

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
21 September 2006 (21.09.2006)

PCT

(10) International Publication Number
WO 2006/099142 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/US2006/008602

(22) International Filing Date: 10 March 2006 (10.03.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/660,243 10 March 2005 (10.03.2005) US

(71) Applicants (for all designated States except US): **THE TRUSTEES OF BOSTON UNIVERSITY** [US/US]; One Sherborn Street, Boston, Massachusetts 02215 (US). **CHILDREN'S MEDICAL CENTER CORPORATION** [US/US]; 55 Shattuck Street, Boston, Massachusetts 02115 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **SEBASTIANI, Paola** [IT/US]; 160 Commonwealth Avenue, Boston, Massachusetts 02116 (US). **BALDWIN, Clinton, T.** [US/US]; 80 Easter Avenue, Arlington, Massachusetts 02476 (US). **STEINBERG, Martin, H.** [US/US]; 6 Cumston Street, Boston, Massachusetts 02118 (US). **RAMONI, Marco, F.** [IT/US]; 160 Commonwealth Avenue, Boston, Massachusetts 02116 (US).

(74) Agents: **EISENSTEIN, Ronald** et al.; Nixon Peabody, LLP, 100 Summer Street, Boston, Massachusetts 02110 (US).

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

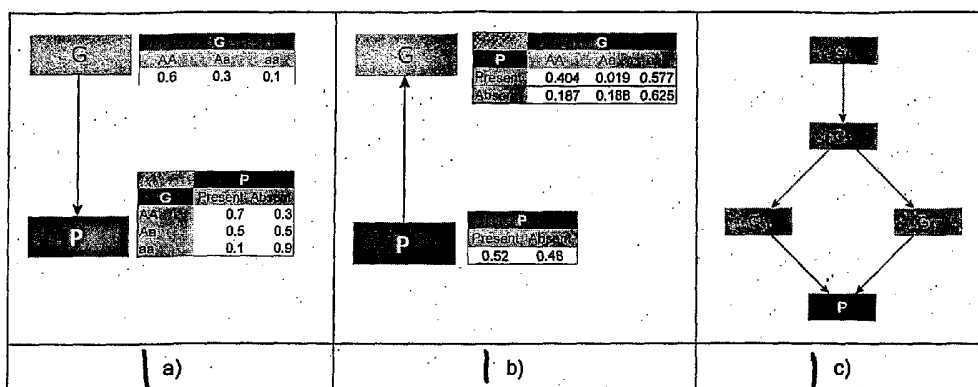
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

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

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

(54) Title: PROGNOSTIC METHOD FOR VASCULAR DISEASES



(57) Abstract: The present invention is directed to a group of genes and genetic polymorphisms that can be used to screen for individuals who are at high risk of developing a vascular disease, disorder or condition, such as cranio-vascular disease, including vascular occlusion, and overt stroke. The method comprises analysis of at least one marker in at least one of the genes of ADCY9 (NMJ)01 116; ANXA2 (NMJ)01 002858; BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NMJ 82918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243). When 5 or fewer of the genes are included in the analysis, it is preferable that one includes markers for at least SELP or BMP6 in the analysis. The invention is further directed to methods of predicting individuals at risk of specific diseases or disorders in a population using nucleic acid polymorphisms and Bayesian network analysis.

WO 2006/099142 A2

PROGNOSTIC METHOD FOR VASCULAR DISEASES

CROSS-REFERENCE TO RELATED APPLICATIONS

- [001] This application claims the benefit under 35 U.S.C. 119(e) of the U.S. provisional patent application Serial No. 60/660,243, filed March 10, 2005, the content of which is herein incorporated by reference in its entirety.

GOVERNMENT SUPPORT

- [002] This invention was made with Government Support under Contract No. ECS-0120309 awarded by the National Science Foundation and Contract No. HL68970 awarded by the National Institutes of Health. The Government has certain rights in the invention.

BACKGROUND

Field of The Invention

- [003] The present invention is directed to diagnostic and prognostic methods for vascular complications, such as stroke, using analysis of nucleotide polymorphisms in a specific group of genes with a Bayesian network (BN) analysis.

Background of the Invention

- [004] Identification of genetic predisposition and risk of individuals to so called complex traits or phenotypes is a problem. Complex traits are generally defined as phenotypes that result from interaction of numerous genes wherein the pathways these genes play a role in are also affected by environmental factors such as diet, air pollutants, toxic chemicals and the like. Also, single gene mutations affecting function on one single gene, sometimes creates a "ripple effect," which is based on a complex network of proteins and consequently affects the phenotypic expression of, for example, disease conditions. So, for example, in an individual with high blood pressure, a risk of stroke is increased, but not all individuals with high blood pressure get a stroke. Also, a single gene mutation in, for example, a fibrillin gene, which causes Marfan syndrome, results in various clinical outcomes, subjecting some of the individuals having such mutation to much more severe cardiovascular problems than others. It would therefore be of great clinical benefit to be able to enhance our

accuracy in predicting these “ripple effects” based on an individual’s genetic background.

- [005] Most of the phenotypes, such as length of an individual, intelligence of an individual, blood pressure in an individual, cholesterol levels in an individual, and occurrence of vascular events such as heart disease, or vascular occlusions, such as stroke, are determined by interaction of several genes that encode proteins and their regulatory sequences that differ slightly in their structure and/or function which consequently affects the metabolic network where the proteins encoded by these genes play a part. Consequently, genetic analysis of complex traits has been a challenge.
- [006] Unlike classical Mendelian disorders (mistakes in one gene result in one disease), the genetic component of complex traits is not simply attributable to single causative genes making it difficult to study by standard genetic and molecular biological approaches. Recent advances in the knowledge of the human genome, coupled with the development of technologies for large scale analysis of gene activity via DNA microarrays, now affords the opportunity to identify genes whose expression is involved in a resultant phenotype. Consequently, researchers have tried to use various statistical methods to infer associations between these genes or so called quantitative trait loci (QTL) and complex traits with varying success.
- [007] The sequencing of the entire human genome promises to transform the study of human health by providing an opportunity to develop genomic knowledge that will eventually boost prevention, diagnosis and treatment of disease. Genome research in the post-sequencing era is now faced with massive, multi-disciplinary challenges in order to realize this promise. Most complex illnesses result (i) from the combined action of gene variants that can be considered “normal”, as they do not destroy the function of the gene that they modify; (ii) from factors provided by the environment, and (iii) from a stochastic component that can be best defined as “chance”. The ensemble of genetic modifiers that enhance the impact of environmental factors on health represents the genetic susceptibility to ailments.
- [008] Traditionally, complex trait related genetic data have been analyzed using standard, and widely endorsed “frequentist” statistical approach to inference, wherein the results are presented in “P-values.” This type of analysis uses only so called deductive probabilities, which calculate, under certain assumptions, all possible

outcomes in the experiment were repeated many times, that can be described, for example, in the form of a bell-shaped curve. The P-value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis, which is assumed true), of obtaining a result equal to or more extreme than what was actually observed (S.N. Goodman, *Ann Intern. Med.* 1999; 130:995-1004). Therefore, by its definition, calculation of a P-value does not take into account any information that can contribute to the outcome outside the experiment (S.N. Goodman, *Ann Intern. Med.* 1999; 130:1005-1013. This presents a serious problem in analysis of complex traits, because the trait or phenotype is often affected by more variables, including genetic variables, that have or can actually be analyzed or that are even known.

[009] Other approaches have been suggested, although researchers in the field have not thus far embraced these techniques as universally as they have accepted the P-value-based statistical methods. For example, U.S. Patent No. 5,580,728 discloses that in non-X-linked disorders, the multiple linked markers enable phenotype determination via Bayesian analysis. This is done using conventional techniques (I. D. Young, *Introduction to Risk Calculation in Genetic Counselling*. Oxford: Oxford University Press, 1991) or rule-based analysis (D. K. Pathak and M. W. Perlin, "Automatic Computation of Genetic Risk," in *Proceedings of the Tenth Conference on Artificial Intelligence for Applications*, San Antonio, Tex., 1994, pp. 164-170).

[0010] U.S. Patent Application No. 20030224383 also discloses methods of determining whether a gene expression pattern is correlated with a disease phenotype using a Bayesian analysis. This application provides genes whose expression is correlated with and determinant of an atherosclerotic phenotype that were identified using the Bayesian analysis.

[0011] Vascular diseases, particularly craniovascular disease, including vascular occlusions, such as overt stroke, are all affected by a network of metabolic pathways and environmental factors that affect the expression of genes regulating these pathways in varying degree. These diseases also affect a large population of individuals. It has also been observed, that some individuals who are affected with certain disease conditions are at higher risk than the general population of developing vascular diseases, such as stroke. Diseases, where the affected population is at increased risk of stroke include, for example, high blood pressure, diabetes mellitus, carotid or other artery disease, atrial fibrillation, high blood cholesterol, lupus, and

sickle cell anemia. It is also known that not all the individuals with these diseases develop vascular diseases, and in some individuals the vascular diseases manifest in less severe form than an overt stroke.

[0012] The potential to detect these individuals at higher risk among these “at risk” populations would increase efficiency of follow-up and allow earlier intervention. It would be useful to develop screening, diagnostic and prognostic methods for sorting out individuals who are at high risk of developing a disease with a complex genetic background. Such sorting would allow preventive interventions.

[0013] For example, sickle cell anemia (SCA) is a paradigmatic single gene disorder caused by homozygosity for a unique mutation on the β globin locus. Phenotypically, SCA is a complex disease with different clinical courses, ranging from early childhood mortality to a virtually unrecognized condition. Overt stroke is a severe complication affecting 6—8% of SCA patients. It has been conjectured that modifier genes interact to determine the susceptibility to stroke, but they remain unknown.

[0014] Stroke is a major vascular complication of SCA, more frequent in subjects under the age of 20. Although recovery may be complete, stroke can cause permanent brain damage, paralysis, loss of various bodily functions and even death.

[0015] Traditional, and most used method of attempting to predict stroke include Trans-Cranial Doppler (TCD) flow studies which can predict the likelihood of stroke in children with SCA, but only 10% of individuals with abnormal TCD values will have stroke in the year following the study, and stroke will occur in approximately 19% individuals with normal TCD1.

[0016] More accurate prognostic methods are therefore needed to enable targeting prophylactic treatments, such as transfusions (2) or hydroxyurea, to individuals at highest risk (3,4).

SUMMARY OF THE INVENTION

[0017] The present invention is directed to a group of genes and genetic polymorphisms that can be used to screen for individuals who are at high risk of developing a vascular condition, such as cranial vascular disease, including vascular occlusion, and overt stroke. The invention is further directed to methods of predicting individuals at risk of specific diseases or disorders in a population using nucleic acid polymorphisms and Bayesian network analysis.

[0018] The invention can be used for prognosis of stroke in any population with at least 30%, more preferably, at least 30-40%, still more preferably at least 40-50%, still more preferably at least about 50-60%, still more preferably at least about 60-70%, still more preferably at least about 70-80%, still more preferably at least about 80-90%, and still more preferably at least about 90-95%, still more preferably at least about 95-98%, and still more preferably about 100% accuracy in predicting the risk of a vascular disease, for example, craniovascular disease, such as vascular occlusion, including stroke, and overt stroke.

[0019] We have identified a group of 11 genes the analysis of at least one of which will assist in prognosis of vascular events, such as cranio vascular events, such as stroke.

[0020] Accordingly, the invention provides a method of analysis that involves use of any of a number of genes from the group. For example, use of at least one, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, or at least 11 of the genes selected from the group consisting of ADCY9, ANXA2, BMP6, CCL2, CSF2, ECE1, ERG, MET, SELP, TEK, and TGFRB3. When using a smaller group of the genes, for example one to four or five genes, it is preferable that at least one of the genes is SELP or BMP6.

[0021] In one embodiment, the invention provides a prognostic method for cranio vascular events in a sickle cell anemia (SCA) patient. The method comprises the steps of analyzing a biological sample of an individual affected with SCA for polymorphisms in at least one, at least 2, at least 3, at least 4, at least 5, still more preferably at least 6, at least 7, at least 8, at least 9, at least 10, or at least 11 genes selected from the group consisting of ADCY9, ANXA2, BMP6, CCL2, CSF2, ECE1, ERG, MET, SELP, TEK, and TGFRB3.

[0022] In one embodiment, the group of genes that is analyzed comprises at least one polymorphism in or around the SELP gene.

[0023] In another embodiment, the group of genes that is analyzed comprises at least one polymorphism in or around the BMP6 gene.

[0024] In one embodiment, the analysis of at least one, 2, 3, 4, 5, 6, 7, 8, 9, 10 or 11 genes selected from the group consisting of ADCY9, ANXA2, BMP6, CCL2, CSF2, ECE1, ERG, MET, SELP, TEK, and TGFRB3 is combined with other genes and/or transcripts in a test wherein at least one other disease susceptibility trait is screened

for. Such disease susceptibilities include but are not limited to multigene traits such as obesity or high blood pressure, and single gene disorders, such as sickle cell anemia and the like. One skilled in the art can readily combine the analysis of the present genes and their polymorphisms with a gene analysis system, such as a microarray or beads.

[0025] Based on this analysis and results in sickle cell anemia patients, genotyping of polymorphic nucleotide markers within or around the identified genes can be used alone or in combination with other genes to predict, and develop predictive kits for vascular disease in general population as well as other specific populations, such as individuals with high blood pressure, diabetes mellitus, carotid or other artery disease, atrial fibrillation, high blood cholesterol, and lupus.

[0026] Accordingly, in one embodiment, the invention is directed to a method of predicting vascular disease, including craniovascular disease, such as vascular occlusion and stroke in an individual, wherein the method comprises the steps of genotyping the individual for at least one polymorphism in a locus comprising of at least 1, 2, 3, 4, 5, 7, 8, 9, 10 or 11, from the group of genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243). The numbers in parenthesis are GenBank identification numbers for the sequences of these genes.

[0027] The method of the present invention comprises the use of at least one of the 11 identified genes to identify individuals who are at risk of developing vascular complications, such as craniovascular complications, such as vascular occlusion, such as stroke.

[0028] In one embodiment, the population to be screened for is newborns.

[0029] In another embodiment, the population to be screened for is newborns/children with SCA. It is known that SCA patients are at increased risk of, for example stroke. Children who develop stroke are at high risk of long time disability, and often the riskiest time of developing a stroke in an SCA patient is under the age of 20. Accordingly, it would be beneficial to screen at least all newborns/children who are diagnosed with SCA to establish, as soon as possible, whether the individual needs frequent monitoring and/or prophylactic intervention.

[0030] In yet another embodiment, the population to be screened is African-American newborns/children. It is known that the African American population, in general, is at higher risk of developing vascular diseases. Accordingly, adding to a prophylactic screening system the analysis of at least one of the predisposing the genes/polymorphisms of the present invention would allow prescribing preventive measures to the predisposed individuals.

[0031] The invention also provides method of preventing vascular events by diagnosing an individual at risk of a vascular event using the method of the invention, i.e. analyzing at least one polymorphisms in of at least one, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, or at least 11 genes selected from the group consisting of ADCY9, ANXA2, BMP6, CCL2, CSF2, ECE1, ERG, MET, SELP, TEK, and TGFRB3. If the individual is identified as belonging to a high risk group, a number of steps can be taken. For example, in the case of an SCA individual having being identified as being at high risk of developing a stroke, such as an over stroke, can be subjected to hyrdoxyurea treatment or transfusion therapy. Other prescribed prophylactic measures may include specific dietary changes, weight loss therapy, and other life-style changes such as quitting smoking. If the individual is identified as belonging to the low or no risk population, no such preventive measures are necessary, at least if/when the individual is asymptomatic at the time of the test.

[0032] In one preferred embodiment, the polymorphisms in these genes are selected from the group consisting of SNP Id. numbers as listed in Table 1. Other polymorphic markers can readily be selected and used by a skilled artisan based on the knowledge in the field and the description in this specification. Preferably at least one, two, three, 4, 5, 6, 7, 8, 9, or 10 polymorphisms per gene are analyzed.

[0033] In one preferred embodiment, the vascular disease risk is evaluated using genotyping of a set of genes comprising at least one of the genes and their corresponding polymorphic markers as shown in Table 2, wherein the combination of genotypes reveal risk of stroke (as shown, for example, in the column "Risk").

[0034] Accordingly, for example, an individual is at low risk of developing a vascular event, for example stroke if he/she is carrying a combination of genotypes selected from the group consisting of genotype AG in the polymorphic marker hCV26910500 or any marker in tight linkage disequilibrium with these alleles in or around ANXA2

gene locus; a genotype TT in the polymorphic marker rs267196 or any marker in tight linkage disequilibrium with these alleles in or around BMP6 gene locus; a genotype TT in the polymorphic marker rs408505 or any marker in tight linkage disequilibrium with these alleles in or around BMP6 gene locus; a genotype CT in the polymorphic marker rs3917733 rs408505 or any marker in tight linkage disequilibrium with these alleles in or around SELP gene locus; a genotype CT in the polymorphic marker rs284875 or any marker in tight linkage disequilibrium with these alleles in or around TGFRBR3 gene locus; and/or a genotype AG in the polymorphic marker rs989554 or any marker in tight linkage disequilibrium with these alleles in or around ERG gene locus.

