

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7137519号

(P7137519)

(45)発行日 令和4年9月14日(2022.9.14)

(24)登録日 令和4年9月6日(2022.9.6)

(51)国際特許分類

F I

C 1 2 Q 1/6827(2018.01)

C 1 2 Q 1/6827

Z Z N A

G 0 1 N 33/50 (2006.01)

G 0 1 N 33/50

P

C 1 2 N 15/09 (2006.01)

C 1 2 N 15/09

Z

請求項の数 2 (全16頁)

(21)出願番号 特願2019-81544(P2019-81544)
 (22)出願日 平成31年4月23日(2019.4.23)
 (65)公開番号 特開2020-174642(P2020-174642
 A)
 (43)公開日 令和2年10月29日(2020.10.29)
 審査請求日 令和3年12月27日(2021.12.27)

(73)特許権者 504217812
 ジェネシスヘルスケア株式会社
 東京都渋谷区恵比寿4丁目20番3号
 恵比寿ガーデンプレイスタワー26階
 (74)代理人 100079108
 弁理士 稲葉 良幸
 (74)代理人 100109346
 弁理士 大貫 敏史
 (74)代理人 100117189
 弁理士 江口 昭彦
 (74)代理人 100134120
 弁理士 内藤 和彦
 (72)発明者 宇田川 ヨスバニ
 東京都渋谷区恵比寿四丁目20番3号
 恵比寿ガーデンプレイスタワー26階
 最終頁に続く

(54)【発明の名称】 結核のリスクを判定する方法

(57)【特許請求の範囲】

【請求項1】

結核に罹患したヒトの遺伝子型データと、結核に罹患していないヒトの遺伝子型データと、を学習データとして用いて機械学習したモデルを用いて、結核と正の相関があるrs2853677(配列番号1)、rs3218097(配列番号5)、及びrs1111875(配列番号6)と、結核と負の相関があるrs1208285(配列番号7)と、結核と正の相関又は負の相関があるrs16853722(配列番号4)、rs2268361(配列番号2)、及びrs2200733(配列番号3)を少なくとも含む一塩基多型セットの遺伝子型情報であって、

rs16853722の遺伝子型がTT、rs2853677の遺伝子型がTC、rs2268361の遺伝子型がAA、rs3218097の遺伝子型がCC、rs2200733の遺伝子型がCC、rs1111875の遺伝子型がAA、rs16853722の遺伝子型がTC、rs2268361の遺伝子型がAG、rs2200733の遺伝子型がTC、及びrs1208285の遺伝子型がTCであるか否かに関する遺伝子型情報に基づいて、結核のリスクを判定する、方法。

【請求項2】

リスクの判定を受ける対象者の体液サンプル、細胞サンプル又は体毛を用いる、請求項1に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は結核のリスクを判定する方法に関する。

10

20

【背景技術】

【0002】

疾患のリスクの判定に用いるために、一塩基多型（以下、「SNP」ともいう。）と疾患との関連性の特定が進められている。NCBI SNP Databaseは、ヒトのSNPをまとめたデータベースであり、SNPにrs番号を付して管理している。本明細書におけるrs番号も、このNCBI SNP Databaseにおける登録番号を意味するものとする。

【0003】

本明細書においてrs番号で特定されるSNPと、当該SNPに関連する疾患、病態又は状態等として非特許文献等で開示されているものとの関係は、以下のとおりである。

rs2853677：肺がん（肺腺がん）に関するSNP（非特許文献1）

rs2268361：多嚢胞性卵巣症候群に関するSNP（非特許文献2）

rs2200733：脳卒中に関するSNP（非特許文献3）

rs16853722：血液中の尿素窒素濃度に関するSNP（非特許文献4）

rs16853722：慢性腎臓病に関するSNP（非特許文献5）

rs3218097：平均赤血球容積に関するSNP（非特許文献6）

rs3218097：赤血球数に関するSNP（非特許文献7）

rs3218097：平均赤血球ヘモグロビン量に関するSNP（非特許文献8）

rs1111875：2型糖尿病に関するSNP（非特許文献9）

rs1208285：肥厚性幽門狭窄症に関するSNP（非特許文献10）

【先行技術文献】

【非特許文献】

【0004】

【文献】Shiraishi K, Kunitoh H, Daigo Y, Takahashi A, Goto K, Sakamoto H, et al. A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat. Genet.* 2012;44: 900-3. Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, et al. Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nat. Genet.* 2012;44: 1020-5.

NINDS Stroke Genetics Network (SiGN), International Stroke Genetics Consortium (ISGC). Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. *Lancet Neurol* 2016;15: 174-184.

Okada Y, Sim X, Go MJ, Wu JY, Gu D, Takeuchi F, et al. Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* 2012;44: 904-9.

Okada Y, Sim X, Go MJ, Wu JY, Gu D, Takeuchi F, et al. Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* 2012;44: 904-9.

Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* 2010;42: 210-5.

Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* 2010;42: 210-5.

Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* 2010;42: 210-5.

Takeuchi F, Serizawa M, Yamamoto K, Fujisawa T, Nakashima E, Ohnaka K, et al. Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes* 20

10

20

30

40

50

09;58: 1690-9.

