAD (m)	Control (M)	Control (m)	LOAD (MM)	LOAD (Mm)	LOAD (mm)	Control (MM)	Control (Mm)	Control (mm)	OR	95% CI lower	95% CI upper
rs1038025	50096812	T/c	c	0.31	0.28	0.37	106	41	39	23	40	27	7	11	17	3	0.66	0.35	1.23
rs1038026	50096902	A/g	g	0.31	0.28	0.37	106	41	39	23	40	27	7	11	17	3	0.66	0.35	1.23
rs1160985	50095252	C/t	t	0.30	0.28	0.37	107	41	39	23	40	27	7	11	17	3	0.65	0.34	1.19
rs1305062	50097361	G/c	c	0.28	0.26	0.31	106	38	43	19	43	24	7	13	17	1	0.81	0.42	1.56
rs34215622	50098378	-/g	g	0.28	0.26	0.34	110	38	40	21	42	26	6	12	17	2	0.66	0.35	1.25
rs34878901	50094317	C/t	t	0.26	0.25	0.28	105	35	44	17	45	23	6	15	15	1	0.86	0.44	1.70
rs7259620	50099628	G/a	a	0.30	0.27	0.37	108	40	39	23	41	26	7	11	17	3	0.63	0.33	1.18
rs741780	50096271	T/c	c	0.30	0.28	0.37	107	41	39	23	40	27	7	11	17	3	0.65	0.34	1.19
rs760136	50095698	A/g	g	0.30	0.28	0.37	107	41	39	23	40	27	7	11	17	3	0.65	0.34	1.19
rs8106922	50093506	A/g	g	0.28	0.26	0.31	109	39	43	19	42	25	7	13	17	1	0.81	0.42	1.55
APOE ε3/ε4																			
SNP ID	Position	Allele	clade B allele	MAF (all)	MAF (cases)	MAF (controls)	LOAD (M)	LOAD (m)	Control (M)	Control (m)	LOAD (MM)	LOAD (Mm)	LOAD (mm)	Control (MM)	Control (Mm)	Control (mm)	OR	95% CI lower	95% CI upper
rs1038025	50096812	T/c	c	0.28	0.25	0.38	68	28	63	33	22	24	2	17	29	2	0.79	0.43	1.45
rs1305062	50097361	G/c	c	0.27	0.24	0.38	69	25	64	32	25	21	2	17	30	1	0.72	0.39	1.35
rs34215622	50098378	-/g	g	0.28	0.25	0.38	70	26	64	32	24	22	2	17	30	1	0.74	0.40	1.38
rs34878901	50094317	C/t	t	0.24	0.20	0.38	69	25	61	31	25	21	2	18	29	1	0.71	0.38	1.34
rs8106922	50093506	A/g	g	0.28	0.25	0.38	70	25	64	32	25	21	2	17	30	1	0.71	0.38	1.33

WO 2010/019550 PCT/US2009/053373

Another polymorphism that distinguished the two clades and, therefore, two groups of ApoE ε3 haplotypes, was a poly-T variant (rs10524523) located in intron 6 of TOMM40. On ApoE ε4 chromosomes, the variant was relatively long, with a narrow, unimodal distribution of lengths (21 – 30 T residues, mean = 26.78, s.d. = 2.60, *n* = 32), whereas on ApoE ε3 chromosomes, a bimodal distribution of lengths was evident with peaks at 15.17 (s.d. = 0.85, *n* = 36) and 33.15 (s.d. = 2.09, *n* = 55) T residues (**Figure 8**). Longer poly-T lengths (T >= 27) segregated almost exclusively into clade A, the higher risk clade, in the AS cohort ($P = 7.6 \times 10^{-46}$, *n* = 210, Fisher's exact test, two-tailed). The case/control ratio for the category containing the two, most common, shorter lengths (15 or 16 T residues) was 1.46 (95% CI = 1.25 – 1.75), and the case/control ratio for the longer length category (28, 29, 33 and 34 T residues) was 2.02 (95% CI = 1.13 – 2.87). This data showed a trend to an association between the longer rs10524523 poly-T length and AD (OR = 1.38, 95% CI = 0.80 – 2.39).

