(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2019-532410
(P2019-532410A)
(43)公表日　令和1年11月7日(2019.11.7)

(51)Int.Cl.		F I			テーマコード（参考）
G16B 25/10	(2019.01)	G16B	25/10		4B063
C12Q 1/6874	(2018.01)	C12Q	1/6874	Z	

審査請求 未請求　予備審査請求 未請求　（全 55 頁）

(21)出願番号　特願2019-513943 (P2019-513943)
(86)(22)出願日　平成29年5月30日 (2017.5.30)
(85)翻訳文提出日　平成31年3月12日 (2019.3.12)
(86)国際出願番号　PCT/EP2017/063073
(87)国際公開番号　WO2018/050299
(87)国際公開日　平成30年3月22日 (2018.3.22)
(31)優先権主張番号　62/394,551
(32)優先日　平成28年9月14日 (2016.9.14)
(33)優先権主張国・地域又は機関
米国 (US)

(71)出願人　596060424
フィリップ・モーリス・プロダクツ・ソシエテ・アノニム
スイス国セアシュ－２０００　ヌシャテル、ケ、ジャンルノー　３
(74)代理人　100094569
弁理士　田中　伸一郎
(74)代理人　100109070
弁理士　須田　洋之
(74)代理人　100067013
弁理士　大塚　文昭
(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)【要約】
喫煙者ステータスなど、対象の生物学的ステータスを予測するように、対象のサンプルを評価するためのシステムおよび方法。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およびＴＬＲ５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。

【特許請求の範囲】
【請求項１】
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、ＡＨＨＲ、ＣＤＫＮ１Ｃ、ＬＲＲＮ３、ＰＩＤ１、ＧＰＲ１５、ＳＡＳＨ１、ＣＬＥＣ１０Ａ、ＬＩＮＣ００５９９、Ｐ２ＲＹ６、ＤＳＣ２、Ｆ２Ｒ、ＳＥＭＡ６ＢおよびＴＬＲ５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、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、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、ＡＫ８、ＦＳＴＬ１、ＲＧＬ１およびＶＳＩＧ４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、遺伝子の発現レベルを検出する、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を含む、試薬のセットと、
前記個人の喫煙者ステータスを予測する前記キットを使用するための説明書と、を備えるキット。

【請求項１０】
前記キットは、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、請求項９に記載のキット。
【請求項１１】
前記喫煙製品の前記代替品は、加熱式たばこ製品である、請求項１０に記載のキット。
【請求項１２】
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ことである、請求項９～１１のいずれかに記載のキット。
【請求項１３】
前記遺伝子シグネチャは更に、ＡＫ８、ＦＳＴＬ１、ＲＧＬ１およびＶＳＩＧ４を含む、請求項９～１２のいずれかに記載のキット。
【請求項１４】
前記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、請求項９～１３のいずれかに記載のキット。
【請求項１５】
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、ＬＲＲＮ３、ＡＨＨＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、ＬＩＮＣ００５９９、Ｐ２ＲＹ６、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＣＴＴＮＢＰ２およびＧＰＲ６３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、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、を含む、コンピュータ実装された方法。
【請求項１６】
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１５に記載のコンピュータ実装された方法。
【請求項１７】
ＬＲＲＮ３、ＡＨＨＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、ＬＩＮＣ００５９９、Ｐ２ＲＹ６、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＣＴＴＮＢＰ２およびＧＰＲ６３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１５～１６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１８】
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１７に記載のコンピュータ実装された方法。
【請求項１９】
前記遺伝子のセットは、ＬＲＲＮ３、ＡＨＨＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、ＬＩＮＣ００５９９、Ｐ２ＲＹ６、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＣＴＴＮＢＰ２およびＧＰＲ６３から成る、請求項１５に記載のコンピュータ実装された方法。
【請求項２０】
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１５～１９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２１】

