

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7022119号

(P7022119)

(45)発行日 令和4年2月17日(2022.2.17)

(24)登録日 令和4年2月8日(2022.2.8)

(51)国際特許分類 F I
 G 1 6 B 25/10 (2019.01) G 1 6 B 25/10

請求項の数 14 (全42頁)

(21)出願番号	特願2019-513943(P2019-513943)	(73)特許権者	596060424
(86)(22)出願日	平成29年5月30日(2017.5.30)		フィリップ・モーリス・プロダクツ・ソ
(65)公表番号	特表2019-532410(P2019-532410 A)		シエテ・アノニム
(43)公表日	令和1年11月7日(2019.11.7)		スイス国セアシュ - 2 0 0 0 ヌシャテ
(86)国際出願番号	PCT/EP2017/063073	(74)代理人	100094569
(87)国際公開番号	WO2018/050299		弁理士 田中 伸一郎
(87)国際公開日	平成30年3月22日(2018.3.22)	(74)代理人	100109070
審査請求日	令和2年6月1日(2020.6.1)		弁理士 須田 洋之
(31)優先権主張番号	62/394,551	(74)代理人	100067013
(32)優先日	平成28年9月14日(2016.9.14)		弁理士 大塚 文昭
(33)優先権主張国・地域又は機関	米国(US)	(74)代理人	100086771
			弁理士 西島 孝喜
		(74)代理人	100109335
			弁理士 上杉 浩

最終頁に続く

(54)【発明の名称】 個人の生物学的ステータスを予測するためのシステム、方法および遺伝子シグネチャ

(57)【特許請求の範囲】

【請求項1】

対象から取得したサンプルから該対象の喫煙ステータスを予測するための、コンピュータ実装された方法であって、

少なくとも一つのハードウェアプロセッサを含むコンピュータシステムによって、前記サンプルと関連付けられるデータセットを受け取ることであって、前記データセットは、全ゲノムより少ない遺伝子のセットに対する定量的な発現データを含み、前記遺伝子のセットは、A H R R、C D K N 1 C、L R R N 3、P I D 1、G P R 1 5、S A S H 1、C L E C 1 0 A、L I N C 0 0 5 9 9、P 2 R Y 6、D S C 2、F 2 R、S E M A 6 BおよびT L R 5を含む、受け取ることと、

前記少なくとも一つのハードウェアプロセッサによって、前記受け取ったデータセットの中の前記遺伝子のセットに対する前記定量的な発現データに基づいてスコアを生成することであって、前記スコアは、A H R R、C D K N 1 C、L R R N 3、P I D 1、G P R 1 5、S A S H 1、C L E C 1 0 A、L I N C 0 0 5 9 9、P 2 R Y 6、D S C 2、F 2 R、S E M A 6 BおよびT L R 5の各々を含む40個より少ない遺伝子に基づき、前記対象の予測される喫煙ステータスを示す、生成することと、

前記スコアに基づいて、前記対象の予測される喫煙ステータスを決定することと、を含む、コンピュータ実装された方法。

【請求項2】

前記遺伝子のセットは更に、A K 8、F S T L 1、R G L 1およびV S I G 4を含む、請

求項 1 に記載のコンピュータ実装された方法。

【請求項 3】

前記遺伝子のセットは更に、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HGおよびPTGFRNを含む、請求項 1～2のいずれかに記載のコンピュータ実装された方法。

【請求項 4】

前記スコアは、前記データセットに適用される分類スキームの結果であり、前記分類スキームは、前記データセットの中の前記定量的な発現データに基づいて決定される、請求項 1～3のいずれかに記載のコンピュータ実装された方法。

【請求項 5】

AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5の各々に対して、倍率変化値を演算することを更に含む、請求項 1～4のいずれかに記載のコンピュータ実装された方法。

【請求項 6】

各演算された倍率変化値のそれぞれが、少なくとも二つの独立した母集団データセットに対する所定の閾値を超えることを要する少なくとも一つの基準を、各倍率変化値が満たすと決定することを更に含む、請求項 5 に記載のコンピュータ実装された方法。

【請求項 7】

前記遺伝子のセットは、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5から成る、請求項 1 に記載のコンピュータ実装された方法。

【請求項 8】

少なくとも一つのプロセッサを備えるコンピュータ化したシステムで実行されるとき、請求項 1～7のいずれかに記載の前記方法の一つ以上の工程を前記プロセッサに実施させる、コンピュータ可読指示を含む、コンピュータプログラム製品。

【請求項 9】

個人の喫煙者ステータスを予測するためのキットであって、40個より少ない遺伝子を有する遺伝子シグネチャにおいて、遺伝子の発現レベルを検出するように構成された試薬のセットであって、前記遺伝子シグネチャは、試験サンプルの中にAHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5の各々を含む、試薬のセットを備えるキット。

【請求項 10】

前記遺伝子シグネチャは、さらに、AK8、FSTL1、RGL1およびVSIG4を含む、請求項 9 に記載のキット。

【請求項 11】

前記遺伝子シグネチャは、さらに、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HGおよびPTGFRNを含む、請求項 9 又は 10 に記載のキット。

【請求項 12】

喫煙製品の代替品の前記個人に対する効果を評価するために使用される、請求項 9 から 11のいずれか一つに記載のキット。

【請求項 13】

前記喫煙製品の代替品の前記代替品は、加熱煙草製品である、請求項 12 に記載のキット。

【請求項 14】

前記代替品の前記個人に対する前記効果は、前記個人を非喫煙者として分類する、請求項 12 又は 13 に記載のキット。

【発明の詳細な説明】

【技術分野】

10

20

30

40

50

【 0 0 0 1 】

関連出願の相互参照

本出願は、米国特許法 1 1 9 条の下、2 0 1 6 年 9 月 1 4 日に出願した米国仮特許出願第 6 2 / 3 9 4 , 5 5 1 号の利益を主張し、全体を参照することによって本明細書に援用する。本出願は、2 0 1 4 年 1 2 月 1 1 日に出願した P C T 出願第 P C T / E P 2 0 1 4 / 0 7 7 4 7 3 号、および 2 0 1 4 年 8 月 1 2 日に出願した P C T 出願第 P C T / E P 2 0 1 4 / 0 6 7 2 7 6 号に関係し、各出願は、全体を参照することによって本明細書に援用される。

【背景技術】

【 0 0 0 2 】

人間は、有害な分子変化を誘発する場合がある、外部からの毒物（例えば、たばこの煙、農薬）に絶えずさらされている。2 1 世紀の毒性学の観点におけるリスク評価は、毒性のメカニズムの解明、および高スループットデータからの曝露反応に関するマーカーの特定を頼りにしている。効率を向上し、曝露反応評価に対してよりデータ駆動型である手法を提供するように、全ゲノムマイクロアレイなど、新技術が毒性試験に取り込まれてきた。マイクロアレイおよび R N A シークエンシングなどの高スループット技術によって、多くの試験済み実験条件下でトランスクリプトームの断片が提供されるため、それらの技術の出現と共に、転写性の遺伝子調節のゲノムスケールでの推論が可能になってきている。

【 0 0 0 3 】

生物医学学会は概して、疾患診断のためのロバストなシグネチャの発見に関心がある。疾患の分子レベルにおける分類が、形態学的分類よりも正確な場合があるという根拠がある。しかしながら、曝露の原発部位（例えば、煙または大気汚染物質曝露の場合は気道）からのサンプル獲得は、大抵侵襲的であり、そのため曝露の評価および監視には都合が悪い。低侵襲の代替法として、全身性バイオマーカーを定着させるように、末梢血サンプリングが一般集団で採用され得る。血液は、含有する多くの異なる細胞亜集団から、分析するのが複雑である。しかしながら、血液は、より直接的に毒物に曝露されるすべての器官の中を循環し、容易にアクセスできるため、マーカー同定を調査するのに非常に関係の深い組織である。その上に、組織学的異常が目に見えないときでさえも、煙曝露への分子反応を検出し得る。

【発明の概要】

【課題を解決するための手段】

【 0 0 0 4 】

個人の喫煙者ステータスを予測するために使用し得る、ロバストな血液に基づく遺伝子シグネチャを特定する、クラウドソーシング法を使用するための演算システムおよび方法が提供される。本明細書に記述する遺伝子シグネチャは、現在喫煙している対象と、喫煙したことがない対象とを区別できるようにすることによって、個人の喫煙者ステータスを正確に予測できる。

【 0 0 0 5 】

ある態様では、本開示のシステムおよび方法は、対象から取得したサンプルを評価するためのコンピュータ実装された方法を提供する。コンピュータ実装された方法は、少なくとも一つのハードウェアプロセッサを含むコンピュータシステムによって、サンプルと関連付けられるデータセットを受け取ることを含む。データセットは、全ゲノムより少ない遺伝子のセットに対する定量的な発現データを含み、遺伝子のセットは、A H H R、C D K N 1 C、L R R N 3、P I D 1、G P R 1 5、S A S H 1、C L E C 1 0 A、L I N C 0 0 5 9 9、P 2 R Y 6、D S C 2、F 2 R、S E M A 6 B および T L R 5 を含む。少なくとも一つのハードウェアプロセッサは、受け取ったデータセットの中の遺伝子のセットに対する定量的な発現データに基づいてスコアを生成し、スコアは、4 0 個より少ない遺伝子に基づき、対象の予測される喫煙ステータスを示す。

【 0 0 0 6 】

ある実装では、遺伝子のセットは更に、A K 8、F S T L 1、R G L 1 および V S I G 4

10

20

30

40

50

を含む。ある実装では、遺伝子のセットは更に、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HGおよびPTGFRNを含む。

【0007】

ある実装では、スコアは、データセットに適用される分類スキームの結果であり、分類スキームは、データセットの中の定量的な発現データに基づいて決定される。ある実装では、コンピュータ実装された方法は更に、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5の各々に対して、倍率変化値を演算することを含む。コンピュータ実装された方法は更に、各演算された倍率変化値のそれぞれが、少なくとも二つの独立した母集団データセットに対する所定の閾値を超えることを要する少なくとも一つの基準を、各倍率変化値が満たすと決定することを含んでもよい。

10

【0008】

ある実装では、遺伝子のセットは、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5から成る。

【0009】

ある態様では、本開示のシステムおよび方法は、個人の喫煙者ステータスを予測するためのキットを提供する。キットは、40個より少ない遺伝子を有する遺伝子シグネチャに、遺伝子の発現レベルを検出する、試薬のセットであって、遺伝子シグネチャは、試験サンプルの中にAHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5を含む、試薬のセットと、個人の喫煙者ステータスを予測するキットを使用するための説明書とを含む。

20

【0010】

ある実装では、キットは、喫煙製品の代替品の個人に対する効果を評価するために使用される。喫煙製品の代替品は、加熱式たばこ製品を含んでもよい。代替品の個人に対する効果は、個人を非喫煙者として分類することであってもよい。ある実装では、遺伝子シグネチャは更に、AK8、FSTL1、RGL1およびVSI4を含む。ある実装では、遺伝子シグネチャは更に、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HGおよびPTGFRNを含む。

30

【0011】

ある態様では、本開示のシステムおよび方法は、対象から取得したサンプルを評価するためのコンピュータ実装された方法を提供する。コンピュータ実装された方法は、少なくとも一つのハードウェアプロセッサを含むコンピュータシステムによって、サンプルと関連付けられるデータセットを受け取ることを含み、データセットは、全ゲノムより少ない遺伝子のセットに対する定量的な発現データを含み、遺伝子のセットは、LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63を含む。少なくとも一つのハードウェアプロセッサは、受け取ったデータセットの中の遺伝子のセットに対する定量的な発現データに基づいてスコアを生成し、スコアは、40個より少ない遺伝子に基づき、対象の予測される喫煙ステータスを示す。

40

【0012】

ある実装では、スコアは、データセットに適用される分類スキームの結果であり、分類スキームは、データセットの中の定量的な発現データに基づいて決定される。

【0013】

ある実装では、少なくとも一つのハードウェアプロセッサは、LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63の各々に対して

50

、倍率変化値を演算する。コンピュータ実装された方法は更に、各演算された倍率変化値のそれぞれが、少なくとも二つの独立した母集団データセットに対する所定の閾値を超えることを要する少なくとも一つの基準を、各倍率変化値が満たすと決定することを含んでもよい。

【0014】

ある実装では、遺伝子のセットは、LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63から成る。

【0015】

ある態様では、本開示のシステムおよび方法は、個人の喫煙者ステータスを予測するためのキットを提供する。キットは、40個より少ない遺伝子を有する遺伝子シグネチャに、遺伝子の発現レベルを検出する、試薬のセットであって、遺伝子シグネチャは、試験サンプルの中にLRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63を含む、試薬のセットと、個人の喫煙者ステータスを予測するキットを使用するための説明書とを備える。

10

【0016】

ある実装では、キットは、喫煙製品の代替品の個人に対する効果を評価するために使用される。喫煙製品の代替品は、加熱式たばこ製品を含んでもよい。代替品の個人に対する効果は、個人を非喫煙者として分類することであってもよい。

20

【0017】

ある態様では、本開示のシステムおよび方法は、生物学的ステータスを予測するために遺伝子シグネチャを取得する、コンピュータ実装された方法を提供する。コンピュータ実装された方法は、通信ポートと、訓練データセットおよび試験データセットを含む少なくとも一つの電子データベースを記憶する、少なくとも一つの非一時的コンピュータ可読媒体と通信する少なくとも一つのコンピュータプロセッサとを含む、コンピュータシステムによって、訓練データセットをネットワークで、複数のユーザー装置へ提供することを含む。訓練データセットは、訓練サンプルのセットを含み、試験データセットは、試験サンプルのセットを含む。各訓練サンプルおよび各試験サンプルは、遺伝子発現データを含み、生物学的ステータスのセットより選択される、既知の生物学的ステータスを有する患者に対応する。コンピュータ実装された方法は更に、ネットワークから、訓練データセットに基づいて分類子を取得することによって各々生成する、候補遺伝子シグネチャを受け取ることを含み、各候補遺伝子シグネチャは、訓練データセットの中で異なる生物学的ステータスを判別するように決定される、遺伝子のセットを含む。試験サンプルの既知の生物学的ステータスを予測するとき、それぞれの候補遺伝子シグネチャの性能に基づいて、それぞれの候補遺伝子シグネチャ各々へ、スコアが割り当てられる。候補遺伝子シグネチャのサブセット（または候補遺伝子シグネチャのセット全体を含んでもよい、候補遺伝子シグネチャの一部）は、割り当てられたスコアに基づいて特定され、少なくとも閾値数の候補遺伝子シグネチャに含まれていた遺伝子は、サブセットの中で特定される。特定された遺伝子は、遺伝子シグネチャとして記憶される。

30

40

【0018】

ある実装では、コンピュータ実装された方法は更に、複数のユーザー装置へ、各候補遺伝子シグネチャの中で許容される遺伝子の最大閾値数を表す数字を提供することを含む。

【0019】

ある実装では、コンピュータ実装された方法は更に、試験データセットの一部をネットワークで、複数のユーザー装置へ提供することを含み、試験データセットの一部は、既知の生物学的ステータスを有する患者に対する遺伝子発現データを含み、患者の既知の生物学的ステータスを含まない。コンピュータ実装された方法は更に、各候補遺伝子シグネチャについて、試験データセットの中の各サンプルの信頼水準を受け取ることを含む。信頼水準は、試験データセットの中のサンプルが、生物学的ステータスのうちの一つに属す

50

ると予測される尤度を示す値であってもよい。スコアは、信頼水準に少なくとも一部基づいてもよい。特に、スコアは、試験データセットの中の信頼水準、および患者の既知の生物学的ステータスより演算される、適合率 - 再現率下面積 (area under the precision recall: AUPR) 測定基準に少なくとも一部基づいてもよい。

【0020】

ある実装では、スコアは、対応する候補遺伝子シグネチャが、試験データセットの中の患者の既知の生物学的ステータスと一致する予測を提供するかに少なくとも一部基づく。対応する候補遺伝子シグネチャが、試験データセットの中の患者の既知の生物学的ステータスと一致する予測を提供するかは、マッシュアップ相関係数 (MCC) を使用して決定されてもよい。

10

【0021】

ある実装では、候補遺伝子シグネチャは、各候補遺伝子シグネチャに対して一位および二位を取得するように、少なくとも二つの異なる測定基準に従ってランク付けされる。各候補遺伝子シグネチャに対する一位および二位は、それぞれの候補遺伝子シグネチャ各々に対してスコアを取得するように平均化されてもよい。

【0022】

ある実装では、生物学的ステータスのセットは喫煙者ステータスを含む。喫煙者ステータスは、現喫煙者および非喫煙者を含んでもよい。

20

【0023】

ある実装では、遺伝子シグネチャは、全ゲノムより少なく、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6BおよびTLR5を含む。加えて、遺伝子シグネチャは更に、AK8、FSTL1、RGL1およびVSI G4を含んでもよい。加えて、遺伝子シグネチャは更に、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HGおよびPTGFRNを含んでもよい。加えて、遺伝子シグネチャは更に、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3およびZNF618を含んでもよい。一部の实装では、遺伝子シグネチャは、10個、15個、20個、25個、30個、35個、40個、または全ゲノムの中の遺伝子の数より少ない、いかなる他の好適な数の遺伝子など、遺伝子の閾値数に限定されてもよい。

30

【0024】

ある実装では、遺伝子シグネチャは、全ゲノムより少なく、LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63を含む。加えて、遺伝子シグネチャは更に、DSC2、TLR5、RGL1、FSTL1、VSI G4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1およびGUCY1B3を含んでもよい。一部の实装では、遺伝子シグネチャは、10個、15個、20個、25個、30個、35個、40個、または全ゲノムの中の遺伝子の数より少ない、いかなる他の好適な数の遺伝子など、遺伝子の閾値数に限定されてもよい。