Feenstra B, Geller F, Krogh C, Hollegaard MV, Gortz S, Boyd HA, et al. Common variants near MBNL1 and NKX2-5 are associated with infantile hypertrophic pyloric stenosis. Nat. Genet. 2012;44: 334-7

【発明の概要】

【発明が解決しようとする課題】

【0005】

本発明は、結核（以下、「本疾患」ともいう。）のリスクを判定する方法を提供することを課題とする。

【課題を解決するための手段】

【0006】

本発明者らは上記課題を解決するために鋭意検討した。その結果、これまで一見すると本疾患との関連性がないと認められる個々の一塩基多型を、一つのまとまったセットとしてみたときに、本疾患との関連性があることを見出した。そして、その関連性を用いることで、本疾患のリスクを判定する本発明を完成するに至った。

【0007】

すなわち、本発明の方法では、本疾患との関連性を見出した、rs2853677、rs2268361、rs2200733、rs16853722、rs3218097、rs1111875、及びrs1208285を少なくとも含む一塩基多型セット（以下、「本SNPセット」ともいう。）の遺伝子型情報に基づいて、本疾患のリスクを判定する。

【0008】

本発明の方法においては、「一塩基多型セット」とは、複数の一塩基多型の一つのまとまったセットを意味し、この一つのセットにより本疾患との関連性が見出されている。

【0009】

また、本発明の方法における「遺伝子型情報」とは、一塩基多型における2つのホモ接合型（AA，BB）と、ヘテロ接合型（AB）に分類して示される、一塩基多型の遺伝子型（Genotype）の情報を意味し、「本SNPセットの遺伝子型情報」とは、本SNPセットにおいて特定される各一塩基多型の遺伝子型情報を一まとまりとしたセットを意味し、言い換えれば、各rs番号で示される塩基配列中の各SNPの多型となる塩基に関する情報のセットである。本SNPセットの遺伝子型情報は、図1において示すとおりである。

【発明の効果】

【0010】

本発明によれば、本疾患のリスクを判定することができる。

【図面の簡単な説明】

【0011】

【図1】本SNPセットの遺伝子型情報を示す。

【図2】本SNPセットの遺伝子型情報にSNP毎の接合型に対応付ける値の関係を示した変換テーブルの一例を示す。

【図3】本SNPセットを用いたモデルのROC曲線とAUCを示す。また、N個のSNPを含む本SNPセットから1つのSNPを任意に抜いたN-1個のSNPを含むSNPセットを、「比較SNPセット」ともいい、各比較SNPセットを表す場合には、比較SNPセット1、比較SNPセット2と、記載する。

【図4】比較SNPセット1を用いたモデルのROC曲線とAUCを示す。

【図5】比較SNPセット2を用いたモデルのROC曲線とAUCを示す。

【図6】比較SNPセット3を用いたモデルのROC曲線とAUCを示す。

【図7】比較SNPセット4を用いたモデルのROC曲線とAUCを示す。

【図8】比較SNPセット5を用いたモデルのROC曲線とAUCを示す。

【図9】比較SNPセット6を用いたモデルのROC曲線とAUCを示す。

【図10】比較SNPセット7を用いたモデルのROC曲線とAUCを示す。

10

20

30

40

50

【図 1 1】比較 S N P セット 8 を用いたモデルの R O C 曲線と A U C を示す。

【図 1 2】比較 S N P セット 9 を用いたモデルの R O C 曲線と A U C を示す。

【図 1 3】比較 S N P セット 1 0 を用いたモデルの R O C 曲線と A U C を示す。

【発明を実施するための形態】

【 0 0 1 2 】

本発明の実施の形態について説明する。以下の実施形態は、本発明を説明するための例示であり、本発明をこの実施形態にのみ限定する趣旨ではない。本発明は、その要旨を逸脱しない限り、様々な形態で実施することができる。

【 0 0 1 3 】

本実施形態において、結核とは、結核菌が原因となって起こる病気である。

10

【 0 0 1 4 】

また、本実施形態において、本疾患は、一般には、本疾患に関する医学会の公表するガイドラインに沿って診断される疾患、医療用医薬品の添付文書において、効能・効果の欄に記載される疾患、あるいは、医薬・医療業界において汎用される用語として理解される疾患の少なくともいずれかを意味するものと解することができる。

【 0 0 1 5 】

本実施形態の方法においては、一見すると本疾患との関連性がないと認められる所定数の一塩基多型セットを用いて、本疾患のリスクを判定する。

【 0 0 1 6 】

本疾患のリスクとは、本疾患の罹りやすさや罹りにくさなどの本疾患に罹る可能性をいう。「リスクを判定する」とは、例えば、現在または将来において本疾患に罹る可能性をいくつかのレベルに分けて出力することや、数値により出力することを含む。本疾患のリスクの判定には、疾患に罹りやすい傾向にあるのか、罹りにくい傾向にあるのかといった、疾患に対する遺伝的要因あるいは遺伝的感受性についての評価が含まれる。

20

【 0 0 1 7 】

なお、本疾患のリスクを判定するにあたっては、本疾患のリスクの判定を受ける対象者が、本疾患のリスクの判定時において、実際に本疾患に罹患しているか（発症しているか）否かは問わない。

【 0 0 1 8 】

本実施形態の方法では、本 S N P セットで特定される各 S N P の遺伝子型を 2 つのホモ接合型（A A , B B ）と、ヘテロ接合型（A B ）に分類した遺伝子型のセットである、本 S N P セットの遺伝子型情報を用いる。そして、本 S N P セットの遺伝子型情報に基づいて、対象者の本疾患のリスクを判定する。