[0035] Similarly, a genotype combination “AG, TT, CT, CC and AG” with the respective markers and gene loci as described above, is another example of a genotype that indicates low risk in developing a vascular event, for example, stroke.

[0036] For example, a genotype combination “GG, TT, CC, CT, CC and AA” with the respective markers and gene loci, is indicative of a high risk of developing a vascular event, such as stroke. Accordingly, an individual carrying the above genotype, is a good candidate for preventive measures, and close follow up.

[0037] A genotype combination of “GG, TT, CC, CC, CC, AA” with respective markers and gene loci, is also indicative of a high risk of developing a vascular event, such as stroke. Accordingly, an individual carrying the above genotype, is a good candidate for preventive measures, and close follow up.

[0038] In addition, a genotype combination of “AA, TT, CC, CC, CC, and AA”, with respective markers and gene loci, is indicative of a high risk of developing a vascular event, such as stroke. Accordingly, an individual carrying the above genotype, is a good candidate for preventive measures, and close follow up.

[0039] For example, a genotype combination of “AA, TT, CT, CC, CC, and AA”, with respective markers and gene loci, is indicative of “medium” risk. Individuals carrying this allele combination, are candidates for follow up but likely not for drastic prophylactic treatments.

[0040] In one embodiment, the invention provides a kit comprising primers for amplifying the nucleic acid sample from an individual, and for genotyping the individual for polymorphisms in a group of genes comprising at least three, preferably at least 1, 2, 3, 4, or at least 5, of the genes selected from the group consisting of

ANXA2 (hCV26910500); BMP6 (rs267196 and rs408505); ERG (rs989554); SELP (rs3917733); and TGFBR3 (rs284875), wherein the polymorphic markers to be genotyped are marked in parenthesis, and necessary buffers and containers. The kit further includes an instruction booklet wherein high and low risk genotypes are indicated, for example, in a table form, like Table 3. The kit can be directed for the use of assessing vascular disease, such as craniovascular disease, such as stroke in a general population. In one preferred embodiment, the kit includes instructions to use it in stroke risk assessment of patients affected with sickle cell anemia. In one preferred embodiment, the kit consists of detection methods, such as primers or probes or nucleic acid array, to analyze the markers as shown in Table 2.

[0041] Other genotypes associated with an increased or decreased risk of vascular complications, particularly stroke, are listed *infra*.

[0042] In one embodiment, one, 2, 3, 4, 5, 6, 7, 8, 9, 10, or all 11, are included in a kit for neonatal screening of disease prognosis/diagnosis.

[0043] In yet another embodiment, the invention is directed to a method of predicting stroke or a stroke related vascular condition in an individual affected with a disease or disorder, wherein the incidence of stroke is known to be increased the method comprising the steps of genotyping the individual for at least one polymorphism in a locus of genes comprising at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least about 10, or at least 11, of the genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243). The numbers in parenthesis are GenBank identification numbers for the sequences of these genes. Also, markers that are in tight linkage disequilibrium and are located within or flanking the above-identified genes are contemplated by the present invention. A skilled artisan can readily detect whether a marker is in linkage disequilibrium with the gene or any of the markers disclosed herein using standard genetic methods.

[0044] Other genes that are either related to a vascular condition or other disease condition may be analyzed simultaneously or in parallel with analyzing the set of these genes.

[0045] In another embodiment, the invention is directed to a method of predicting stroke and related conditions in an individual affected with sickle cell anemia, the method comprising the steps of genotyping the individual for at least one polymorphism in a gene locus comprising at least 4, 5, 6, 7, 8, 9, 10, or at least 11 of the genes, wherein the genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243). The numbers in parenthesis are GenBank identification numbers for the sequences of these genes.

[0046] In one embodiment, the present invention provides methods that allow prognosis of overt stroke in individuals affected with SCA with vastly increased predictive accuracy as compared to the conventional methods. The method of the invention can also be used to detect associations between any nucleic acid polymorphisms and a complex trait to develop methods for diagnosis and prognosis. The method comprises genotyping a nucleic acid sample from an individual affected with SCA for at least one, preferably at least two, 3, 4, 5, 6, 7, 8, 9, or at least 10 markers in a group of genes comprising at least 4, 5, 6, 7, 8, 9, 10, or at least 11 of the genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243). The numbers in parenthesis are GenBank identification numbers for the sequences of these genes. Thereby providing at least about 30%-40%, still more preferably at least about 40%-50%, still more preferably at least about 50%-60%, still more preferably at least about 60%-70%, still more preferably at least about 70%-80%, still more preferably at least about 80%-90%, still more preferably at least about 90%-95%, still more preferably at least about 96%, still more preferably at least about 97%, still more preferably at least about 98%, still more preferably at least about 99%-100% accuracy in predicting the risk of stroke in the individual.

[0047] Using Bayesian networks (BNs), we analyzed 108 single nucleotide polymorphisms (SNPs) on 39 candidate genes in 1398 SCA subjects. We found that 31 SNPs on 12 genes interact with fetal hemoglobin to modulate the risk of stroke.

This network of interactions includes three genes in the TGF- β pathway and SELP, already associated with stroke in the general population. We validated this model in a different population by predicting the occurrence of stroke in 114 subjects with 98.2% accuracy.

[0048] Accordingly, the present invention provides a method for prognosis of stroke in a patient affected with SCA, comprising analysis of at least one nucleic acid polymorphism in at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or at least 11 genes, wherein the genotypes of the individuals are analyzed using Bayesian network analysis. In one embodiment, when fewer than 5 genes are analyzed, at least one of the genes is SELP or BMP6.

[0049] In one preferred embodiment, one analyzes at least one, preferably at least two, more preferably at least three or more polymorphic markers per gene locus.

[0050] In one preferred embodiment, the invention provides a method of prognosis of overt stroke in an individual affected with sickle cell anemia comprising analysis of a nucleic acid sample from the individual, analyzing at least one, preferably at least two, 3, 4, 5, 6, 7, 8, 9, or 10 polymorphisms in a group of genes comprising at least about 5, preferably 6, 7, 8, 9, 10, or all 11 of the gene loci selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes), wherein the allelic data from the polymorphic markers are analyzed using Bayesian networks.

[0051] In one preferred embodiment, the polymorphic markers are selected from the group consisting of rs437115, rs2238432, rs2238426, rs2072338, rs2283497, hCV26910500, rs267196, rs267201, rs408505, rs449853, rs4586, rs25882, rs212528, rs212531, rs989554, rs38850, rs38859, rs2420378, rs3917733, rs3753306, rs489347, rs284875, rs2148322, rs2765888, and rs2007686.

[0052] In another embodiment, the invention provides a group of genetic polymorphisms useful in determining a risk of vaso-occlusive (or vascular occlusive) condition in an individual, the group comprising at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or at least 11 of the genes selected from the genes consisting of ADCY9 (NM_001116);

ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes).

[0053] In yet another embodiment, the invention provides a method for prognosis of developing a phenotype, which is a result of interaction of multiple genes and environmental stress factors. The method comprises the steps of identifying genes that are part of networks that affect the phenotype and identifying informative polymorphic markers in these genes or loci close to these genes. One also needs to identify a group of individuals who have the phenotype and a control group who do not have this phenotype. For example, a group of diabetic or lupus or sickle cell patients who have suffered a stroke and a group of individuals, preferably at advanced age or family history without stroke, who have not suffered a stroke. Consequently, the nucleic acid samples from these individuals are subjected to genotyping and the genotypes are analyzed using Bayesian networks to allow deduction of alleles that show association to the phenotype. Once such alleles are identified, a prognostic kit can be produced, such as for example, the kit described above.

[0054] It is to be understood that the gene groups presented in this specification can be expanded or changed as a result of analyzing other populations than the sickle cell anemia patients. Therefore, the invention is not limited to only analysis of the genes listed in the preferred groups.

BRIEF DESCRIPTION OF FIGURES

[0055] Figures 1A-1C show examples of BN structures. Figure 1A shows a simple BN with two nodes representing a SNP (G) and a phenotype (P). The probability distribution of G represents the genotype distribution in the population, while the conditional probability distribution of P describes the distribution of the phenotype given each genotype. Figure 1B shows the association between G and P can be reversed using Bayes theorem. Figure 1C shows a BN linking four SNPs (G1—G4) to a phenotype P. The phenotype is independent of the other SNPs, once we know the SNPs G3 and G4. The joint probability distribution of the network is fully specified by the 5 distributions representing the distribution of G1 (2 parameters), of G2 given

G1 (6 parameters), of G3 given G2 (6 parameters), of G4 given G2 (6 parameters), and of P given G3 and G4 (9 parameters). While the full probability distribution requires $81 \times 2 - 1 = 161$ parameters, this network requires only 29.

- [0056] Figure 2 shows the BN describing the joint association of 69 SNPs with stroke. Nodes represent SNPs or clinical factors, and the numbers after each gene name distinguishes different SNPs on the same gene. SNP are represented as nodes without an asterisk and their “rs” numbers are shown in Supplementary Table 2. Clinical variables (HbF.G: fetal hemoglobin g/dL; HbF.P: fetal hemoglobin percent; HbG: total hemoglobin concentration; THALASSEMIA: heterozygosity or homozygosity for -3.7 kb α thalassemia deletion) are represented by nodes with an asterisk. Twenty five SNPs on the genes ADCY9, ANXA2, BMP6, CCL2, CSF2, ECE1, ERG, MET, SELP, TGFBR3 and TEK are directly associated with the phenotype and have the largest independent effect on the risk of stroke. Note the association of stroke with several SNPs on ADCY9, BMP6, MET, SELP, TGFBR3, which usually reduces the possibility of false positives(18).
- [0057] Figure 3 shows box plot of the predictive probability of stroke (risk in 5 years) in an independent set of 7 stroke and 107 non-stroke subjects. The plot shows a clear split of the predictive probabilities between these two outcomes: the predictive probabilities of stroke in the seven stroke subjects are above 0.6, while the predictive probabilities of stroke in the 107 non-stroke subjects are almost 0 and for only 2 subjects the probability is above 0.5. The predictive probabilities are reported in Supplementary Table 4.
- [0058] Figure 4 shows Table 1. Table 1 shows the summary information for the genes and the clinical variable HbF directly associated with stroke. From left to right: official gene symbol, chromosomal position, SNP ID from dbSNP/Celera databases; Bayes factor of the model associating the SNP to stroke versus the model of independence. The accuracy of a single gene is the proportion of individuals whose phenotype is correctly predicted using only the SNPs in this gene. Single gene contribution is the loss of predictive accuracy when all SNPs of this gene are removed. Both single gene accuracy and contribution were measured on the independent test set of 114 patients.
- [0059] Figure 5 shows Supplementary Table 1, which shows the extensive clinical information that was available for the 1398 African Americans with SCA – 92

subjects with reported overt stroke and 1306 subjects without – enrolled in the Cooperative Study of Sickle Cell Disease(9). Values in brackets refer to the independent set of 107 non-stroke patients and 7 stroke patients, and double dashes denote data not available. HbF.G:HbF g/dL; HbF.P:HbF percent; HbG:total hemoglobin concentration; α thalassemia; heterozygosity or homozygosity for 3.7 kb α thalassemia deletion. Age refers to the beginning of the study, and the age of the stroke patients in the independent population is the age of the event.

[0060] Figure 6 shows Table 2. Table 2 shows risk of stroke in 5 years, and 95% credible intervals (within brackets) given particular genotypes of the SNPs in some genes directly associated to stroke in the network in Figure 2. We used an exact probabilistic algorithm(15) to compute the odds for stroke predicted by the network in Figure 2, given the genotypes in the table. The last column (N) reports the frequency of subjects for each genotype.

[0061] Figure 7 shows Supplementary Table 2 shows the correspondence tables between the single nucleotide polymorphisms (SNP) of the network in Figure 2 and their RS number. SNPs in the network are named with the name of their gene plus a progressive index number, if necessary. The last column reports the chromosome where the gene is located.

[0062] Figure 8 shows Supplementary Table 3 which shows the conditional probability tables quantifying the network are estimated from the data in Figure 2. These network parameters were estimated from the data obtained from the original cohort study, as described in the Methods section of the Example. Each table contains the conditional probability distributions of each node in the network. For example, the first table reports the conditional probability distributions over the genotypes of the single nucleotide polymorphism (SNP) TGFNR3.9 given its parents in the network TGFNR3.8, TGFNR3.10, and TGFNR3.2.

[0063] Figure 9 shows the Supplementary Table 4 which shows the results of the predictive validation. Observed values, prediction and posterior probabilities of stroke for each patient with (1) or with (0) stroke in the independent set of 114 subjects. These data are plotted in the box plots in Figure 3.

[0064] Figure 10 shows Supplementary Figure 1, which shows how the predictive accuracy decreased using the logistic regression model using stepwise regression. Predictive validation of this regression model on the same independent set produced

10 errors in the non-stroke subjects (false positive rate = 0.09) and 3 errors in the stroke subjects (false negative rate = 0.43), with an overall accuracy of 88%. Figure 10 shows a box plot of the predictive probability of stroke (risk in 5 years) in an independent set of 7 stroke patients and 107 non-stroke patients obtained through logistic regression. Compared to the predictive probabilities obtained by the BN model plotted in Figure 3, the greater overlapping of the two distributions shows a significantly decreased discriminatory power of the logistic regression model, consistent with its lower predictive accuracy.

[0065] Figure 11, shows Supplementary Table 5, which shows the summary description of the logistic regression model. Estimates, standard errors and p-values of the genotype effects that were found significantly associated with the stroke using logistic regression. HbF.G describes the effect of fetal hemoglobin (g/dL). The p-values were computed using asymptotic normal theory. Box plots of the predictive probabilities are shown in Supplementary Figure 1 (see, Figure 10).

DETAILED DESCRIPTION OF THE INVENTION

[0066] The present invention provides methods for prognosis for vascular disease risk. The invention further provides a group of genes the genotype of which can be used to assess risk of vascular diseases in a population, when at least one, preferably at least two, 3, 4, 5, 6, 7, 8, 9, or 10 polymorphic markers in a group comprising at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least about 8, at least 9, at least about 10, or at least 11 genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes).