While there were only trends toward association of TOMM40 haplotypes or individual polymorphisms with LOAD for the AS cohort, there was a significant association between poly-T length category of rs10524523 and age of LOAD onset. This was tested using the DS cohort of autopsy-confirmed ApoE ε3/ε4 subjects for whom there was disease onset data. Longer poly-T alleles (>= 27 T residues) were significantly associated with onset of disease at a much younger age (70.5 years +/- 1.2 *versus* 77.6 years +/- 2.1, $P = 0.02$, *n* = 34) (Fig. 5).

This polymorphism, therefore, significantly impacted age of disease onset for individuals who carry an ApoE ε3 allele. Three other poly-T length polymorphisms located in intron 6 (rs34896370, rs56290633 and rs10602329) also distinguish clades A and B, but these polymorphisms were not associated with age of disease onset. Similarly, there was no relationship between haplotypes of clade-distinguishing SNPs and age of LOAD, or for the single SNP, rs8106922, which had been significantly associated with AD risk in three genome-wide association studies (Abraham et al. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics* **1**, 44 (2008); Carrasquillo et al. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet* **41**, 192-198 (2009); Takei et al. Genetic association study on in and around the ApoE in late-onset Alzheimer disease in Japanese. *Genomics* **93**, 441-448 (2009)) (data not shown).

We conclude that longer poly-T tracts at rs10524523 are significantly correlated with earlier age of onset of LOAD. The length of this variant is relatively homogeneous, and relatively long, on ApoE ε4 chromosomes, whereas there are two categories of poly-T lengths linked to ApoE ε3. ApoE ε2 chromosomes also appear to carry variable-length poly-T repeats

similar to ε3 chromosomes, but further investigation is needed to verify this preliminary finding and to determine if the poly-T repeat impacts the very late age of disease onset for carriers of ApoE ε2.

While it is possible that there are other variants that influence age of onset of LOAD for individuals who are not homozygous for ApoE ε4, the length of the poly-T polymorphism in TOMM40 intron 6 appears to be the most powerful genetic predictor in this linkage region and should be validated prospectively. These data suggest that ApoE genotype-stratified age of onset curves (Corder et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921-3 (1993); Li et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Arch Neurol* **65**, 45-53 (2008)) are, in reality, sets of curves with each curve reflecting a specific interaction of linked polymorphisms in ApoE and TOMM40. Therefore, these data add resolution to the prediction of age of LOAD onset, within a 5-7 year window, for individuals over 60 years of age. The study to validate the association of ApoE genotypes and TOMM40 haplotypes or rs10524523 with age of disease onset is currently being planned. This study will be a prospective, 5 year, population-based study conducted in several ethnic groups, and will be combined with a prevention or delay of disease onset drug trial.

METHODS

The two cohorts analyzed in this study were from the Arizona Alzheimer's Disease Research Center (ADRC), Phoenix, Arizona and the Duke Bryan ADRC, Durham, North Carolina. All subjects were of European descent. The Arizona and Duke studies were approved by institutional review boards and appropriate informed consent was obtained from all participants. Age and gender data for the cases and controls in each cohort are shown in **Table 2**. For the Duke cohort, the age of disease onset was determined retrospectively and disease diagnosis was confirmed by autopsy.

Samples were plated on 96 well plates for long-range PCR and DNA sequencing at Polymorphic DNA Technologies (Alameda, CA).

Long-range PCR was performed using Takara LA Taq Polymerase (Takara Mirus Bio). The reaction mix and PCR conditions were the same as those recommended by the manufacturer. PCR was conducted in a 50 μL volume with 2.5 U of LA Taq and 200-400 ng human genomic DNA. Thermocycling was carried out with the following conditions: 94°C, 1 min for 1 cycle; 94°C, 30 sec; 57°C, 30 sec; 68°C, 9 min for 14 cycles; 94°C, 30 sec; 57°C, 30 sec; 68°C, 9 min +15 sec/cycle for 16 cycles; 72°C, 10 min for 1 cycle. Primers for long-range PCR are shown in **Table 4**.