30

【 0 0 1 9 】

本実施形態の方法で用いる本 S N P セットは、本疾患との関連性が従来認められていなかった S N P を含むセットである。すなわち、通常は、本 S N P セットに含まれる S N P を個別に分析したとしても、本疾患のリスクを判定することはできない。しかしながら、本実施形態の方法では、本 S N P セットに含まれる各 S N P の遺伝子型情報を一まとまりのセットとして分析することにより、本疾患のリスクを判定することができる。また、本 S N P セットを分析した場合と、比較 S N P セットを分析した場合とを比較すると、本 S N P セットを分析した場合の方が統計的に有意な結果が得られている。すなわち、本実施形態の方法において、本 S N P セットを分析して本疾患のリスクを判定することで、高精度が高い、あるいは予測能力が高いリスクの判定方法を提供することができる。

40

【 0 0 2 0 】

以下、本 S N P セットに含まれる各 S N P に関連して、r s 番号と、各 S N P が存在する染色体番号（性染色体の場合には、X か Y で示す）と、各 S N P の染色体上の位置と、r s 番号に対応する塩基配列と、を列記する。なお、各 r s 番号で示される塩基配列中において、S N P は [] で囲って示す。例えば、[A / G] と表記した場合には、その塩基配列の位置において A 又は G の一塩基多型があることを示す。また、各 S N P に関する塩基配列や疾患などの情報は、例えば、r s 番号に基づいて N C B I S N P D a t a b a

50

seを検索することで得られる。それらの情報は当該Databaseにより参照可能であり、また、本明細書で援用する。なお、以下に記す染色体上の位置は、assembly genomeのバージョンGRCh37に対応するものである。

【0021】

rs2853677

染色体番号 5

染色体上の位置 1287194

塩基配列 TCATTTTGTCTACTAGAGACCCG[T/C]CTG
GTGCACTCTGATTCTCCACTTG(配列番号1)

【0022】

rs2268361

染色体番号 2

染色体上の位置 49201612

塩基配列 TGAAAGATCTCAATCCCAGGGCACT[A/G]ACA
AGATGCCCTTCTGCTCTCACAGCAT(配列番号2)

【0023】

rs2200733

染色体番号 4

染色体上の位置 111710169

塩基配列 GTGGTACTTGGGTTTGTATTTGAT[T/C]AGA
GAAAATTAGAACAGGTAAATTT(配列番号3)

【0024】

rs16853722

染色体番号 3

染色体上の位置 169150632

塩基配列 TTAAAGGCTACAGTAGATTGTTCTA[T/C]TCT
GTGCCAAGATTCTGCTCCAGCT(配列番号4)

【0025】

rs3218097

染色体番号 6

染色体上の位置 41905275

塩基配列 TCCCTGCCTTGAACTCCATGGTGCC[T/C]GGC
CTTAAAATCAGAAACAGAACTG(配列番号5)

【0026】

rs1111875

染色体番号 10

染色体上の位置 94462882

塩基配列 TCCGTACCATCAAGTCAATTTCTCT[A/G]GAC
GTCTGAACCTGCCTCAGGGTC(配列番号6)

【0027】

rs1208285

染色体番号 6

染色体上の位置 134158800

塩基配列 CCATATAAGCCCAAGCACTCACGCAA[T/C]CTT
TGTCCTCCAGTTGACTTAAGA(配列番号7)

【0028】

本実施形態の方法において、本SNPセットを構成する各SNPはrs番号により特定される塩基配列を参照することによって特定可能であるが、本明細書において記載するrs番号が他のrs番号と併合され、新たなrs番号が付与された場合には、本明細書において該当するrs番号は、併合後のrs番号及び併合される他のrs番号をも意味する。

10

20

30

40

50

また、本明細書において記載する r s 番号が複数の r s 番号の併合により付与された番号である場合には、本明細書において該当する r s 番号は、その他の元となる r s 番号をも意味する。

【 0 0 2 9 】

また、SNPに関する各 r s 番号で示される上記塩基配列は、特定の塩基配列として示しているが、人種の相違等によって、当該塩基配列において該当するSNP以外の部分における塩基配列は変更されてもよい。

【 0 0 3 0 】

本実施形態の方法は、いずれの人種の被検者に対しても用いることができるが、特に、アジア人に好適に用いることができる。アジア人の中でも日本人等の東アジア人の被検者により好適に用いることができる。また、本実施形態の方法は、いずれの性別の被検者に対しても用いてもよい。

【 0 0 3 1 】

以下、本SNPセットの遺伝子型情報を分析することにより本疾患のリスクを判定する方法の一態様について説明する。但し、判定方法は、以下に限定されない。

【 0 0 3 2 】

はじめに、対象者の試料を用いて、試料中の本SNPセットに含まれる各SNPの遺伝子型を特定する。SNPの検出に用いる試料としては、染色体DNAを含む試料であれば特に制限されない。このような試料としては、例えば、唾液、血液、尿等の体液サンプル；口腔粘膜などの細胞サンプル；毛髪等の体毛などが挙げられる。SNPの検出には、これらの試料から常法により単離した染色体DNAを直接使用してもよいし、単離した染色体DNAを増幅して、増幅後の染色体DNAを使用してもよい。