[0067] The gene or gene groups the analysis of which is useful for the prognostic/diagnostic analysis comprises at least the groups of 2, 3, 4, 5, 6, 7, 8, 9, and 10, shown in the following Table format:

Groups of 10

1	2	3	4	5	6	7	8	9	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 10

12	13	14	15	16	17	18	19	20	21	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2

Groups of 10

12	13	14	15	16	17	18	19	20	21	22
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 9

1	2	3	4	5	6	7	8	9	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 9

12	13	14	15	16	17	18	19	20	21	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 9

23	24	25	26	27	28	29	30	31	32	33
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 9

23	24	25	26	27	28	29	30	31	32	33
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 8

1	2	3	4	5	6	7	8	9	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 8

12	13	14	15	16	17	18	19	20	21	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2

Groups of 8

12	13	14	15	16	17	18	19	20	21	22
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 8

23	24	25	26	27	28	29	30	31	32	33
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 8

34	35	36	37	38	39	40	41	42	43	44
----	----	----	----	----	----	----	----	----	----	----

Groups of 8

34	35	36	37	38	39	40	41	42	43	44
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 7

1	2	3	4	5	6	7	8	9	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 7

12	13	14	15	16	17	18	19	20	21	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 7

23	24	25	26	27	28	29	30	31	32	33
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 7

34	35	36	37	38	39	40	41	42	43	44
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 7

45	46	47	48	49	50	51	52	53	54	55
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 6

1	2	3	4	5	6	7	8	9	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2

Groups of 6

12	13	14	15	16	17	18	19	20	21	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 6

23	24	25	26	27	28	29	30	31	32	33
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3

Groups of 6

23	24	25	26	27	28	29	30	31	32	33
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 6

34	35	36	37	38	39	40	41	42	43	44
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 6

45	46	47	48	49	50	51	52	53	45	55
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9

Groups of 6

45	46	47	48	49	50	51	52	53	45	55
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 6

56	57	58	59	60	61	62	63	64	65	66
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 5

1*	2*	3*	4	5*	6*	7*	8*	9*	10*	11*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2

Groups of 5

1*	2*	3*	4	5*	6*	7*	8*	9*	10*	11*
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2

Groups of 5

12*	13*	14*	15*	16	17*	18*	19*	20*	21	22*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2

Groups of 5

23*	24*	25*	26	27	28*	29*	30*	31*	32	33*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 5

34*	35*	36*	37	38	39*	40*	41*	42*	43	44*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 5

45*	46*	47*	48	49	50*	51*	52*	53*	54	55*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 5

56*	57*	58*	59	60*	61*	62*	63*	64*	65	66*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2

Groups of 5

56*	57*	58*	59	60*	61*	62*	63*	64*	65	66*
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 5

67*	68*	69*	70	71	72*	73*	74*	75*	76*	77*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 4

1*	2*	3*	4	5	6*	7*	8*	9*	10	11*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6

Groups of 4

12*	13*	14*	15	16	17	18*	19*	20*	21*	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2

Groups of 4

23*	24*	25*	26*	27	28	29*	30*	31*	32	33
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2

Groups of 4

34*	35*	36*	37	38	39	40*	41*	42*	43	44
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 4

45*	46*	47*	48	49	50	51*	52*	53*	54	55
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 4

56*	57*	58*	59	60	61*	62*	63*	64*	65	66
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 4

67*	68*	69*	70	71	72	73*	74*	75*	76	77*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 4

78*	79*	80*	81	82	83	84*	85*	86*	78*	88
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 3

1*	2*	3*	4	5	6	7*	8*	9*	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2

Groups of 3

12	13*	14*	15	16	17	18	19*	20*	21	22*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6

Groups of 3

23	24*	25*	26	27*	28	29	30*	31*	32*	33
----	-----	-----	----	-----	----	----	-----	-----	-----	----

Groups of 3

23	24*	25*	26	27*	28	29	30*	31*	32*	33
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2

Groups of 3

34	35*	36*	37*	38	39	40	41*	42*	43	44
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2

Groups of 3

45	46*	47*	48	49	50	51	52*	53*	54	55
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 3

56	57*	58*	59	60	61	62*	63*	64*	65	66
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3

Groups of 3

56	57*	58*	59	60	61	62*	63*	64*	65	66
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 3

67*	68*	69*	70	71	72*	73	74*	75*	76	77
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Group 3

78	79*	80*	81	82	83	84	85*	86*	87	88*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 3

89	90*	91*	92*	93	94	95	96*	97*	98*	99
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9

Groups of 3

89	90*	91*	92*	93	94	95	96*	97*	98*	99
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

Groups of 2

1	2*	3*	4	5	6	7	8*	9*	10	11
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9

Groups of 2

12*	13	14*	15	16	17	18*	19	20*	21	22
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2

Groups of 2

23	24	25*	26	27	28*	29	30	31*	32	33*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6

Groups of 2

34	35	36*	37	38*	39	40	41	42*	43*	44
----	----	-----	----	-----	----	----	----	-----	-----	----

Groups of 2

34	35	36*	37	38*	39	40	41	42*	43*	44
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2

Groups of 2

45	46	47*	48*	49	50	51	52	53*	54	55
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ECE1	ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2

Groups of 2

56	57	58*	59	60	61	62	63*	64*	65	66
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
ERG	MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1

Groups of 2

67	68*	69*	70	71	72	73*	74	75*	76	77
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
MET	SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG

Groups of 2

78*	79	80*	81	82	83*	84	85	86*	87	88
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
SELP	TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET

Groups of 2

89	90	91*	92	93*	94	95	96	97*	98	99*
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
TEK	TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP

Groups of 2

100	101	102*	103*	104	105	106	107	108*	109*	110
ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK	TGFRB3
TGFRB3	ADCY9	ANXA2	BMP6	CCL2	CSF2	ECE1	ERG	MET	SELP	TEK

[0068] When one analyzes groups of genes with 5 or fewer, methods of analysis using the groups marked with an asterisk (*) are preferred.

[0069] In one embodiment, the groups are selected from the following combinations of two:

[0070] BMP6 & ADCY9; BMP6 & SELP; BMP6 & TGFBR3; BMP6 & CCL2; BMP6 & CSF2; and BMP6 & ANXA2.

[0071] In one embodiment, the groups are selected from the following combinations of three:

[0072] BMP6 & ADCY9 & SELP; BMP6 & ADCY9 & TGFBR3; BMP6 & ADCY9 & CCL2; BMP6 & ADCY9 & CSF2; BMP6 & ADCY9 & ANXA2; BMP6 & SELP & TGFBR3; BMP6 & SELP & CCL2; BMP6 & SELP & CSF2; BMP6 & SELP & ANXA2; BMP6 & TGFBR3 & CCL2; BMP6 & TGFBR3 & CSF2; BMP6 & TGFBR3 & ANXA2; BMP6 & CCL2 & CSF2; BMP6 & CCL2 & ANXA2; and BMP6 & CSF2 & ANXA2.

[0073] In one embodiment, the groups are selected from the following combinations of four:

[0074] BMP6 & SELP & ADCY9 & TGFBR3; BMP6 & SELP & ADCY9 & CCL2; BMP6 & SELP & ADCY9 & CSF2; and BMP6 & SELP & ADCY9 & ANXA2.

[0075] In one embodiment, the groups are selected from the following combinations of five:

[0076] BMP6 & SELP & ADCY9 & TGFBR3 & CCL2; BMP6 & SELP & ADCY9 & TGFBR3 & CSF2; and BMP6 & SELP & ADCY9 & TGFBR3 & ANXA2.

[0077] In one embodiment, the groups are selected from the following combination of six: BMP6 & SELP & ADCY9 & TGFBR3 & CCL2 & CSF2.

[0078] The group of these risk predictive genes may form a part of a diagnostic or prognostic method or a kit.

[0079] For example, a diagnostic method or kit for sickle cell anemia may include as a sub-part a method or kit for determining the prognosis of the affected individual to have severe vascular problems such as craniovascular problems, such as stroke.

[0080] The methods of the present invention provide an immense improvement in prognostic accuracy. For example, in an individual affected with sickle cell anemia, analysis of the 11 genes for 25 polymorphic markers and their genotypes for the markers as shown in Figure 8, provides over 98% prognostic accuracy. This will

allow the high-risk patients to be closely monitored and/or subjected to preventive/prophylactic therapies.

[0081] Unlike the analysis of phenotypic expression data in the U.S. patent application publication No. 20030224383, the present invention presents the first analysis of exact detected genotypes using Bayesian networks. Also, unlike the linked markers used in the U.S. Patent No. 5,580,728 for analysis of X-linked diseases, the present invention relies genetic loci that are not genetically linked.

[0082] For example, analysis of only 5 of the genes using only 6 polymorphic markers, as shown in Table 3, results in a vastly improved prognostic test compared to the traditional TCD test, where only 10% of the individuals having abnormal TCD will have a stroke within a year, and 19% of individuals with normal TCD will develop a stroke. Moreover, analysis of groups even fewer than 5, can give superior results when compared to the traditional diagnostic methods. For example, if one analyzes a group of fewer than 5 genes, and includes either SELP and/or BMP6 into the analysis, a significant predictive power results.

[0083] Any polymorphism within or around these genes can be used in the analysis.

[0084] Particularly, one skilled in the art can easily determine a linkage of a new polymorphic marker alleles with the alleles that have been shown to be associated with low or high risk of vascular disease. Such analysis can be performed, for example, using linkage disequilibrium analysis well known to one skilled in the art. By August 2005, over 350 computer programs were available to calculate Linkage-Disequilibrium. It is well known that Linkage-Disequilibrium measures cosegregation in a population (essentially a huge pedigree). Briefly, when two alleles cosegregate they are said to be in linkage disequilibrium with each others. This happens, when the two alleles are so close to each others that essentially no recombinations occur between the two alleles. Should one identify additional alleles in or around the genes identified in this study, that are shown to be cosegregated within a population in the same allele as the one determined in this study to be associated with either high, low or medium risk for vascular disease, such as stroke, such alleles can be used in combination with or instead of the alleles specifically listed in Figure 8. One can easily determine the allele frequencies of the alleles in or around the presently identified genes in a population. A comparison of allele frequency of any of these alleles in a general population to a population with vascular

disease such as stroke, will allow one to determine all risk allele and genotype combinations in all populations, using the genes and genotypes as identified herein. From the table in Figure 8, one can see some specific genotypes associated with high, medium and low risk of vascular disease, such as stroke.

[0085] In one embodiment, the polymorphism alters the gene product or expression of the gene product. The resulting gene product may be functionally impaired. Such functional impairment may result from altered size of the gene product, for example, resulting from deletions or insertions in the gene, shortened protein half life, or shortened transcript half life.

[0086] Any genetic polymorphisms can be used in the methods of the present invention, including, but not limited to short, di-, tri-, or tetranucleotide repeats, (STR), one or more nucleic acid insertions or deletions, or single nucleic acid polymorphisms (SNPs) or combinations thereof well known to one skilled in the art. The gene variants are often in the form of single nucleotide SNPs. In one preferred embodiment, SNPs are used. SNPs represent subtle variations in a gene's coding sequence or the associated regulatory regions resulting in a mild to moderate impact on the function or concentration of the encoded protein. The inheritance of unique combinations of genetic variants can have a dominant impact that fosters the pathogenesis of any complex trait. In principle, one would like to identify all variants of all genes and assay them for their contribution towards the genesis of any given complex trait. Even if one were able to identify all variants, one would be limited by ones ability to assay and analyze such a vast number of SNPs. Accordingly, practically, one must take an approach that falls somewhere between an analysis restricted to known candidate genes identified on the basis of clinical and biological knowledge (functional candidate genes) and an investigation of the entire genomic complement of genes. See Nussbaum R L MRaWH. Genetics in Medicine. New York: W. B. Saunders Company, 2001. Science 1996; 272:689-93.

[0087] Such an approach should involve prioritization based on programmatic qualification mechanisms.

[0088] One can detect polymorphisms using any known methods well known to one skilled in the art. For example, one can use restriction enzyme analysis combined with gel electrophoresis and/or mass spectrometry, or Southern blot and other nucleic acid hybridization techniques, or immuno-PCR.

[0089] For analyzing the genotypes of the present invention, it may be appropriate to use oligonucleotides specific for alternative alleles. Such oligonucleotides which detect single nucleotide variations in target sequences may be referred to by such terms as "allele-specific oligonucleotides", "allele-specific probes", or "allele-specific primers". The design and use of allele-specific probes for analyzing polymorphisms is described in, e.g., *Mutation Detection A Practical Approach*, ed. Cotton et al. Oxford University Press, 1998; Saiki et al., *Nature* 324, 163-166 (1986); Dattagupta, EP235,726; and Saiki, WO 89/11548.

[0090] In another embodiment, a probe or primer may be designed to hybridize to a segment of target DNA such that the template nucleic acid containing the allele to be genotyped aligns with either the 5' most end or the 3' most end of the probe or primer. In a specific preferred embodiment which is particularly suitable for use in a oligonucleotide ligation assay (U.S. Pat. No. 4,988,617), the most nucleotide of the probe aligns with the position to be genotyped in the target sequence.

[0091] Oligonucleotide probes and primers may be prepared by methods well known in the art. Chemical synthetic methods include, but are limited to, the phosphotriester method described by Narang et al., 1979, *Methods in Enzymology* 68:90; the phosphodiester method described by Brown et al., 1979, *Methods in Enzymology* 68:109, the diethylphosphoamidate method described by Beaucage et al., 1981, *Tetrahedron Letters* 22:1859; and the solid support method described in U.S. Pat. No. 4,458,066.

[0092] In another embodiment of the invention, the polymorphic marker detection reagent of the invention is labeled with a fluorogenic reporter dye that emits a detectable signal. While the preferred reporter dye is a fluorescent dye, any reporter dye that can be attached to a detection reagent such as an oligonucleotide probe or primer is suitable for use in the invention. Such dyes include, but are not limited to, Acridine, AMCA, BODIPY, Cascade Blue, Cy2, Cy3, Cy5, Cy7, Dabcyl, Edans, Eosin, Erythrosin, Fluorescein, 6-Fam, Tet, Joe, Hex, Oregon Green, Rhodamine, Rhodol Green, Tamra, Rox, and Texas Red.

[0093] In yet another embodiment of the invention, the detection reagent may be further labeled with a quencher dye such as Tamra, especially when the reagent is used as a self-quenching probe such as a TaqMan (U.S. Pat. Nos. 5,210,015 and 5,538,848) or Molecular Beacon probe (U.S. Pat. Nos. 5,118,801 and 5,312,728), or

other stemless or linear beacon probe (Livak et al., 1995, PCR Method Appl. 4:357-362; Tyagi et al., 1996, Nature Biotechnology 14: 303-308; Nazarenko et al., 1997, Nucl. Acids Res. 25:2516-2521; U.S. Pat. Nos. 5,866,336 and 6,117,635).

[0094] The detection reagents used in the methods of the invention may also contain other labels, including but not limited to, biotin for streptavidin binding, hapten for antibody binding, and oligonucleotide for binding to another complementary oligonucleotide such as pairs of zipcodes.

[0095] One may use nucleic acid arrays, nucleic acids attached to beads and other solid-phase hybridization based methods to detect the polymorphisms. Such systems are well known to one skilled in the art.

[0096] The polymorphisms can also be analyzed using methods amenable for automation such as the different methods utilizing polymerase chain reaction and/or primer extension analysis. For example, primer extension analysis can be preformed using any method known to one skilled in the art including PYROSEQUENCING™ (Uppsala, Sweden); Mass Spectrometry including MALDI-TOF, or Matrix Assisted Laser Desorption Ionization – Time of Flight; genomic nucleic acid arrays (Shalon et al., Genome Research 6(7):639-45, 1996; Bernard et al., Nucleic Acids Research 24(8):1435-42, 1996); solid-phase mini-sequencing technique (U.S. Patent No. 6,013,431, Suomalainen et al. Mol. Biotechnol. Jun;15(2):123-31, 2000); ion-pair high-performance liquid chromatography (Doris et al. J. Chromatogr. A May 8;806(1):47-60, 1998); and 5' nuclease assay or real-time RT-PCR (Holland et al. Proc Natl Acad Sci USA 88: 7276–7280, 1991), or primer extension methods described in the U.S. Patent No. 6,355,433. Nucleic acids sequencing, for example using any automated sequencing system and either labeled primers or labeled terminator dideoxynucleotides can also be used to detect the polymorphisms. Systems for automated sequence analysis include, for example, Hitachi FMBIO® and Hitachi FMBIO® II Fluorescent Scanners (Hitachi Genetic Systems, Alameda, CA); Spectrumedix® SCE 9610 Fully Automated 96-Capillary Electrophoresis Genetic Analysis System (SpectruMedix LLC, State College, PA); ABI PRISM® 377 DNA Sequencer; ABI® 373 DNA Sequencer; ABI PRISM® 310 Genetic Analyzer; ABI PRISM® 3100 Genetic Analyzer; ABI PRISM® 3700 DNA Analyzer (Applied Biosystems, Headquarters, Foster City, CA); Molecular Dynamics FluorImager™ 575 and SI Fluorescent Scanners and Molecular Dynamics FluorImager™ 595 Fluorescent

Scanners (Amersham Biosciences UK Limited, Little Chalfont, Buckinghamshire, England); GenomyxSC™ DNA Sequencing System (Genomyx Corporation (Foster City, Calif.); Pharmacia ALF™ DNA Sequencer and Pharmacia ALFexpress™ (Amersham Biosciences UK Limited, Little Chalfont, Buckinghamshire, England); and immuno-PCR.

[0097] One embodiment of the present invention is directed to methods of predicting a risk of specific phenotype, wherein the phenotype is a result of a complex network of genetic interaction, in an individual using nucleic acid polymorphisms and Bayesian networks. The invention is based on a finding that using genotyping of polymorphic markers in a large number of genes known to associate with a specific phenotype. The surprising predictive power of this approach is shown in the context of sickle cell anemia patients, a subpopulation of whom is at high risk of a stroke. Having shown the feasibility of this method, the invention is generally directed to using the association of nucleic acid polymorphisms to specific phenotypic outcomes in Bayesian networks instead of the traditionally accepted p-values. The particular genes and associated polymorphisms identified in this study are useful in assessing risk of vascular diseases, such as stroke in a population.

[0098] Traditionally, genetic data has been analyzed using standard, and widely endorsed “frequentist” statistical approach to inference, wherein the results are presented in “P-values.” This analysis uses only the so called deductive probabilities, which calculate, under certain assumptions, all possible outcomes in the experiment were repeated many times (think about a bell-shaped curve). The P-value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis, which is assumed true), of obtaining a result equal to or more extreme than was actually observed (S.N. Goodman, Ann Intern. Med. 1999; 130:995-1004). Therefore, by its definition, calculation of a P-value does not take into account any information that can contribute to the outcome outside the experiment (S.N. Goodman, Ann Intern. Med. 1999; 130:1005-1013. This presents a serious problem in analysis of complex traits, because the trait or phenotype is often affected by more genetic variations that have been analyzed.