40

【0025】

ある実装では、遺伝子シグネチャは、全ゲノムより少なく、AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1およびTBX21を含む。一部の

50

実装では、遺伝子シグネチャは、10個、15個、20個、25個、30個、35個、40個、または全ゲノムの中の遺伝子の数より少ない、いかなる他の好適な数の遺伝子など、遺伝子の閾値数に限定されてもよい。

【0026】

ある態様では、本開示のシステムおよび方法は、対象から取得したサンプルを評価するためのコンピュータ実装された方法を提供する。コンピュータ実装された方法は、少なくとも一つのハードウェアプロセッサを含むコンピュータシステムによって、サンプルと関連付けられるデータセットを受け取ることを含む。データセットは、全ゲノムより少ない遺伝子のセットに対する定量的な発現データを含み、遺伝子のセットは、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3およびZNF618を含む。少なくとも一つのハードウェアプロセッサは、受け取ったデータセットに基づいてスコアを生成し、スコアは、対象の予測される喫煙ステータスを示す。

10

【0027】

ある実装では、スコアは、データセットに適用される分類スキームの結果であり、分類スキームは、データセットの中の定量的な発現データに基づいて決定される。

20

【0028】

ある実装では、コンピュータ実装された方法は更に、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3およびZNF618の各々に対して、倍率変化値を演算することを含む。コンピュータ実装された方法は更に、各演算された倍率変化値のそれぞれが、少なくとも二つの独立した母集団データセットに対する所定の閾値を超えることを要する少なくとも一つの基準を、各倍率変化値が満たすと決定することを含んでもよい。

30

【0029】

ある実装では、遺伝子のセットは、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GUCY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3およびZNF618から成る。

40

【0030】

ある態様では、本開示のシステムおよび方法は、個人の喫煙者ステータスを予測するためのキットを提供する。キットは、試験サンプルの中の遺伝子シグネチャに遺伝子の発現レベルを検出する、試薬のセットであって、遺伝子シグネチャは、AHHR、CDKN1C、LRRN3、PID1、GPR15、SASH1、CLEC10A、LINC00599、P2RY6、DSC2、F2R、SEMA6B、TLR5、AK8、FSTL1、RGL1、VSIG4、C15orf54、CTTNBP2、RANK1、GSE1、GU

50

CY1A3、LOC200772、MARC2、MIR4697HG、PTGFRN、ASGR2、B3GALT2、CYP4F22、FUCA1、GPR63、GUCY1B3、MB21D2、NLK、NR4A1、P2RY1、PF4、PTGFR、SH2D1B、ST6GALNAC1、TMEM163、TPPP3およびZNF618を含む、試薬のセットと、個人の喫煙者ステータスを予測するキットを使用するための説明書とを備える。

【0031】

ある実装では、キットは、喫煙製品の代替品の個人に対する効果を評価するために使用される。喫煙製品の代替品は、加熱式たばこ製品を含んでもよい。代替品の個人に対する効果は、個人を非喫煙者として分類することであってもよい。

10

【0032】

ある態様では、本開示のシステムおよび方法は、対象から取得したサンプルを評価するためのコンピュータ実装された方法を提供する。コンピュータ実装された方法は、少なくとも一つのハードウェアプロセッサを含むコンピュータシステムによって、サンプルと関連付けられるデータセットを受け取ることを含み、データセットは、全ゲノムより少ない遺伝子のセットに対する定量的な発現データを含み、遺伝子のセットは、AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1およびTBX21を含む。少なくとも一つのハードウェアプロセッサは、受け取ったデータセットの中の遺伝子のセットに対する定量的な発現データに基づいてスコアを生成し、スコアは、40個より少ない遺伝子に基づき、対象の予測される喫煙ステータスを示す。

20

【0033】

ある実装では、スコアは、データセットに適用される分類スキームの結果であり、分類スキームは、データセットの中の定量的な発現データに基づいて決定される。

【0034】

ある実装では、コンピュータ実装された方法は更に、AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1およびTBX21の各々に対して、倍率変化値を演算することを含む。コンピュータ実装された方法は更に、各演算された倍率変化値のそれぞれが、少なくとも二つの独立した母集団データセットに対する所定の閾値を超えることを要する少なくとも一つの基準を、各倍率変化値が満たすと決定することを含んでもよい。

30

【0035】

ある実装では、遺伝子のセットは、AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1およびTBX21から成る。

【0036】

ある態様では、本開示のシステムおよび方法は、個人の喫煙者ステータスを予測するためのキットを提供する。キットは、試験サンプルの中の遺伝子シグネチャに遺伝子の発現レベルを検出する、試薬のセットであって、遺伝子シグネチャは、AHHR、P2RY6、KLRG1、LRRN3、COX6B2、CTTNBP2、DSC2、F2R、GUCY1B3、MT2、NGFRAP1、REEP6、SASH1およびTBX21を含み、遺伝子シグネチャは、40個より少ない遺伝子を含む、試薬のセットと、個人の喫煙者ステータスを予測するキットを使用するための説明書とを備える。

40

【0037】

ある実装では、キットは、喫煙製品の代替品の個人に対する効果を評価するために使用される。喫煙製品の代替品は、加熱式たばこ製品を含んでもよい。代替品の個人に対する効果は、個人を非喫煙者として分類することであってもよい。

【図面の簡単な説明】

【0038】

50

開示の更なる特徴、その性質および様々な利点は、全体を通して同様の参照文字が同様の一部を指す添付の図面と併せて、

以下の発明を実施するための形態を考慮することにより明らかになるであろう。

【0039】

【図1】図1は、クラウドソーシングを使用して、遺伝子シグネチャの特定を遂行するための、コンピュータ化したシステムのブロック図である。

【0040】

【図2】図2は、本明細書に記載するコンピュータ化したシステムのいずれかに、構成要素のいずれかを実装するために使用される場合がある、例示的なコンピューティング装置のブロック図である。

10

【0041】

【図3】図3は、個人の生物学的ステータスを予測するため、遺伝子シグネチャを特定するように、クラウドソーシングを使用するプロセスのフローチャートである。

【0042】

【図4】図4Aおよび4Bは、ヒトデータ(図4A)および種に依存しないデータ(図4B)に対する、異なるチーム間の共起を示す表である。

【0043】

【図5】図5は、対象の予測される喫煙ステータスを示すスコアを評価するための、プロセスのフローチャートである。

【0044】

【図6】図6は、異なる研究について、サンプル群/クラス、サイズおよび特性を要約する表である。

20

【0045】

【図7】図7Aは、ヒトおよびマウスの全血遺伝子発現データから、化学物質の曝露反応マーカーを特定することと、新規血液サンプルを曝露または非曝露群の一部として予測分類するために、これらのマーカーを演算モデルでシグネチャとして活用することとを示す図である。

【0046】

図7Bは、(i)喫煙者と現非喫煙者とを識別(課題1)し、続いて(ii)現非喫煙者を、喫煙経験者および喫煙未経験者と分類する(課題2)、ロバストでスパースなヒト(サブチャレンジ1、SC1)および種に依存しない(サブチャレンジ2、SC2)血液を基にした遺伝子シグネチャ分類モデルの開発を示す図である。

30

【0047】

【図8】図8は、血液遺伝子発現データの訓練データセット、試験データセットおよび検証データセットの公開を示す図である。

【0048】

【図9】図9Aは、喫煙者と非喫煙者との明らかな分離を示す箱ひげ図である。

【0049】

図9Bは、喫煙群に対して0日および5日の譲渡の間に有意な差を示さないが、0日のそれぞれのベースラインと比較しすると、C e s s群およびS w i t c h群に対して有意な減少を示す、二つの箱ひげ図を含む。

40

【0050】

【図10】図10は、クラス予測のために、遺伝子シグネチャ分類モデルのクラス予測性能を示す、二つの表を含む。

【0051】

【図11A】図11Aおよび11Bは、試験および検証データセットに対する、参加者による血液サンプルクラス予測を示す、箱ひげ図である。

【図11B】同上。

【0052】

【図12】図12は、検証データセットに対する、閉じ込められた0日目と5日目との間

50

の集団の対数オッズ比を示す、箱ひげ図を含む。

【0053】

【図13】図13は、群ノクラスごと、およびpMRTPもしくは候補MRTPへの曝露時、またはpMRTPもしくは候補MRTPへの切り替え後に分けられた集団の対数オッズ分布を示す、箱ひげ図である。

【0054】

【図14A】図14および15は、MLを基にしたクラス予測で、長さ2から18のシグネチャの可能な全組み合わせの性能を検討する、MCCおよびAUPRスコアのプロットである。

【図14B】同上。

【図14C】同上。

【図15A】同上。

【図15B】同上。

【図15C】同上。

【発明を実施するための形態】

【0055】

個人の生物学的ステータスを予測するために使用し得る、ロバストな遺伝子シグネチャを特定するための、演算システムおよび方法を本明細書に記載する。特に、生物学的ステータスは、個人の喫煙曝露反応ステータスに対応してもよい。本明細書に記載する遺伝子シグネチャは、現在喫煙している対象を、喫煙したことがない対象、または喫煙をやめた対象と区別することができる。本明細書に記載する実施例は、主に喫煙者ステータスまたは喫煙曝露反応ステータスに関係する一方、当業者は、本開示のシステムおよび方法は、個人の生物学的ステータスを予測するため遺伝子シグネチャを特定するように、クラウドソーシング手法の使用に適用できることを理解するであろうし、生物学的ステータスは、喫煙曝露反応ステータス、喫煙者ステータス、疾患ステータス、生理学的状態、化学物質への曝露状態、または個人の生物学的データと関連付けられる、個人のいかなる他の好適なステータスもしくは状態を指してもよい。

【0056】

本明細書で使用する通り、個人の生物学的ステータスは、疾病で、または一つ以上の毒物、薬物、環境変化（例えば、温度、微小重力、圧力および放射など）、もしくはそれらのいかなる好適な組み合わせへの曝露に応じて生成されてもよい、様々な分子変化を表してもよい。基準は、予測分類モデルに対して定義され、予測分類モデルの開発および訓練のために、コンピュータ分析で使用される。クラスを識別する特徴が抽出され、クラス予測用の分類モデルに埋め込まれる。本明細書に使用される通り、分類子は、クラス予測に使用される、判別特徴および規則を含む。

【0057】

本明細書に記載するクラウドソーシング手法は、個人の一つ以上の化学物質への曝露ステータスを予測するよう、ロバストな遺伝子シグネチャを特定するのに使用されてもよい。下の実施例1に関して記載する研究は、個人の煙への曝露を予測するために、遺伝子シグネチャを特定する一つのようなクラウドソーシング手法の例示的図解を伴う。下に記載する実施例1の研究では、集団（例えば、複数のチャレンジ参加者）から取得される、ヒトの血液を基とする喫煙曝露反応遺伝子シグネチャの遺伝子リスト、および集団から取得される、種に依存しない血液を基とする喫煙曝露反応遺伝子シグネチャの遺伝子リストの両方を特定する。本明細書に記載する遺伝子シグネチャは、個人が煙に曝露されていたか否かを予測するように、新規の人（ヒトシグネチャ）またはヒトおよび齧歯類（種に依存しないシグネチャ）の血液遺伝子発現サンプルデータに適用されてもよい、一つ以上の分類モデルに適用されてもよい。本明細書に記載するシステムおよび方法は、個人が一つ以上の化学物質に曝露されてきたか否かを予測するために、遺伝子シグネチャおよび一つ以上の分類モデルを特定するよう拡張されてもよい。下の実施例1に関して記載する研究は、血液を基とする遺伝子シグネチャの特定に関係する一方、当業者は、本開示のシステ

10

20

30

40

50

ムおよび方法が、血液のみに基づかない遺伝子シグネチャを特定するように、クラウドソーシング手法の使用に適用可能であることを理解するであろう。代わりに、本開示は、例えば、タンパク質およびメチル化変化など、組織および他の特徴に基づく、遺伝子シグネチャの特定に適用可能である。

【0058】

本開示のシステムおよび方法は、毒物への曝露を予測できるマーカーを特定するように使用されてもよい。実際に、新規サンプルに適用される、ロバストなマーカーに基づく分類モデルによって、(i) 対象が化学物質に曝露していたか、またはしていなかったかの予測が可能になり、(ii) 製品の試験または離脱中に、曝露反応の大きさを経過観察することが可能になってもよい。

10

【0059】

本明細書で使用する通り、「ロバスト」な遺伝子シグネチャは、研究、臨床検査、サンプル源および他の人口統計学的因子にわたって、強い性能を維持するものである。ロバストなシグネチャは、大きな個人差を含む母集団データの1セットであってさえも検出可能であるべきことが重要である。データセットにわたるロバスト性は、シグネチャの性能についての過度の楽観的な報告を避けるためにも、適切に検査されるべきである。

【0060】

システム生物学は、生物システムが、外部刺激（例えば、薬物、栄養および温度）および遺伝子改変（例えば、変異、エピジェネティック修飾）に反応または適応する、メカニズムの詳細な理解を生み出すことを目的とする。新しいメカニズムに関する洞察は、オミクスまたはハイコンテンツスクリーニングなど、先進技術を使用して生成する、大量の分子および機能データの分析および統合を通じて獲得される。毒性学の分野に適用される場合、システム毒性学と呼ばれる全体手法によって、生体異物（例えば、農薬、化学物質）によりトリガーされる生物システムの動揺を定量化し、毒性作用様式を解明し、関連するリスクを検討することが可能になる。システム毒性学は、短期的な知見から長期的な成果を推定し、実験系より特定される潜在的风险をヒトへ翻訳する将来性を有し、それを応用することがリスク評価および意思決定の新しい標準になり得ると示唆する。予測される毒物学的成果およびリスク見積に対する推定および翻訳だけでなく、システム毒性学データの分析も、先進的な演算方法論の開発に必要とされる。新規演算手法の性能および信頼性の向上を実証するために、研究者は、それらの技法を最先端の方法に対して評価するが、偏った検討をもたらす、いわゆる「自己評価の罠」に陥る場合がしばしばある。さらに、システム生物学/毒性学で生成し分析するデータの氾濫が、公表される結果および結論の審査を、査読者にとって退屈なものにする。再評価者は、原則として公共のリポジトリに記憶されている未加工データにアクセスし得るものの、自身で全体の分析を再現するのはしばしば困難である。そのため、外部の第三者が関与する、方法およびデータの独立した客観的検討または検証の必要性が明確に存在する。本開示のシステムおよび方法は、この必要性に対処し、研究者からの提出を受け取り、優良技法を特定し、生物学的ステータスを予測するため、ロバストな遺伝子シグネチャを作り出すように、それらの成果を集約するクラウドソーシング手法を提供する。

20

30

【0061】

図1は、本明細書に開示するシステムおよび方法を実装するために使用される場合がある、コンピュータネットワークおよびデータベース構造の例を描写する。図1は、図解の実装に従い、クラウドソーシングを使用して、遺伝子シグネチャの特定を遂行するための、コンピュータ化したシステム100のブロック図である。システム100は、サーバ104と、コンピュータネットワーク102上でサーバ104に接続される二つのユーザー装置108aおよび108b（概して、ユーザー装置108）とを含む。サーバ104はプロセッサ105を含み、各ユーザー装置108は、プロセッサ110aまたは110bおよびユーザーインターフェース112aまたは112bを含む。本明細書で使用する通り、「プロセッサ」または「コンピューティング装置」という用語は、本明細書に記載するコンピュータ化された技法のうちの一つ以上を実施するために、ハードウェア、ファーム

40

50

ウェアおよびソフトウェアで構成される、一つ以上のコンピュータ、マイクロプロセッサ、論理装置、サーバまたは他の装置を指す。プロセッサおよび処理装置はまた、入力、出力および現在処理しているデータを記憶するための一つ以上のメモリ装置を含んでもよい。本明細書に記載するプロセッサおよびサーバのうちのいずれかを実装するように使用されてもよい、図解のコンピューティング装置 200 について、図 2 を参照して下に詳細に記載する。本明細書で使用する通り、「ユーザーインターフェース」は、一つ以上の入力装置（例えば、キーボード、タッチスクリーン、トラックボール、音声認識システムなど）および/または一つ以上の出力装置（例えば、視覚表示、スピーカ、触覚ディスプレイ、印刷装置など）のいかなる好適な組み合わせを含むが、これらに限定されない。本明細書で使用する通り、「ユーザー装置」は、本明細書に記載する、一つ以上のコンピュータ

10

化された作用または技法を実施するためのハードウェア、ファームウェアおよびソフトウェアで構成される、一つ以上の装置のいかなる好適な組み合わせを含むが、これらに限定されない。ユーザー装置の例としては、パーソナルコンピュータ、ノートパソコンおよびモバイルデバイス（例えば、スマートフォン、タブレットコンピュータなど）を含むが、これらに限定されない。図面を複雑にするのを避けるために、一つのサーバ、一つのデータベースおよび二つのユーザー装置のみを図 1 に示すが、当業者は、システム 100 が複数のサーバ、および任意の数のデータベースまたはユーザー装置をサポートする場合がありますことを理解するであろう。