Table 4. Forward and reverse sequencing primers are listed. The shaded row indicates the forward and reverse primers used for long-range PCR of R2 (**Figure 2**)

Forward Primers				Reverse Primers			
Sequence	UCSC Primer Coordinate (of 3'-end of primer)	Primer Position in Cloned PCR Product (of 3'-end of primer)	SEQ ID NO:	Sequence	UCSC Primer Coordinate (of 3'-end of primer)	Primer Position in Cloned PCR Product (of 3'-end of primer)	SEQ ID NO:
AACTCAGAGGCCAGAGATTCTAAGT	50,092,429	25	1	AACAGCCTAATCCCAGCACATTTAC	50,101,560	9,156	2
CAGGAAACAGCTATGAC	50,092,292	-112	3	CCCACTGGTTGTTGA	50,093,034	630	4
GTGTGATGGTGATTCAAC	50,093,038	634	5	GAATAGGGGCCTTTCA	50,093,282	878	6
CTGCAGGTATGAAAG	50,093,287	883	7	CAATCTCCTAGGGTGC	50,093,512	1108	8
GTCTCTGCAGATGTG	50,093,601	1197	9	CGGAAGTTGCAGTAAG	50,093,706	1302	10
TACTGCAACTTCCGC	50,093,722	1318	11	AAGGTCAAGGTTACACT	50,094,318	1914	12
TCTCTGTTGCCCACG	50,094,289	1885	13	ACAAGCCTAGGTGACAT	50,094,790	2386	14
CCCAACTAATTTTTGTATTCG	50,094,609	2205	15	CCTGTAATCCCAGCTAT	50,095,002	2598	16
ACATTTGTGGCCTGTAC	50,095,129	2725	17	TCATCTCTCTGTGAACCTAA	50,095,324	2920	18
CCACATGGGCTTGTGT	50,095,603	3199	19	GGCAAAATGACGATCAGT	50,095,804	3400	20
CCCAGATGCCCAAATC	50,096,082	3678	21	GCAGCACCAGCTAGT	50,096,218	3814	22
AACTCTGAGTGGATGTG	50,096,471	4067	23	GATGGTCTCAATCTCCTTA	50,096,620	4216	24
CTATAGTCCCAACTACTGA	50,096,730	4326	25	TTTTTTCCAAGCATAAAACATAGTA	50,096,863	4459	26
AGTCCCCGCTACTTA	50,097,080	4676	27	GGGGATGGACAAAGCT	50,097,268	4864	28
ACCACAGGTGTATGCC	50,097,451	5047	29	TGAAAAGCCCTCTAGAC	50,097,898	5494	30
GAACAGATTCATCCGCA	50,097,864	5460	31	CACCCACGATCCAGTT	50,098,141	5737	32
TGTGGATAGCAACTGGAT	50,098,148	5744	33	CAAAGCCACACTGAAACTT	50,098,231	5827	34
GGGATTCTGAGTAGCA	50,098,469	6065	35	CAGAATCCTGCGT	50,098,526	6122	36
TGCTGCCTTAAGTCCG	50,098,937	6533	37	ACACTTGAGAAAACGG	50,098,797	6393	38
CTGGGGTCAGCTGAT	50,099,350	6946	39	ACAAAGTCCTCTATAGCC	50,099,077	6673	40
TGAAACATCTGGGATTTATAAC	50,099,679	7275	41	TAACCTGGGGTTGGTT	50,099,429	7025	42
CTGGAAACCACAATACC	50,099,990	7586	43	AAGTTCCTTTGCTCATCAG	50,099,829	7425	44
ATCTCGGCTCACTGTA	50,100,261	7857	45	GCAAGAGGGAGACTGT	50,100,207	7803	46
GTCAAAAGACCTCTATGC	50,100,739	8335	47	TGTGCCTGGATGAATGTA	50,100,567	8163	48
AGGACTCCACGAGT	50,101,197	8793	49	TGAGCTCATCCCCGT	50,100,960	8556	50
				CCGTGTTCCATTTATGAG	50,101,328	8924	51
				GTAAAACGACGGCCAG	50,101,681	9277	52