[0099] Bayesian inference is usually presented as a method for determining how scientific belief should be modified by actual observed data. Although Bayesian methodology has been one of the most active areas of statistical development in the

past 20 years, medical researchers have been reluctant to embrace what they perceive as a subjective approach to data analysis. It has been little understood that Bayesian methods have a data-based core, which can be used as a calculus of evidence. This core is the Bayes factor, also called s likelihood ratio. The minimum Bayes factor is objective and can be used in lieu of the P-values as a measure of evidential strength. Unlike P-values, Bayes factors have a sound theoretical foundation and an interpretation that allows their use in both inference and decision making. Interestingly, Bayes factors show that P-values greatly overstate the evidence against the null hypothesis. Most importantly, Bayes factors require the addition of background knowledge to be transformed into inferences, i.e. probabilities that a give conclusion is right or wrong. They make the distinction clear between experimental evidence and inferential conclusions while providing a framework in which to combine prior with current evidence. (Goodman, S.N. Ann Intern. Med. 1999; 130:1005-1013). Thus, the data obtained from genotyping combined with the known phenotype allows one to form an inference of the association between several genotypes and the phenotype.

[00100] Here we show, using a model population, the enormous power of the Bayesian network to predict a phenotypic outcome based on complex interactions of genes. Our results open a new way to analyze genomic data using Bayesian networks, which is much more powerful than the traditionally used P-values. We have shown that genotyping polymorphisms present in genes that affect a complex network of metabolic pathways in a population will reveal potent predictive sets of alleles that can be used to predict phenotypic outcomes in individuals. This method also allows detection of the “ripple effects”, i.e. the variable phenotypes that are associated with even diseases caused by single gene mutations in individuals with different genetic makeup.

[00101] This method can be used to identify predictive allelic combinations, i.e. allele combinations in several different genes, that when present in one individual, predispose that individual to a certain clinical outcome.

[00102] Therefore, in one embodiment, the invention provides a method of identifying allele sets or combinations that are present in several different genetic loci to predict development of particular phenotype. The method comprises identifying a phenotype and obtaining nucleic acid samples from the individuals having the phenotype,

selecting genes or any genetic loci that affect or are associated with various aspects of the phenotype, selecting informative polymorphisms in these genes or loci, genotyping a population of individuals with and without the phenotypic expression, and comparing the phenotypes and genotypes using the Bayesian network analysis, preferably using a “bottom-up” search strategy, known the K2 algorithm(30), which results in a combination of predictive alleles that are associated with the phenotype, and validating the identified allele combinations.

[00103] For example, to build the BN one can use a Bayesian approach that was developed by Cooper and Herskovitz(30) and is implemented in, for example, the program Bayesware Discoverer (www.bayesware.com). The program searches for the most probable network of dependency given the data. To find such a network, the Discoverer explores a space of different network models, scores each model by its posterior probability conditional on the available data, and returns the model with maximum posterior probability. This probability is computed by Bayes' theorem as

$$[00104] \quad p(M | D) \propto p(D | M)p(M),$$

[00105] where $p(D|M)$ is the probability that the observed data are generated from the network model M , and $p(M)$ is the prior probability encoding knowledge about the model M before seeing any data. We assumed that all models were equally likely a priori, so that $p(M)$ is uniform and $p(M|D)$ becomes proportional to $p(D|M)$, a quantity known as marginal likelihood. The marginal likelihood averages the likelihood functions for different parameters values and it is calculated as

$$[00106] \quad p(D | M) = \int p(D | \theta) p(\theta) d\theta$$

[00107] where $p(D|\theta)$ is the traditional likelihood function and $p(\theta)$ is the parameter prior density. For categorical data in which $p(\theta)$ follows a Dirichlet distribution, the integral $p(D | M) = \int p(D | \theta) p(\theta) d\theta$ has a closed form solution(16) that is computed in product form as: $p(D | M) = \prod_i p(D_i | M_i)$ where M_i is the model describing the dependency of the i^{th} variable on its parent nodes, and D_i are the observed data of the i^{th} variable(16). The factorization of the marginal likelihood implies that a model can be learned locally, by selecting the most probable set of parents for each variable, and then joining these local structures into a complete network, in a procedure that closely resembles standard path analysis. This modularity

property allows one to assess, locally, the strength of local associations represented by rival models. This comparison is based on the Bayes factor that measures the odds of a model M_i versus a model $\tilde{M}_i$ by the ratio of their posterior probabilities

$p(M_i | D_i) / p(\tilde{M}_i | D_i)$ or, equivalently, by the ratio of their marginal likelihoods

$\rho = p(D_i | M_i) / p(D_i | \tilde{M}_i)$. Given a fixed structure for all the other associations,

the posterior probability $p(D_i | M_i)$ is $p(D_i | M_i) = \rho / (1 + \rho)$ and a large Bayes

factor ρ implies that the probability $p(D_i | M_i)$ is close to 1, meaning that there is

very strong evidence for the associations described by the model M_i versus the

alternative model. Note that, when one explores different dependency models for the i^{th} variable, the posterior probability of each model depends on the same data.

[00108] To reduce the search space, one can use a bottom-up search strategy known as the K2 algorithm(30). The space of candidate models to be explored was specified by imposing a search order on the database variables in which older SNPs (more uniformly distributed in the population) were tested as children of more recent SNPs (asymmetrically distributed SNPs).

[00109] Preferably, when using this approach, one focuses on models in which the phenotype is the root node of the network, and the genotypes can be either conditionally dependent or marginally independent of it. As in traditional regression models, in which the phenotype is dependent on the genotypes, this inverted dependency structure can represent the association of independent as well as interacting SNPs with the phenotype(16). However, this structure is also able to capture more complex models of dependency(19) because, in this model, the marginal likelihood measuring the association of each SNP with the phenotype is functionally independent of the association of other SNPs with the phenotype. In contrast, in regression structures, the presence of an association between a SNP and the phenotype affects the marginal likelihood measuring the association between the phenotype and other SNPs, reducing the set of SNPs that can be detected as associated with the phenotype.

[00110] The BN induced by this search procedure can be quantified by the conditional probability distribution of each node given the parents nodes. The conditional

probabilities can be estimated, for example as $p(x_{ik} | \pi_{ij}) = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}}$

[00111] where x_{ik} represents the state of the child node, π_{ij} represents a combination of states of the parents nodes, n_{ijk} is the sample frequency of (x_{ik}, π_{ij}) and n_{ij} is the sample frequency of π_{ij} . The parameters α_{ijk} and $\alpha_{ij} = \sum_k \alpha_{ijk}$ encode the prior distribution with the constrain $\sum_j \alpha_{ij} = \alpha$ for all j , as suggested in(16). We chose $\alpha = 8$ in(16) by sensitivity analysis(16).

[00112] One can further validate the predictions. To assess the robustness of the network to sampling variability, one can first use, for example, a 5-fold cross validation in which the original data set is partitioned into five non overlapping subsets that are used for learning the network dependency. Each network is then used to predict the phenotypes of the individuals not included in the learning process, and the accuracy is measured by the frequency of individuals for whom the correct phenotypes is predicted with probability larger than 0.5. This model can be adopted readily to other at risk populations.

[00113] For example, the calculation of the predictive probability of stroke, given evidence in the network, was carried out using the “clique algorithm” described in(16) that is implemented in Discoverer. The predictive accuracy of the models was carried out using an independent set of 114 sickle cell anemia subjects, including 7 subjects with stroke who were not part of the multicenter study and 107 subjects randomly selected from the original database for which we had a complete medical history and were not used to build the BN model. We used the model to predict the phenotypes of these subjects and assess the predictive accuracy by the frequency of individuals for whom the correct phenotypes were predicted with probability larger than 0.5.

[00114] With the data accumulating relating to genes, their expression, proteins and their interactions with each others, one can easily select a collection of, for example, about 25-100 genes that have something to do, for example, with cardiovascular system, or fat metabolism or immune system.

[00115] For example, when analyzing individuals having other medical condition than sickle cell anemia, which medical condition is associated with higher than usual risk of vascular diseases, such as stroke, the subject gene collections may include all of, a selection of, or exactly those genes that are listed in Figure 2, or they may include additional genes that are not listed in Figure 2. Where the subject collections include

such additional genes, in certain embodiments the % number of additional genes that are present in the subject collections does not exceed about 50%, usually does not exceed about 25%. In many embodiments where additional "non-Figure 2" genes are included, a great majority of genes in the collection are vascular disease phenotype determinative genes, where by great majority is meant at least about 75%, usually at least about 80% and sometimes at least about 85, 90, 95% or higher, including embodiments where 100% of the genes in the collection are vascular disease phenotype determinative genes.

[00116] Once a set of genes have been identified using the Bayesian analysis of the genotypes in the individuals, databases, such as dbSNP database at <http://www.ncbi.nlm.nih.gov/SNP> or SNPs available from Celera Discovery System™ database can be used to identify polymorphic markers in these genes or loci.

[00117] Genotyping of individuals for these polymorphic markers is a routine analysis, and can be performed with or without amplification of the nucleic acid sample, and manually or automatically differentiating the alleles for each loci in each individual. Systems such as gel electrophoresis, mass spectrometry, hybridization of nucleic acids on nucleic acid containing solid supports, such as DNA chips, comprising polymorphic alleles, can be used to dissect the genotypes.

[00118] The genotypic and phenotypic data are consequently analyzed with a Bayesian approach. Several computer programs exist that can be used to implement this analysis. For example, www.bayesware.com provides a widely used software, Bayesware Discoverer, to analyze data. The modular representation of Bayesian network of the phenotype and any give genotype allows capturing complex dependency models - able to integrate associations between polymorphisms and phenotype, associations between polymorphisms due to linkage disequilibrium or evolutionary patterns(17), and interaction processes linking polymorphisms, phenotypes and modulating factors(18) – with a small number of parameters. The reduction in the number of parameters allows one to learn large dependency networks of genes affecting the phenotype from comparatively small data sets, and well established techniques exist to induce Bayesian networks from data in almost automated manner (16).

[00119] In one embodiment, the invention is directed to a method of predicting cerebrovascular diseases, including any vaso-occlusive events, such as stroke in an individual, wherein the method comprises the steps of genotyping the individual for at least one polymorphism in a locus of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11, where the genes comprise the group of genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes).

[00120] In one embodiment, the invention is directed to a method of predicting stroke in an individual wherein the method comprises the steps of genotyping the individual for at least one polymorphism in a locus of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or 11, of genes, wherein the genes comprise the group of genes consisting the group of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes).

[00121] In yet another embodiment, the invention is directed to a method of predicting stroke or a stroke related condition in an individual affected with a disease or disorder, wherein the incidence of stroke is known to be increased the method comprising the steps of genotyping the individual for at least one polymorphism in a locus of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or at least 11 of the genes, wherein the genes comprise the group of genes consisting the group of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes).

[00122] In another embodiment, the invention is directed to a method of predicting stroke and related conditions in an individual affected with sickle cell anemia, the method comprising the steps of genotyping the individual for at least one

polymorphism in a locus of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or at least 11 of the genes, wherein the genes comprise the group of genes consisting the group of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243) (the numbers in parenthesis are GenBank identification numbers for the sequences of these genes).

[00123] In SCA subjects, the products of modifier genes may interact to determine the likelihood of stroke and other complications. Recent studies have reported significant associations of SNPs in VCAM15, IL4R and ADRB26 with stroke in SCA subjects. Furthermore, the risk of stroke is reduced in subjects with α thalassemia(7), and increased fetal hemoglobin (HbF) levels have been associated with decreased risks for other complications(8). To identify the genetic basis of stroke in SCA, we selected 80 candidate genes involved in vaso-regulation, inflammation, cell adhesion, coagulation, hemostasis, cell proliferation, oxidative biology and other functions. On these genes, we analyzed 108 SNPs in 1398 African Americans with SCA – 92 subjects with reported overt stroke and 1306 subjects without – enrolled in the Cooperative Study of Sickle Cell Disease(9). For each subject, extensive clinical information was available (Supplementary Table 1).

[00124] The genetic dissection of a complex trait requires the ability to disentangle the web of interactions among genes, environment and phenotype(10-12). To model these relationships, we have undertaken a multivariate analysis using BNs, multivariate dependency models that account for simultaneous associations and interactions among multiple genes and their interplay with clinical and physiological factors. BNs have been already applied to the analysis of several types of genomic data — from gene expression(13) to protein-protein interactions(14) and pedigree analysis(15) — and their modular nature makes them an ideal tool for the analysis of large association studies. Furthermore, BNs can be used for prognosis: a network capturing the relationship between genotypes and phenotype can be used to compute the probability that a new individual with particular genotype will manifest the phenotype(15,16).

[00125] A BN is a directed acyclic graph in which nodes represent random variables and arcs define directed stochastic dependencies quantified by probability distributions. Figure 1 depicts three BNs. Figure 1a is a simple network describing the

dependency of a phenotypic character P on a single SNP G . The graph decomposes the joint probability distribution of the two variables into the product of the marginal distribution of G – the parent node - and the conditional distribution of P – the child node – given G . The marginal and conditional probability distributions are sufficient to define the association between P and G because their product determines the joint probability distribution. This property persists when we invert the direction of the arc in the graph (Figure 1b), and when we expand the graphical structure to include several variables (Figure 1c): the overall association is measured by the joint probability distribution that is still defined by the product of each child-parent conditional distribution. This modular nature of a BN is due to the conditional independences among the variables encoded by the directed acyclic graph(16): The graph specifies the set of parents of each node as those having an arc pointing directly to it, and each node becomes independent of its predecessors given the parent nodes. This modular representation captures complex dependency models – able to integrate associations between SNPs and phenotype, associations between SNPs due to linkage disequilibrium or evolutionary patterns(17), and interaction processes linking SNPs, phenotype and modulating factors(18) – with a small number of parameters. The reduction in the number of parameters allows us to learn large dependency networks from comparatively small data sets, and well-established techniques exist to induce BNs from data in an almost automated manner(16).

EXAMPLE

[00126] We focused on those networks that describe the dependencies of genotypes on the phenotype because, as noted in Hoh and Ott18, the analysis conditional on the phenotype reduces the complexity of the search, and can lead to discovery of larger sets of associations between SNPs and phenotype. This modeling strategy describing the diagnostic rather than prognostic associations is commonly used in data mining to build predictive models in large data sets(19). Figure 2 shows the overall dependency network we discovered linking 69 SNPs in 20 genes, HbF levels, total hemoglobin concentration (HbG), coincidence of α thalassemia to the phenotype. Thirty one SNPs on 12 genes interact with HbF to modulate the risk of stroke. Twenty five of these SNPs are directly associated with the phenotype, meaning that they have the largest independent effect on the prediction of the risk of stroke. The strength of the

dependency of each of these nodes is summarized by the odds of the model with the dependency versus the model without the dependency (Table 1). The conditional probability tables quantifying the network are estimated from the data (Supplementary Table 3).

[00127] Table 1 shows the summary information for the genes and the clinical variable HbF directly associated with stroke. From left to right: official gene symbol, chromosomal position, SNP ID from dbSNP/Celera databases; Bayes factor of the model associating the SNP to stroke versus the model of independence. The accuracy of a single gene is the proportion of individuals whose phenotype is correctly predicted using only the SNPs in this gene. Single gene contribution is the loss of predictive accuracy when all SNPs of this gene are removed. Both single gene accuracy and contribution were measured on the independent test set of 114 patients.

[00128] The network dissects the genetic basis of stroke into 11 genes whose variants have a direct effect on the disease that is modulated by HbF levels, and 9 genes whose variants are indirectly associated with stroke. An example is the cluster of 5 SNPs on the gene EDN1 that is associated with SNPs on ANXA2 and BMP6 genes. Both EDN1 and BMP6 are on chromosome 6, respectively 6p24.1 and 6p24.3, and their association points to a possible chromosomal region associated with an increased risk of stroke. A regulatory role of ANXA2 in cell surface plasmin generation was recently identified(20) and the role of EDN1 as a potent vasoconstrictor and mitogen secreted in response to hypoxia was suggested(21) and supported the hypothesis that EDN1 antagonists may be useful in the prevention and treatment of sickle vaso-occlusive crises. Our model suggests that variants of BMP6 are the strongest risk factors, while variants of EDN1 are associated with stroke via BMP6 and ANXA2 but they are not as relevant for risk prediction. Stroke is also directly associated with variants on TGFBR3, and indirectly to variants on TGFBR2, which play an essential, non-redundant role in TGF- β signaling pathway(22). BMP6 is part of the TGF- β superfamily, and the simultaneous association of three genes with functional role in the TGF- β signaling pathway suggests a role of this pathway in an increased risk of stroke. This conjecture is further supported by the association with CSF2, a protein necessary for the survival, proliferation, and differentiation of leukocyte progenitors. Several studies have already reported an association between variants in SELP and

stroke in the general population(23), and our analysis confirms its strong, although per se insufficient, prognostic role.

[00129] Table 2 shows risk of stroke in 5 years, and 95% credible intervals (within brackets) given particular genotypes of the SNPs in some genes directly associated to stroke in the network in Figure 2. We used an exact probabilistic algorithm(15) to compute the odds for stroke predicted by the network in Figure 2, given the genotypes in the table. The last column (N) reports the frequency of subjects for each genotype.