【0062】

コンピュータ化したシステム 100 は、個人の生物学的ステータスを予測するために遺伝子シグネチャを特定するとき、クラウドの英知を活用するように使用されてもよい。上に記載した通り、システム生物学を研究する科学者は、偏った検討をもたらす自己評価の罠にしばしば陥る。本明細書に記載するクラウドソーシング手法は、チャレンジを設計し、科学界へ公開し（例えば、遺伝子発現に関するデータ、および既知の生物学的ステータスデータベース 106 を、ユーザー装置 108 で利用可能にすることによって）、独立した科学者またはグループから提出を受け取り（例えば、ユーザー装置 108 a および 108 b から）、優良な結果または予測を集約することによって、これらのバイアスを避けるのに役立つ。幅広い参加を保証するために、チャレンジは、個人の生物学的ステータスまたは喫煙者ステータスを予測するために、血液を基とする遺伝子シグネチャを特定するなど、共通の関心である科学的諸問題に関係する論題に対処することを目的とする。

20

30

【0063】

チャレンジによって、個体群から取得された血液サンプルデータと関連付けられるあるデータが、科学界で利用可能になる。特に、遺伝子発現および既知の生物学的ステータスデータベース 106（概して、データベース 106）は、個人のセットの既知の生物学的ステータスを表すデータ、および遺伝子発現データ（患者のセットからの血液サンプルから取得される）を含む、データベースである。個人（その血液サンプルデータがデータベース 106 に記憶されている）のセットの中の各個人は、無作為に訓練サンプルまたは試験サンプルとして割り当てられてもよい。一部の実装では、個人の訓練または試験サンプルとしての割り当ては、完全には無作為でなくてもよい。この場合、異なる生物学的ステータスを持つ、類似の数の個人が、訓練および試験データセットの各々の中にあることを保証するなど、一つ以上の基準が、割り当て中に使用されてもよい。概して、いかなる好適な方法が、個人を訓練または試験サンプルとして割り当てるように使用されてもよく、一方で、生物学的ステータスの分布が、訓練データセットおよび試験データセットにおいて少々類似していることを保証する。

40

【0064】

各訓練サンプルおよび試験サンプルは、既知である個人の生物学的ステータス（例えば、既知である個人の喫煙者ステータス）だけでなく、個人の血液サンプルから測定される遺伝子発現レベルも含む。訓練サンプルは訓練データセットを構成し、試験サンプルは試験データセットを構成する。全体の訓練データセットが、データベース 106 からユーザー装置 108 へ提供され、一方試験データセットの一部分のみがユーザー装置 108 へ提供

50

される。特に、試験サンプルから測定される遺伝子発現レベルは、ユーザー装置 108 へ提供されるが、試験サンプルに対応する既知の生物学的ステータスは、ユーザー装置 108 から隠されたままである。

【0065】

ユーザー装置 108 にいる科学者は、測定される遺伝子発現レベルと、訓練データセットの中の個人の生物学的ステータスとの間のいかなる依存性、関連または相関を特定するよう試みるように、訓練サンプルを分析してもよい。特定される相関は、候補遺伝子シグネチャおよび分類子の形態を有してもよい。候補遺伝子シグネチャは、異なる生物学的ステータス（例えば、喫煙者対非喫煙者）と関連付けられるサンプルに対して、異なった形で発現される遺伝子のリストを含む。科学者は、フィルター、ラッパーおよび埋め込み法など、いかなる特徴選択技法を使用して候補遺伝子シグネチャを特定するよう、いかなる好適な演算技法を使用してよい。抽出される特徴は、判別分析、サポートベクターマシン、線形回帰、ロジスティック回帰、決定木、ナイーブベイズ、k 最近傍、K 平均、ランダムフォレストまたはいかなる他の好適な技法など、機械学習の手法を使用して訓練される分類モデルに組み合わされる。分類子は、サンプルをクラスに割り当てるように、候補遺伝子シグネチャの中の遺伝子の発現レベルを使用する、決定規則またはマッピングを含み、個人の予測される生物学的ステータスを指してもよい。このように、各ユーザー装置 108 にいる各科学者は、訓練データセットに基づいて、候補遺伝子シグネチャおよび分類子を特定する。

10

【0066】

ユーザー装置 108 にいる科学者は、それらの候補遺伝子シグネチャおよび分類子を使用して、試験データセットの中の試験サンプルの生物学的ステータスを予測する。各試験サンプルに対して取得される結果だけでなく候補遺伝子シグネチャも、ユーザー装置 108 からネットワーク 102 を介してサーバ 104 へ提供される。科学者からの提出は匿名であってもよい。一例では、各試験サンプルの結果は、対応する試験サンプルが、予測される生物学的ステータスの資格があるという、尤度または確率に対応する信頼水準を含む。信頼水準については、図 3 の工程 308 に関係して詳細に記載する。別の例では、結果は、信頼水準ではなくむしろ、各試験サンプルに対して予測される生物学的ステータスのみを含む。

20

【0067】

サーバ 104 はその後、各試験サンプルに対して取得された結果と、各試験サンプルの既知の生物学的ステータスとを比較することによって、最良の候補遺伝子シグネチャを特定してもよい。概して、優良候補遺伝子シグネチャは、既知の生物学的ステータスにぴったり合致する結果を有する。サーバ 104 はその後、個人の生物学的ステータスを予測するのに使用されてもよい、ロバストな遺伝子シグネチャを取得するように、優良候補遺伝子シグネチャを集約する。このプロセスについては、図 3 の工程 314、316 および 318 に関係してより詳細に記載する。

30

【0068】

図 1 のシステム 100 の構成要素は、いくつものやり方のうちのいずれかで配設され、分散され、組み合わせられてもよい。例えば、ネットワーク 102 を介して接続される複数の処理装置および記憶装置に渡って、システム 100 の構成要素を分散するコンピュータ化されたシステムが使用されてもよい。そのような実装が、共通のネットワークリソースへのアクセスを共有する、無線および有線通信システムを含む複数の通信システムに渡り、分散コンピューティングに適切である場合がある。一部の実装では、システム 100 は、構成要素のうちの一つ以上が、インターネットまたは他の通信システムを介して接続される、異なる処理および記憶サービスによって提供される、クラウドコンピューティング環境に実装される。サーバ 104 は、例えば、クラウドコンピューティング環境でインスタンス化された、一つ以上の仮想サーバであってもよい。一部の実装では、サーバ 104 は、データベース 106 と組み合わせられて、一つの構成要素となる。

40

【0069】

50

図3は、個人の生物学的ステータスを予測するため、遺伝子シグネチャを特定するように、クラウドソーシングを使用する方法300のフローチャートである。方法300は、サーバ104によって実行されてもよく、遺伝子発現データおよび既知の生物学的ステータスを含む訓練データセットを、ユーザー装置のセットへ提供し(工程302)、遺伝子発現データを含む試験データセットを、ユーザー装置のセットへ提供し(工程304)、訓練データセットの中の異なる生物学的ステータスを判別するように決定される、遺伝子のセットを含む候補遺伝子シグネチャを受け取り(工程306)、各候補遺伝子シグネチャに対して、試験データセットの中の各サンプルに対する信頼水準を受け取る(工程308)工程を含む。方法300は更に、信頼水準と試験データセットの中の既知の生物学的ステータスとの比較に基づいて、第一性能測定基準に従い候補遺伝子シグネチャをランク付けること(工程310)と、各候補遺伝子シグネチャに対して、試験データセットの中の各サンプルを、予測される生物学的ステータスに割り当てるように、信頼水準を使用すること(工程312)と、予測される生物学的ステータスが、試験データセットの中の既知の生物学的ステータスに合致するかに基づいて、第二性能測定基準に従い候補遺伝子シグネチャをランク付けること(工程314)と、工程310および314で割り当てられたランクに基づいて、第三性能測定基準に従い候補遺伝子シグネチャをランク付けること(工程316)と、最上位にランク付けられた候補遺伝子シグネチャにおける、少なくとも閾値数の候補遺伝子シグネチャに含まれる遺伝子を特定すること(工程318)とを含む。

10

【0070】

工程302で、遺伝子発現データを含む訓練データセット、および訓練サンプルのセットに対する既知の生物学的ステータスが、ユーザー装置108のセットへ提供される。図1に関して記載するように、工程302で提供される訓練データセットは、個人の既知の生物学的ステータスだけでなく、個人の血液サンプルから測定される遺伝子発現レベルを含む、訓練サンプルを含む。ユーザー装置108にいる科学者が、訓練データセットを受け取り、測定された遺伝子発現レベルと、既知の生物学的ステータスとの間にマッピングを提供する分類子を訓練するように、訓練データセットを使用する。工程304で、遺伝子発現データを含む試験データセットが、ユーザー装置108のセットへ提供される。図1に関して記載するように、工程304で提供される試験データセットは、個人の血液サンプルから測定される遺伝子発現レベルを含むのみの試験サンプルを含むが、個人の既知の生物学的ステータスは含まない。換言すれば、試験サンプルの既知の生物学的ステータスは、ユーザー装置108にいる科学者には隠されたままである。

20

30

【0071】

工程306で、訓練データセットの中の異なる生物学的ステータスを判別するように決定される、遺伝子のセットを含む候補遺伝子シグネチャを受け取る。ユーザー装置108にいる各科学者または科学者の各チームは、候補遺伝子シグネチャをサーバ104へ提供してもよく、科学者は、候補遺伝子シグネチャの中の遺伝子発現レベルの組み合わせが、一つ以上の基準(訓練データセットの中の生物学的ステータス、またはサンプルの曝露反応ステータスなど)の判別点であると決定してきた。訓練データセットを提供するユーザー装置は、科学者が候補遺伝子シグネチャを提供するユーザー装置と同じであってもよく、または異なってもよい。

40

【0072】

工程308で、各候補遺伝子シグネチャに対して、試験データセットの中の各試験サンプルに対する信頼水準を受け取る。信頼水準は、0と1との間の値であってもよく、対応する試験サンプルがある特定の生物学的ステータスに属する尤度を表す。一例では、二つの生物学的ステータス(例えば、第一生物学的ステータスおよび第二生物学的ステータス)が存在するとき、信頼水準は、ある特定の試験サンプルが第一生物学的ステータスに属するという尤度を指す、値 p に対応してもよい。この場合、値 $1-p$ は、ある特定の試験サンプルが第二生物学的ステータスに属するという尤度を指してもよい。概して、二つより多い生物学的ステータスが存在するとき、複数の信頼水準が、各試験サンプルおよび各候補遺伝子シグネチャに提供されてもよい。

50

【 0 0 7 3 】

工程 3 1 0 で、サーバ 1 0 4 は、信頼水準（工程 3 0 8 で受信した）と試験データセットの中の既知の生物学的ステータスとの比較に基づく第一性能測定基準に従い、候補遺伝子シグネチャ（工程 3 0 6 で受信した）をランク付ける。工程 3 1 0 で遂行したランク付けで、各候補遺伝子シグネチャを一位の値に割り当てさせる。

【 0 0 7 4 】

候補遺伝子シグネチャの性能を検討する一手段は、行に予測される生物学的ステータス、および列に実際の生物学的ステータスを含む表に、予測結果を表示することである。下に示す表 1 は、予測結果を表示するための一手段の例である。表の第一行は、第一生物学的ステータスを実際に有する個人（例えば、真の喫煙者）の数、およびサンプルが第一生物学的ステータス（例えば、予測される喫煙者）と関連付けられると予測された、第二生物学的ステータスを実際に有する個人（例えば、現非喫煙者）の数を示す。表の第二行は、第一生物学的ステータスを実際に有する個人（例えば、真の喫煙者）の数、およびサンプルが第二生物学的ステータス（例えば、予測される非喫煙者）と関連付けられると予測された、第二生物学的ステータスを実際に有する個人（例えば、現非喫煙者）の数を示す。

10

【表 1】

表 1

	実際の生物学的ステータス 1	実際の生物学的ステータス 2
予測される生物学的ステータス 1	真陽性	偽陽性
予測される生物学的ステータス 2	偽陰性	真陰性

20

30

完璧な予測子は、第一生物学的ステータスを実際に有する個人のすべてを、第一生物学的ステータス（真陽性が 1 0 0 % で、偽陰性が 0 % であろう）を有すると正確に予測するであろうし、第二生物学的ステータスを実際に有するすべての個人が、第二生物学的ステータス（真陰性が 1 0 0 % で、偽陽性が 0 % であろう）を有すると正確に予測されるであろう。本明細書に記載する通り、個人は、喫煙ステータス（例えば、現喫煙者、現非喫煙者、喫煙経験者、喫煙未経験者など）など、複数の生物学的ステータスに分類されてもよいが、概して、当業者は、本明細書に記載するシステムおよび方法が、いかなる分類スキームにも適用可能であることを理解するであろう。

【 0 0 7 5 】

予測子（例えば、分類子および候補遺伝子シグネチャ）の強さを検討するために、予測結果表の中の値に基づく様々な測定基準が使用されてもよい。第一例では、一つの測定基準は、「感度」または「再現率」と本明細書で称され、第一生物学的ステータスを実際に有する個人のセットのうち、第一生物学的ステータス（例えば、現喫煙者）と正確に分類された個人の割合である。換言すれば、感度（または再現率）測定基準は、真陽性の数を真陽性と偽陰性との合計で割り算したもので、すなわち、 $TP / (TP + FN)$ に等しい。1 という感度値は、第一生物学的ステータスに実際に属する全サンプルが、第一生物学的ステータスに属すると正しく予測されたことを示すが、他のサンプルが何個、第一生物学的ステータスに属すると誤って予測されたか（FP）に関する情報は提供しない。

40

【 0 0 7 6 】

50

第二例では、一つの測定基準は、「特異性」と本明細書で称され、第二生物学的ステータスを実際に有する個人のセットのうち、第二生物学的ステータス（例えば、現非喫煙者）と正確に分類された個人の割合である。換言すれば、特異性測定基準は、真陰性の数を真陰性と偽陽性との合計で割り算したものの、すなわち、 $TN / (TN + FP)$ に等しい。1 という特異性値は、第二生物学的ステータスに実際に属する全サンプルが、第二生物学的ステータスに属すると正しく予測されたことを示すが、第二生物学的ステータスを有すると誤って予測された、第一生物学的ステータスを有するサンプルの数（FN）に関する情報は提供しない。

【0077】

第三例では、一つの測定基準は、「適合率」と本明細書で称され、第一生物学的ステータスを有すると予測された個人のセットのうち、第一生物学的ステータス（例えば、現喫煙者）と正確に分類された個人の割合である。換言すれば、適合率測定基準は、真陽性の数を真陽性と偽陰性との合計で割り算したものの、すなわち、 $TP / (TP + FP)$ に等しい。1 という適合率値は、ある特定のクラス（例えば、生物学的ステータス）に属すると予測された全サンプルが、実際にそのクラスに属することを示すが、第二生物学的ステータスを有すると誤って予測された、第一生物学的ステータスを有するサンプルの数（FN）に関する情報は提供しない。

【0078】

強力な予測子とみなされるには、感度および特異性の両方、感度および適合率の両方、または感度、特異性および適合率において高い値が望ましい場合がある。本明細書では、候補遺伝子シグネチャの性能を検討するために、感度、特異性および精度測定基準が使用されてもよい一方、概して、陰性試験の予測値（ $TN / (TN + FN)$ ）など、本開示の範囲を逸脱することなく、いかなる他の測定基準がまた使用されてもよい。

【0079】

例では、第一性能測定基準は、曲線下面積（area under a curve: AUC）測定基準に関係している。特に、曲線は、受信者動作特性（ROC）曲線または適合率 - 再現率（precision-recall: PR）曲線に対応してもよい。ROC 曲線の軸は、感度（または真陽性率: $TP / (TP + FN)$ ）および偽陽性率（ $FP / (FP + TN)$ ）に対応する。PR 曲線の軸は、感度（ $TP / (TP + FN)$ ）および適合率（ $TP / (TP + FP)$ ）に対応する。一例では、PR 曲線下面積（AUPR）は、ある特定の候補遺伝子シグネチャに一位を取得させるように、第一性能測定基準として使用される。別の例では、ROC 曲線下面積が、第一性能測定基準として使用される。PR 曲線および/または ROC 曲線が連続してもよい一方、本開示は離散値を使用してもよく（閾値が異なるため）、一つ以上の補間法が曲線下面積を演算するのに使用されてもよい。

【0080】

工程 312 で、各候補遺伝子シグネチャに対して、サーバ 104 は、試験データセットの中の各サンプルを、予測される生物学的ステータスへ割り当てるように、信頼水準を使用する。特に、科学者からの各提出に対して、各試験サンプルは、提出の中にある信頼水準に基づいて、予測される生物学的ステータスに割り当てられる。一例では、二つの生物学的ステータス（第一生物学的ステータスおよび第二生物学的ステータス）が存在するとき、信頼水準は、試験サンプルが第一生物学的ステータスに属するという尤度である、値 p を有してもよい。その上に、値 $1 - p$ は、試験サンプルが第二生物学的ステータスに属するという尤度に対応してもよい。概して、科学者は、複数の生物学的ステータスが存在するとき、複数の信頼水準を提出してもよく、ある特定の候補遺伝子シグネチャに対する予測される生物学的ステータスは、最高の信頼水準を有する生物学的ステータスに対応してもよい。

【0081】

工程 314 で、サーバは、予測される生物学的ステータス（工程 312 で取得した）が、試験データセットの中の既知の生物学的ステータスに合致するかに基づく第二性能測定基準に従い、候補遺伝子シグネチャをランク付ける。工程 314 で遂行したランク付けで、

10

20

30

40

50

各候補遺伝子シグネチャを二位の値に割り当てさせる。

【 0 0 8 2 】

別の例では、第二性能測定基準は、マシューズ相関係数 (MCC) 測定基準に対応してもよい。MCC測定基準は、すべての真/偽陽性率と真/偽陰性率とを組み合わせ、それゆえ単一の値である妥当な測定基準を提供する。MCCは、複合性能スコアとして使用されてもよい、性能測定基準である。MCCは、-1と+1との間の値であり、本質的に既知の二項分類と予測される二項分類との間の相関係数である。MCCは、以下の式を使用して演算される場合がある。