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、遺伝子の発現レベルを検出する、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、
前記個人の喫煙者ステータスを予測する前記キットを使用するための説明書と、を備えるキット。
【請求項２２】
前記キットは、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、請求項２１に記載のキット。
【請求項２３】
前記喫煙製品の前記代替品が、加熱式たばこ製品である、請求項２２に記載のキット。
【請求項２４】
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ことである、請求項２１～２３のいずれかに記載のキット。
【請求項２５】
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であって、
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、前記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であって、
前記訓練データセットは、訓練サンプルのセットを含み、前記試験データセットは、試験サンプルのセットを含み、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、提供することと、
前記ネットワークから、前記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であって、各候補遺伝子シグネチャは、前記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、受け取ることと、
前記試験サンプルの前記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、前記それぞれの候補遺伝子シグネチャ各々へスコアを割り当てることと、
前記割り当てられたスコアに基づいて、前記候補遺伝子シグネチャのサブセットを特定することと、
前記サブセットの中で、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を特定することと、
前記特定された遺伝子を前記遺伝子シグネチャとして記憶することと、を含む、方法。
【請求項２６】
各候補遺伝子シグネチャに許容される遺伝子の最大閾値数を表す数字を、前記複数のユーザー装置へ提供することを更に含む、請求項２５に記載の方法。
【請求項２７】
前記試験データセットの一部分を前記ネットワークで、前記複数のユーザー装置へ提供することを更に含み、前記試験データセットの前記一部分は、既知の生物学的ステータスを有する患者に対する前記遺伝子発現データを含み、前記患者の前記既知の生物学的ステータスを含まない、請求項２５または２６に記載の方法。
【請求項２８】
各候補遺伝子シグネチャについて、前記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更に含む、請求項２７に記載の方法。

【請求項２９】
前記信頼水準は、前記試験データセットの中のサンプルが、前記生物学的ステータスのうちの一つに属すると予測される尤度を示す値である、請求項２８に記載の方法。
【請求項３０】
前記スコアは、前記信頼水準に少なくとも一部基づく、請求項２８または２９に記載の方法。
【請求項３１】
前記スコアは、前記試験データセットの中の前記信頼水準、および患者の前記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く、請求項３０に記載の方法。
【請求項３２】
前記スコアは、対応する前記候補遺伝子シグネチャが、前記試験データセットの中の前記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、請求項２５～３１のいずれかに記載の方法。
【請求項３３】
前記対応する候補遺伝子シグネチャが、前記試験データセットの中の前記患者の既知の生物学的ステータスと一致する前記予測を提供するかは、マシューズ相関係数（ＭＣＣ）を使用して決定される、請求項３２に記載の方法。
【請求項３４】
前記候補遺伝子シグネチャは、各候補遺伝子シグネチャに一位および二位を取得させるように、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、請求項２５～３３のいずれかに記載の方法。
【請求項３５】
各候補遺伝子シグネチャに対する前記一位および前記二位は、それぞれの候補遺伝子シグネチャ各々に前記スコアを取得させるように平均化される、請求項３４に記載の方法。
【請求項３６】
前記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、請求項２５～３５のいずれかに記載の方法。
【請求項３７】
前記喫煙者ステータスは、現喫煙者および非喫煙者を含む、請求項３６に記載の方法。
【請求項３８】
前記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、請求項２５～３７のいずれかに記載の方法。
【請求項３９】
前記遺伝子シグネチャは更に、ＡＫ８、ＦＳＴＬ１、ＲＧＬ１およびＶＳＩＧ４を含む、請求項３８に記載の方法。
【請求項４０】
前記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、請求項３９に記載の方法。
【請求項４１】
前記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を含む、請求項４０に記載の方法。
【請求項４２】
前記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、請求項２５～３７のいずれかに記載の方法。

【請求項４３】

前記遺伝子シグネチャは更に、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．を含む、請求項４２に記載の方法。

【請求項４４】

前記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、請求項２５～３７のいずれかに記載の方法。

【請求項４５】

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２５～４４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４６】

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、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、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、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、を含む、コンピュータ実装された方法。

【請求項４７】

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４６に記載のコンピュータ実装された方法。

【請求項４８】

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更に含む、請求項４６～４７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４９】

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４８に記載のコンピュータ実装された方法。

【請求項５０】

前記遺伝子のセットは、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、請求項４６～４９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５１】

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４６～５０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５２】

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、

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、前記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、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、試薬のセットと、