【 0 0 3 3 】

SNPの検出は、通常の遺伝子多型解析方法によって行うことができる。例えば、DNAチップ法（DNAマイクロアレイ）、サンガー法を用いた従来型のシーケンサーや次世代シーケンサー（NGS；Next Generation Sequencer）などを用いたシーケンス解析、PCR（Polymerase Chain Reaction）、ハイブリダイゼーション、インベーター法などが挙げられるが、これらに限定されない。

【 0 0 3 4 】

DNAチップ法では、SNP部位を含む多数のDNA断片（プローブ）を基板上に配置したDNAチップを用い、染色体DNAをチップ上のプローブとハイブリダイズさせて、結合部位を蛍光又は電流により検出することにより、染色体DNAの配列を解析する。SNP解析に用いられるDNAチップとしては、SNP部位を含む塩基配列を検出可能なオリゴヌクレオチドプローブが配置されたチップが挙げられる。

【 0 0 3 5 】

また、シーケンス解析は通常のサンガー法により行うことができる。例えば、多型を示す塩基の5'側の数十塩基の位置に設定したプライマーを使用してシーケンス反応を行い、その解析結果から、該当する位置がどの種類の塩基であるかを決定することができる。なお、シーケンス反応の前に、あらかじめSNP部位を含む断片をPCRなどによって増幅しておくことが好ましい。効率の観点からはNGS技術を使用してもよい。

【 0 0 3 6 】

また、SNPの検出は、例えば従来のPCRによる増幅の有無を調べることによって行うことができる。例えば、多型を示す塩基を含む領域に対応する配列を有し、かつ、3'末端が各多型に対応するプライマーをそれぞれ用意する。それぞれのプライマーを使用してPCRを行い、増幅産物の有無によってどのタイプの多型であるかを決定することができる。また、LAMP法（Loop-Mediated Isothermal Amplification；特許第3313358号明細書）、NASBA法（Nucleic Acid Sequence-Based Amplification；特許2843586号明細書）、ICAN法（Isothermal and Chimeric primer-initiated Amplification of Nucleic ac

10

20

30

40

50

ids ; 特許第3433929号公報)などによって増幅の有無を調べることもできる。その他、単鎖増幅法やNGSを用いた解析法を用いてもよい。

【0037】

また、SNP部位を含むDNA断片を増幅し、増幅産物の電気泳動における移動度の違いによってどのタイプの多型であるかを決定することもできる。このような方法としては、例えば、PCR-SSCP(single-strand conformation polymorphism)法(Genomics, 1992 Jan 1; 12(1): 139-146.)などが挙げられる。具体的には、まず、目的のSNPを含むDNAを増幅し、増幅したDNAを一本鎖DNAに解離させる。次いで、解離させた一本鎖DNAを非変性ゲル上で分離し、分離した一本鎖DNAのゲル上での移動度の違いによってどのタイプの多型であるかを決定することができる。

10

【0038】

さらに、多型を示す塩基が制限酵素認識配列に含まれる場合は、制限酵素による切断の有無によって解析することもできる(RFLP(Restriction Fragment Length Polymorphism)法)。この場合、まず、DNA試料を制限酵素により切断する。次いで、DNA断片を分離し、検出されたDNA断片の大きさによってどのタイプの多型であるかを決定することができる。

【0039】

また、ハイブリダイゼーションの有無を調べることによって多型の種類を解析することも可能である。すなわち、各塩基に対応するプローブを用意し、いずれのプローブにハイブリダイズするかを調べることによってSNPがいずれの塩基であるかを調べることもできる。

20

【0040】

このようにして、本SNPセットの各SNPに関して、対象者の遺伝子型データを決定することができる。なおここで、「対象者の遺伝子型データ」とは、対象者の有する遺伝子型情報をいう。

【0041】

次いで、本SNPセットの遺伝子型情報に基づいて、本疾患のリスクを判定する。リスクの判定には、任意のモデルを用いることができる。モデルとしては、特に制限されないが、例えば、本SNPセットの遺伝子型情報を用いて、対象者の遺伝子型データから算出される特徴量を入力とし、本疾患のリスクを出力とするロジスティック回帰モデルを用いることができる。当該ロジスティック回帰モデルは、予め、本疾患に罹患したヒトの遺伝子型データと、本疾患に罹患していないヒトの遺伝子型データと、を学習データとして用いてパラメータを機械学習している。

30

【0042】

また、疾患のリスクを判定するモデルとしては、ロジスティック回帰モデルに代えて、多層パーセプトロン、CNN(Convolutional Neural Network)及びRNN(Recurrent Neural Network)などのニューラルネットワーク、ガウシアンカーネル等の任意のカーネル関数を用いるサポートベクターマシン、回帰木としてモデル化したランダムフォレスト、重回帰分析、隠れマルコフモデルなどを利用したモデル、統計モデルや確率モデルなど種々の他のモデルを採用することもできる。また、種々のモデルを組み合わせて総合的な判定を行うモデルを採用することもできる。

40

【0043】

次いで、モデルを用いた本疾患のリスク判定の一例について説明する。まず、本疾患のリスクの判定をする対象者の遺伝子型データを、モデルに入力可能な特徴量に変換する。本実施形態の方法における特徴量は、例えば、本SNPセットの各SNPについて、対象者の遺伝子型データがホモ接合型(AA)、ホモ接合型(BB)、又はヘテロ接合型(AB)のいずれであるかを示すパラメータである。遺伝子型は、相同染色体のSNPが共にG(グアニン)であることを示す“GG”や、一方がG(グアニン)で、他方がA(アデニ