【数1】

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

10

式中、TPは真陽性、FPは偽陽性、TNは真陰性、FNは偽陰性である。しかしながら、概して、性能測定基準のセットに基づいて、複合性能測定基準を生成するためのいかなる好適な技法が、候補遺伝子シグネチャの性能およびその対応する予測を評価するために、使用されてもよい。+1というMCC値は、モデルが完全な予測を取得することを示し、0というMCC値は、モデル予測が無作為と何ら変わらず遂行されることを示し、-1というMCC値は、モデル予測が完全に不正確であることを示す。MCCは、クラス予測のみが可能なやり方で、分類子機能をコード化すると、容易に演算することができる利点を有する。概して、TP、FP、TNおよびFNは、本開示に従って第二性能測定基準として使用されてもよい。

20

【 0 0 8 3 】

工程316で、サーバ104は、工程310および314で割り当てたランクに基づく第三性能測定基準に従い、候補遺伝子シグネチャをランク付ける。特に、工程310の一位は、未加工の信頼水準と試験サンプルの既知の生物学的ステータスとの比較に基づいて取得され、工程314の二位は、予測される生物学的ステータス(信頼水準から評価された)と試験サンプルの既知の生物学的ステータスとの比較に基づいて取得される。一位および二位は、第三性能測定基準を取得するように、平均化され(または何らかの手段で組み合わせられ)てもよい。

30

【 0 0 8 4 】

工程318で、サーバ104は、最上位にランク付けられたN個の候補遺伝子シグネチャのうち、少なくとも閾値数(例えば、M)の候補遺伝子シグネチャに含まれる、遺伝子のセットを特定する。例では、第三性能測定基準に従い最高位にランク付けられたN個の候補遺伝子シグネチャが決定される。これらN個の候補遺伝子シグネチャのうち少なくともM個に現れるいずれかの遺伝子が、工程318で特定される遺伝子に含まれ、MはNより小さい。一部の実装では、(N, M) = (3, 2)、(4, 3)、(4, 2)、(5, 4)、(5, 3)、(5, 2)、(6, 5)、(6, 4)、(6, 3)、(6, 2)、またはNおよびMに対するいかなる他の好適な組み合わせであり、式中、Nは2から候補遺伝子シグネチャの総数に及ぶ整数であり、Mは2からNに及ぶ整数である。

40

【 0 0 8 5 】

実施例1 - はじめに

【 0 0 8 6 】

個人の喫煙者ステータスを正確に予測するために、ロバストな遺伝子シグネチャを取得するようクラウドソーシング方法が使用される、実施例の研究について本明細書に記載する。実施例の研究の一つの目的は、喫煙および禁煙ステータスを予測する、ヒトおよび種に依存しない血液曝露反応マーカーおよびモデルを特定するための演算方法を基準に従って評価することによって、血液中で化学物質への曝露反応のマーカーを特定することである。

【 0 0 8 7 】

実施例1 - 研究対象母集団およびデザイン

50

【0088】

全血サンプルは、臨床研究および生体内研究にPAXgene（商標）チューブに収集するか、またはバイオバンクのリポジトリから購入する。異なる研究に対するサンプル群/クラス、サイズおよび特性は、図6に示す表に要約する。手短に言えば、ヒトの血液サンプルは、(i)英国ロンドンのQueen Ann Street Medical Center (QASMC)で行われ、識別子NCT01780298でClinicalTrials.govに登録された臨床症例対照研究、(ii)バイオバンクのリポジトリ(米国メリーランド州ベルツビルのBioServe Biotechnologies Ltd.) (データセットBLD-SMK-01)から取得される。これら両方の出所からのサンプルは、よく定義された組み入れ基準で選択された喫煙者(S)、喫煙経験者(FS)および喫煙未経験者(NS)(図6)、ならびに(iii)無作為化、対照、非盲検、3並行群間および単一施設研究に対応する、臨床のZRHR曝露低減(Reduced exposure: REX)C-03-EUおよび-04-JP研究を含む。REX研究は、5日間閉じ込められて従来のはたきを使用し続ける(喫煙者)のと比較して、喫煙する健康な対象が、候補のリスク低減たばこ製品(「MRTP(modified risk tobacco product)」)または禁煙(「Cess(cessation)」)へ切り替えるときの、選択した煙成分への曝露の減少を実証するのを目的とする。概して、MRTPは加熱式たばこ製品であってもよい。本明細書で使用する通り、加熱式たばこ製品は、使用中にたばこを燃焼させず、たばこまたはたばこを含む混合物を加熱することにより、エアロゾルを発生する製品を含む。マウスの血液サンプルは、メスのC57BL/6およびApoE-/-マウスでそれぞれ7か月および8か月間行った、二つの独立したたばこの煙(「CS」)吸引研究から取得される。研究は、以下、偽(空気に曝露)、3R4F(基準のたばこ3R4FからのCSに曝露)、試作品/候補MRTP(ニコチン濃度が3R4Fに合致する、試作品/候補MRTPからの主流エアロゾルに曝露)、禁煙(Cess)、および2か月の3R4Fへの曝露後に試作品/候補MRTPへ切り替え(Switch)の五つの群に無作為化されたマウスを含む。血液サンプルは異なる時点で収集される。

10

20

【0089】

実施例1 - 血液トランスクリプトミクスデータセット

【0090】

トランスクリプトミクスデータセットは、PAXgene（商標）チューブの中に収集される全血サンプルから生成される。

30

【0091】

ヒトおよびマウスの血液サンプルからのデータ生成

【0092】

全RNAは、PAXgene Bloodキットを使用して分離する。RNAサンプルの濃度および純度は、UV分光光度計(米国マサチューセッツ州ウォルサムにあるThermo Fisher ScientificのNanoDrop(登録商標)1000またはNanoDrop 8000)を使用して、230nm、260nmおよび280nmにおける吸光度を測定することによって決定される。RNAの完全性は更に、Agilent 2100 Bioanalyzer(米国カリフォルニア州サンタクララのAgilent Technologies)を使用して調べる。6つより多いRNA完全性番号を持つRNAのみが、更なる分析のために処理される。

40

【0093】

全RNAは、製造業者の説明書(Qiagen)に従い、PAXgene（商標）チューブの中でサンプルから分離される。抽出されるRNAの品質と、Ovation(登録商標)Whole Blood ReagentおよびOvation RNA Amplification System V2(オランダ、AC LeekのNuGEN)を使用するターゲット調製、および断片化(例えば、断片化しビオチン化した最終製品のサイズ分布を、電気泳動図を使用して監視)の後のcDNAの品質とを、Agilent 21

50

00 Bioanalyzer (米国カリフォルニア州サンタクララ) を使用して調べる。cDNAの品質を、SpectraMax (登録商標) 384 Plus マイクロプレートリーダー (米国カリフォルニア州サニーベールのMolecular Devices) で測定する。cDNA品質を、Fragment Analyzer (米国アイオワ州アンケニーのAdvanced Analytical) を使用して、断片化されていないcDNAのサイズを評価することによって決定する。断片化およびラベリングの後、製造業者のガイドラインに従い、cDNA断片をGeneChip (登録商標) Human Genome U133 Plus 2.0 Array (Affymetrix) にハイブリダイズする。未加工のトランスクリプトミクスデータを、マイクロアレイ画像分析から取得する。QASMC研究のために、血液トランスクリプトミクスデータがAROS Applied Biotechnology AS (デンマーク、オルフス) によって生

10

【0094】

データ処理

【0095】

各データセットからの未加工データ (CELファイル) は、凍結のロバストマイクロアレイ分析であるfRMA v1.1を使用して、R環境 (v3.1.2) で処理および正規化される。凍結したパラメータベクトルのヒト (hgu133plus2frmavecs v1.3.0) を、fRMAおよびGNUSE機能が使用する。brainarrayのヒト用特注cdfファイル (hgu133plus2hsentrezgcdf v16.0.0) を、アフィメトリクスプローブからentrez遺伝子IDまでが、マッピングし、一つの遺伝子の関係性に一つのプローブセットをもたらすために使用する。

20

【0096】

データは、本明細書に記載する基準に対する次のカットオフのうちの一つを通さなかった、全CELファイルを除去する、品質検査工程を通過する。第一に、所与のプローブセットjに対して、正規化非スケール化標準誤差 (Normalized Unscaled Standard Error: NUSE) は、他のアレイと比べて、所与のアレイi上への発現見積りの適合率の尺度を提供する。問題のあるアレイは、標準誤差 (SE) 中央値よりも高いSEとなる。NUSE中央値が1を超える、またはアレイが広い四分位範囲 (IQR) を有するいずれかの場合、アレイは品質が低いと疑われる。1.05より高いNUSE値を持つアレイは除去される。第二に、相対対数発現 (Relative Log Expression: RLE) は、各アレイについて、すべてのjアレイ上の所与のプローブに対する強度レベルの中央値に対して、そのプローブの強度レベルを比較する。アレイ特有のRLE分布は、ある特定のアレイが、優勢的に低くまたは高度に発現された特徴を有するかを決定するのに使用される。ゼロに近くないRLE中央値は、上方制御される遺伝子の数が、下方制御される遺伝子の数とおおよそ等しくはならないことを示し、RLEの広いIQRは、遺伝子の大部分が異なった形で発現することを示す。RLE中央値 > 0.1 (絶対値で) を持つアレイを、外れ値とみなし除去する。第三に、すべてのアレイデータセットの絶対RLE中央値 (Median Absolute RLEs: MARLEs) の絶対偏差中央値を0.01の平方根で割り算したものよりも大きい、MARLE (または中央値 (MARLE) / (1.4826 * mad (MARLEs)) > 1 / sqrt (0.01)) を持つアレイを、品質の悪いチップを有するとみなし除去する。

30

40

【0097】

Brainarrayの特注のマウスおよびヒト用CDFファイルを、Entrez Gene IDマッピングへのAffymetrixプローブに使用し、一つの遺伝子関係に対して一つのプローブセットがもたらされる (それぞれHGU133Plus2_Hs_ENTREZG v16.0、Mouse4302_Mm_ENTREZG v16.0)。品質検査で、最低限の品質基準に合格しない、CELファイルを除外する。データセットの取り扱いを促進するために、ヒトおよびマウスの遺伝子発現データセットには、両方にヒト遺伝子記号が提供される。マウス遺伝子は、NCBI/HCOPマッピングファ

50

イルを使用して、ヒト遺伝子に対応付けられる。マウス遺伝子が複数のヒト遺伝子に位置する場合、大文字で書かれたマウス遺伝子に合致するヒト遺伝子のみが保持される。

【0098】

実施例1 - チャレンジ概要

【0099】

チャレンジのために、喫煙者(S)および現非喫煙者(NCS)の対象血液からの遺伝子発現プロフィールを、図1に關係して記載するネットワーク102上などで、科学界へ提供する。遺伝子発現プロフィールのセットは、均等に訓練セットおよび試験セットに分割される。訓練データセット(喫煙者、喫煙経験者、喫煙未経験者クラスという対象の生物学的ステータスについて完全な情報を持つ)は、試験データセット(対象の生物学的ステータスについての情報は持たない)を公開する前に公開される。135名の登録科学者を、61チームのグループに分ける。61チーム中の23チームがチャレンジ規則に一致した提出を行い、23チーム中の12チームが適格な提出を行っている。図7Aは、チャレンジの目的が、ヒトおよびマウスの全血遺伝子発現データから、化学物質への曝露反応マーカーを特定し、新規血液サンプルを曝露または非曝露群の一部として予測分類するために、これらのマーカーを演算モデルでシグネチャとして活用することであることを示す。

10

【0100】

データは、ヒトおよび齧歯類におけるCS曝露および禁煙に關係する、独立した臨床研究および生体内研究で収集される、血液サンプルから取得される。実験群はまた、試作品/候補MRTPに曝露される個人、または一定期間CSに曝露された後、試作品/候補MRTPに切り替える個人も含む。参加者には、血液サンプルから生成される対象の遺伝子発現プロフィールに基づいて、喫煙曝露を予測するモデルを開発するように依頼する。具体的には、以下の二つの課題を解決するよう、参加者に依頼する。(1)喫煙者の対象対現非喫煙者の対象を特定する。(2)現非喫煙者と予測される各対象に対して、対象が喫煙経験者(FS)または喫煙未経験者(NS)のどちらの対象かを特定する。スコアリングに対して適格であるためには、チームは、両方の課題に対して、予測(例えば、各試験サンプルに対する信頼水準)および候補遺伝子シグネチャ(最大40個の遺伝子を含む)の提出を要する。チャレンジが終了すると、匿名化された予測を、専門家の外部委員会で確立されるパイプラインに従ってスコア化する。チャレンジにおける最高の遂行者は、喫煙者と現非喫煙者とを識別するように、ほぼ完ぺきな予測を実現した。

20

30

【0101】

チャレンジの目標および規則

【0102】

参加者には、(i)喫煙者と現非喫煙者とを識別(課題1)し、続いて(ii)現非喫煙者を、喫煙経験者および喫煙未経験者として分類する(図7Bの課題2)、ロバストでスパースなヒト(サブチャレンジ1、SC1)および種に依存しない(サブチャレンジ2、SC2)血液を基にした遺伝子シグネチャ分類モデルを開発するように依頼する。第一の制約として、予測モデルは、モデルを再訓練/洗練させる必要も、サンプルクラスを予測するように、訓練および試験データセットを組み合わせる半教師付き手法を使用する必要もなく、単一の個人血液サンプルがどのクラスに属するかを予測する能力によって、誘導的(伝達的とは対照的に)であるように要求される。第二の制約として、シグネチャは40個以下の遺伝子を含み得る。

40

【0103】

訓練、試験および検証データセットとして公開されるデータ

【0104】

図8は、血液遺伝子発現データの訓練データセット、試験データセットおよび検証データセットを公開する方法を示す。血液サンプル処理および遺伝子発現データ生成の後、独立した研究からのデータを、訓練、試験および検証データセットに分割する。訓練データセットからのデータおよびクラスラベルを、血液を基とする遺伝子シグネチャ分類モデルの開発および訓練に提供する。血液サンプルのクラス予測のために、訓練済みモデルを、無

50

作為化された試験および検証遺伝子発現データセットに盲検的に適用する。

【0105】

具体的には、QASMC臨床（図7BのデータセットH1）研究、およびマウスC57BL/6の吸引（図7BのデータセットM1a）研究からの正規化された遺伝子発現データおよびクラスラベルを、訓練データセットとして提供する。ヒトBLD-SMK-01およびマウスApoE-/-データ（それぞれ図7BのデータセットH2およびM2a）を、試験データセットとして使用する。REXC-03-EU（図7BのデータセットH3）/04-JP（図7BのデータセットH4）臨床研究、ならびにマウスC57BL/6（図7BのデータセットM1b）およびApoE-/-（図7BのデータセットM2b）吸引研究からのデータを、検証データセットとして公開する。試験および検証セットからのサンプルデータを完全に無作為化し、クラスラベル予測のために順次公開された、クラスのバランスが取れた二つのサブセットに分ける（図8）。試験データセットからのサンプルは、参加者の予測をスコア化し、各サブチャレンジにおけるチーム成績を評価するのに使用する。検証セットは、参加者がサンプルを、喫煙者または現非喫煙者のどちらにより近いと予測したかを検討するのに使用する。ヒトデータのみ、ならびにヒトおよびマウスのデータを、SC1およびSC2それぞれのために公開する（図7B）。

10

【0106】

予測遺伝子シグネチャ分類モデル

【0107】

選択バイアスを避けるために、または全体のアレイに基づく遺伝子シグネチャの性能に通常影響する、次元の呪いを低減するために、二つの公の独立したデータセットを、フィルタリングおよび遺伝子選択を導くように使用する。独立した研究からの最高倍率変化の遺伝子を合同で、二つの研究のうちのN個の最高倍率変化（絶対値で）の交点における、遺伝子に基づく線形判別モデルの検討（各々N-1）で使用する。最高のNは、5重交差検証（100回繰り返される）によって選ばれ、11遺伝子シグネチャにつながる。

20

【0108】

チャレンジのために、参加者は、際立った特徴（遺伝子）を特定し、サンプルを分類するように、様々な特徴選択手法および機械学習手法を使用する。ランダムフォレスト、部分最小二乗判別分析、線形判別分析（LDA）およびロジスティック回帰は、両方のサブチャレンジにおける上位三つの優良なチームが使用する分類方法である。試験および検証データセットからの各サンプルについて、参加者には、サンプルがクラス1（例えば、喫煙者）に属していた信頼値P（0と1の間）と、サンプルがクラス2（例えば、現非喫煙者）に属していた信頼値に対応する、信頼値1-Pとを提供するように要求する。Pおよび1-Pは不等であることが要求される。

30

【0109】

性能評価のスコアリング

【0110】

試験データセットに存在し、検証データセットに存在しないサンプルは、各サブチャレンジにおけるチーム成績を評価するのに使用する。匿名化された参加者のクラス予測を、マッシュアップ相関係数および適合率-再現率曲線下面積測定基準を使用して、スコア化する。全体のチーム成績は、測定基準および課題（課題1：喫煙者対現非喫煙者、課題2：喫煙経験者対喫煙未経験者）に渡って演算される平均ランクに基づく。スコアリング結果および最終ランク付けは、当該分野の専門家から成る外部の独立したスコアリング審査委員会によって審査され、承認される。本公表用の検証データセットに関するチーム成績を検討するために、REX研究からの喫煙者および喫煙経験者（Cess）サンプルを使用して、同じスコアリング方式が適用される。