[00130] By decomposing the overall distribution into interrelated modules, the network summarizes all relevant dependencies without losing the multigenic nature of stroke. Using the network in Figure 2, we can compute the probability distribution of the phenotype, given the genotype of any of SNP and, conversely, compute the conditional distribution of any genotype, given values of other variables in the network. In this way, the model is able to describe the determinant effects of genetic variants on stroke, to predict the odds for stroke of new individuals given their genotypes, and to find the most probable combination of genetic variants leading to stroke.

[00131] Table 2 reports the risk of stroke predicted by the network in Figure 2 for some genotypes, and it shows the impossibility of predicting this risk using individual SNPs. For example, homozygosity (TT) of BMP6.10 is, by itself, associated with both negligible and very large risk, and only the simultaneous consideration of other SNPs can determine the actual risk of stroke. This situation is confirmed by the analysis of single-gene accuracy and contribution (Table 2), highlighting the small effect of individual genes. On the other hand, the 98.5% predictive accuracy reached by the model in 5-fold cross validation shows the determinant role of the simultaneous presence of all SNPs and their interplay with clinical variables for the correct prediction of stroke susceptibility.

[00132] We validated our results in a different population by predicting the occurrence of stroke in 114 subjects not included in the original study: 7 subjects with reported stroke and 107 subjects without, a proportion consistent with the phenotype distribution in the original cohort study. Our model predicted the correct outcome for all 7 stroke subjects, and for 105 of 107 non-stroke subjects, with 100% true positive rate and 98.14% true negative rate, for an overall predictive accuracy of 98.2%. Figure 3 shows the clear difference between the predictive probabilities of stroke,

underlying the fact that the predictions were not only correct but also inferred with high confidence. The subjects with reported stroke were not part of the original cohort study. The 107 non-stroke subjects were part of the original study but were not used to build the BN. For these subjects, the average follow up was 5.2 ± 2 years, with an average age at the beginning of the study 22 ± 2.6 years (Supplementary Table 1). Because the risk of stroke in SCA patients decreases rapidly after the age of 1024, these non-stroke subjects provide a reliable test set.

[00133] For comparison, we also built a logistic regression model using stepwise regression. The model captured as significant only five SNPs (located on SELP and BMP6) of the 25 SNPs identified by the BN model, and HbF (Supplementary Table 5), with a consequent decrease of predictive accuracy (Supplementary Figure 1). Predictive validation of this regression model on the same independent set produced 10 errors in the non-stroke subjects (false positive rate = 0.09) and 3 errors in the stroke subjects (false negative rate = 0.43), with an overall accuracy of 88%.

[00134] Another interesting feature highlighted by the network is that, although SNPs on the same gene tend to neatly cluster together, some dependencies extend across different genes and, sometimes, across chromosomes. These patterns seem to support the emerging view¹⁷ that physical distance between two polymorphisms is not the only arbiter of their association: their time of origin – as reflected by their distribution – and their evolutionary history shape a web of relationships far more complex than the one allowed by physical distances alone. Dependencies between SNPs across different chromosomes can also be explained by their interactions to determine other vaso-occlusive complications of SCA, such as osteonecrosis, priapism and acute chest syndrome^{2,8}. This explanation is consistent with the design of our study, which included subjects with at least one vaso-occlusive complication of SCA. For example, the web of interactions linking the SNPs on TGFBR2 – on chromosome 3 – and BMP6 – on chromosome 6 – can be explained by their simultaneous association to osteonecrosis²⁵, a known complication of SCA affecting 30% of the patient population.

[00135] Several authors have recently identified in the inadequacy of traditional analysis methods a critical impediment to the discovery of the genetic basis of complex traits in large association studies^(12,18,26,27). Our results show the promise of this approach for the discovery, representation, and prognostic use of the genetic

basis of complex traits. The predictive accuracy of our model is also a step toward the development of accurate prognostic tests to identify subjects at risk of stroke and to help the selection of better treatment options. Understanding the genetic networks modulating the likelihood of stroke may provide additional insights into the pathogenesis of the disease and suggest novel therapeutic targets. Although further investigation is required to establish the causative role of these genetic markers, our results support the emerging hypothesis that stroke in SCA patients is a complex trait caused by the interaction of multiple genes(6). The presence among the risk factors of genes already associated with stroke, such as SELP, suggests that some genetic factors predisposing to stroke are shared by both SCA patients and stroke victims in the general population, and that our model may offer some insights into the genetic basis of the third leading cause of death in the United States.

[00136] Methods

[00137] Data Collection Between Oct. 1978 and Sept. 1988, the first phase of the Cooperative Study of Sickle Cell Disease (CSSCD) enrolled 4,082 African American patients from 23 clinical centers across the United States⁹. Newborns and all patients who had visited a participating clinic for any medical reason between 1975 and 1978 were eligible for participation. Except for newborns who were enrolled throughout the study period, enrollment was closed in May 1981. Subjects were observed for an average of 5.2 ± 2 years. Five years after the start of the CSSCD, blood samples were obtained for the determination of α thalassemia and the β -globin gene cluster haplotype. DNA from this sample was also deposited in an NIH-controlled repository and this DNA was used for SNP genotyping. We limited our genotyping to samples from SCA subjects (homozygosity for the HbS gene) with or without coincident α thalassemia, who presented at least one vaso-occlusive complication of SCA. The CSSCD database provided the clinical information about the sub-phenotypes of SCA including overt stroke, osteonecrosis, acute chest syndrome, painful episodes, leg ulceration, renal failure, priapism and proliferative retinopathy. Information about other clinical features – including fetal hemoglobin (HbF) levels, systolic and diastolic blood pressure and gender – were also collected. The DNA samples were used to genotype 235 SNPs in 80 candidate genes.

- [00138] Strokes were classified by the investigator at each center based on the available clinical and imaging studies. For our analysis, we considered only subjects with a confirmed history of or incident complete non-hemorrhagic stroke, documented by imaging studies. Ninety-five percent of the subjects classified as having infarctive stroke underwent computer tomography (CT) scan, brain scan, and/or magnetic resonance imaging (MRI) at the time of the event. MRI information was not collected before December 1986. A detailed description of stroke in subjects followed in the CSSCD was reported previously²⁸. The study was reviewed and approved by the Institutional Review Board of Boston University School of Medicine. Samples from stroke subjects of the independent validation set were obtained from the Medical College of Georgia, with the approval of its Institutional Review Board.
- [00139] Genotyping DNA samples were used for SNP genotyping by high-throughput mass spectrometry. SNPs with population frequency information and heterozygosity values greater than 0.2 in the candidate genes were selected from dbSNP (<http://www.ncbi.nlm.nih.gov/SNP>) and the Celera database. The amplification primers used in all reactions were compared to the SNP database to ensure that there were no hidden SNPs in the amplification priming site that would result in inaccurate genotyping. The Sequenom Mass Spectrometry HME assay was used for genotyping⁽²⁹⁾. The detection method consisted of PCR amplification of the region containing the SNP, treatment with shrimp alkaline phosphatase, hybridization with a primer upstream of the polymorphism, and extension of the primer with a combination of a normal and a di-deoxy dNTP that corresponds to the SNP. Salt was then removed from the sample and the primer extension products analyzed on a Bruker Biflex II Mass Spectrometer. For assay design, the 5' amplification primer was tagged with the sequence ACTTAGGTTTTCCCAGTCACGAC (SEQ ID NO:1) and the 3' amplification primer tagged with AGCGGATAACAATTTCACACAGG (SEQ ID NO: 2). The T_m of the unique portion of the amplification primers ranged from 560-580C and the product size ranged from 80-150 bp. The mass of the detecting primer was from 4,000 to 8,000 Da. Multiplex groups of 5-8 SNPs, with similar sequence context, were assembled into single reactions for analysis. Information about phenotypes, clinical features and genotypes were assembled into a large database and a unique ID was assigned to each sample to anonymize the data.

[00140] Statistical Analysis Of the 235 SNPs genotyped in CSSCD subjects, 116 did not enter the statistical analysis either because monomorphic or because of failure of the primers. Of the remaining 118 SNPs, we included in the analysis the SNPs with less than 30% missing genotypes satisfying Hardy Weinberg equilibrium in the non-stroke subjects, for a total of 108 SNPs in 39 candidate genes. Continuous variables were discretized into four bins with equal frequencies.

[00141] To build the BN we used a popular Bayesian approach that was developed by Cooper and Herskovitz(30) and is implemented in the program Bayesware Discoverer (www.bayesware.com). The program searches for the most probable network of dependency given the data. To find such a network, the Discoverer explores a space of different network models, scores each model by its posterior probability conditional on the available data, and returns the model with maximum posterior probability. This probability is computed by Bayes' theorem as

$$[00142] \quad p(M | D) \propto p(D | M)p(M),$$

[00143] where $p(D|M)$ is the probability that the observed data are generated from the network model M , and $p(M)$ is the prior probability encoding knowledge about the model M before seeing any data. We assumed that all models were equally likely a priori, so that $p(M)$ is uniform and $p(M|D)$ becomes proportional to $p(D|M)$, a quantity known as marginal likelihood. The marginal likelihood averages the likelihood functions for different parameters values and it is calculated as

$$[00144] \quad p(D | M) = \int p(D | \theta)p(\theta)d\theta$$

[00145] where $p(D|\theta)$ is the traditional likelihood function and $p(\theta)$ is the parameter prior density. For categorical data in which $p(\theta)$ follows a Dirichlet distribution, the integral $p(D | M) = \int p(D | \theta)p(\theta)d\theta$ has a closed form solution¹⁶ that is computed in product form as: $p(D | M) = \prod_i p(D_i | M_i)$ where M_i is the model describing the dependency of the i^{th} variable on its parent nodes, and D_i are the observed data of the i^{th} variable(16). The factorization of the marginal likelihood implies that a model can be learned locally, by selecting the most probable set of parents for each variable, and then joining these local structures into a complete network, in a procedure that closely resembles standard path analysis. This modularity property allows us to assess, locally, the strength of local associations represented by rival models. This

comparison is based on the Bayes factor that measures the odds of a model M_i versus a model $\tilde{M}_i$ by the ratio of their posterior probabilities $p(M_i | D_i) / p(\tilde{M}_i | D_i)$ or, equivalently, by the ratio of their marginal likelihoods $\rho = p(D_i | M_i) / p(D_i | \tilde{M}_i)$. Given a fixed structure for all the other associations, the posterior probability $p(D_i | M_i)$ is $p(D_i | M_i) = \rho / (1 + \rho)$ and a large Bayes factor ρ implies that the probability $p(D_i | M_i)$ is close to 1, meaning that there is very strong evidence for the associations described by the model M_i versus the alternative model. Note that, when we explore different dependency models for the i^{th} variable, the posterior probability of each model depends on the same data.

[00146] To reduce the search space, we used a bottom-up search strategy known as the K2 algorithm(30). The space of candidate models to be explored was specified by imposing a search order on the database variables in which older SNPs (more uniformly distributed in the population) were tested as children of more recent SNPs (asymmetrically distributed SNPs). Simulations results we have carried out suggest that this heuristic leads to better networks with largest marginal likelihood. We focused on models in which the phenotype is the root node of the network, and the genotypes can be either conditionally dependent or marginally independent of it. As in traditional regression models, in which the phenotype is dependent on the genotypes, this inverted dependency structure can represent the association of independent as well as interacting SNPs with the phenotype(16). However, this structure is also able to capture more complex models of dependency(19) because, in this model, the marginal likelihood measuring the association of each SNP with the phenotype is functionally independent of the association of other SNPs with the phenotype. In contrast, in regression structures, the presence of an association between a SNP and the phenotype affects the marginal likelihood measuring the association between the phenotype and other SNPs, reducing the set of SNPs that can be detected as associated with the phenotype.

[00147] The BN induced by this search procedure was quantified by the conditional probability distribution of each node given the parents nodes. The conditional

probabilities were estimated as
$$p(x_{ik} | \pi_{ij}) = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}}$$

[00148] where x_{ik} represents the state of the child node, π_{ij} represents a combination of states of the parents nodes, n_{ijk} is the sample frequency of (x_{ik}, π_{ij}) and n_{ij} is the sample frequency of π_{ij} . The parameters α_{ijk} and $\alpha_{ij} = \sum_k \alpha_{ijk}$ encode the prior distribution with the constrain $\sum_j \alpha_{ij} = \alpha$ for all j , as suggested in(16). We chose $\alpha = 8$ in(16) by sensitivity analysis(16).

[00149] Predictive Validation To assess the robustness of the network to sampling variability, we first used 5-fold cross validation in which the original data set was partitioned into five non overlapping subsets that were used for learning the network dependency. Each network was then used to predict the phenotypes of the individuals not included in the learning process, and the accuracy was measured by the frequency of individuals for whom the correct phenotypes was predicted with probability larger than 0.5. The calculation of the predictive probability of stroke, given evidence in the network, was carried out using the “clique algorithm” described in(16) that is implemented in Discoverer. The predictive accuracy of the models was carried out using an independent set of 114 sickle cell anemia subjects, including 7 subjects with stroke who were not part of the multicenter study and 107 subjects randomly selected from the original database for which we had a complete medical history and were not used to build the BN model. We used the model to predict the phenotypes of these subjects and assess the predictive accuracy by the frequency of individuals for whom the correct phenotypes were predicted with probability larger than 0.5.

[00150] Logistic Regression We built a logistic regression model to identify the SNPs associated with the phenotype using the stepwise procedure implemented in the R program, which uses the Akaike information criterion in the forward step. We then selected the significant regressors (p-value < 0.05).

REFERENCES

[00151] The references cited herein and throughout the specification and herein incorporated by reference in their entirety.

1. Adams, R.J. et al. Stroke and conversion to high risk in children screened with transcranial Doppler ultrasound during the STOP study. Blood 103, 3689-94 (2004).

2. Steinberg, M.H., Forget, B.G., Higgs, D.R. & Nagel, R.L. Disorders of Hemoglobin: Genetics, Pathophysiology, and Clinical Management, (Cambridge University Press, Cambridge, 2001).
3. Ware, R.E., Zimmerman, S.A. & Schultz, W.H. Hydroxyurea as an alternative to blood transfusions for the prevention of recurrent stroke in children with sickle cell disease. *Blood* 94, 3022-6 (1999).
4. Adams, R.J. et al. Prevention of a first stroke by transfusions in children with sickle cell anemia and abnormal results on transcranial Doppler ultrasonography. *N Engl J Med* 339, 5-11 (1998).
5. Taylor, J.G.t. et al. Variants in the VCAM1 gene and risk for symptomatic stroke in sickle cell disease. *Blood* 100, 4303-9 (2002).
6. Hoppe, C. et al. Gene interactions and stroke risk in children with sickle cell anemia. *Blood* 103, 2391-6 (2004).
7. Adams, R.J. et al. Alpha thalassemia and stroke risk in sickle cell anemia. *Am J Hematol* 45, 279-82 (1994).
8. Platt, O.S. et al. Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N Engl J Med* 330, 1639-44 (1994).
9. Gaston, M. et al. Recruitment in the Cooperative Study of Sickle Cell Disease (CSSCD). *Control Clin Trials* 8, 131S-140S (1987).
10. Gabriel, S.B. et al. Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nat Genet* 31, 89-93 (2002).
11. Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. A vision for the future of genomics research. *Nature* 422, 835-47 (2003).
12. Carlson, C.S., Eberle, M.A., Kruglyak, L. & Nickerson, D.A. Mapping complex disease loci in whole-genome association studies. *Nature* 429, 446-52 (2004).
13. Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science* 303, 799-805 (2004).
14. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-453 (2003).
15. Lauritzen, S.L. & Sheehan, N.A. Graphical models for genetic analysis. *Statist Sci* 18, 489-514 (2004).

16. Cowell, R.G., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. Probabilistic Networks and Expert Systems., (Springer Verlag, New York, 1999).
17. Chakravarti, A. Population genetics--making sense out of sequence. *Nat Genet* 21, 56-60 (1999).
18. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet* 4, 701-9 (2003).
19. Hand, D.J., Mannila, H. & Smyth, P. Principles of Data Mining, (MIT Press, Cambridge, MA, 2001).
20. Ling, Q. et al. Annexin II regulates fibrin homeostasis and neoangiogenesis in vivo. *J Clin Invest* 113, 38-48 (2004).
21. Angerio, A.D. & Lee, N.D. Sick cell crisis and endothelin antagonists. *Crit Care Nurs Q* 26, 225-9 (2003).
22. Brown, C.B., Boyer, A.S., Runyan, R.B. & Barnett, J.V. Requirement of type III TGF-beta receptor for endocardial cell transformation in the heart. *Science* 283, 2080-2 (1999).
23. Zee, R.Y. et al. Polymorphism in the P-selectin and interleukin-4 genes as determinants of stroke: a population-based, prospective genetic analysis. *Hum Mol Genet* 13, 389-96 (2004).
24. Alexander, N., Higgs, D., Dover, G. & Serjeant, G.R. Are there clinical phenotypes of homozygous sickle cell disease? *Br J Haematol* 126, 606-11 (2004).
25. Steinberg, M.H. et al. Association of polymorphisms in genes of the transforming growth factor-beta pathway with sickle cell osteonecrosis. *Blood* 102, 262A-263A (2003).
26. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 33 Suppl, 228-37 (2003).
27. Beaumont, M.A. & Rannala, B. The Bayesian revolution in genetics. *Nat Rev Genet* 5, 251-261 (2004).
28. Ohene-Frempong, K. et al. Cerebrovascular accidents in sickle cell disease: rates and risk factors. *Blood* 91, 288-94 (1998).