前記個人の喫煙者ステータスを予測する前記キットを使用するための説明書と、を備えるキット。

【請求項５３】

前記キットは、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、請求項５２に記載のキット。

【請求項５４】

前記喫煙製品の前記代替品は、加熱式たばこ製品である、請求項５３に記載のキット。

【請求項５５】

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ことである、請求項５２～５４のいずれかに記載のキット。

【請求項５６】

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、ＡＨＨＲ、Ｐ２ＲＹ６、ＫＬＲＧ１、ＬＲＲＮ３、ＣＯＸ６Ｂ２、ＣＴＴＮＢＰ２、ＤＳＣ２、Ｆ２Ｒ、ＧＵＣＹ１Ｂ３、ＭＴ２、ＮＧＦＲＡＰ１、ＲＥＥＰ６、ＳＡＳＨ１およびＴＢＸ２１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、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、を含む、コンピュータ実装された方法。

【請求項５７】
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５６に記載のコンピュータ実装された方法。
【請求項５８】
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更に含む、請求項５６～５７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５９】
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５８に記載のコンピュータ実装された方法。
【請求項６０】
前記遺伝子のセットは、ＡＨＨＲ、Ｐ２ＲＹ６、ＫＬＲＧ１、ＬＲＲＮ３、ＣＯＸ６Ｂ２、ＣＴＴＮＢＰ２、ＤＳＣ２、Ｆ２Ｒ、ＧＵＣＹ１Ｂ３、ＭＴ２、ＮＧＦＲＡＰ１、ＲＥＥＰ６、ＳＡＳＨ１およびＴＢＸ２１から成る、請求項５６に記載のコンピュータ実装された方法。
【請求項６１】
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５６～６０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６２】
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、
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、前記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、前記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、
前記個人の喫煙者ステータスを予測する前記キットを使用するための説明書と、を備えるキット。
【請求項６３】
前記キットは、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、請求項６２に記載のキット。
【請求項６４】
前記喫煙製品の前記代替品は、加熱式たばこ製品である、請求項６３に記載のキット。
【請求項６５】
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ことである、請求項６３～６４のいずれかに記載のキット。

【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
本出願は、米国特許法１１９条の下、２０１６年９月１４日に出願した米国仮特許出願第６２／３９４，５５１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、２０１４年１２月１１日に出願したＰＣＴ出願第ＰＣＴ／ＥＰ２０１４／０７７４７３号、および２０１４年８月１２日に出願したＰＣＴ出願第ＰＣＴ／ＥＰ２０１４／０６７２７６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】
【０００２】
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。２１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ＲＮＡ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。
【０００３】
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】
【０００４】
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。
【０００５】
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およびＴＬＲ５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。
【０００６】
ある実装では、遺伝子のセットは更に、ＡＫ８、ＦＳＴＬ１、ＲＧＬ１およびＶＳＩＧ４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、ＰＩＤ

１、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。

【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。

【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ＶＳＩＧ４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。

【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。

【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、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、ＳＥＭＡ

６Ｂ、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。
【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。
【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分）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。
【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分をネットワークで、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分は、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、マシューズ相関係数（ＭＣＣ）を使用して決定されてもよい。

【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。

【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ＶＳＩＧ４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。

【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、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を含んでもよい。一部の実装では、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。

【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。

【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。

【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。

【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。

【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、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、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。

【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。

【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。

【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。

【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】

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。
【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。
【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。
【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、Ｃｅｓｓ群およびＳｗｉｔｃｈ群に対して有意な減少を示す、二つの箱ひげ図を含む。
【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／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そ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。

【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ト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リスク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。

【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、キーパッド、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る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１０８ａおよび１０８ｂ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。

【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。

【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、現喫煙者対現非喫煙者）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に、いかなる好適な演算技法を使用してもよい。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。

【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。

【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。

【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、組み合わされてもよい。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し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渡る、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と組み合わされて、一つの構成要素となる。

【００６９】

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。

【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に関係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に関係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。

【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。

【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。

【００７３】

工程３１０で、サーバ１０４は、信頼水準（工程３０８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３０６で受信した）をランク付ける。工程３１０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。

【００７４】

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、真の現喫煙者）の数、およびサンプルが第一生物学的ステータス（例えば、予測される現喫煙者）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、真の現喫煙者）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。

【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

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１００％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１００％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。

【００７５】

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、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。

【００７６】

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、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、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、各候補遺伝子シグネチャを二位の値に割り当てさせる。

【００８２】

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】

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

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。

【００８３】

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。

【００８４】

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の少なくとも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。