40

【0111】

チャレンジ後分析

【0112】

血液サンプルが喫煙者群または3R4F群のどちらに属するかに対応する信頼値を、対数

50

オッズ ($\log(P/(1-P))$) として変換する。個々の上位3チームに対する(検証データセットを使用して再スコア化される)、または資格のある全チームの中央値として集約される、対数オッズの分布を、クラスごとに箱ひげ図に可視化する。対を成す(長軸方向のREX研究の0日目対5日目)ウェルチのt検定を、主要な比較(すなわち、対応する喫煙者/3R4F群と比較されるすべての群)に対して遂行した。すべての統計および図式の視覚化は、Rソフトウェアv3.1.2を使用して行われる。

【0113】

実施例1 - 結果

【0114】

本実施例の事例研究では、MRTP評価に係るシステム毒性学における、方法およびデータの独立検証の結果を報告する。研究の一つの目的は、喫煙曝露ステータスまたは禁煙ステータスを予測する能力を持つ、血液を基とするヒトおよび種に依存しない遺伝子発現シグネチャ分類モデルの開発のために、演算方法を検討することである(図7)。参加者は、喫煙者/3R4Fおよび現非喫煙者(喫煙経験者/Cessおよび喫煙未経験者/Sham)のデータと、試作品/候補MRTPに曝露されたマウス、または従来のCSへの曝露後に、候補MRTPに切り替えたヒト対象およびマウスからのデータとを含む、独立した遺伝子発現データセットに、訓練済みモデルを盲検的に適用した。各サンプルに対して、参加者は、煙に曝露された群、または現在煙に曝露されていない群のどちらに、サンプルが属するかの信頼値を提出する。

【0115】

ヒト喫煙曝露遺伝子シグネチャ分類モデルの使用時、5日間禁煙して候補MRTPに切り替えた群のサンプルと、喫煙者(S)群のサンプルとの関連が減少

【0116】

ヒト喫煙曝露反応遺伝子シグネチャ分類モデルを、喫煙者、喫煙経験者および喫煙未経験者を含んだ、QASMCデータセットで訓練する。特定されたシグネチャは、以下の11遺伝子 LRRN3、SASH1、TNFRSF17、DDX43、RGL1、DST、PALLD、CDKN1C、IFI44L、IGJおよびLPAR1のセットを含む。喫煙者と現非喫煙者とを識別する、シグネチャの能力を試験するために、モデルを試験データセット(BLD-SMK-01)に適用し、サンプルが喫煙者群に属していた可能性を持つLDAスコアを、各サンプルに対して演算する。サンプルと喫煙者群または現非喫煙者群との関連を定量化するように、サンプルが喫煙者群(P)およびNCs群(1-P)に属する可能性を演算し、対数オッズ($P/(1-P)$)として変換する。群/クラスごとの対数オッズ分布を、箱ひげ図に可視化する(図9A、ウェルチのt検定により、 p -値 $3 * < 0.001$ 対 S群)。喫煙者クラスに対する対数オッズ分布の中央値は、おおよそ +3.0 であり、一方、喫煙経験者クラスおよび喫煙未経験者クラスに対して、中央値はそれぞれおおよそ -3.8 および -5.8 である。喫煙者クラスと現非喫煙者クラスとの中央値の差が大きくなればなるほど、遺伝子シグネチャ分類モデルはより判別可能になる。箱ひげ図は、片側の喫煙者と、他方側の現非喫煙者として定義される喫煙経験者および喫煙未経験者との間に、明確な分別を示す(図9A)。

【0117】

同じモデルおよび手順を、SwitchまたはCess対象のデータが、喫煙者または現非喫煙者どちらにより近いと分類されたかを決定するように、検証データセット(REXC-03-EUおよびREXC-04-JP)に直接適用する(図9A)。特に、Switchは候補MRTPに切り替えた対象であり、Cessは5日間閉じ込められて喫煙をやめた対象である。5日間みの禁煙または切り替えの後、これらの群に係る対数オッズは、喫煙者群と比較すると有意に減少し、一方、Cess群とSwitch群の間には差異が見られない(図9A)。喫煙群に対して、0日と5日との間に有意な差(対数オッズ比)は見られず、一方、Cess群およびSwitch群について、0日目のそれぞれのベースラインと比較すると、有意な減少が観察された(図9B、対となるt-試験 p -値 $3 * < 0.001$)。

10

20

30

40

50

【 0 1 1 8 】

クラウドソーシングによるデータ検証で、5日の禁煙群および候補M R T Pへの切り替え群の血液サンプルが喫煙者群に属するという、信頼低下の予測を確認

【 0 1 1 9 】

ヒト喫煙曝露反応遺伝子シグネチャ分類モデルを訓練した後、参加者は、無作為化された試験および検証データセットにモデルを適用し、対象が喫煙者群に属する信頼値（確率）を、各対象に対して演算した。チャレンジが終了した後、喫煙者、喫煙経験者および喫煙未経験者のみを含む試験データセット上で、スコアリングを遂行した。参加者の予測提出物が、検証コホートのみに対して再度スコア化され、チーム225、264および257を、SC1の上位3チームとして特定する（図10に示す表）。クラス予測用の遺伝子シグネチャ分類モデルのクラス予測性能を、喫煙者およびC e s s（性能評価では喫煙経験者とみなされる）の真のクラスラベルを、至適基準として使用して評価し、A U P R曲線値は、優良な上位3チームに対して、少なくとも0.90であると判明する（図10に示す表）。

10

【 0 1 2 0 】

図11は、試験および検証データセットに対する、参加者によるヒトおよびマウスの血液サンプルクラス予測を示す。特に、参加者は、煙に曝露される（ヒトはSまたはマウスは3 R 4 F）ヒト対象およびマウスと、現在煙に曝露されていない（N C S）（喫煙経験者F S / C e s sおよび喫煙未経験者N S / S h a m）ヒト対象およびマウスとを識別するように、ヒト（図11A）および種に依存しない（図11B）血液を基とする喫煙曝露遺伝子シグネチャを訓練した。各サンプルについて、参加者に、サンプルがS / 3 R 4 F群に属するという信頼値P、およびサンプルがN C S群に属するという信頼値1 - Pを提供するように依頼する。信頼値を、対数オッズ（ $\log(P / (1 - P))$ ）として変換し、参加資格のある全12チームに対する各サンプルの中央値を演算することによって集約し、箱ひげ図のようなクラスごとの分布として表示する（図11A）。全ての結果が、試験データセットに対して、喫煙者と現非喫煙者（喫煙経験者および喫煙未経験者）との明確な識別を示す。検証データセットについて、モデルを使用して取得された、5日間のC e s sおよびS w i t c h群と喫煙者群とのサンプルの関連が低減するという知見が、類似の結果を生み出した、個々のまたは集約された参加者の予測によって明白に確認された（図11A）。ウェルチのt検定のp - 値は、S / 3 R 4 F群に対して、 $* < 0.05$ 、 $2 * < 0.01$ 、 $3 * < 0.001$ である。経験者 / 未経験者クラスへのこの信頼値の低下は、シグネチャ遺伝子発現に改変が生じたこと、および5日間の禁煙または候補M R T Pへの切り替え後に、血球の中で既に改変が検出可能であることを反映している。

20

30

【 0 1 2 1 】

ヒトおよび齧歯類種にかかわらず、血液サンプルクラス予測に対して特定された最優良の喫煙曝露モデルを基準に従って評価する、クラウドソーシングによる技法

【 0 1 2 2 】

SC2では、参加者に、ヒトおよび齧歯類データの両方に直接適用可能であったクラス予測のために、種に依存しない喫煙曝露反応遺伝子シグネチャモデルを開発するように依頼する。検証データセットを使用する、参加者の予測提出の再スコアリングによって、チーム219、250および264を、SC2の上位3チームとして特定する（図10の表）。SC1に対して、優良チームによってまたは全チームの値の集約後に取得される信頼値を、クラスごとに対数オッズ分布として可視化する（図11B）。C S / 3 R 4 Fに曝露されるコホートと、曝露されない（喫煙未経験者 / S h a mおよび喫煙経験者 / C e s s）コホートとの明確な分別が、箱ひげ図上でヒトおよびマウスの両方に対して観察でき、モデルは、種とかわりなく血液サンプルを分類できることを示している（図10、図11Bに示す表）。独立した二つのマウスの生体内研究からの検証サンプルに、モデルを盲検的に適用するとき、試作品M R T P（p M R T P）または候補M R T Pに曝露される群に対応するサンプルは、マウスおよびヒトのデータセットに対して、S h a mおよび喫煙未経験者対照群それぞれに類似するレベルを持つ、対数オッズ値を有する（図11B）。

40

50

【 0 1 2 3 】

図 1 2 は、検証データセットに対する、閉じ込められた 0 日目と 5 日目との間の、集団の対数オッズ比を示す。対数オッズ比は、C e s s 群および S w i t c h 群に対して、0 日目と 5 日目との間で有意に異なるが、予想通り、喫煙者群に対しては有意に異なるとはいえない (対となる t - 試験の p - 値 $3 * < 0 . 0 0 1$)。

【 0 1 2 4 】

図 1 3 は、群 / クラスごと、および p M R T P もしくは候補 M R T P への曝露時、または p M R T P もしくは候補 M R T P への切り替え後ごとに分けられた集団の対数オッズ分布を示す。具体的には、2 か月の C S 曝露から p M R T P へ切り替わった後、クラスを各時点で分けると、対数オッズ値の斬新的減少が、時間と共に観察され (例えば、p M R T P への 1 か月、3 か月および 4 か月の曝露に対応する S w i t c h 3、S w i t c h 5 および S w i t c h 7)、時間と共に血球の中に生じる漸進的な遺伝子発現の変化を示す。

10

【 0 1 2 5 】

喫煙曝露ステータスを示す、血液中のヒトおよび種に依存しない応答マーカーは、共有性を示し、チーム全体で高度に不変であった、コア遺伝子サブセットを含んでいた。

【 0 1 2 6 】

喫煙曝露コア遺伝子サブセットは、上位 3 チームおよび P M I シグネチャで、少なくとも二つの共起を持つ遺伝子を抽出することで特定される (図 4)。サイクリン依存性キナーゼ阻害因子 1 C (C D K N 1 C)、ロイシンリッチリピート神経 3 型 (L R R N 3)、ならびに S A M および S H 3 ドメイン含有 1 (S A S H 1) をコードする遺伝子は、ヒトシグネチャに最も頻繁に出現する遺伝子であり (図 4 A)、アリアル炭化水素受容体リプレッサー (A H R R)、P 2 Y 6 受容体 (p y r i m i d i n e r g i c r e c e p t o r : P 2 R Y 6) をコードする遺伝子は、種に依存しないシグネチャで最も高い共起を有する (図 4 B)。両方のコア遺伝子サブセット間の比較により、L R R N 3、S A S H 1、A H R R および P 2 R Y 6 をコードする四つの遺伝子の共通セットが明らかになる (図 4)。

20

【 0 1 2 7 】

実施例 1 - 上位 6 チームのヒトを基とする喫煙曝露コンセンサスシグネチャからの遺伝子の全組み合わせの性能分析、遺伝子シグネチャの長さ、遺伝子発現の共線性レベルおよび分類方法の影響

30

【 0 1 2 8 】

方法

【 0 1 2 9 】

コンセンサスシグネチャからの遺伝子の可能な全組み合わせを考慮する。18 個の遺伝子を基とするヒトの喫煙曝露コンセンサスシグネチャの抽出は、この分析に要するコンピュータを利用した計算により課される限定のため、上位 6 チーム (資格のある 12 チームではなく) に限定される。D S C 2、F S T L 1、G P R 6 3、G S E 1、G U C Y 1 A 3、R G L 1、C T T N B P 2、F 2 R、S E M A 6 B、C D K N 1 C、C L E C 1 0 A、G P R 1 5、L I N C 0 0 5 9 9、P 2 R Y 6、P I D 1、S A S H 1、A H R R および L R R N 3 を含んでいた、血液中の 18 個の遺伝子を基とするコンセンサスシグネチャを、上位 6 チームのシグネチャに少なくとも二つの共起を持つ遺伝子の選択によって特定する。遺伝子シグネチャのサイズおよび共線性レベルの分類性能への影響を調査する。五重交差検証による訓練 (10 回の繰り返しによる)、および S C 1 からの試験データセットをそれぞれ使用して、分析を行う。チャレンジで最も幅広く適用される機械学習 (M L) 方法は、ランダムフォレスト (R F)、線形カーネル (s v m L i n e a r) によるサポートベクターマシン、部分最小二乗判別分析 (P L S)、ナイーブベイズ (N B)、k 最近傍 (k N N)、線形判別分析 (L D A) およびロジスティック回帰 (L R) を含む。長さ 2 から 18 の 18 個の遺伝子の可能な全組み合わせ (すなわち、262, 125 の遺伝子セット) が生成される。七つの M L 方法の各々を各遺伝子セットに適用すると、総計 1, 834, 875 の試験済み分類戦略をもたらす。遺伝子セット内における遺伝子の共線

40

50

性レベルは、その遺伝子セットに制限される発現マトリクスの第一主成分の相違率として反映される。1, 834, 875個の遺伝子セット - ML予測（「上位」と呼ぶ）の性能は、MCCおよびAUPRスコアの演算によって検討する。これら「上位」遺伝子セットの性能を、異なった形で発現する遺伝子（*differentially expressed gene*: DEG、つまり偽陽性率（*false discovery rate*）、すなわち $FDR \leq 0.5$ ）、またはHG-U133_Plus_2チップ上に表される全遺伝子の中から無作為に選択される遺伝子セット（2~18個の遺伝子）の性能と比較する。サンプリングプロセスを、各遺伝子セットサイズに対して1,000回繰り返し、総計17,000個の無作為「DEG」または「全遺伝子」の遺伝子セットをもたらす。

10

【0130】

結果：上位6チームからの18個の遺伝子を基とするコンセンサスシグネチャの遺伝子セットの組み合わせは、情報価値があり、喫煙曝露ステータスのクラス予測については、「DEG」および「全遺伝子」由来の遺伝子セットをしのぐ。

【0131】

遺伝子シグネチャサイズおよび共線性レベルの、喫煙曝露ステータスのクラス予測性能への影響は、上位6チームの予測からの18個の遺伝子を基とするコンセンサスシグネチャを使用して探求する。MCCおよびAUPRスコアを、MLを基にしたクラス予測で、長さ2から18のシグネチャの可能な全組み合わせの性能を検討するように計算する（図14および15）。図14および15は、MCCスコア（図14）およびAUPRスコア（図15）の結果を表示する。両図面で、パネルAは、交差検証および試験データセットに対する、スコア対遺伝子シグネチャサイズを描写する。特徴は、(i)「上位」遺伝子（すなわち、シグネチャの一部として、参加者が頻繁に選択する遺伝子）、(ii)「DEG」、つまり、異なった形で発現する遺伝子のリスト、(iii)「全遺伝子」、つまり、測定された全遺伝子のリストより選択される。両図面で、パネルBは、スコア対シグネチャの中の遺伝子間の類似性の係数を描写する。以下の七つの異なる機械学習、ランダムフォレスト（RF）、線形カーネル（*svmlinear*）によるサポートベクターマシン、部分最小二乗判別分析（PLS）、ナイーブベイズ（NB）、k近傍（kNN）、線形判別分析（LDA）およびロジスティック回帰（LR）の分類子を試験する。両図面で、パネルCは、CVおよび試験セットデータにおけるスコアの分布に加えて、「上位」（上）、「DEG」（中間）および「全遺伝子」（下）の選択に対する差異の分布を描写する。

20

30

【0132】

図14および15でデータが示す通り、予測性能は、訓練セット（交差検証、CV）（CVでは、サイズ2に対して $MCC = 0.57$ 、およびサイズ18に対して $MCC = 0.91$ ）、および試験セット（試験では、サイズ2に対して $MCC = 0.42$ 、およびサイズ18に対して $MCC = 0.77$ ）の両方で、最大18個の遺伝子を含め、遺伝子セットサイズと共に増大し、よりセットが長くなると共に徐々に安定した（図14A）。「上位」遺伝子セットの中の遺伝子の共線性レベル（遺伝子セットの発現マトリクスから演算される第一主成分により表わされる相違率が反映される）が、50%から60%の間で動いたとき、予測性能は最大に到達し、その後、共線性の増大と共に減少した（図14B）。「上位」遺伝子セットが、異なるチームからのシグネチャ遺伝子から構成され、既に非常に多様であったことを考慮すると、ある程度共線的な遺伝子を組み合わせることで、予測が強化される場合がある。DEGからの遺伝子セット内の遺伝子の共線性が増加すると共に、性能は低下した（図14B）。概して、「上位」、「DEG」および「全遺伝子」からの遺伝子セットにより、それぞれ最高、中程度および最低の性能が与えられた（図14）。加えて、CVに由来する性能は、試験セットに対して演算された性能をしのいだ（図14）。様々なML方法により取得された性能測定基準は、類似のパターンを示し（図14B）、そのため、結果の可視化を促進するように集約された（図14Aおよび図14C）。全体として、18個の遺伝子を基とするコンセンサスシグネチャからの血液遺伝子は、組み合わせると、情報価値があり、喫煙曝露ステータスに対して高い予測力を有したと、

40

50

結果は示した。

【0133】

実施例1 - 議論

【0134】

本実施例の研究で取得された結果によって、候補M R T Pに曝露された対象、または従来のC S曝露に続き、候補M R T Pに切り替えた対象からの血液サンプルが、煙に曝露される群、または現在煙に曝露されていない群に属するという、予測通りの信頼がもたらされる。