29. Chiu, N.H. et al. Mass spectrometry of single-stranded restriction fragments captured by an undigested complementary sequence. *Nucleic Acids Res* 28, E31 (2000).
30. Cooper, G.F. & Herskovitz, G.F. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn* 9, 309-347 (1992).

CLAIMS

We claim:

1. A method of predicting vascular disease in an individual, wherein the method comprises the steps of genotyping a biological sample from an individual for at least one polymorphism per gene locus, where the gene loci comprise at least 3 genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFBR3 (NM_003243), wherein when 5 or fewer genes are selected, at least SELP or BMP6 are included into the group of analyzed loci, and wherein the risk of vascular disease is increased when the individual carries at least one allele of a group of two alleles that is different from the allele that cosegregates within a control population associated with low risk for vascular disease.
2. The method of claim 1, wherein the at least one allele is present in a genotype selected from the group consisting of genotype GG or AA in the polymorphic marker hCV26910500 or any marker in tight linkage disequilibrium with these alleles in or around ANXA2 gene locus; a genotype TT in the polymorphic marker rs267196 or any marker in tight linkage disequilibrium with these alleles in or around BMP6 gene locus; a genotype CC in the polymorphic marker rs408505 or any marker in tight linkage disequilibrium with these alleles in or around BMP6 gene locus; a genotype CT or CT in the polymorphic marker rs3917733 rs408505 or any marker in tight linkage disequilibrium with these alleles in or around SELP gene locus; a genotype CC in the polymorphic marker rs284875 or any marker in tight linkage disequilibrium with these alleles in or around TGFRBR3 gene locus; and/or a genotype AA in the polymorphic marker rs989554 or any marker in tight linkage disequilibrium with these alleles in or around ERG gene locus.
3. The method of claim 1 or 2, wherein the cerebro-vascular disease is a stroke.
4. The method of claims 1, 2, or 3, wherein the individual has sickle cell anemia.
5. The method of claim 1 or 2, wherein the polymorphic markers comprise at least 5-11 markers selected from the group consisting of rs437115, rs2238432, rs2238426, rs2072338,

rs2283497, hCV26910500, rs267196, rs267201, rs408505, rs449853, rs4586, rs25882, rs212528, rs212531, rs989554, rs38850, rs38859, rs2420378, rs3917733, rs3753306, rs489347, rs284875, rs2148322, rs2765888, and rs2007686.

6. A method of selecting an individual for close medical follow-up and/or prophylactic treatment for stroke comprising the steps of genotyping a biological sample from an individual for at least one polymorphism per gene locus, where the gene loci comprise at least 3 genes selected from the group consisting of ADCY9 (NM_001116); ANXA2 (NM_001002858); BMP6 (NM_001718); CCL2 (NM_002982); CSF2 (NM_000758); ECE1 (NM_001397 or its transcription variant NM_182918); ERG (NM_004449); MET (NM_000245); SELP (NM_003005); TEK (NM_000459); and TGFRBR3 (NM_003243), wherein when 5 or fewer genes are selected, at least SELP or BMP6 are included into the group of analyzed loci, and wherein the individual is selected for for close medical follow-up and/or prophylactic treatment of stroke if the individual carries at least one allele of a group of two alleles that is different from the allele that cosegregates within a control population associated with low risk for stroke.

7. The method of claim 6, wherein at least one of the alleles is present in a genotype selected from the group consisting of genotype GG or AA in the polymorphic marker hCV26910500 or any marker in tight linkage disequilibrium with these alleles in or around ANXA2 gene locus; a genotype TT in the polymorphic marker rs267196 or any marker in tight linkage disequilibrium with these alleles in or around BMP6 gene locus; a genotype CC in the polymorphic marker rs408505 or any marker in tight linkage disequilibrium with these alleles in or around BMP6 gene locus; a genotype CT or CT in the polymorphic marker rs3917733 rs408505 or any marker in tight linkage disequilibrium with these alleles in or around SELP gene locus; a genotype CC in the polymorphic marker rs284875 or any marker in tight linkage disequilibrium with these alleles in or around TGFRBR3 gene locus; and/or a genotype AA in the polymorphic marker rs989554 or any marker in tight linkage disequilibrium with these alleles in or around ERG gene locus.

8. The method of claims 5 or 6, wherein the individual has sickle cell anemia.

9. A kit for selecting an individual with sickle cell anemia for a close medical follow-up group for development of stroke comprising primers for amplifying the nucleic acid sample from an individual, and a system for genotyping the individual for polymorphisms in a group of genes comprising at least three of the genes selected from the group consisting of ANXA2 (hCV26910500); BMP6 (rs267196 and rs408505); ERG (rs989554); SELP (rs3917733); and TGFBR3 (rs284875), and an instruction booklet wherein high and low risk genotypes are indicated in a Table 3 form and instructions on the individuals based on the genotype combinations that need to be closely followed or prophylactically treated for stroke.

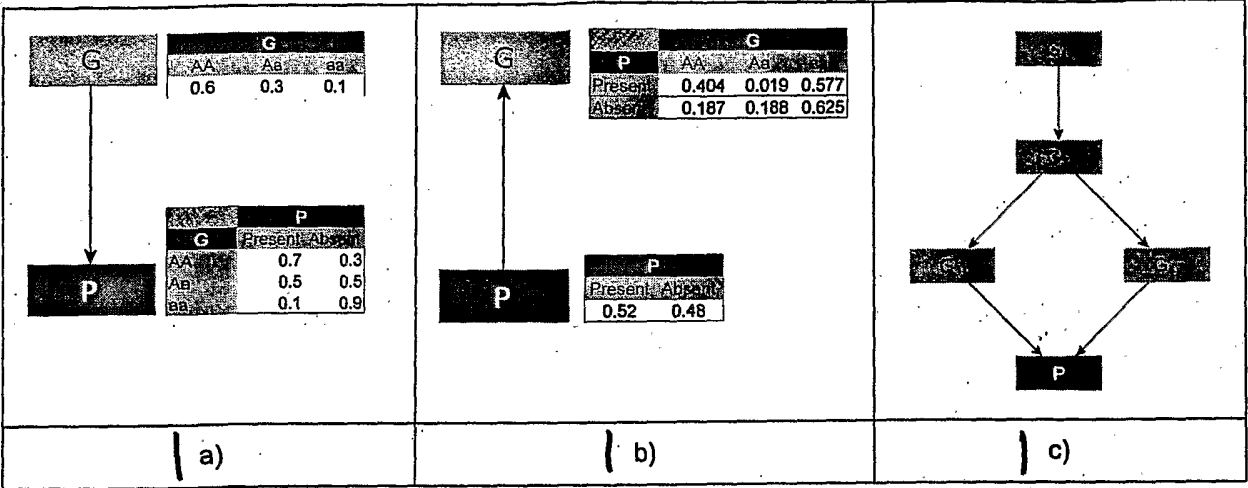


Figure 1A-1C

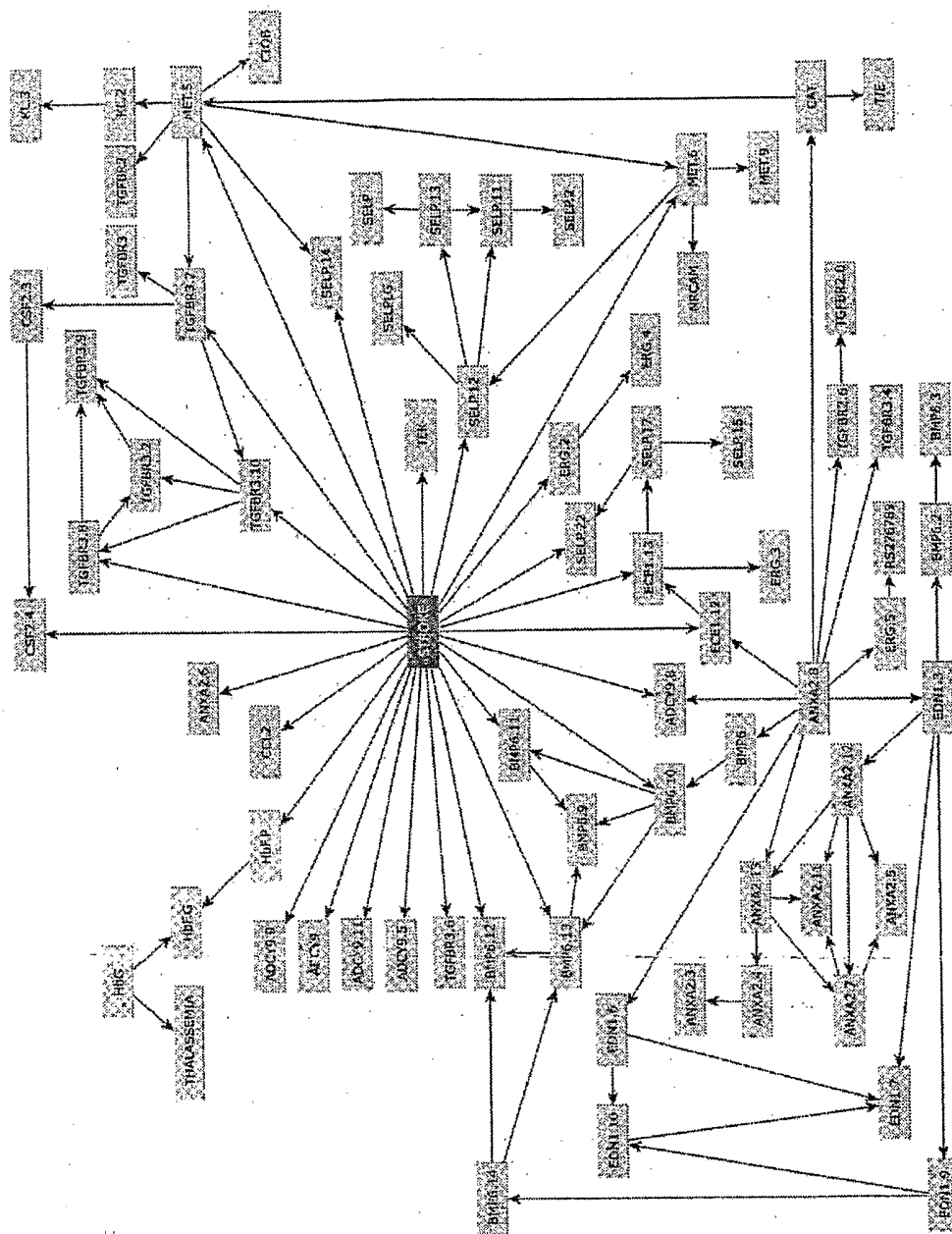


Figure 2:

FIGURE 2

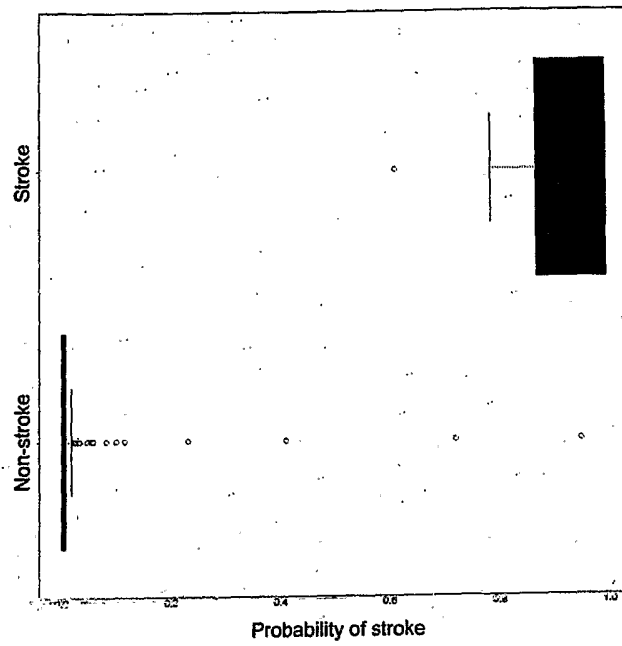


Figure 3

Gene Symbol	Position	SNP Id	Bayes Factor	Single Gene	
				Accuracy	Contribution
ADCY9	16p13.3	rs437115	3	71.93%	2%
		rs2238432	98		
		rs2238426	3381		
		rs2072338	638		
		rs2283497	10		
ANXA2	15q22.2	hCV26910500	1.68E+08	43.86%	2%
BMP6	6p24.3	rs267196	2.31E+16	83.33%	5%
		rs267201	1.92E+103		
		rs408505	4.06E+101		
		rs449853	2.20E+57		
CCL2	17q11.2	rs4586	844	55.14%	1%
CSF2	5q23.3	rs25882	1.19E+198	50.88%	1%
ECE1	1p36.12	rs212528	1.55E+04	13.15%	0.20%
		rs212531	2.34E+80		
ERG	21q22.2	rs989554	62	42.98%	1%
MET	7q31.2	rs38850	68	23.68%	1%
		rs38859	1.58E+39		
SELP	1q24.2	rs2420378	1.90E+10	80.70%	7%
		rs3917733	2.84E+34		
		rs3753306	2.32E+65		
TEK	9p21.2	rs489347	2	8%	1%
TGFB3	1p22.1	rs284875	443992	50.88%	2%
		rs2148322	68988		
		rs2765888	41968		
		rs2007686	1739		
HbF.P			482	72.81%	1%

Table 1

Figure 4

	Stroke			Non-stroke		
	Percent	Mean	StE	Percent	Mean	StE
Male	46% (57%)			44% (54%)		
α thalassemia	22% (-)			29% (29%)		
HbG		8.37 (-)	0.16 (-)		8.42 (8.38)	0.04 (0.13)
HbF.P		6.30 (1.18)	0.94 (1.25)		7.88 (9.31)	0.24 (1.99)
HbF.G		0.45 (0.95)	0.06 (0.69)		0.69 (0.79)	0.02 (0.18)
Age		22.46 (12.88)	1.44 (4.00)		21.9 (22.4)	0.34 (1.29)

Supplementary Table 1
Figure 5

Risk	ANXA2.6 hCV26910500	BMP6.10 rs267196	BMP6.12 rs408505	SELP.14 rs3917733	TGFBR3.10 rs284875	ERG.2 rs989554	N
0.007 (0;0.03)	AG	TT	TT	CT	CT	AG	1
0.06 (0;0.38)	AG	TT	TT	CT	CC	AG	4
0.185 (0.09;0.30)	AA	TT	CT	CC	CC	AA	50
0.727 (0.61;0.83)	AA	TT	CC	CC	CC	AA	64
0.868 (0.70;0.97)	GG	TT	CC	CC	CC	AA	21
0.968 (0.79;1)	GG	TT	CC	CT	CC	AA	8

Table 2

Figure 6

Name	RS Number	Chromosome
ADCY9	rs437115	16
ADCY9.11	rs2238432	
ADCY9.5	rs2238426	
ADCY9.8	rs2072338	
ADCY9.9	rs2283497	
	rs7163836	15
	rs1027899	
	hCV9036132	
	hCV1571628	
	rs8041032	
	rs931967	
	hCV26910500	
	rs1033028	
	hCV16131594	
BMP6	rs199205	6
BMP6.10	rs267196	
BMP6.11	rs267201	
BMP6.12	rs408505	
BMP6.13	rs449853	
BMP6.14	rs1225934	
BMP6.2	rs270387	
BMP6.3	rs270393	
BMP6.9	rs267192	
CAT	rs1001179	11
CCL2	rs4586	17
CIQB	rs291988	1
CSF2.3	rs25881	5
CSF2.4	rs25882	
	rs212528	1
	rs212531	
EDN1.10	rs5370	
EDN1.3	rs2070699	
EDN1.6	rs5369	
EDN1.7	hCV3107155	
EDN1.9	rs1630736	
ERG.2	rs989554	21
ERG.3	rs2836354	
ERG.4	rs2252210	
ERG.5	rs2836430	

Supplementary Table 2

Figure 7 (Page 1 of 2)