PCR products were run on a 0.8% agarose gel, visualized by crystal violet dye, compared to size standards, cut out of the gel, and extracted with purification materials included with the TOPO XL PCR Cloning kit (Invitrogen). Long-range PCR products were cloned into a TOPO XL PCR cloning vector. This system uses a TA cloning vector and is recommended for inserts of up to 10 kb. Per the manufacturer's instructions, electro-competent cells (from the same kit) were transformed by the vector, plated in the presence of antibiotic, and incubated. Ten clones from each plate were picked and cultured in a 96-well format.

Diluted cultures were transferred to a denaturing buffer that was part of the TempliPhi DNA Sequencing Template Amplification kit (GE HealthCare/Amersham Biosciences). This buffer causes the release of plasmid DNA but not bacterial DNA. Cultures were heated, cooled, spun, and transferred to fresh plates containing the TempliPhi enzyme and other components. This mixture was incubated at 30°C for 18 hours to promote amplification of the plasmid templates. These products were then spun and heated to 65°C to destroy the enzyme.

Plasmid templates were used in DNA sequencing reactions using the Big Dye, version 3.1 sequencing kit (Applied Biosystems). For each reaction, an appropriate sequencing primer (**Table 4**) was used that was designed to anneal to a unique location of the template. Cycle sequencing was carried out with an annealing temperature of 50°C, an elongation temperature of 60°C, and a denaturation temperature of 96°C, for a total of 30 cycles. Sequencing reaction products were run on an ABI 3730XL DNA sequencer with a 50 cm capillary array using standard run mode.

A proprietary sequencing analysis program called 'Agent' (developed by Celera) was used to align sequencing reads to the appropriate reference sequence, and produce 'contigs' associated with each clone. The system provides estimated quality scores for all bases for which there is any variation for any of the samples. The sequencing report for each sample was analyzed for the presence of SNPs that were correlated in one haplotype pattern for one subset of clones and in a different haplotype pattern for the remaining clones. A reference file for the region of interest was prepared by listing the known variations for that region publicly available from NCBI dbSNP. A genotype file for the region of interest was created by searching each subject's haplotype report for all variations between the known reference sequence and the consensus haplotype sequences.

The magnitude of the length-reading error for the poly-T variants (e.g., rs10524523) was estimated by examining the observed lengths from the 10 clones that were prepared for samples that had a single haplotype. For a typical sample with short poly-T length of 16, the standard deviation for the 10 clones was 0.97. For a typical sample with longer poly-T length, e.g., 27, the standard deviation was 1.58.

Phylogenetic analysis was conducted. A multiple sequence alignment of the sequences was performed using the ClustalW2 (version 2.0.10) program using default parameters. Manual adjustment of the alignments was completed using Genedoc (version 2.7.000). Phylogenetic trees were constructed using Bayesian, maximum likelihood and distance-based reconstructions. The phylogenetic tree construction software used was Paup* (version 4.0b10), ClustalX2 (neighbor-joining methods, version 2.0.10) and Mr. Bayes (version 3.1.2).

Tree-bisection and reconnection branch swapping were used in all methods. The best fitting model of sequence evolution was estimated using the Modeltest program (version 3.7) which provided estimates for the following key determinants: rate matrix, shape of the gamma distribution and proportion of invariant sites. Bootstrap analysis was performed using 1000 replicates to determine statistical support for specific tree morphology.

Haplotype networks were also constructed from the sequence data using the program TCS (version 1.21 (Clement et al. TCS: a computer program to estimate gene genealogies. *Mol Ecol* **9**, 1657-9 (2000))) to compare the phylogenetic trees to cladograms estimated using statistical parsimony. The phylogenetic trees and haplotype networks were constructed twice, with gaps treated as missing data for the first instance and as a fifth character for the second instance. Nucleotide diversity in the region of interest was calculated using DnaSP (version 5.00.02 (Librado et al. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-2 (2009))).