【0135】

結果により、喫煙者および現非喫煙者は明確に分別される。チャレンジ参加者は、ヒトおよびマウス種にかかわらず、喫煙曝露ステータス予測に対して非常に良い性能を示す、種に依存しない血液を基とする遺伝子シグネチャモデルの開発に成功した。ヒトの試験データセットでは、喫煙経験者群は、喫煙未経験者群に非常に近いものの、喫煙者群と喫煙未経験者群との中間に残り、喫煙経験者の遺伝子シグネチャの中の遺伝子発現は、喫煙未経験者の発現レベルに戻るほど、完全には反転しない場合があることを示した。変化の復帰は、対象一人ひとりで異なる、喫煙歴および禁煙期間に依存する可能性があり、この群に対する予測のより高い可変性も説明している。喫煙経験者の血球については、DNAメチル化レベル（例えば、F 2 R L 3 遺伝子）が、生涯喫煙量（p a c k y e a r）および止めてからの時間に依存する場合がある。

【0136】

マウスデータセットでは、C e s s 群の発現レベルが、S h a m 群のレベルに到達し、シグネチャ遺伝子発現の復帰が、より遺伝的かつ実験的に均質である、マウス株の血球で変化することを示唆している。興味深いことに、この復帰は、禁煙期間に基づいて群を分けるときに観察されるように、時間と共に徐々に生じる。これは、遺伝子シグネチャ分類手法が、二項分類に有用であるだけでなく、製品試験または使用中止時に血液中で生じる変化の大きさおよび動態に従うように、より定量的（例えば、L D A スコアまたは関連する信頼値など、モデルパラメータの大きさ）にも使用され得ることを示唆する。実際に、これは、検証用のヒトのR E X データセットからのS w i t c h 群およびC e s s 群の場合であり、有意な対数オッズは、喫煙者群と比較すると、喫煙未経験者群の値の方へと減少する。この知見は、喫煙曝露シグネチャ遺伝子により反映される分子変化が、候補M R T Pへ切り替えるか、または従来のたばこを止めてたった5日後に、血球の中に生じることを示す。これらの結果は、臨床の「たばこ一日当たり削減」閉じ込め研究において一週間後に測定した、曝露の用量反応性のバイオマーカーの減少と一致する。マウスの検証データセットについて、切り替え後の候補M R T Pまたはp M R T Pへのより長い（数か月）曝露により説明することができ、従来のC Sと比較して、M R T Pの血球へのより低い生物学的効果を反映していたため、3 R 4 F 群と、試作品/候補M R T P群またはS w i t c h 群（S h a mに類似のレベル）との間の対数オッズの差は、より一層重要である。

【0137】

血液を基とする喫煙曝露反応分類モデルを、開発および訓練するのに使用する演算方法が異なるとしても、成績上位チームによって取得されるサンプル分類性能は高い。チームに渡り高度に一致するコア遺伝子シグネチャが特定され、ヒトのみ、またはヒトおよびマウス（種に依存しないシグネチャ）において、喫煙曝露ステータスを予測する、特定のロバストな血液マーカーを共に構成した遺伝子を選択するのに、煙曝露により誘導される遺伝子発現の変化は、十分に情報価値があり、一致していることを示す。

【0138】

喫煙者および非喫煙者からの細胞特有の白血球の報告済みDNAメチル化分析に類似する、血液細胞型特有のトランスクリプトーム分析は、各血液細胞型の喫煙曝露反応シグネチャへの寄与をより良く理解するのに役立つ場合がある。一部の遺伝子は、特定の血液細胞亜集団に関係してもよい。全体として、コアシグネチャの一部である、これらの喫煙曝露関連遺伝子は、従来のたばこの影響と比較して、候補M R T Pなどの新製品の影響を監視

10

20

30

40

50

し、場合により定量化するように活用され得る、ロバストな血液マーカーのセットを構成する。

【 0 1 3 9 】

実施例 1 に関係して記載する研究は、クラウドの力が、システム毒性学において、演算方法を検討し、データを検証するのに活用されてもよいことを示す。古典的な査読プロセスを補完するのに加えて、製品リスク評価データの独立した公平な検討は、科学的な結論の中で信頼を確認し提供するように使用されてもよく、意思決定する規制当局を支援する場合がある。本明細書に記載する例は、大部分が、個人の喫煙者ステータスを予測するために、ロバストな遺伝子シグネチャを特定するクラウドソーシング手法の使用に関する一方、本開示のシステムおよび方法が、喫煙者ステータス、疾患ステータス、生理学的状態、曝露状態、または個人の生物学的状態と関連付けられる、個人のいかなる他の好適なステータスもしくは状態を含め、個人の生物学的ステータスを予測するために、遺伝子シグネチャを取得するように適用されてもよいことを、当業者は理解するであろう。

10

【 0 1 4 0 】

下の表 2 は、実施例 1 に従って行われた研究からの結果を含む。特に、表 2 に示す結果は、ヒトの喫煙シグネチャから引き出され、第一列に遺伝子のセットを一覧として示す。第二列は、そのシグネチャの中に対応する遺伝子を含んでいた、チームまたは参加者の数（全 12 中）を一覧として示す。第三列は、そのシグネチャの中に対応する遺伝子を含んでいた、上位 3 チーム（試験データセットに従い評価）の数を一覧として示す。第四列は、そのシグネチャの中に対応する遺伝子を含んでいた、上位 3 チーム（検証データセットに従い評価）の数を一覧として示す。第五列は、第三列および第四列の値の平均を一覧として示す。

20

【表 2 - 1】

表 2

スコアリング 試験セット	合計(12 チーム中)	試験セット上位 3 つの合計	検証セット上位 3 つ の合計	試験+検証セッ トの平均
LRRN3	9	3	3	3
AHRR	9	3	3	3
CDKN1C	9	3	3	3
PID1	8	3	3	3
SASH1	7	3	3	3
GPR15	7	3	3	3
P2RY6	6	3	3	3
LINC00599	6	2	3	2.5
CLEC10A	6	3	2	2.5
SEMA6B	5	2	3	2.5
F2R	5	2	2	2
DSC2	5	1	0	0.5
TLR5	5	0	1	0.5
RGL1	4	1	2	1.5
FSTL1	4	1	0	0.5
VSIG4	4	0	0	0
AK8	4	0	0	0
CTTNBP2	3	2	2	2
GUCY1A3	3	1	1	1
GSE1	3	1	0	0.5
MIR4697HG	3	0	0	0

30

40

50

【表 2 - 2】

PTGFRN	3	0	0	0
LOC200772	3	0	0	0
FANK1	3	0	0	0
C15orf54	3	0	0	0
MARC2	3	0	0	0
GPR63	2	2	1	1.5
TPPP3	2	1	1	1
ZNF618	2	1	1	1
PTGFR	2	1	0	0.5
GUCY1B3	2	0	1	0.5
P2RY1	2	0	0	0
TMEM163	2	0	0	0
ST6GALNAC1	2	0	0	0
SH2D1B	2	0	0	0
CYP4F22	2	0	0	0
PF4	2	0	0	0
FUCA1	2	0	0	0
MB21D2	2	0	0	0
NLK	2	0	0	0
B3GALT2	2	0	0	0
ASGR2	2	0	0	0
NR4A1	2	0	0	0
RTN1	1	1	1	1
MAFB	1	1	1	1
ARHGEF10L	1	1	1	1
CLDN23	1	1	1	1
TGFBI	1	1	1	1
LOC284837	1	1	1	1
SYCE1L	1	1	1	1
SEZ6L	1	1	1	1
KLF4	1	1	1	1
NOD1	1	1	1	1
FAM225A	1	1	1	1
CRACR2B	1	1	0	0.5

【0141】

一部の実施形態では、喫煙曝露反応ステータスを決定するのに使用される遺伝子シグネチャは、成績上位三つの遺伝子シグネチャのうち少なくとも二つに現れる遺伝子に対応する、表2に一覧として示す遺伝子を含む。試験データセット（例えば、表2の第三列に示す）に従って評価するとき、これは、LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63を含む。検証データセット（例えば、表2の第四列に示す）に従って評価するとき、これは、LRRN3、AHRR、CDKN1C、PID1、SASH1、GPR15、P2RY6、LINC00599、CLEC10A、SEMA6B、F2R、RGL1およびCTTNBP2を含む。試験および検証データセットの平均（例えば、表2の第五列に示す）に従って評価するとき、これは、LR

R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1、G P R 1 5、P 2 R Y 6、L I N C 0 0 5 9 9、C L E C 1 0 A、S E M A 6 B、F 2 RおよびC T T N B P 2を含む。

【 0 1 4 2 】

一部の実施形態では、喫煙曝露反応ステータスを決定するのに使用される遺伝子シグネチャは、12個の候補遺伝子シグネチャのうち少なくともM個に現れる遺伝子に対応する、表2に一覧として示す遺伝子を含み、Mは1、2、3、4、5、6、7、8または9である。例えば、Mが9のとき、遺伝子シグネチャは、第二列に少なくとも9の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R RおよびC D K N 1 Cを含む。別の例として、Mが8のとき、遺伝子シグネチャは、第二列に少なくとも8の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 CおよびP I D 1を含む。別の例として、Mが7のとき、遺伝子シグネチャは、第二列に少なくとも7の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1およびG P R 1 5を含む。別の例として、Mが6のとき、遺伝子シグネチャは、第二列に少なくとも6の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1、G P R 1 5、P 2 R Y 6、L I N C 0 0 5 9 9およびC L E C 1 0 Aを含む。別の例として、Mが5のとき、遺伝子シグネチャは、第二列に少なくとも5の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1、G P R 1 5、P 2 R Y 6、L I N C 0 0 5 9 9、C L E C 1 0 A、S E M A 6 B、F 2 R、D S C 2およびT L R 5を含む。別の例として、Mが4のとき、遺伝子シグネチャは、第二列に少なくとも4の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1、G P R 1 5、P 2 R Y 6、L I N C 0 0 5 9 9、C L E C 1 0 A、S E M A 6 B、F 2 R、D S C 2、T L R 5、R G L 1、F S T L 1、V S I G 4およびA K 8を含む。別の例として、Mが3のとき、遺伝子シグネチャは、第二列に少なくとも3の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1、G P R 1 5、P 2 R Y 6、L I N C 0 0 5 9 9、C L E C 1 0 A、S E M A 6 B、F 2 R、D S C 2、T L R 5、R G L 1、F S T L 1、V S I G 4、A K 8、C T T N B P 2、G U C Y 1 A 3、G S E 1、M I R 4 6 9 7 H G、P T G F R N、L O C 2 0 0 7 7 2、F A N K 1、C 1 5 o r f 5 4およびM A R C 2を含む。別の例として、Mが2のとき、遺伝子シグネチャは、第二列に少なくとも2の値を持つこれらの遺伝子、すなわち、L R R N 3、A H R R、C D K N 1 C、P I D 1、S A S H 1、G P R 1 5、P 2 R Y 6、L I N C 0 0 5 9 9、C L E C 1 0 A、S E M A 6 B、F 2 R、D S C 2、T L R 5、R G L 1、F S T L 1、V S I G 4、A K 8、C T T N B P 2、G U C Y 1 A 3、G S E 1、M I R 4 6 9 7 H G、P T G F R N、L O C 2 0 0 7 7 2、F A N K 1、C 1 5 o r f 5 4、M A R C 2、G P R 6 3、T P P P 3、Z N F 6 1 8、P T G F R、G U C Y 1 B 3、P 2 R Y 1、T M E M 1 6 3、S T 6 G A L N A C 1、S H 2 D 1 B、C Y P 4 F 2 2、P F 4、F U C A 1、M B 2 1 D 2、N L K、B 3 G A L T 2、A S G R 2およびN R 4 A 1を含む。別の例として、Mが1のとき、遺伝子シグネチャは、上の表2に一覧として示すすべての遺伝子を含む。

【 0 1 4 3 】

下の表3は、実施例1に従って行われた研究からの結果を含む。特に、表2に示す結果は、種に依存しない喫煙シグネチャから引き出され、第一列に遺伝子のセットを一覧として示す。第二列は、そのシグネチャの中に対応する遺伝子を含んでいた、チームまたは参加者の数(全12中)を一覧として示す。第三列は、そのシグネチャの中に対応する遺伝子を含んでいた、上位3チーム(試験データセットに従い評価)の数を一覧として示す。第四列は、そのシグネチャの中に対応する遺伝子を含んでいた、上位3チーム(検証データセットに従い評価)の数を一覧として示す。第五列は、第三列および第四列の値の平均を一覧として示す。

10

20

30

40

50

【表 3 - 1】

表 3

スコアリング 試験セット	合計 (12チ ーム 中)	試験セット上位 3つの合計	検証セット上位3 つの合計	試験+検証セッ トの平均
AHRR	5	3	3	3
P2RY6	4	3	3	3
COX6B2	2	2	2	2
DSC2	2	2	2	2
KLRG1	3	2	2	2
LRRN3	3	2	2	2
SASH1	2	2	2	2
TBX21	2	2	2	2
ADORA3	1	1	1	1
AF529169	1	1	1	1
AKAP5	1	1	1	1
ASGR2	1	1	1	1
B3GALT2	1	1	1	1
BCL3	1	1	1	1
BIRC2	1	1	1	1
CCR4	1	1	1	1
CDKN1C	1	1	1	1
CLEC10A	1	1	1	1
CLEC5A	1	1	1	1
CNNM1	1	1	1	1
COL6A3	1	1	1	1
COX6C	1	1	1	1
CRACR2B	1	1	1	1
CTNNAL1	1	1	1	1
CTTNBP2	2	1	1	1
DCAF8	1	1	1	1
EIF5A2	1	1	1	1
ELOVL7	1	1	1	1
ENDOU	1	1	1	1
ERI1	1	1	1	1
ESAM	1	1	1	1
EVA1B	1	1	1	1
F2R	2	1	1	1
FANK1	1	1	1	1
FKRP	1	1	1	1

10

20

30

40

50

【表 3 - 2】

FSTL1	1	1	1	1
GGT7	1	1	1	1
GLCCI1	1	1	1	1
GNAZ	1	1	1	1
GNPDA2	1	1	1	1
GP1BA	1	1	1	1
GPR63	1	1	1	1
GSE1	1	1	1	1
GUCY1B3	2	1	1	1
HES1	1	1	1	1
HPGD	1	1	1	1
HSPB6	1	1	1	1
IRF7	1	1	1	1
JARID2	1	1	1	1
KCNQ10T1	1	1	1	1
KISS1R	1	1	1	1
LIMS1	1	1	1	1
LRRK1	1	1	1	1
LTBP1	1	1	1	1
MBTD1	1	1	1	1
MCEMP1	1	1	1	1
MKNK1	1	1	1	1
MPP2	1	1	1	1
MRAS	1	1	1	1
MT2	2	1	1	1
NDUFA3	1	1	1	1
NGFRAP1	2	1	1	1
NR4A1	1	1	1	1
PF4	1	1	1	1
PGRMC1	1	1	1	1
PHACTR3	1	1	1	1
PID1	1	1	1	1
PTGFR	1	1	1	1
R3HDM4	1	1	1	1
RBM43	1	1	1	1
REEP6	2	1	1	1
REXO2	1	1	1	1
RUNDC3A	1	1	1	1
SAMD11	1	1	1	1
SDR16C5	1	1	1	1
SIAH1A	1	1	1	1

10

20

30

40

50

【表 3 - 3】

SLPI	1	1	1	1
SPINK2	1	1	1	1
STAR	1	1	1	1
SYTL4	1	1	1	1
TCEAL8	1	1	1	1
TLR2	1	1	1	1
TMEM163	1	1	1	1
TRIB3	1	1	1	1
UBE2B	1	1	1	1
VCAN	1	1	1	1
VSIG4	1	1	1	1
WDFY1	1	1	1	1
ZFP704	1	1	1	1

10

【 0 1 4 4 】

一部の実施形態では、喫煙曝露反応ステータスを決定するのに使用される遺伝子シグネチャは、成績上位三つの遺伝子シグネチャのうち少なくとも二つに現れる遺伝子に対応する、表 3 に一覧として示す遺伝子を含む。表 3 に示すように、これが試験データセット（例えば、表 3 の第三列に示す）、検証データセット（例えば、表 3 の第四列に示す）、または試験データセットおよび検証データセットの平均（例えば、表 3 の第五列に示す）に従って評価されるかにかかわらず、これは、A H R R、P 2 R Y 6、C O X 6 B 2、D S C 2、K L R G 1、L R R N 3、S A S H 1 および T B X 2 1 を含む。

20

【 0 1 4 5 】

一部の実施形態では、喫煙曝露反応ステータスを決定するのに使用される遺伝子シグネチャは、12個の提出された遺伝子シグネチャのうち少なくともM個に現れる遺伝子に対応する、表 3 に一覧として示す遺伝子を含み、Mは1、2、3、4または5である。例えば、Mが5のとき、遺伝子シグネチャは、第二列に少なくとも5の値を持つそれらの遺伝子、すなわち、A H R R を含む。別の例として、Mが4のとき、遺伝子シグネチャは、第二列に少なくとも4の値を持つそれらの遺伝子、すなわち、A H R R および P 2 R Y 6 を含む。別の例として、Mが3のとき、遺伝子シグネチャは、第二列に少なくとも3の値を持つそれらの遺伝子、すなわち、A H R R、P 2 R Y 6、K L R G 1 および L R R N 3 を含む。別の例として、Mが2のとき、遺伝子シグネチャは、第二列に少なくとも2の値を持つそれらの遺伝子、すなわち、A H R R、P 2 R Y 6、K L R G 1、L R R N 3、C O X 6 B 2、D S C 2、S A S H 1、T B X 2 1、C T T N B P 2、F 2 R、G U C Y 1 B 3、M T 2、N G F R A P 1 および R E E P 6 を含む。別の例として、Mが1のとき、遺伝子シグネチャは、上の表 3 に一覧として示すすべての遺伝子を含む。