KL2	rs560014	13
KL3	rs650439	
MET5	rs38850	7
MET6	rs38859	
MET9	rs42383	
NRCAM	rs401433	7
rs276789	rs276789	16
SELP	rs6128	1
SELP11	rs3917803	
SELP12	rs2420378	
SELP13	rs3917771	
SELP14	rs3917733	
SELP15	rs6131	
SELP17	rs3917691	
SELP2	rs6135	
SELP22	rs3753306	
SELPLG	rs765267	12
TEK	rs489347	9
TGFBR2	rs1019856	
TGFBR2.0	rs876687	
TGFBR2.6	rs3773658	
TGFBR3	rs1555889	1
TGFBR3.10	rs284875	
TGFBR3.2	rs2038931	
TGFBR3.4	rs901917	
TGFBR3.6	rs2148322	
TGFBR3.7	rs2765888	
TGFBR3.8	rs2007686	
TGFBR3.9	rs284874	
TIE	rs1556581	1

Figure 7 (Page 2 of 2)

FIGURE 8
Page 1 of 21

TGFB3.8	TGFB3.10	TGFB3.2	TGFB3.9		
			TT	CC	CT
TT	TT	TT	0.333	0.333	0.333
TT	TT	CC	0.333	0.333	0.333
TT	TT	CT	0.333	0.333	0.333
TT	CT	TT	0.333	0.333	0.333
TT	CT	CC	0.076	0.076	0.848
TT	CT	CT	0.333	0.333	0.333
TT	CC	TT	0.333	0.333	0.333
TT	CC	CC	0.052	0.896	0.052
TT	CC	CT	0.076	0.848	0.076
CT	TT	TT	0.333	0.333	0.333
CT	TT	CC	0.333	0.333	0.333
CT	TT	CT	0.333	0.333	0.333
CT	CT	TT	0.333	0.333	0.333
CT	CT	CC	0.042	0.042	0.916
CT	CT	CT	0.043	0.043	0.914
CT	CC	TT	0.333	0.333	0.333
CT	CC	CC	0.007	0.666	0.327
CT	CC	CT	0.013	0.171	0.816
CC	TT	TT	0.333	0.333	0.333
CC	TT	CC	0.927	0.006	0.067
CC	TT	CT	0.848	0.076	0.076
CC	CT	TT	0.076	0.076	0.848
CC	CT	CC	0.266	0.009	0.725
CC	CT	CT	0.827	0.001	0.172
CC	CC	TT	0.861	0.017	0.122
CC	CC	CC	0.171	0.354	0.475
CC	CC	CT	0.261	0.059	0.68

FIGURE 8
Page 2 of 21

SELP.11	SELP.2		
	CC	TT	CT
TT	0.105	0.655	0.24
AA	0.577	0.16	0.263
AT	0.116	0.175	0.709

TGFB3.10	TGFB3.8	TGFB3.2		
		TT	CC	CT
TT	TT	0.333	0.333	0.333
TT	CT	0.333	0.333	0.333
TT	CC	0.017	0.911	0.072
CT	TT	0.157	0.686	0.157
CT	CT	0.01	0.911	0.079
CT	CC	0.006	0.642	0.352
CC	TT	0.013	0.93	0.057
CC	CT	0.013	0.659	0.328
CC	CC	0.075	0.538	0.387

SELP.12	SELP.13	SELP.11		
		TT	AA	AT
TT	AA	0.624	0.043	0.333
TT	TT	0.105	0.836	0.059
TT	AT	0.103	0.103	0.794
AT	AA	0.648	0.145	0.207
AT	TT	0.023	0.704	0.273
AT	AT	0.06	0.105	0.835
AA	AA	0.737	0.038	0.225
AA	TT	0.18	0.461	0.359
AA	AT	0.225	0.098	0.677

SELP.13	SELP		
	AA	GG	AG

FIGURE 8
Page 3 of 21

SELP.13	SELP		
	AA	GG	AG
AA	0.36	0.176	0.464
TT	0.205	0.332	0.463
AT	0.277	0.226	0.497

STROKE	TGFB3.10	TGFB3.8		
		TT	CT	CC
0	TT	0.024	0.024	0.952
0	CT	0.006	0.12	0.874
0	CC	0.019	0.226	0.755
1	TT	0.333	0.333	0.333
1	CT	0.19	0.62	0.19
1	CC	0.026	0.113	0.861

SELP.12	SELPLG		
	AA	GG	AG
TT	0.218	0.439	0.344
AT	0.029	0.657	0.314
AA	0.029	0.712	0.259

SELP.12	SELP.13		
	AA	TT	AT
TT	0.218	0.691	0.091
AT	0.059	0.369	0.572
AA	0.319	0.168	0.513

ANXA2.12	ANXA2.13	ANXA2.7	ANXA2.11		
			AA	AG	GG
GG	AA	GG	0.333	0.333	0.333
GG	AA	GT	0.333	0.333	0.333
GG	AA	TT	0.333	0.333	0.333

FIGURE 8
Page 4 of 21

ANXA2.12	ANXA2.13	ANXA2.7	ANXA2.11		
			AA	AG	GG
GG	AG	GG	0.333	0.333	0.333
GG	AG	GT	0.333	0.333	0.333
GG	AG	TT	0.333	0.333	0.333
GG	GG	GG	0.333	0.333	0.333
GG	GG	GT	0.076	0.848	0.076
GG	GG	TT	0.006	0.067	0.927
CC	AA	GG	0.333	0.333	0.333
CC	AA	GT	0.03	0.333	0.637
CC	AA	TT	0.002	0.021	0.977
CC	AG	GG	0.175	0.333	0.492
CC	AG	GT	0.02	0.861	0.119
CC	AG	TT	0.032	0.362	0.606
CC	GG	GG	0.887	0.087	0.026
CC	GG	GT	0.307	0.613	0.08
CC	GG	TT	0.2	0.41	0.39
CG	AA	GG	0.333	0.333	0.333
CG	AA	GT	0.333	0.333	0.333
CG	AA	TT	0.333	0.333	0.333
CG	AG	GG	0.848	0.076	0.076
CG	AG	GT	0.012	0.494	0.494
CG	AG	TT	0.002	0.051	0.947
CG	GG	GG	0.043	0.043	0.914
CG	GG	GT	0.032	0.853	0.115
CG	GG	TT	0.013	0.356	0.631

BMP6.13	BMP6.10	BMP6.11	BMP6.9		
			TT	CC	CT
TT	AA	CC	0.992	0.004	0.004
TT	AA	TT	0.076	0.076	0.848
TT	AA	CT	0.03	0.03	0.94

FIGURE 8
Page 5 of 21

BMP6.13	BMP6.10	BMP6.11	BMP6.9		
			TT	CC	CT
TT	AT	CC	0.776	0.007	0.217
TT	AT	TT	0.023	0.489	0.489
TT	AT	CT	0.124	0.063	0.813
TT	TT	CC	0.883	0.077	0.04
TT	TT	TT	0.08	0.688	0.232
TT	TT	CT	0.067	0.24	0.693
CC	AA	CC	0.651	0.016	0.333
CC	AA	TT	0.076	0.848	0.076
CC	AA	CT	0.043	0.478	0.478
CC	AT	CC	0.301	0.107	0.592
CC	AT	TT	0.014	0.835	0.151
CC	AT	CT	0.014	0.127	0.859
CC	TT	CC	0.164	0.626	0.21
CC	TT	TT	0.01	0.957	0.033
CC	TT	CT	0.032	0.658	0.31
CT	AA	CC	0.478	0.004	0.518
CT	AA	TT	0.03	0.637	0.333
CT	AA	CT	0.207	0.019	0.774
CT	AT	CC	0.238	0.107	0.655
CT	AT	TT	0.008	0.252	0.74
CT	AT	CT	0.006	0.056	0.938
CT	TT	CC	0.439	0.202	0.36
CT	TT	TT	0	0.849	0.151
CT	TT	CT	0.02	0.387	0.593

BMP6.13	STROKE	BMP6.14	BMP6.12		
			TT	CT	CC
TT	0	AA	0.103	0.794	0.103
TT	0	AC	0.02	0.423	0.557
TT	0	CC	0.038	0.122	0.84

FIGURE 8
Page 6 of 21

BMP6.13	STROKE	BMP6.14	BMP6.12		
			TT	CT	CC
TT	1	AA	0.333	0.333	0.333
TT	1	AC	0.103	0.103	0.794
TT	1	CC	0.013	0.013	0.974
CC	0	AA	0.672	0.237	0.091
CC	0	AC	0.351	0.556	0.093
CC	0	CC	0.181	0.37	0.449
CC	1	AA	0.043	0.333	0.624
CC	1	AC	0.01	0.139	0.851
CC	1	CC	0.003	0.037	0.96
CT	0	AA	0.395	0.211	0.395
CT	0	AC	0.078	0.779	0.143
CT	0	CC	0.049	0.344	0.607
CT	1	AA	0.333	0.333	0.333
CT	1	AC	0.103	0.103	0.794
CT	1	CC	0.043	0.043	0.914

STROKE	CSF2.3	CSF2.4		
		CC	CT	TT
0	TT	0.966	0.008	0.026
0	CT	0.075	0.846	0.079
0	CC	0.022	0.236	0.742
1	TT	0.86	0.07	0.07
1	CT	0.382	0.492	0.126
1	CC	0.087	0.152	0.761

ANXA2.4	ANXA2.3		
	AA	AT	TT
AA	0.061	0.257	0.682
GG	0.272	0.47	0.258
AG	0.098	0.512	0.39

FIGURE 8
Page 7 of 21

ANXA2.12	ANXA2.7	ANXA2.5		
		AA	TT	AT
GG	GG	0.333	0.333	0.333
GG	GT	0.157	0.686	0.157
GG	TT	0.018	0.373	0.609
CC	GG	0.049	0.731	0.22
CC	GT	0.079	0.483	0.437
CC	TT	0.21	0.305	0.486
CG	GG	0.076	0.591	0.333
CG	GT	0.07	0.708	0.222
CG	TT	0.087	0.49	0.422

TGFBR3.7	STROKE	TGFBR3.10		
		TT	CT	CC
TT	0	0.115	0.208	0.677
TT	1	0.19	0.19	0.62
CC	0	0.009	0.163	0.828
CC	1	0.005	0.016	0.979
CT	0	0.03	0.241	0.729
CT	1	0.333	0.333	0.333

MET.6	STROKE	SELP.12		
		TT	AT	AA
TT	0	0.12	0.268	0.612
TT	1	0.333	0.333	0.333
CT	0	0.033	0.243	0.724
CT	1	0.228	0.07	0.702
CC	0	0.009	0.195	0.796
CC	1	0.107	0.005	0.888

FIGURE 8
Page 8 of 21

MET.6	MET.9		
	AA	AT	TT
TT	0.503	0.318	0.179
CT	0.027	0.515	0.458
CC	0.012	0.15	0.838

TGFB3.7	TGFB3		
	AA	GG	AG
TT	0.333	0.587	0.08
CC	0.016	0.763	0.221
CT	0.065	0.456	0.478

TGFB3.7	CSF2.3		
	TT	CT	CC
TT	0.122	0.249	0.629
CC	0.042	0.361	0.597
CT	0.049	0.495	0.456

ANXA2.12	ANXA2.13	ANXA2.7		
		GG	GT	TT
GG	AA	0.333	0.333	0.333
GG	AG	0.333	0.333	0.333
GG	GG	0.017	0.072	0.911
CC	AA	0.005	0.059	0.936
CC	AG	0.018	0.439	0.543
CC	GG	0.124	0.424	0.452
CG	AA	0.333	0.333	0.333
CG	AG	0.019	0.119	0.862
CG	GG	0.009	0.372	0.619

KL.2	KL.3		
	TT	AT	AA

FIGURE 8
Page 9 of 21

KL.2	KL.3		
	TT	AT	AA
TT	0.061	0.061	0.878
CT	0.017	0.32	0.663
CC	0.111	0.39	0.498

ANXA2.13	ANXA2.4		
	AA	GG	AG
AA	0.692	0.067	0.241
AG	0.371	0.098	0.531
GG	0.25	0.251	0.498

BMP6.10	STROKE	BMP6.14	BMP6.3		
			TT	CC	CT
AA	0	AA	0.794	0.103	0.103
AA	0	AC	0.1	0.362	0.538
AA	0	CC	0.438	0.09	0.473
AA	1	AA	0.333	0.333	0.333
AA	1	AC	0.333	0.333	0.333
AA	1	CC	0.333	0.333	0.333
AT	0	AA	0.02	0.691	0.289
AT	0	AC	0.001	0.372	0.627
AT	0	CC	0.26	0.166	0.574
AT	1	AA	0.333	0.333	0.333
AT	1	AC	0.333	0.333	0.333
AT	1	CC	0.333	0.333	0.333
TT	0	AA	0.005	0.898	0.097
TT	0	AC	0.026	0.556	0.418
TT	0	CC	0.15	0.333	0.517
TT	1	AA	0.043	0.914	0.043
TT	1	AC	0.066	0.868	0.066
TT	1	CC	0.154	0.803	0.043

FIGURE 8
Page 10 of 21

SELP.17	STROKE	SELP.22		
		CC	CT	TT
TT	0	0.199	0.199	0.602
TT	1	0.333	0.333	0.333
CC	0	0.354	0.492	0.155
CC	1	0.986	0.007	0.007
CT	0	0.099	0.528	0.373
CT	1	0.97	0.015	0.015

EDN1.3	EDN1.6	EDN1.10	EDN1.7		
			CC	TT	CT
AA	CT	AA	0.333	0.333	0.333
AA	CT	CC	0.333	0.333	0.333
AA	CT	AC	0.333	0.333	0.333
AA	TT	AA	0.333	0.333	0.333
AA	TT	CC	0.333	0.333	0.333
AA	TT	AC	0.333	0.333	0.333
AA	CC	AA	0.333	0.333	0.333
AA	CC	CC	0.883	0.01	0.107
AA	CC	AC	0.848	0.076	0.076
AC	CT	AA	0.333	0.333	0.333
AC	CT	CC	0.664	0.036	0.3
AC	CT	AC	0.914	0.043	0.043
AC	TT	AA	0.333	0.333	0.333
AC	TT	CC	0.914	0.043	0.043
AC	TT	AC	0.333	0.333	0.333
AC	CC	AA	0.962	0.019	0.019
AC	CC	CC	0.376	0.031	0.593
AC	CC	AC	0.786	0.002	0.212
CC	CT	AA	0.333	0.333	0.333
CC	CT	CC	0.339	0.001	0.66
CC	CT	AC	0.914	0.015	0.071

FIGURE 8
Page 11 of 21

EDN1.3	EDN1.6	EDN1.10	EDN1.7		
			CC	TT	CT
CC	TT	AA	0.333	0.333	0.333
CC	TT	CC	0.963	0.003	0.034
CC	TT	AC	0.848	0.076	0.076
CC	CC	AA	0.821	0.021	0.158
CC	CC	CC	0.144	0.377	0.479
CC	CC	AC	0.325	0.036	0.639

MET.6	NRCAM		
	CG	CC	GG
TT	0.549	0.41	0.041
CT	0.508	0.223	0.269
CC	0.456	0.185	0.359

SELP.17	SELP.15		
	TT	CC	CT
TT	0.038	0.84	0.122
CC	0.129	0.412	0.46
CT	0.015	0.598	0.387

MET.5	KL.2		
	TT	CT	CC
TT	0.103	0.449	0.449
CT	0.026	0.163	0.811
CC	0.008	0.116	0.876

MET.5	TGFB2		
	GG	CG	CC
TT	0.103	0.218	0.679
CT	0.026	0.209	0.765
CC	0.005	0.119	0.876

FIGURE 8
Page 12 of 21

MET.5	STROKE	TGFBR		
		TT	CC	CT
TT	0	0.061	0.47	0.47
TT	1	0.333	0.333	0.333
CT	0	0.013	0.769	0.218
CT	1	0.19	0.62	0.19
CC	0	0.016	0.856	0.128
CC	1	0.016	0.979	0.005

MET.5	STROKE	SELP.14		
		TT	CC	CT
TT	0	0.061	0.606	0.333
TT	1	0.333	0.333	0.333
CT	0	0.023	0.908	0.069
CT	1	0.62	0.19	0.19
CC	0	0.002	0.851	0.147
CC	1	0.286	0.406	0.308

STROKE	MET.5	MET.6		
		TT	CT	CC
0	TT	0.606	0.333	0.061
0	CT	0.051	0.656	0.293
0	CC	0.009	0.151	0.84
1	TT	0.333	0.333	0.333
1	CT	0.19	0.62	0.19
1	CC	0.005	0.048	0.947

ECE1.13	SELP.17		
	TT	CC	CT
TT	0.195	0.195	0.61
CT	0.014	0.76	0.226
CC	0.016	0.754	0.23

FIGURE 8
Page 13 of 21

MET.5	CIQB		
	TT	CT	CC
TT	0.103	0.449	0.449
CT	0.108	0.218	0.674
CC	0.019	0.259	0.722

EDN1.9	BMP6.14		
	AA	AC	CC
AA	0.158	0.279	0.563
AG	0.047	0.334	0.619
GG	0.025	0.242	0.733

EDN1.9	EDN1.6	EDN1.10		
		AA	CC	AC
AA	CT	0.333	0.333	0.333
AA	TT	0.157	0.686	0.157
AA	CC	0.014	0.927	0.059
AG	CT	0.006	0.78	0.214
AG	TT	0.076	0.848	0.076
AG	CC	0.017	0.698	0.285
GG	CT	0.001	0.734	0.265
GG	TT	0.009	0.95	0.041
GG	CC	0.066	0.609	0.325