50

ン)であることを示す“AG”などヌクレオチドにより表記されることが一般的であるため、対象者の遺伝子型データを、本SNPセットの遺伝子型情報を用いるモデルに入力可能なパラメータに変換する。しかし、モデルが、このようなパラメータへの変換の必要がないものである場合には、上記変換は必要とされない。

【0044】

対象者の遺伝子型データの特徴量への変換は、例えば、本SNPセットに含まれるSNP 1つ1つに関して、対象者の遺伝子型データに値を付すことにより行うことができる。例えば、各SNPについて、対象者の遺伝子型データがホモ接合型(AA)、ホモ接合型(BB)、又はヘテロ接合型(AB)のいずれに該当するのかに応じて、そのSNPに値(例えば、0又は1)を対応づける。これにより、対象者の遺伝子型データの特徴量に変換することができる。なお、以下では、各SNPに対応させる値を0又は1とした場合を例に説明するが、SNPに対応させる値は0又は1の2つの値に限られるものではない。

10

【0045】

接合型に対応づける値はSNPごとに決めることができる。例えば、あるSNPは、対象者の遺伝子型データがホモ接合型(AA)である場合に値1を対応付け、ホモ接合型(BB)及びヘテロ接合型(AB)である場合に値0を対応付けるようにし、他のSNPは、対象者の遺伝子型データがヘテロ接合型(AB)である場合に値1を対応付け、ホモ接合型(AA)及びホモ接合型(BB)である場合に値0を対応付けるようにしてもよい。そのほか、対象者の遺伝子型データがヘテロ接合型(AB)及びホモ接合型(BB)である場合に値1を対応付け、ホモ接合型(AA)である場合に値0を対応付けるようにしてもよい。

20

【0046】

上記のように、対象者の遺伝子型データの特徴量に変換することができる。この特徴量への変換において対応付けに使用する値は、任意に決定することができる。例えば、上記非特許文献に基づいて、もともと各SNPが関連する疾患に関係の高い遺伝子型に対して値1を対応付けるようにし、かつ、各SNPが関連する疾患に関係の低い遺伝子型に対して値0を対応付けるようにすることができる。

【0047】

このような、SNPごとの接合型とその接合型に対応付ける値の関係は、図1のような本SNPセットの遺伝子型情報をもとに、例えば、図2のような変換テーブルとして表すこともできる。図2の変換テーブルでは、網掛けをした遺伝子型と一致する場合にはそのSNPに対応付ける値を1とし、一致しない場合には対応付ける値を0とする。なお、図1及び2の具体的な遺伝子型の表記において、Aはアデニン、Gはグアニン、Cはシトシン、Tはチミンを示す。但し、特徴量の変換テーブルの形式は、図2に限定されるものではない。

30

【0048】

最後に、本SNPセットの遺伝子型情報に基づいて、対象者の本疾患のリスクを判定する。より具体的には、本SNPセットの遺伝子型情報に基づく変換テーブルを用いて、対象者の遺伝子型データを、モデルに入力可能に変換した特徴量として算出し、当該特徴量を所定の判定モデルに入力し、対象者の本疾患のリスクを判定することができる。

40

【0049】

判定モデルにおいて、特徴量には、本SNPセットのSNPごとに、本疾患のリスクと正の相関があることを表す重みづけや、本疾患のリスクと負の相関があることを表す重みづけをすることができる。例えば、rs2853677、rs3218097、及びrs1111875に対応付けた値(特徴量)については、本疾患のリスクと正の相関があることを表す重みづけを行い、rs1208285に対応付けた値(特徴量)については、本疾患のリスクと負の相関があることを表す重みづけを行い、rs16853722、rs2268361、及びrs2200733に対応付けた値(特徴量)については、その遺伝子型に応じて本疾患のリスクと正の相関又は負の相関があることを表す重みづけを行うことができる。

【0050】

50

例えば、特徴量に対して重みづけを行う場合には、rs16853722の遺伝子型がTT、rs2853677の遺伝子型がTC、rs2268361の遺伝子型がAA、rs3218097の遺伝子型がCC、rs2200733の遺伝子型がCC、及びrs1111875の遺伝子型がAAである場合に、本疾患のリスクと正の相関があることを表す重みづけを行い、rs16853722の遺伝子型がTC、rs2268361の遺伝子型がAG、rs2200733の遺伝子型がTC、及びrs1208285の遺伝子型がTCである場合に、本疾患のリスクと負の相関があることを表す重みづけを行うことができる。また、特徴量として値0を対応付けた各SNPの遺伝子型の場合には、本疾患のリスクと相関がない或いは無視しうる程度に低いものと評価することができる。

【0051】

このような、本疾患のリスクとの相関を表す重みづけは、本疾患に罹患したヒトの遺伝子型データと、本疾患に罹患していないヒトの遺伝子型データと、を学習データとして用いてパラメータを機械学習することにより特定される。この際、あるモデルにおいて、あるSNPが本疾患のリスクと正の相関があることを表す重み付けがなされとした場合、他のモデルにおいても同様にそのSNPは本疾患のリスクと正の相関があることを表す重み付けがなされることが通常である。すなわち、モデルの種類等によって、あるSNPにおいて本疾患のリスクとの相関関係が逆になるような事態は想定し難い。なお、重みづけの具体的な値はモデルによって異なり、特に制限されるものではない。