【００８５】

実施例１－はじめに

【００８６】

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。

【００８７】

実施例１－研究対象母集団およびデザイン

【００８８】

全血サンプルは、臨床研究および生体内研究中に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たばこを使用し続ける（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。

【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。

【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。

【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 ２１００ 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）３８４Ｐｌｕｓ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）によって生み出される。

【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。

【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。

【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に関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。

【０１００】
データは、ヒトおよび齧歯類におけるＣＳ曝露および禁煙に関係する、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。

【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。

【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Ｃ５７

ＢＬ／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。ＲＥＸ　Ｃ－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）。

【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。

【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との間）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。

【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、マシューズ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。

【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評価に関係す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（ＲＥＸ　Ｃ－０３－ＥＵおよびＲＥＸ　Ｃ－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日間のみの禁煙または切り替えの後、これらの群に関係する対数オッズは、喫煙者群と比較すると有意に減少し、一方、Ｃｅｓｓ群とＳｗｉｔｃｈ群と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）。

【０１１８】

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ＭＲＴＰ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

【０１１９】
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Ｃｅｓｓ（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、ＡＵＰＲ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）。

【０１２０】
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３Ｒ４Ｆ）ヒト対象およびマウスと、現在煙に曝露されていない（ＮＣＳ）（喫煙経験者ＦＳ／Ｃｅｓｓおよび喫煙未経験者ＮＳ／Ｓｈａｍ）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／３Ｒ４Ｆ群に属するという信頼値Ｐ、およびサンプルがＮＣＳ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ＣｅｓｓおよびＳｗｉｔｃｈ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／３Ｒ４Ｆ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ＭＲＴＰ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。

【０１２１】
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

【０１２２】
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）。ＣＳ／３Ｒ４Ｆに曝露されるコホートと、曝露されない（喫煙未経験者／Ｓｈａｍおよび喫煙経験者／Ｃｅｓｓ）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ＭＲＴＰ（ｐＭＲＴＰ）または候補ＭＲＴＰ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、Ｓｈａｍ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）。

【０１２３】
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日目との間の、集

団の対数オッズ比を示す。対数オッズ比は、Ｃｅｓｓ群およびＳｗｉｔｃｈ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＊＜０．００１）。

【０１２４】

図１３は、群／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ＣＳ曝露からｐＭＲＴＰ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、ｐＭＲＴＰへの１か月、３か月および４か月の曝露に対応するＳｗｉｔｃｈ３、Ｓｗｉｔｃｈ５およびＳｗｉｔｃｈ７）、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。

【０１２５】

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。

【０１２６】

喫煙曝露コア遺伝子サブセットは、上位３チームおよびＰＭＩ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１Ｃ（ＣＤＫＮ１Ｃ）、ロイシンリッチリピート神経３型（ＬＲＲＮ３）、ならびにＳＡＭおよびＳＨ３ドメイン含有１（ＳＡＳＨ１）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４Ａ）、アリール炭化水素受容体リプレッサー（ＡＨＲＲ）、Ｐ２Ｙ６受容体（ｐｙｒｉｍｉｄｉｎｅｒｇｉｃ　ｒｅｃｅｐｔｏｒ：Ｐ２ＲＹ６）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４Ｂ）。両方のコア遺伝子サブセット間の比較により、ＬＲＲＮ３、ＳＡＳＨ１、ＡＨＲＲおよびＰ２ＲＹ６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）。

【０１２７】

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

【０１２８】

方法

【０１２９】

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。ＤＳＣ２、ＦＳＴＬ１、ＧＰＲ６３、ＧＳＥ１、ＧＵＣＹ１Ａ３、ＲＧＬ１、ＣＴＴＮＢＰ２、Ｆ２Ｒ、ＳＥＭＡ６Ｂ、ＣＤＫＮ１Ｃ、ＣＬＥＣ１０Ａ、ＧＰＲ１５、ＬＩＮＣ００５９９、Ｐ２ＲＹ６、ＰＩＤ１、ＳＡＳＨ１、ＡＨＲＲおよびＬＲＲＮ３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ＳＣ１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（ＭＬ）方法は、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最近傍（ｋＮＮ）、線形判別分析（ＬＤＡ）およびロジスティック回帰（ＬＲ）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ＭＬ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＜＝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。