ANXA2.8	ANXA2.12	ANXA2.13		
		AA	AG	GG
GG	GG	0.333	0.333	0.333
GG	CC	0.333	0.333	0.333
GG	CG	0.157	0.686	0.157
AG	GG	0.333	0.333	0.333
AG	CC	0.087	0.692	0.221
AG	CG	0.076	0.333	0.591

FIGURE 8
Page 14 of 21

ANXA2.8	ANXA2.12	ANXA2.13		
		AA	AG	GG
AA	GG	0.017	0.017	0.966
AA	CC	0.052	0.323	0.625
AA	CG	0.001	0.208	0.791

STROKE	BMP6.10	BMP6.11		
		CC	TT	CT
0	AA	0.775	0.077	0.148
0	AT	0.145	0.055	0.8
0	TT	0.081	0.528	0.391
1	AA	0.333	0.333	0.333
1	AT	0.333	0.333	0.333
1	TT	0.326	0.208	0.465

ECE1.13	ERG.3		
	TT	CT	CC
TT	0.195	0.092	0.713
CT	0.231	0.447	0.322
CC	0.282	0.473	0.246

BMP6.2	BMP6.3		
	CC	GG	CG
GG	0.872	0.029	0.099
AA	0.23	0.381	0.389
AG	0.32	0.057	0.623

HBG		HBF.P		HBF.G			
				[0.251 0.737) [0.737 0.857) [0.857 0.963) [0.963 1.699]			
[1.337	1.497)	[0.251	1.219)	0.877	0.031	0.051	0.041
[1.337	1.497)	[1.219	1.406)	0.284	0.567	0.079	0.07

FIGURE 8
Page 15 of 21

HBG		HBF.P		HBF.G			
				[0.251 0.737) [0.737 0.857) [0.857 0.963) [0.963 1.699]			
[1.337	1.497)	[1.406	1.575)	0.041	0.368	0.524	0.067
[1.337	1.497)	[1.575	2.481]	0.117	0.079	0.402	0.402
[1.496	1.527)	[0.251	1.219)	0.788	0.071	0.082	0.059
[1.496	1.527)	[1.219	1.406)	0.091	0.837	0.021	0.051
[1.496	1.527)	[1.406	1.575)	0.062	0.113	0.803	0.022
[1.496	1.527)	[1.575	2.481]	0.068	0.055	0.227	0.65
[1.527	1.562)	[0.251	1.219)	0.75	0.068	0.091	0.091
[1.527	1.562)	[1.219	1.406)	0.046	0.802	0.061	0.091
[1.527	1.562)	[1.406	1.575)	0.001	0.052	0.855	0.092
[1.527	1.562)	[1.575	2.481]	0.052	0.102	0.031	0.815
[1.562	1.734]	[0.251	1.219)	0.62	0.216	0.109	0.055
[1.562	1.734]	[1.219	1.406)	0.076	0.635	0.213	0.076
[1.562	1.734]	[1.406	1.575)	0.055	0.068	0.557	0.32
[1.562	1.734]	[1.575	2.481]	0.033	0.033	0.025	0.909

STROKE	CAT	MET.5		
		TT	CT	CC
0	AA	0.103	0.333	0.564
0	AG	0.005	0.157	0.838
0	GG	0.005	0.075	0.92
1	AA	0.333	0.333	0.333
1	AG	0.061	0.061	0.878
1	GG	0.005	0.017	0.978

STROKE	ECE1.12	ECE1.13		
		TT	CT	CC
0	CC	0.654	0.293	0.053
0	CT	0.008	0.68	0.312
0	TT	0.001	0.065	0.934
1	CC	0.083	0.458	0.458

FIGURE 8
Page 16 of 21

STROKE	ECE1.12	ECE1.13		
		TT	CT	CC
1	CT	0.083	0.271	0.646
1	TT	0.005	0.134	0.861

EDN1.3	EDN1.9		
	AA	AG	GG
AA	0.724	0.065	0.211
AC	0.03	0.719	0.251
CC	0.007	0.129	0.864

CAT	TIE		
	AG	AA	GG
AA	0.686	0.157	0.157
AG	0.207	0.046	0.747
GG	0.232	0.015	0.753

EDN1.3	ANXA2.12		
	GG	CC	CG
AA	0.138	0.431	0.431
AC	0.035	0.679	0.286
CC	0.008	0.77	0.222

TGFB2.6	TGFB2.0		
	CT	CC	TT
CT	0.645	0.063	0.292
CC	0.104	0.755	0.141
TT	0.176	0.023	0.801

ERG.5	RS276789		
	TT	CC	CT
CC	0.092	0.609	0.299

FIGURE 8
Page 17 of 21

ERG.5	RS276789		
	TT	CC	CT
AA	0.031	0.704	0.265
AC	0.05	0.526	0.424

BMP6	STROKE	BMP6.10		
		AA	AT	TT
GG	0	0.262	0.69	0.048
GG	1	0.333	0.333	0.333
CG	0	0.027	0.318	0.655
CG	1	0.039	0.039	0.922
CC	0	0.055	0.326	0.619
CC	1	0.005	0.005	0.99

EDN1.3	BMP6.2		
	GG	AA	AG
AA	0.211	0.211	0.578
AC	0.26	0.277	0.463
CC	0.154	0.36	0.485

ERG.2	ERG.4		
	AC	CC	AA
GG	0.298	0.618	0.084
AG	0.433	0.507	0.06
AA	0.505	0.405	0.09

STROKE	HBF.P			
	[0.251 1.219)	[1.219 1.406)	[1.406 1.575)	[1.575 2.480]
0	0.234	0.257	0.253	0.256
1	0.469	0.146	0.198	0.187

HBG	THALASSEMIA	
	SS	SS_ALPHA

FIGURE 8
Page 18 of 21

HBG	THALASSEMIA	
	SS	SS_ALPHA
[1.337 1.496)	0.756	0.244
[1.496 1.527)	0.756	0.244
[1.527 1.562)	0.646	0.354
[1.562 1.734]	0.619	0.381

ANXA2.8	CAT		
	AA	AG	GG
GG	0.242	0.242	0.516
AG	0.045	0.045	0.91
AA	0.003	0.075	0.922

ANXA2.8	BMP6		
	GG	CG	CC
GG	0.242	0.242	0.516
AG	0.045	0.198	0.757
AA	0.006	0.097	0.897

ANXA2.8	ERG.5		
	CC	AA	AC
GG	0.242	0.516	0.242
AG	0.045	0.91	0.045
AA	0.006	0.881	0.113

ANXA2.8	STROKE	ECE1.12		
		CC	CT	TT
GG	0	0.19	0.62	0.19
GG	1	0.333	0.333	0.333
AG	0	0.079	0.242	0.679
AG	1	0.333	0.333	0.333
AA	0	0.005	0.128	0.867

FIGURE 8
Page 19 of 21

ANXA2.8	STROKE	ECE1.12		
		CC	CT	TT
AA	1	0.048	0.048	0.904

ANXA2.8	EDN1.3		
	AA	AC	CC
GG	0.242	0.242	0.516
AG	0.045	0.299	0.656
AA	0.009	0.159	0.832

ANXA2.8	TGFBR2.6		
	CT	CC	TT
GG	0.516	0.242	0.242
AG	0.096	0.096	0.808
AA	0.19	0.018	0.792

STROKE	ANXA2.8	ADCY9.8		
		AA	GG	AG
0	GG	0.19	0.62	0.19
0	AG	0.734	0.024	0.242
0	AA	0.769	0.009	0.222
1	GG	0.333	0.333	0.333
1	AG	0.333	0.333	0.333
1	AA	0.894	0.005	0.101

ANXA2.8	EDN1.6		
	CT	TT	CC
GG	0.333	0.333	0.333
AG	0.333	0.333	0.333
AA	0.333	0.333	0.333

FIGURE 8
Page 20 of 21

STROKE	ADCY9.5		
	GG	AG	AA
0	0.034	0.212	0.754
1	0.014	0.431	0.555

ANXA2.8	TGFB3.4		
	GG	AA	AG
GG	0.242	0.242	0.516
AG	0.198	0.655	0.147
AA	0.029	0.659	0.312

STROKE	ADCY9.11		
	CT	CC	TT
0	0.334	0.593	0.073
1	0.17	0.67	0.16

STROKE	ERG.2		
	GG	AG	AA
0	0.064	0.403	0.533
1	0.108	0.212	0.68

STROKE	ADCY9		
	CC	TT	CT
0	0.218	0.272	0.51
1	0.097	0.347	0.556

STROKE	TEK		
	CG	GG	CC
0	0.409	0.091	0.5
1	0.472	0.024	0.504

FIGURE 8
Page 21 of 21

STROKE	ADCY9.9		
	AA	AC	CC
0	0.227	0.478	0.295
1	0.285	0.295	0.42

STROKE	TGFB3.6		
	CC	AA	AC
0	0.369	0.164	0.467
1	0.545	0.014	0.441

STROKE	CCL2		
	TT	CC	CT
0	0.096	0.468	0.436
1	0.014	0.66	0.326

STROKE	ANXA2.6		
	GG	AG	AA
0	0.11	0.431	0.459
1	0.316	0.149	0.535

HBG			
[1.337 1.496)	[1.496 1.527)	[1.527 1.562)	[1.562 1.734]
0.236	0.257	0.253	0.254

ANXA2.8		
GG	AG	AA
0.003	0.014	0.983

Case	Observed	Predicted	Probability
111	1	1	0.999
114	1	1	0.999
112	1	1	0.998
113	1	1	0.997
110	1	1	0.95
108	1	1	0.787
109	1	1	0.612
27	0	1	0.95
92	0	1	0.721
53	0	0	0.411
103	0	0	0.231
58	0	0	0.115
87	0	0	0.1
5	0	0	0.081
82	0	0	0.058
15	0	0	0.057
99	0	0	0.056
72	0	0	0.046
83	0	0	0.046
74	0	0	0.033
3	0	0	0.031
48	0	0	0.026
4	0	0	0.022
21	0	0	0.022
55	0	0	0.018
62	0	0	0.013
73	0	0	0.013
32	0	0	0.012
93	0	0	0.011
6	0	0	0.011
97	0	0	0.01
16	0	0	0.009
71	0	0	0.009
100	0	0	0.008
85	0	0	0.007
59	0	0	0.006
78	0	0	0.006
36	0	0	0.004
50	0	0	0.004
18	0	0	0.003
31	0	0	0.003
76	0	0	0.003

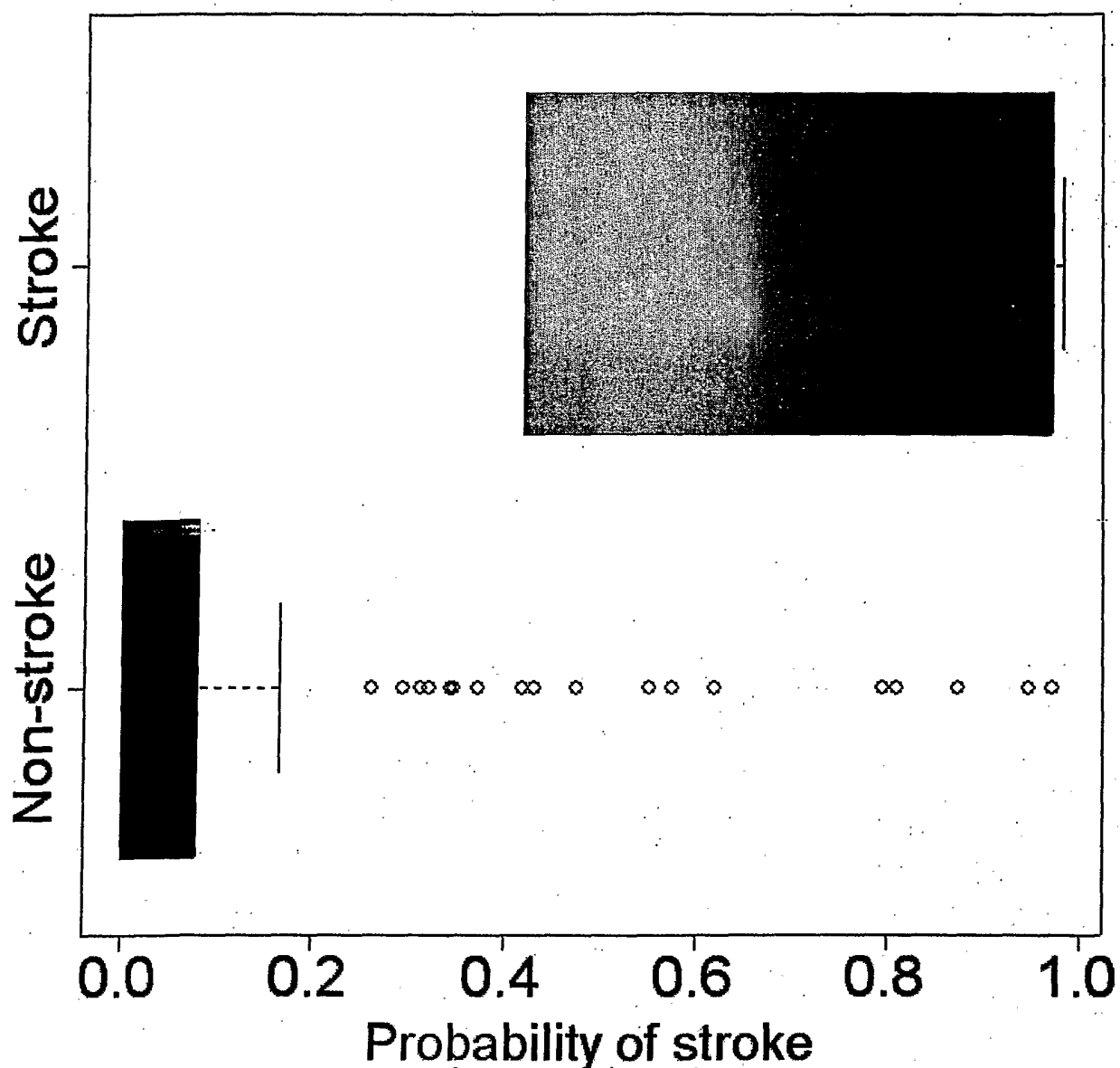
Supplementary Table 4
Figure 9 (Page 1 of 3)

88	0	0	0.003
101	0	0	0.003
14	0	0	0.002
28	0	0	0.002
33	0	0	0.002
34	0	0	0.002
45	0	0	0.002
1	0	0	0.001
19	0	0	0.001
35	0	0	0.001
57	0	0	0.001
75	0	0	0.001
84	0	0	0.001
94	0	0	0.001
107	0	0	0.001
2	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	0	0	0
11	0	0	0
12	0	0	0
13	0	0	0
17	0	0	0
20	0	0	0
22	0	0	0
23	0	0	0
24	0	0	0
25	0	0	0
26	0	0	0
29	0	0	0
30	0	0	0
37	0	0	0
38	0	0	0
39	0	0	0
40	0	0	0
41	0	0	0
42	0	0	0
43	0	0	0
44	0	0	0
46	0	0	0
47	0	0	0
49	0	0	0
51	0	0	0
52	0	0	0
54	0	0	0
56	0	0	0
60	0	0	0
61	0	0	0
63	0	0	0
64	0	0	0

Figure 9 (Page 2 of 3)

65	0	0	0
66	0	0	0
67	0	0	0
68	0	0	0
69	0	0	0
70	0	0	0
77	0	0	0
79	0	0	0
80	0	0	0
81	0	0	0
86	0	0	0
89	0	0	0
90	0	0	0
91	0	0	0
95	0	0	0
96	0	0	0
98	0	0	0
102	0	0	0
104	0	0	0
105	0	0	0
106	0	0	0

Figure 9 (Page 3 of 3)



Supplementary Figure 1
Figure 10

Effect	Estimate	St. Error	P-value
Intercept	3.56	0.94	0.0002
HbF.G	-4.6	1.03	< 0.0000
SELP.12 (TT)	2.41	0.76	0.0006
SELP.14 (CT)	1.98	0.39	< 0.0000
SELP.14 (TT)	7	1.1	< 0.0000
BMP6.11 (CT)	-0.83	0.41	0.0436
BMP6.11 (TT)	-1.87	0.49	0.0001
BMP6.12 (CT)	-4.18	0.82	< 0.0000
BMP6.13 (CT)	-4.6	0.82	< 0.0000
BMP6.13 (TT)	-3.06	0.5	< 0.0000

Supplementary Table 5
Figure 11