【0052】

ここで、本SNPセットのなかで本疾患のリスクと正の相関があることを表す重みづけを行うSNPのまとまりを「正相関SNPセット」といい、本疾患のリスクと負の相関があることを表す重みづけを行うSNPのまとまりを「負相関SNPセット」という。本SNPセットは、正相関SNPセットと負相関SNPセットとを含むものであり、このような本SNPセットの遺伝子型情報に基づくことにより、対象者の本疾患のリスクを、リスクが上昇する要因とリスクが低下する要因の両面を総合して判定することができる。

【0053】

上記のようにして得られる判定結果は、本疾患の専門医が本疾患を診断する際の補助としても用いられる。また、上記のようにして判定した本疾患のリスクと、対象者からのアンケート結果とに基づいて、本疾患のリスクの判定結果は補正されしてもよい。また、本疾患のリスクと、対象者からのアンケート結果とに基づいて、対象者に対して、生活改善に関するアドバイスを出力してもよい。

【0054】

本発明は、プライマーやプローブなどの検査試薬を提供することもできる。このようなプローブとしては、上記SNP部位を含み、ハイブリダイズの有無によってSNP部位の塩基の種類を判定できるプローブが挙げられる。また、プライマーとしては、上記SNP部位を増幅するためのPCRに用いることのできるプライマー、又は上記SNP部位をシーケンス解析するために用いることのできるプライマーが挙げられる。本実施形態の検査試薬はこれらのプライマーやプローブに加えて、PCR用のポリメラーゼやバッファー、ハイブリダイゼーション用試薬などを含むものであってもよい。

【実施例】

【0055】

以下、本実施形態を実施例によりさらに具体的に説明する。但し、本実施形態はこれらの実施例に限定されない。

【0056】

本SNPセットと本疾患との関連性を、以下のように検証した。

【0057】

遺伝子解析サービスの利用者7万3千人以上から、利用者の同意のもと、唾液試料と、各種疾患の罹患情報を収集した。罹患情報とは、例えば、本疾患に罹患している場合に1、罹患していない場合に0となる数値である。唾液試料から、利用者ごとの遺伝子型データを特定し、利用者の遺伝子型データと各種罹患情報とを対応付けたデータベースを構築した。このデータベースの中から、本疾患に罹患している被検者134名と、罹患してい

10

20

30

40

50

ないコントロール 134 名との症例対照セットを構築した。

【0058】

次いで、被検者及びコントロールの本 SNP セットの各 SNP の遺伝子型を、2 つのホモ接合型 (AA, BB) と、ヘテロ接合型 (AB) に分類した。そして、遺伝子型が図 2 に示す網掛けをした変換テーブルの遺伝子型と一致する場合には、 x_i の値を 1 とし、一致しない場合には 0 とし、 $x_1 \sim x_N$ を以下の数式 (1) で表されるロジスティック回帰モデルの説明変数とした。例えば、rs2853677 の場合は、遺伝子型が “TC” であるときに x_1 の値を 1 とし、遺伝子型が “TT” 又は “CC” であるときに x_1 の値を 0 とした。なお、本実施例では $N = 10$ である。また、以下の数式で表されるロジスティック回帰モデルの目的変数は、本疾患に罹患している確率を表す 0 から 1 の間の値 p (罹患情報) とした。

【数 1】

$$p = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^N \beta_i x_i)}}$$

$\alpha = 0.1$

【表 1】

rs ID	i	β_i
rs2853677	1	0.10516368
rs2268361 (AA)	2	0.18479968
rs2268361 (AG)	3	-0.37676396
rs2200733 (CC)	4	0.30807118
rs2200733 (TC)	5	-0.34566316
rs16853722 (TC)	6	-0.40194082
rs16853722 (TT)	7	0.07840097
rs3218097	8	0.22527664
rs1111875	9	0.33414788
rs1208285	10	-0.10340577

【0059】

1. AUC によるモデルの検証

本 SNP セットを用いた判定方法の精度について説明する。上記データベースから、テスト用に、利用者の遺伝子型情報と罹患情報とを対応付けたデータセットを作成した。データセットにおける各利用者の本 SNP セットの各 SNP の遺伝子型を、ホモ接合型 (AA, BB) と、ヘテロ接合型 (AB) に分類し、分類した各遺伝子型が図 2 に示す網掛けをした遺伝子型と一致する場合には、 x_i の値を 1 と評価し、一致しない場合には 0 と評価して、 $x_1 \sim x_N$ を特徴量として算出した。

【0060】

利用者毎の本 SNP セットに関する特徴量を上記ロジスティック回帰モデル (以下、「判定モデル」ともいう。) に入力し、各利用者が本疾患に罹患しているか否かを予測し、その偽陽性率と真陽性率を算出し、ROC (Receiver Operating Characteristic) 曲線と AUC (Area Under the Curve) をそれぞれ求めた。より具体的には、判定モデルについて 5 分割交差検証を行い、5 つの ROC 曲線 (ROC fold 1 から ROC fold 5) を求めて、その平均 (Mean ROC) と標準偏差 ($\pm 1 \text{ std. dev.}$) を求めた。図 3 中の破線 (Luck