【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。

【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、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ＭＲＴＰに曝露された対象、または従来のＣＳ曝露に続き、候補ＭＲＴＰ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、Ｆ２ＲＬ３遺伝子）が、生涯喫煙量（ｐａｃｋ　ｙｅａｒ）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、Ｃｅｓｓ群の発現レベルが、Ｓｈａｍ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、ＬＤＡ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ＲＥＸデータセットからのＳｗｉｔｃｈ群およびＣｅｓｓ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ＭＲＴＰ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ＭＲＴＰまたはｐＭＲＴＰ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ＣＳと比較して、ＭＲＴＰの血球へのより低い生物学的効果を反映していたため、３Ｒ４Ｆ群と、試作品／候補ＭＲＴＰ群またはＳｗｉｔｃｈ群（Ｓｈａｍ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、充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ＭＲＴＰなどの新製品の影響を監視し、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。

【０１３９】

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。

【０１４０】

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。

【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

【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の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、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、Ｐ２ＲＹ６、

ＬＩＮＣ００５９９、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２ＲおよびＣＴＴＮＢＰ２を含む。

【０１４２】

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の少なくとも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の値を持つそれらの遺伝子、すなわち、ＬＲＲＮ３、ＡＨＲＲおよびＣＤＫＮ１Ｃ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の値を持つそれらの遺伝子、すなわち、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１ＣおよびＰＩＤ１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の値を持つそれらの遺伝子、すなわち、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１およびＧＰＲ１５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の値を持つそれらの遺伝子、すなわち、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、Ｐ２ＲＹ６、ＬＩＮＣ００５９９およびＣＬＥＣ１０Ａ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の値を持つそれらの遺伝子、すなわち、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、Ｐ２ＲＹ６、ＬＩＮＣ００５９９、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＤＳＣ２およびＴＬＲ５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、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、Ｐ２ＲＹ６、ＬＩＮＣ００５９９、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＤＳＣ２、ＴＬＲ５、ＲＧＬ１、ＦＳＴＬ１、ＶＳＩＧ４およびＡＫ８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、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、Ｐ２ＲＹ６、ＬＩＮＣ００５９９、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＤＳＣ２、ＴＬＲ５、ＲＧＬ１、ＦＳＴＬ１、ＶＳＩＧ４、ＡＫ８、ＣＴＴＮＢＰ２、ＧＵＣＹ１Ａ３、ＧＳＥ１、ＭＩＲ４６９７ＨＧ、ＰＴＧＦＲＮ、ＬＯＣ２００７７２、ＦＡＮＫ１、Ｃ１５ｏｒｆ５４およびＭＡＲＣ２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、ＬＲＲＮ３、ＡＨＲＲ、ＣＤＫＮ１Ｃ、ＰＩＤ１、ＳＡＳＨ１、ＧＰＲ１５、Ｐ２ＲＹ６、ＬＩＮＣ００５９９、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＤＳＣ２、ＴＬＲ５、ＲＧＬ１、ＦＳＴＬ１、ＶＳＩＧ４、ＡＫ８、ＣＴＴＮＢＰ２、ＧＵＣＹ１Ａ３、ＧＳＥ１、ＭＩＲ４６９７ＨＧ、ＰＴＧＦＲＮ、ＬＯＣ２００７７２、ＦＡＮＫ１、Ｃ１５ｏｒｆ５４、ＭＡＲＣ２、ＧＰＲ６３、ＴＰＰＰ３、ＺＮＦ６１８、ＰＴＧＦＲ、ＧＵＣＹ１Ｂ３、Ｐ２ＲＹ１、ＴＭＥＭ１６３、ＳＴ６ＧＡＬＮＡＣ１、ＳＨ２Ｄ１Ｂ、ＣＹＰ４Ｆ２２、ＰＦ４、ＦＵＣＡ１、ＭＢ２１Ｄ２、ＮＬＫ、Ｂ３ＧＡＬＴ２、ＡＳＧＲ２およびＮＲ４Ａ１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。

【０１４３】

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。

【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

【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
KCNQ1OT1	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

【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

【０１４４】

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の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、ＡＨＲＲ、Ｐ２ＲＹ６、ＣＯＸ６Ｂ２、ＤＳＣ２、ＫＬＲＧ１、ＬＲＲＮ３、ＳＡＳＨ１およびＴＢＸ２１を含む。