After construction of the phylogenetic trees, the haplotype network, and completion of the analysis of nucleotide diversity in the region of interest, the results from the different methods were compared and reconciled to a consensus tree. Groups of sequences sharing a recent disease mutation were presumed to segregate more closely on the phylogenetic tree, however, sporadic cases due to phenocopies, dominance and epistasis can introduce noise into the phenotype-haplotype relationship (Tachmazidou et al. Genetic association mapping via evolution-based clustering of haplotypes. *PLoS Genet* **3**, e111 (2007)).

However, sporadic cases due to phenocopies, dominance and epistasis can introduce noise into the phenotype-haplotype relationship. This phylogenetic analysis focused on a high-level aggregation of clades in order to minimize these effects. The clades determined at the first split in the phylogenetic tree were used to test the hypothesis that TOMM40 subject-haplotypes from clade 'B' were associated with onset of AD at a later age than subject-haplotypes from clade 'A', (each subject contributed two haplotypes to the AD age of onset association signal). The number of tests of association that are performed using this approach was orders of magnitude less than in typical genome-wide association studies since the phylogenetic analysis identified categories of evolutionarily-related subject-haplotypes. If the tests of association

confirmed that the different clades classified the subject-haplotype data by age of onset, further statistical analysis was done to identify the variants that separated the sequences into each clade. Effectively, this analysis assessed the significance of each variant as a factor that influences age of onset using a series of one-degree of freedom tests guided by the tree structure. The phylogenetic analyses were conducted using single nucleotide and insertion/deletion polymorphisms. The statistical tests of association were adjusted with a Bonferroni correction for the number of polymorphic sites included in the analysis.

Haplotype reports from the Polymorphic analysis software and reports from DnaSP software (version 5.00.02 (Librado et al. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-2 (2009))) were used for subsequent statistical analyses. We analyzed individual TOMM40 SNP variants, TOMM40 haplotypes and length of poly-T repeats for association with LOAD risk for the AS cohort and LOAD age of onset for the DS cohort. Differences in the proportions of specific TOMM40 alleles associated with each ApoE allele or ApoE genotype were compared using Fisher's exact test (two-tailed). Starting with 30 parsimony-informative sites and $\alpha = 0.05$, a Bonferroni correction for the significance of a specific allelic association would require a P value of 0.001. Odds ratios (OR) were calculated as the (number of minor alleles in cases/number of minor alleles in controls)/(number of major alleles in cases/number of major alleles in controls) and reported with 95% confidence interval. Means for defined LOAD age of onset groups were compared by *t* tests, two-tailed. A standard *F* test on group variances was performed to determine whether the *t* test was calculated assuming equal or unequal variances. Statistical analysis was completed using JMP software (version 8, SAS Institute, Cary, NC).

Accession Codes: GenBank: TOMM40, translocase of outer mitochondrial membrane 40 homolog, 10452; ApoE, apolipoprotein E, 348

The foregoing is illustrative of the present invention, and is not to be construed as limiting thereof. The invention is defined by the following claims, with equivalents of the claims to be included therein.

1. A method of determining increased risk for development of Alzheimer's disease in a subject, comprising:

(a) detecting from a biological sample containing DNA taken from said subject the presence or absence of a genetic variant of the TOMM40 gene associated with increased or decreased risk of Alzheimer's disease, wherein said variant is a deletion/insertion polymorphism in intron 6 or intron 9 of the TOMM40 gene; and

(b) determining said subject is at increased or decreased risk of Alzheimer's disease when said genetic variant is present or absent.

2. The method of claim 1, further comprising detecting whether said subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject.

3. The method of claim 1, further comprising detecting whether said subject is an Apo E3/E3 or E3/E4 subject.

4. The method of any of claims 1-3 wherein said genetic variant of the TOMM40 is a poly-T insertion variant at genomic location 50,094,889 of NCBI Build 36.3 (rs10524523).