）は、本疾患に罹患しているか否かをランダムに出力する場合であり、予測能力が無いモデルのROC曲線に対応している。

【0061】

また、同様にして、本SNPセットから1つのSNPを除いた各比較SNPセットに対して、上記と同様にそれぞれロジスティック回帰モデル（以下、「比較判定モデル」ともいう。）を作成した。そして、各比較SNPに関する特徴量を各比較判定モデルに入力し、各利用者が本疾患に罹患しているか否かを予測し、偽陽性率と真陽性率を算出し、ROC曲線とAUCをそれぞれ求めた。その結果を図4以降に示す。

【0062】

本SNPセットを用いて本疾患を判定した場合、AUCは 0.82 ± 0.10 であり、ランダムな出力の場合（ $AUC = 0.5$ ）と比べて有意に高く、本SNPセットを用いる判定モデルの予測能力が高いことが確認できる。

10

【0063】

一方、各比較SNPセットを用いる比較判定モデルの場合、AUCは本SNPセットを用いる場合よりも低い。従って、各比較SNPセットを用いる比較判定モデルのAUCは、ランダムな出力の場合（ $AUC = 0.5$ ）よりも高いものの、本SNPセットを用いる判定モデルのAUC（ 0.82 ± 0.10 ）よりも総じて低いことが確認できる。

【0064】

よって、本SNPセットに含まれるSNP全てを用いて判定することで、本SNPセットから1つのSNPを除いた各比較SNPセットを用いる場合よりも、本疾患に罹患しているか否かを高精度で予測できることが分かった。

20

【0065】

2. ウィルコクソンの順位和検定による検証

本SNPセットを用いる判定モデルが、各比較SNPセットを用いる比較判定モデルよりも有意に優れたモデルであることを確かめるために、ノンパラメトリック検定の一種であるウィルコクソンの順位和検定を行った。具体的には、本SNPセットを用いる判定モデルのAUCと、各比較SNPセットを用いる比較判定モデルのAUCとに差が無いという帰無仮説を設定し、有意水準を 0.01 としてウィルコクソンの順位和検定を行った。

【0066】

その結果、p値は以下のとおり最大でも 4.33×10^{-18} であり、帰無仮説が棄却されることが確認された。すなわち、本SNPセットを用いる判定モデルのAUCと、各比較SNPセットを用いる比較判定モデルのAUCとは、統計的に有意な差があり、本SNPセットを用いる判定モデルは、各比較SNPセットを用いる比較判定モデルよりも優れたモデルであるといえる。

30

検証から外したSNP

p 値

rs2853677	4.08×10^{-18}
rs2268361 (AA)	4.08×10^{-18}
rs2268361 (AG)	3.96×10^{-18}
rs2200733 (CC)	3.96×10^{-18}
rs2200733 (TC)	4.08×10^{-18}
rs16853722 (TC)	4.33×10^{-18}
rs16853722 (TT)	4.08×10^{-18}
rs3218097	3.96×10^{-18}
rs1111875	3.96×10^{-18}
rs1208285	3.96×10^{-18}

40

【0067】

上記のとおり、本実施形態の方法は、本疾患に罹患しているか否かを予測する精度が、ランダムな予測の場合の精度よりも有意に高いという効果を有する。また、本実施形態の方法は、本SNPセットの遺伝子型情報に基づく本疾患の判定の結果と、比較SNPセットの遺伝子型情報に基づく本疾患の判定の結果との間に、有意な差があるという効果を有

50

する。当該効果は、本 S N P セットの遺伝子型情報と本疾患の間に、これまで見出されていなかった潜在的な相関性があることに基づくものであると考えられる。上記で例示したロジスティック回帰モデルやその他のモデルは、本 S N P セットの遺伝子型情報を前提として、本疾患に罹患したヒトと本疾患に罹患していないヒトの遺伝子型に関するデータと罹患情報を学習データとして用いてパラメータを機械学習することなどにより得られるものである。すなわち、いずれのモデルも上記潜在的な相関性を表す一つの表現型に過ぎず、本実施形態の方法の実施において使用されるモデルの種類は特に限定されるものではない。

【産業上の利用可能性】

【 0 0 6 8 】

本発明の方法は、医療やヘルスケアに関連する分野において、本疾患のリスクを判定し、その予防および／または治療に貢献するものである。

10

20

30

40

50

【図面】

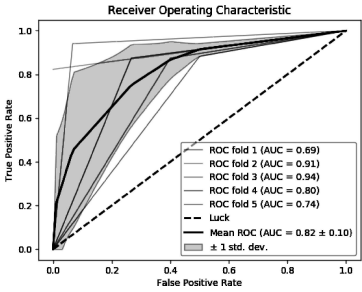
【図 1】

検証した SNP	Genotype		
	AA	AB	BB
rs2853677	TT	TC	CC
rs2268361	AA	AG	GG
rs2268361	AA	AG	GG
rs2200733	TT	TC	CC
rs2200733	TT	TC	CC
rs16853722	TT	TC	CC
rs16853722	TT	TC	CC
rs3218097	TT	TC	CC
rs1111875	AA	AG	GG
rs1208285	TT	TC	CC

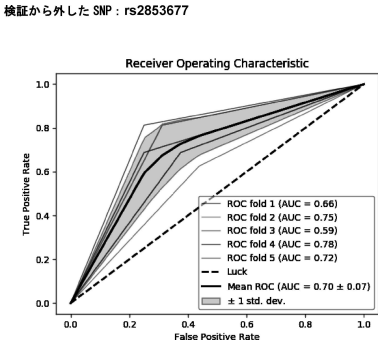
【図 2】

検証した SNP	Genotype		
	AA	AB	BB
rs2853677	TT	TC	CC
rs2268361	AA	AG	GG
rs2268361	AA	AG	GG
rs2200733	TT	TC	CC
rs2200733	TT	TC	CC
rs16853722	TT	TC	CC
rs16853722	TT	TC	CC
rs3218097	TT	TC	CC
rs1111875	AA	AG	GG
rs1208285	TT	TC	CC