30

【 0 1 4 6 】

一部の実施形態では、本明細書に記載する遺伝子シグネチャは、10、11、12、13、14、15、20、25、30、35、40、または全ゲノムの中の遺伝子の数より少ない、いかなる他の好適な数など、遺伝子の最大数を有するように制限される。本明細書に記載する遺伝子シグネチャは、全ゲノムと比較して、比較的少数の遺伝子に制限される。より長い遺伝子シグネチャが、訓練データセットに過剰適合する場合、より長い遺伝子シグネチャは、より短い遺伝子シグネチャよりうまく機能しない場合がある。この場合、より長い遺伝子シグネチャは、訓練データセットに偶発誤差またはノイズを記述する場合がある。より短い遺伝子シグネチャは、試験データセットでクラスを予測するように使用されるとき、過剰適合したより長い遺伝子シグネチャをしのご場合がある。表 2 および 3 に関して記載する遺伝子シグネチャを含む、本明細書に記載する遺伝子シグネチャのいずれも、ある特定の最大数の遺伝子を有するように制限されてもよい。

40

50

【0147】

図5は、本開示の図解の実施形態に従って、対象から取得したサンプルを評価するためのプロセス500のフローチャートである。プロセス500は、サンプルと関連付けられるデータセットを受け取る工程であって、データセットは、LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINCO0599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63に対する定量的な発現データを含む、工程(工程502)と、受け取ったデータセットに基づいてスコアを生成する工程であって、スコアが、対象の予測される喫煙ステータスを示す、工程(工程504)とを含む。一部の実施形態では、工程502で受け取ったデータセットは更に、次のDSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1およびGUCY1B3のうちのいずれの数に対する定量的な発現データも含む。一部の実施形態では、工程502で受け取ったデータセットは更に、上の表2および3に関して記載した遺伝子シグネチャのうちのいずれか、または本明細書に記載するいかなる他の遺伝子シグネチャに対する、定量的な発現データを含む。

10

【0148】

工程504で生成するスコアは、データセットに適用される分類スキームの結果であり、分類スキームは、データセットの中の定量的な発現データに基づいて決定される。特に、本明細書に記載する例では、個人に対して予測される分類を決定するように、機械学習技法を使用して訓練された分類子が、502で受け取られたデータセットに適用されてもよい。

20

【0149】

本明細書に記載する遺伝子シグネチャは、対象から取得したサンプルを評価するための、コンピュータ実装された方法で使用されてもよい。特に、サンプルと関連付けられるデータセットが取得されてもよく、データセットは、コア遺伝子シグネチャのために、LRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINCO0599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63に対する定量的な発現データを含んでもよい。概して、表2および3に関して記載した遺伝子シグネチャのうちのいずれも、コア遺伝子シグネチャとして使用されてもよい。コア遺伝子シグネチャは、ゲノム全体における遺伝子の数より少ない、いくつかの遺伝子を含み、全体として共にみなされるとき、喫煙ステータスなど、生物学的状態の予測について情報価値のある遺伝子のセットを含む。受け取ったデータセットの中の遺伝子シグネチャに基づいて、スコアを生成してもよく、スコアは対象の予測される喫煙ステータスを示す。特に、スコアは、本明細書に記載するクラウドソーシング手法を使用して構築された、分類子に基づいてもよい。データセットは更に、追加マーカーDSC2、TLR5、RGL1、FSTL1、VSIG4、AK8、GUCY1A3、GSE1、MIR4697HG、PTGFRN、LOC200772、FANK1、C15orf54、MARC2、TPPP3、ZNF618、PTGFR、P2RY1、TMEM163、ST6GALNAC1、SH2D1B、CYP4F22、PF4、FUCA1、MB21D2、NLK、B3GALT2、ASGR2、NR4A1およびGUCY1B3のいかなる好適な組み合わせに対して、定量的な発現データを含んでもよく、拡張遺伝子シグネチャに含まれてもよい。データセットは更に、上の表2および3に関して記載した遺伝子シグネチャのうちのいずれに対する、定量的な発現データを含んでもよい。

30

40

【0150】

一部の実施形態では、データセットは、マーカーLRRN3、AHHR、CDKN1C、PID1、SASH1、GPR15、LINCO0599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63のセットのいかなる数のいか

50

なるサブセットも含む。サブセットは、これらの特定される遺伝子のすべてより少ない数を含んでもよい。一つ以上の基準が、コアセットの中のマーカー：LRRN3、AHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63のうちの少なくとも三つ（または4、5、6、7、8、9、10、11もしくは12など、いかなる他の好適な数）、ならびに表2または3に関して記載した遺伝子シグネチャの中のマーカーのいずれかのうちの少なくとも二つ（または2、3、4、5、6、7、8、9、10、11もしくは12など、いかなる他の好適な数）を含むなど、シグネチャの中に含まれるようにマーカーに適用されてもよい。上に記載した通り、一部の実施形態では、シグネチャは、ゲノム全体の中の遺伝子の数より少ない、いくつかの遺伝子に限定され、10、11、12、13、14、15、20、25、30、35、40、または全ゲノムの中の遺伝子の数より少ない、いかなる他の好適な数など、遺伝子の最大数に限定されてもよい。概して、これらのマーカーの組み合わせを使用するいかなるシグネチャも、本開示の範囲を逸脱することなく、喫煙ステータスなど、対象の生物学的ステータスを予測するために使用されてもよい。

【0151】

一部の実施形態では、本明細書に記載するシグネチャ中の遺伝子は、個人の喫煙者ステータスを予測するためのキットを組み立てる際に使用される。特に、キットは、試験サンプル中の遺伝子シグネチャの遺伝子発現レベルを検出する試薬のセットと、個人の喫煙者ステータスを予測するキットを使用するための説明書とを含む。キットは、禁煙、または、HTPなど、喫煙製品の代替品の個人への効果を評価するように使用されてもよい。

【0152】

図2は、図1および図2に関して記載するプロセスなど、本明細書に記載するプロセスのいずれかを遂行する、またはコア遺伝子シグネチャ、拡張遺伝子シグネチャ、もしくは本明細書に記載するいかなる他の遺伝子シグネチャを記憶する、コンピューティング装置のブロック図である。特に、コンピュータ可読媒体上に記憶された遺伝子シグネチャは、LRRN3、AHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63に対する発現データを含む。別の実施形態では、コンピュータ可読媒体は、LRRN3、AHR、CDKN1C、PID1、SASH1、GPR15、LINC00599、P2RY6、CLEC10A、SEMA6B、F2R、CTTNBP2およびGPR63から成る群より選択される、少なくとも4つ、5つ、6つ、7つ、8つ、9つ、10個、11個または12個のマーカーに対する発現データを含む、遺伝子シグネチャを含む。別の例では、コンピュータ可読媒体は、本明細書に記載する遺伝子シグネチャ、またはマーカーのセットのいずれかに関係するデータを含む。

【0153】

ある実装では、構成要素およびデータベースは、いくつかのコンピューティング装置200上に実装されてもよい。コンピューティング装置200は、少なくとも一つの通信インターフェースユニットと、入力/出力コントローラ210と、システムメモリと、一つ以上のデータ記憶装置とを備える。システムメモリは、少なくとも一つのランダムアクセスメモリ(RAM202)と、少なくとも一つの読み取り専用メモリ(ROM204)とを含む。これら要素のすべては、コンピューティング装置200の動作を促進するように、中央処理装置(CPU206)と通信する。コンピューティング装置200は、多くの異なるやり方で構成されてもよい。例えば、コンピューティング装置200は、従来のスタンドアロンコンピュータであってもよく、または代替的に、コンピューティング装置200の機能が、複数のコンピュータシステムおよびアーキテクチャにわたって分散してもよい。コンピューティング装置200は、モデリング動作、スコアリング動作および集約動作のうちの一部またはすべてを遂行するように構成されてもよい。図2では、コンピューティング装置200は、ネットワークまたはローカルネットワークを介して、他のサーバまたはシステムにリンクされる。

10

20

30

40

50

【 0 1 5 4 】

コンピューティング装置 200 は、分散アーキテクチャで構成されてもよく、データベースおよびプロセッサは、別個のユニットまたは場所に収容される。いくつかのそのようなユニットは、主要な処理機能を遂行し、最低でも汎用コントローラーまたはプロセッサ、およびシステムメモリを包含する。そのような態様では、これらのユニットの各々は、通信インターフェースユニット 208 を介して、他のサーバ、クライアントまたはユーザーのコンピュータ、および他の関係する装置との主要通信リンクとして機能を果たす、通信ハブまたは通信ポート（図示せず）に取り付けられる。通信ハブまたは通信ポートは、それ自体最低限の処理能力を有してもよく、主に通信ルーターとして機能を果たす。様々な通信プロトコルが、システムの一部であってもよく、Ethernet（登録商標）、SAP、SAS（商標）、ATP、BLUETOOTH（登録商標）、GSM（登録商標）およびTCP/IPを含むが、これらに限定されない。

10

【 0 1 5 5 】

CPU 206 は、一つ以上の従来のマイクロプロセッサなどのプロセッサ、およびCPU 206 からの作業負荷をオフロードするための数値演算コプロセッサなど、一つ以上の補助コプロセッサを備える。CPU 206 は、通信インターフェースユニット 208 および入力/出力コントローラー 210 と通信し、CPU 206 は、これらを通して他のサーバ、ユーザー端末またはユーザー装置などの他の装置と通信する。通信インターフェースユニット 208 および入力/出力コントローラー 210 は、例えば、他のプロセッサ、サーバまたはクライアント端末との同時通信のために、複数の通信チャネルを含んでもよい。相互に通信する装置は、継続的に相互に送信する必要はない。それどころか、そのような装置は、必要に応じて相互に送信することのみが必要であり、実際には大部分の時間でデータの交換を止めてもよく、装置間の通信リンクを確立するために、いくつかの工程の遂行を要してもよい。

20

【 0 1 5 6 】

CPU 206 はまた、データ記憶装置と通信もする。データ記憶装置は、磁気、光学または半導体メモリの適切な組み合わせを備えてもよく、例えば、RAM 202、ROM 204、フラッシュドライブ、コンパクトディスクなどの光学ディスク、またはハードディスクもしくはハードドライブを含んでもよい。CPU 206 およびデータ記憶装置は各々、例えば、単一のコンピュータ内、もしくは他のコンピューティング装置内に完全に位置しているてもよく、またはUSBポート、シリアルポートケーブル、同軸ケーブル、Ethernet（登録商標）タイプのケーブル、電話線、無線周波数トランシーバー、もしくは他の類似の無線もしくは有線媒体、もしくは前述の組み合わせなどの通信媒体によって相互に接続されてもよい。例えば、CPU 206 は、通信インターフェースユニット 208 を介して、データ記憶装置に接続されてもよい。CPU 206 は、一つ以上のある特定の処理機能を遂行するように構成されてもよい。

30

【 0 1 5 7 】

データ記憶装置は、例えば、(i) コンピューティング装置 200 のためのオペレーティングシステム 212、(ii) 本明細書に記載するシステムおよび方法に従って、かつ特にCPU 206 に関して詳細に記載するプロセスに従って、CPU 206 に指示するように適合された、一つ以上のアプリケーション 214（例えば、コンピュータプログラムコード、またはコンピュータプログラム製品）、または(iii) プログラムが必要とする情報を記憶するように利用される場合がある、情報を記憶するように適合するデータベース（複数可） 216 を記憶してもよい。一部の態様では、データベース（複数可）は、実験データおよび発行された文献モデルを記憶するデータベースを含む。

40

【 0 1 5 8 】

オペレーティングシステム 212 およびアプリケーション 214 は、例えば、圧縮され未コンパイルで暗号化されたフォーマットで記憶されてもよく、コンピュータプログラムコードを含んでもよい。プログラムの命令は、ROM 204 からまたはRAM 202 からなど、データ記憶装置ではなくコンピュータ可読媒体から、プロセッサの主メモリへと読み

50

込まれてもよい。プログラム中で命令シーケンスを実行することによって、CPU 206に本明細書に記載するプロセス工程を遂行させる一方、本開示のプロセスの実施のために、ソフトウェア命令の代わりに、またはソフトウェア命令と組み合わせて配線で接続された回路が使用されてもよい。それゆえ、記載するシステムおよび方法は、ハードウェアとソフトウェアとのいかなる特定の組み合わせにも限定されない。

【0159】

好適なコンピュータプログラムコードが、本明細書に記載する通りの、一つ以上の機能を遂行するために提供されてもよい。プログラムはまた、オペレーティングシステム212、データベース管理システム、および入力/出力コントローラ210を介して、プロセッサが、コンピュータ周辺装置（例えば、ビデオディスプレイ、キーボード、コンピュータマウスなど）と連動することが可能になる「装置ドライバー」などのプログラム要素を含んでもよい。

10

【0160】

「コンピュータ可読媒体」という用語は、本明細書で使用する場合、実行のために、コンピューティング装置200のプロセッサ（または本明細書に記載する装置のいかなる他のプロセッサ）に命令を提供する、またはその提供に關与する任意の非一時的媒体を指す。そのような媒体は、不揮発性媒体および揮発性媒体を含むが、これらに限定されない、多くの形態を取ってもよい。不揮発性媒体としては、例えば、光学、磁気もしくは光磁気ディスク、またはフラッシュメモリなどの集積回路メモリが挙げられる。揮発性媒体としては、通常主メモリを構成する、ダイナミックランダムアクセスメモリ（DRAM）が挙げられる。コンピュータ可読媒体のよくある形態としては、例えば、フロッピー（登録商標）ディスク、フレキシブルディスク、ハードディスク、磁気テープ、いかなる他の磁気媒体、CD-ROM、DVD、いかなる他の光学媒体、パンチカード、紙テープ、いかなる他の孔パターン付きの物理的媒体、RAM、PROM、EPROMもしくはEEPROM（電氣的消去可能なプログラマブル読み取り専用メモリ）、FLASH-EEPROM、いかなる他のメモリチップもしくはカートリッジ、またはコンピュータが読み取ることができるいかなる他の非一時的媒体が挙げられる。

20

【0161】

様々な形態のコンピュータ可読媒体が、実行のために、一つ以上の命令の一つ以上のシーケンスを、CPU 206（または、本明細書に記載する装置のいかなる他のプロセッサ）に運ぶのに關与してもよい。例えば、命令は最初、リモートコンピュータ（図示せず）の磁気ディスク上に置かれてもよい。リモートコンピュータは、命令をそのダイナミックメモリへロードし、Ethernet（登録商標）接続、ケーブル回線、またはモデムを使用する電話線さえも通して、命令を送る場合がある。コンピューティング装置200（例えば、サーバ）に対してローカルである通信装置は、それぞれの通信回線上でデータを受け取り、プロセッサ用のシステムバス上にデータを位置付けてもよい。システムバスは、プロセッサが命令を取得し実行する主メモリに、データを運ぶ。主メモリが受け取った命令は、任意選択により、プロセッサによって実行の前または後のいずれかに、メモリに記憶されてもよい。加えて、命令は、ワイヤレス通信または様々なタイプの情報を運ぶデータストリームの例示的な形態である、電気信号、電気磁気信号または光学信号として、通信ポートを介して受け取られてもよい。

30

40

【0162】

本明細書で参照する各参考文献は、参照することによって、そのそれぞれの全体が本明細書に組み込まれる。

【0163】

本開示の実装を、特定の実施例を参照して具体的に示し記載してきたが、本開示の範囲を逸脱することなく、添付の特許請求の範囲によって定義される通り、形態および詳細の様々な変更が本開示の実装になされてもよいことは、当業者によって理解されるべきである。よって、本開示の範囲は、添付の特許請求の範囲によって示され、したがって、特許請求の範囲の均等物の意味および範囲内に入る、すべての変化を受け入れることが意図され