5. A method of determining a prognosis or the risk for developing Alzheimer's disease in a patient comprising obtaining a patient profile, wherein said obtaining a patient profile comprises:

detecting the presence or absence of at least one ApoE 2, ApoE 3, or ApoE 4 allele in a biological sample of said patient, and

detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) variant_located in intron 6 or intron 9 of the TOMM40 gene associated with increased or decreased risk of Alzheimer's disease, and then

converting said patient profile into said prognosis, wherein the presence of said ApoE 2, ApoE 3 or ApoE 4 allele and the presence of said at least one TOMM40 DIP polymorphism variant identifies said patient as a patient at risk for developing Alzheimer's disease.

6. The method of claim 5, wherein said DIP is rs10524523, rs10602329 or DIP3.

7. The method of claim 5, wherein said DIP is an insertion polymorphism.

8. The method of claim 5, wherein said DIP is a poly-T deletion/insertion polymorphism.

9. The method of claim 5, wherein said DIP is rs10524523.

10. The method of any of claims 5-9, further comprising detecting whether said subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject.

11. A method for stratifying a subject into a subgroup of a clinical trial of a therapy for the treatment of Alzheimer's disease, said method comprising:

detecting the presence or absence of at least one ApoE 2, ApoE 3, or ApoE 4 allele in a biological sample of said patient,

detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) variant located in intron 6 or intron 9 of the TOMM40 gene associated with increased or decreased risk of Alzheimer's disease, and

stratifying said subject into said subgroup for said clinical trial of said therapy based upon the presence or absence of said at least one ApoE 2, ApoE 3, or ApoE 4 allele and/or TOMM40 DIP variant.

12. The method of claim 11, wherein said DIP is an insertion polymorphism.

13. The method of claim 11, wherein said DIP is a poly-T deletion/insertion polymorphism.

14. The method of claims 11, wherein said DIP is rs10524523, rs10602329 or DIP3.

15. The method of claim 11, wherein said DIP is rs10524523.

16. The method of any of claims 11-15, further comprising detecting whether said subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject.

17. A kit for determining if a subject is at increased risk of developing late onset Alzheimer's disease comprising:

(A) at least one reagent that specifically detects ApoE 3, ApoE 4, and/or ApoE 2, wherein said reagent is selected from the group consisting of antibodies that selectively bind ApoE isoforms, and oligonucleotide probes that selectively bind to DNA encoding ApoE alleles;

(B) at least one reagent that specifically detects the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) variant located in intron 6 or intron 9 of the TOMM40 gene associated with increased or decreased risk of Alzheimer's disease; and

(C) instructions for determining that the subject is at increased risk of developing late onset Alzheimer's disease by

(i) detecting the presence or absence of an ApoE 3, ApoE4, and/or ApoE 2 isoform in said subject with said at least one reagent;

(ii) detecting the presence or absence of at least one TOMM40 deletion/insertion polymorphism (DIP) variant located in intron 6 or intron 9 of the TOMM40 gene associated with increased or decreased risk of Alzheimer's disease; and

(iii) observing whether or not the subject is at increased risk of developing late onset Alzheimer's disease by observing if the presence of ApoE 3, ApoE 4, and/or ApoE 2 and said TOMM40 DIP is or is not detected with said at least one reagent, wherein the presence of ApoE 3, ApoE 4 and/or ApoE 2 isoform and said TOMM40 DIP variant indicates said subject is at increased risk of developing late onset Alzheimer's disease.

18. The kit of claim 17, wherein said at least one reagent and said instructions are packaged in a single container.

19. The kit of claim 17 or claim 18, wherein said DIP is an insertion polymorphism.

20. The kit of claim 17 or claim 18, wherein said DIP is a poly-T deletion/insertion polymorphism.

21. The kit of claim 17 or claim 18, wherein said DIP is rs10524523, rs10602329 or DIP3.

22. The kit of claim 17 or claim 18, wherein said DIP is rs10524523.

23. The kit of any of claims 17-22, wherein said determining step further comprises detecting whether said subject is an Apo E2/E2, E2/E3, E2/E4, E3/E3, E3/E4, or E4/E4 subject.