【図 3】

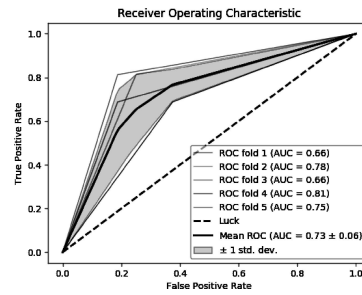


【図 4】



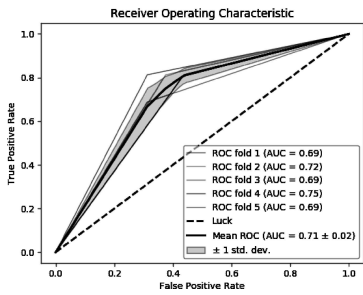
【図 5】

検証から外した SNP : rs2268361 (AA)



【図 6】

検証から外した SNP : rs2268361 (AG)



10

20

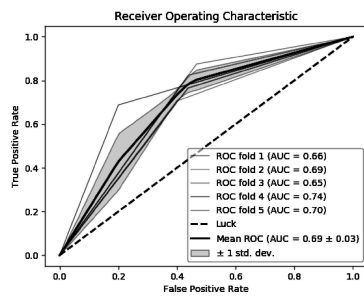
30

40

50

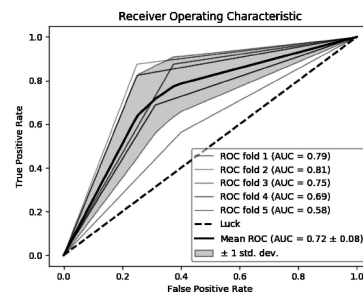
【 7 】

検証から外した SNP : rs2200733 (CC)



【 8 】

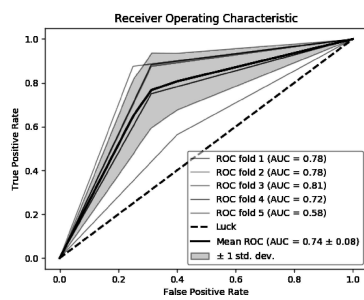
検証から外した SNP : rs2200733 (TC)



10

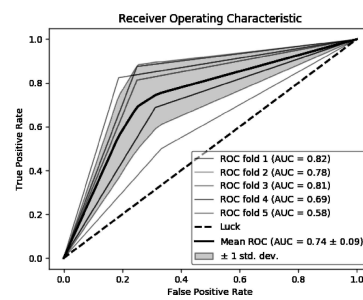
【 9 】

検証から外した SNP : rs16853722 (TC)



【 10 】

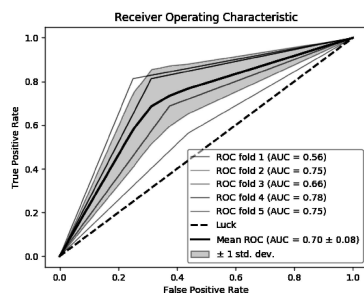
検証から外した SNP : rs16853722 (TT)



20

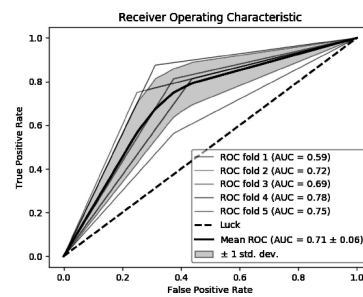
【 11 】

検証から外した SNP : rs3218097



【 12 】

検証から外した SNP : rs1111875



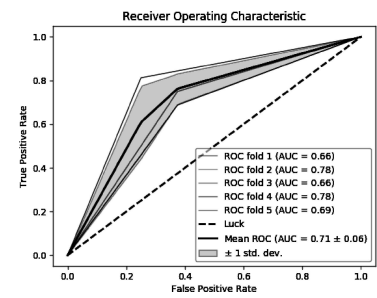
30

40

50

【 図 1 3 】

検証から外した SNP : rs1208285



10

【 配列表 】

0007137519000001.app

20

30

40

50

フロントページの続き

ジェネシスヘルスケア株式会社内

(72)発明者 黄 鶴

東京都渋谷区恵比寿四丁目 2 0 番 3 号 恵比寿ガーデンプレイスタワー 2 6 階 ジェネシスヘルスケア株式会社内

(72)発明者 佐藤 バラン 伊里

東京都渋谷区恵比寿四丁目 2 0 番 3 号 恵比寿ガーデンプレイスタワー 2 6 階 ジェネシスヘルスケア株式会社内

審査官 平林 由利子

(56)参考文献 国際公開第 2 0 1 9 / 0 2 7 0 0 5 (W O , A 1)

(58)調査した分野 (Int.Cl., D B 名)

C 1 2 Q 1 / 0 0 - 3 / 0 0

G 0 1 N 3 3 / 4 8 - 3 3 / 9 8

C 1 2 N 1 5 / 0 0 - 1 5 / 9 0

C A p l u s / M E D L I N E / E M B A S E / B I O S I S (S T N)