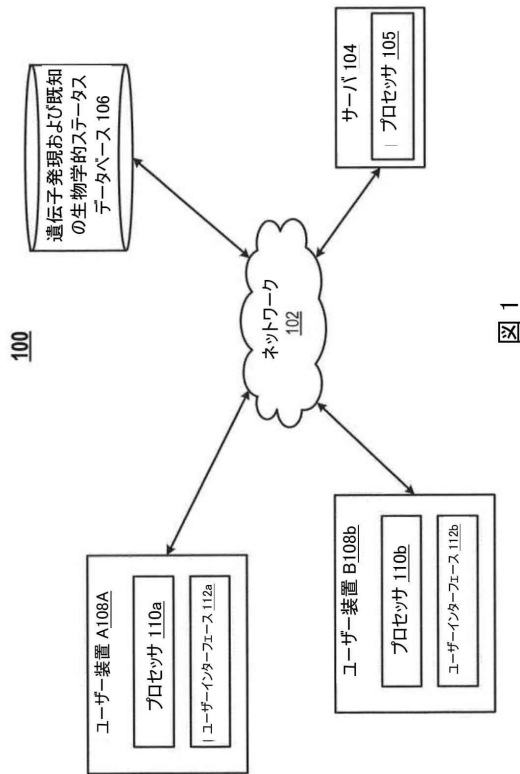
50

る。

【図面】

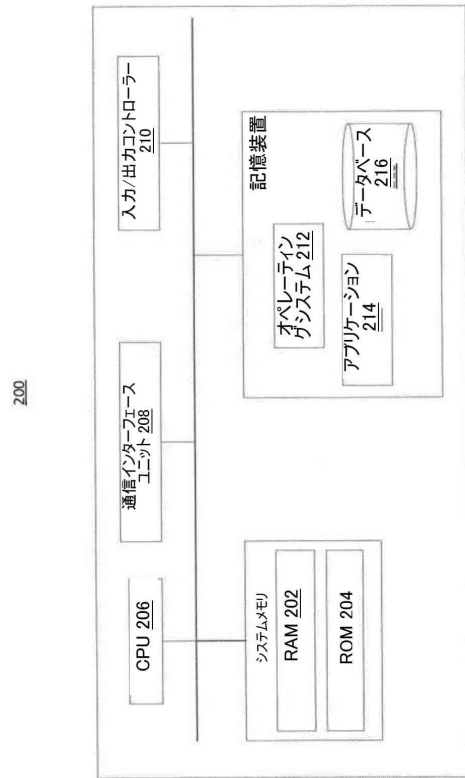
【図 1】

【図 1】



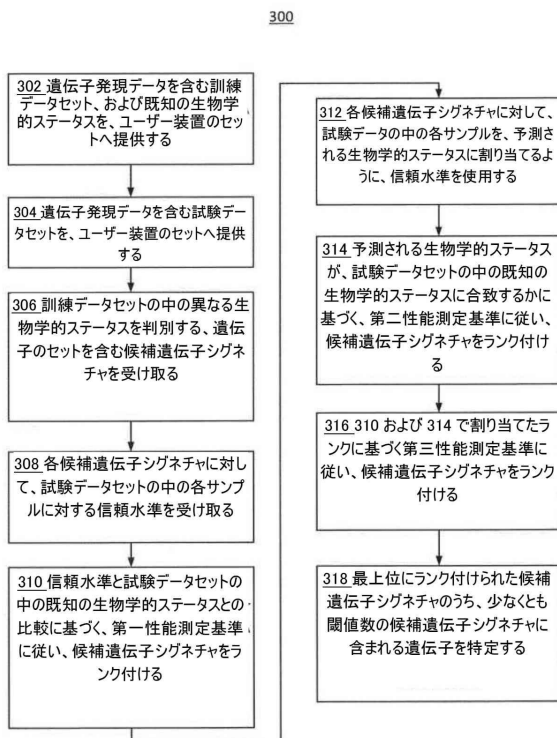
【図 2】

【図 2】



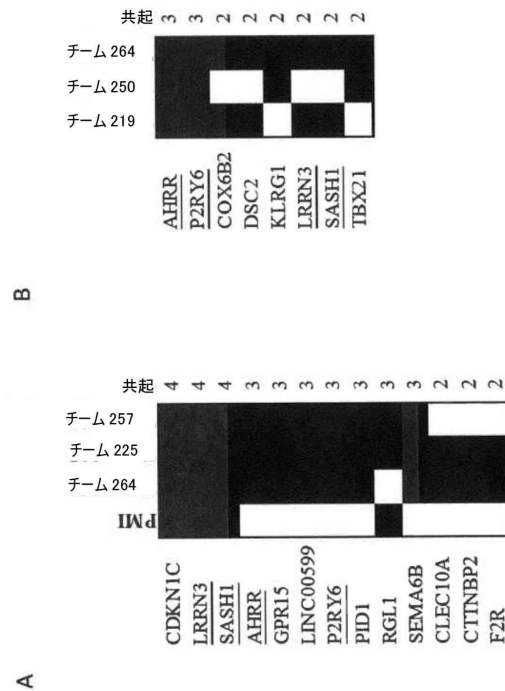
【図 3】

【図 3】



【図 4】

【図 4】



10

20

30

40

50

【 図 9 】

【 図 9 】

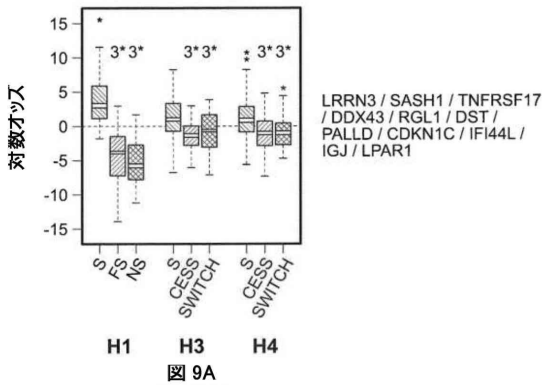


図 9A

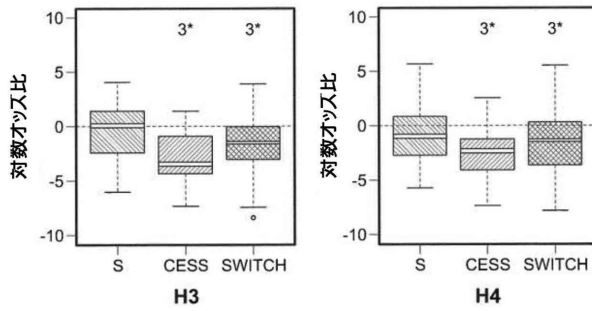


図 9B

【 図 10 】

【 図 10 】

SC2	平均ランク			
	チーム	AUPR	MCC	平均ランク
SC1	219	0.99	0.87	1
	250	0.96	0.81	2
	264	0.96	0.75	3
	225	0.73	0.38	4
	247	0.62	0.20	5.5
	221	0.46	0.39	5.5
SC1	225	0.94	0.24	1
	264	0.92	0.18	2
	257	0.91	0.16	3
	259	0.90	0.15	4
	269	0.90	0.10	6.5
	222	0.89	0.13	7
	250	0.89	0.12	7
	247	0.90	0.09	8
	283	0.90	0.08	8
	290	0.87	0.11	8.5
	221	0.85	0.06	11
	215	0.82	-0.07	12
PMI	0.93	0.24		

図 10

【 図 11 A 】

【 図 11 A 】

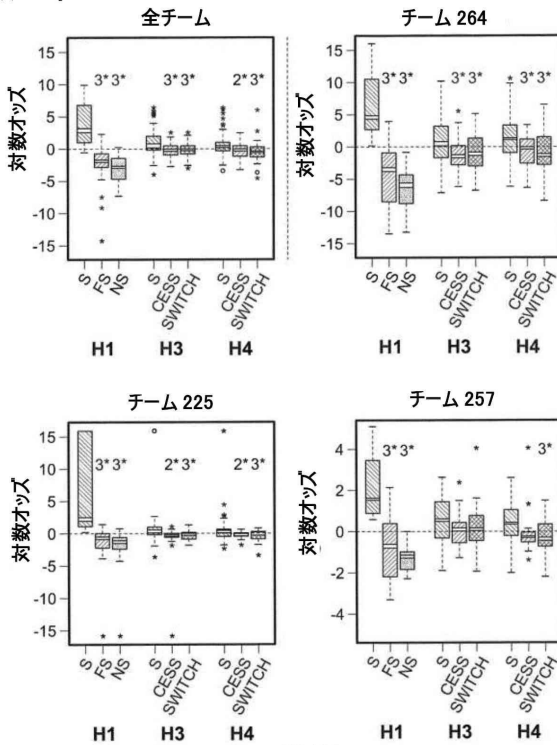


図 11A

【 図 11 B 】

【 図 11 B 】

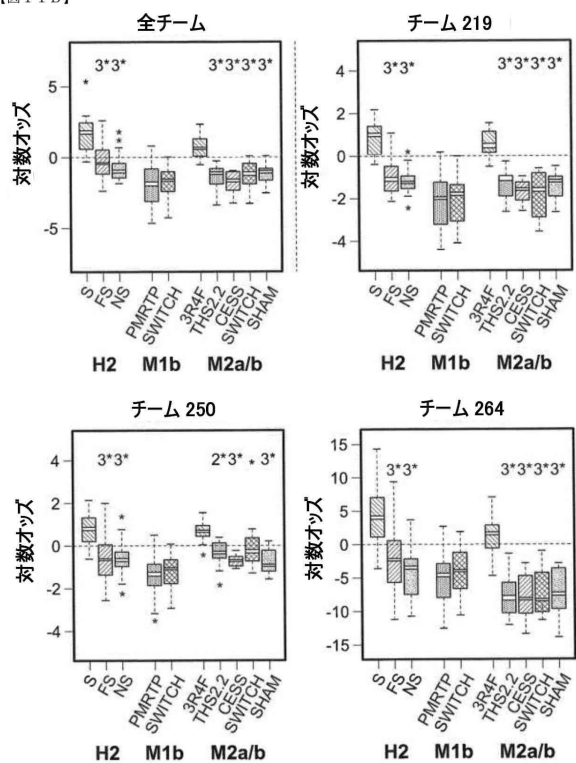


図 11B

10

20

30

40

50

【 図 1 2 】

【 図 1 2 】

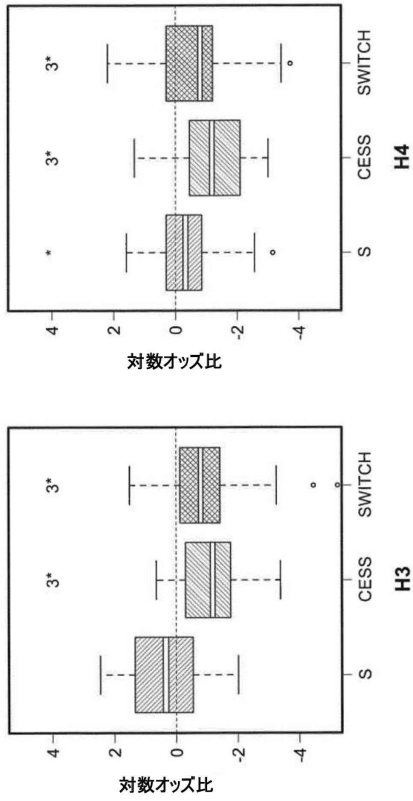


図 12

【 図 1 3 】

【 図 1 3 】

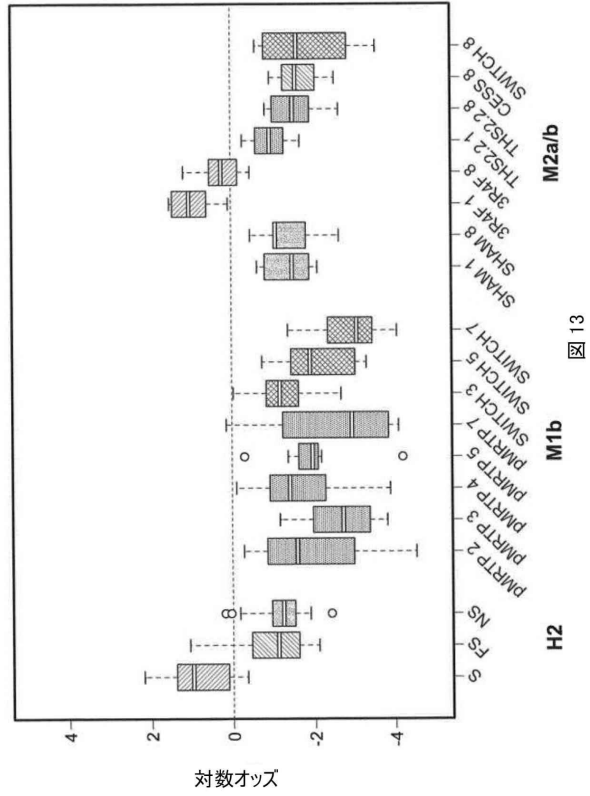


図 13

【 図 1 4 A 】

【 図 1 4 A 】

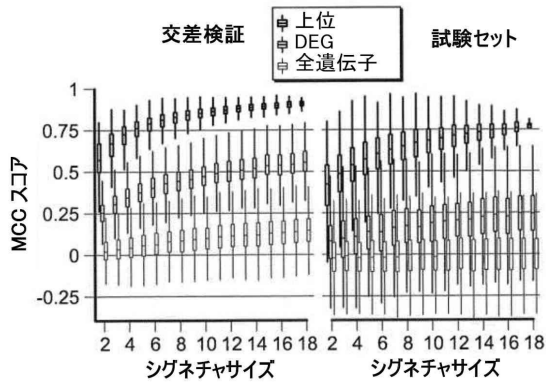


図 14A

【 図 1 4 B 】

【 図 1 4 B 】

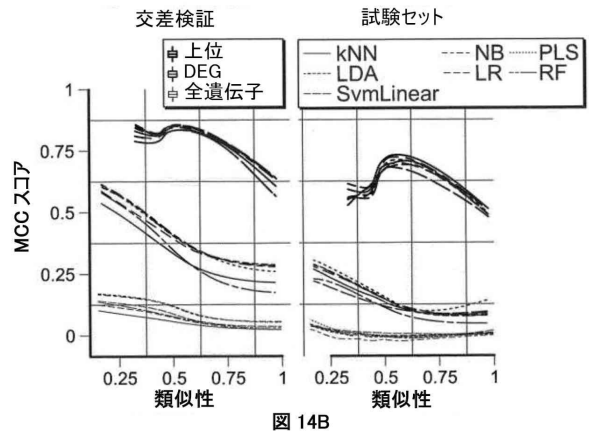


図 14B

10

20

30

40

50

【図 14 C】

【図 14 C】

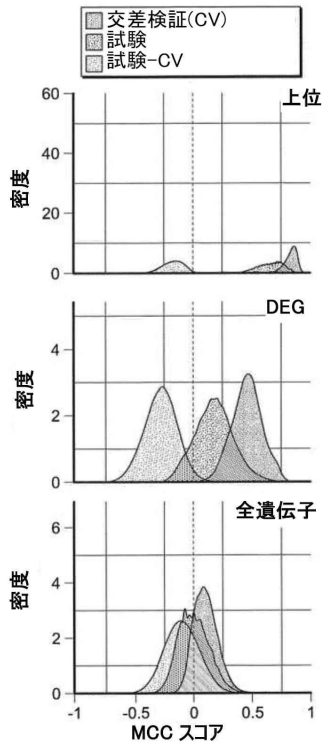


図 14C

【図 15 A】

【図 15 A】

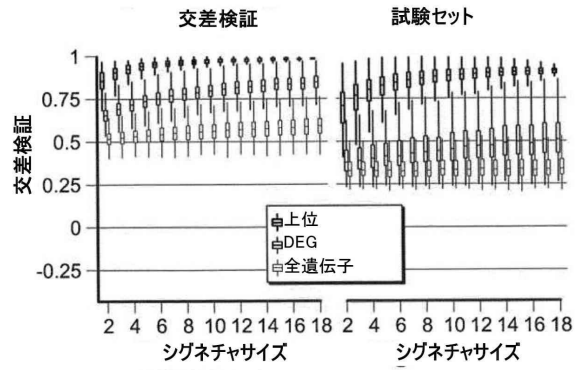


図 15A

10

20

【図 15 B】

【図 15 B】

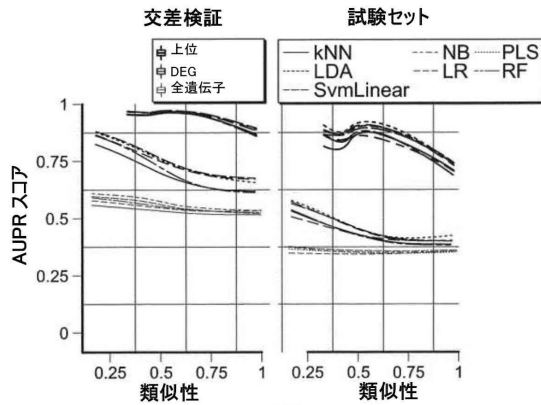


図 15B

【図 15 C】

【図 15 C】

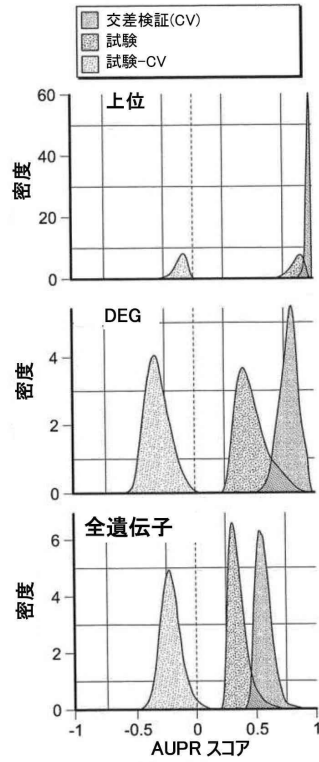


図 15C

30

40

50

フロントページの続き

- (74)代理人 100120525
弁理士 近藤 直樹
- (74)代理人 100139712
弁理士 那須 威夫
- (74)代理人 100158551
弁理士 山崎 貴明
- (72)発明者 プーサン, カリーヌ
スイス国 ツェーハー - 2000 ヌーシャテル, ケ ジャンルノー 3
- (72)発明者 ベルカストロ, ヴィンチェンツォ
スイス国 ツェーハー - 1400 イヴェルドン - レ - バン, リュ ドゥ フール 4
- (72)発明者 マルティン, フロリアン
スイス国 2000 ヌーシャテル, ケ ジャンルノー 3
- (72)発明者 プー, ステファニ
スイス国 ツェーハー - 2068 オトリヴ, ヴェルジェ クロツテュ 3
- (72)発明者 パイチ, マヌエル クロード
スイス国 ツェーハー - 2034 プスー, シュマン ガブリエル 5
- 審査官 山内 裕史
- (56)参考文献 特表2014 - 531202 (JP, A)
特開2006 - 314315 (JP, A)
特表2012 - 518997 (JP, A)
特表2011 - 516046 (JP, A)
- (58)調査した分野 (Int.Cl., DB名)
G16B 5/00 - 99/00