【０１４５】

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の少なくとも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、ＡＨＲＲ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、ＡＨＲＲおよびＰ２ＲＹ６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、ＡＨＲＲ、Ｐ２ＲＹ６、ＫＬＲＧ１およびＬＲＲＮ３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、ＡＨＲＲ、Ｐ２ＲＹ６、ＫＬＲＧ１、ＬＲＲＮ３、ＣＯＸ６Ｂ２、ＤＳＣ２、ＳＡＳＨ１、ＴＢＸ２１、ＣＴＴＮＢＰ２、Ｆ２Ｒ、ＧＵＣＹ１Ｂ３、ＭＴ２、ＮＧＦＲＡＰ１およびＲＥＥＰ６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。

【０１４６】

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ぐ場合がある。表２および３に関係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。

【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、ＬＩＮＣ００５９９、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に関係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。

【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。

【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、ＬＩＮＣ００５９９、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に関係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に関係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。

【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、ＬＩＮＣ００５９９、Ｐ２ＲＹ６、ＣＬＥＣ１０Ａ、ＳＥＭＡ６Ｂ、Ｆ２Ｒ、ＣＴＴＮＢＰ２およびＧＰＲ６３のセットのいかなる数のい

か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、ＡＨＨＲ、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に関係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に関係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、ＡＨＨＲ、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、ＡＨＨＲ、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／出力コントローラー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ー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。

【０１５４】
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。
【０１５５】
ＣＰＵ２０６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ＣＰＵ２０６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。ＣＰＵ２０６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、ＣＰＵ２０６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。
【０１５６】
ＣＰＵ２０６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、ＲＡＭ２０２、ＲＯＭ２０４、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。ＣＰＵ２０６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、ＣＰＵ２０６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。ＣＰＵ２０６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。
【０１５７】
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ＣＰＵ２０６に関して詳細に記載するプロセスに従って、ＣＰＵ２０６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。
【０１５８】
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、ＲＯＭ２０４からまたはＲＡＭ２０２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、ＣＰＵ２０６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／出力コントローラー２１０を介して、プロセッサが、コンピュータ周辺装置（例えば、ビデオディスプレー、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。

【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、またはその提供に関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（電気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。

【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、ＣＰＵ２０６（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）に運ぶのに関与してもよい。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。

【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

れる。

【図１】

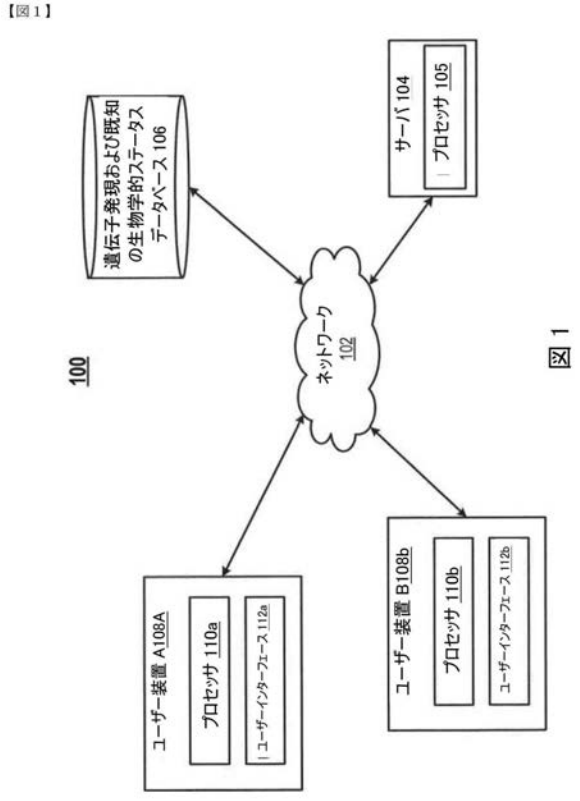

【図1】
100
遺伝子発現および既知の生物学的ステータスデータベース 106
サーバ 104
プロセッサ 105
ネットワーク 102
ユーザー装置 A108A
プロセッサ 110a
ユーザーインターフェース 112a
ユーザー装置 B108b
プロセッサ 110b
ユーザーインターフェース 112b

図 1

【図２】

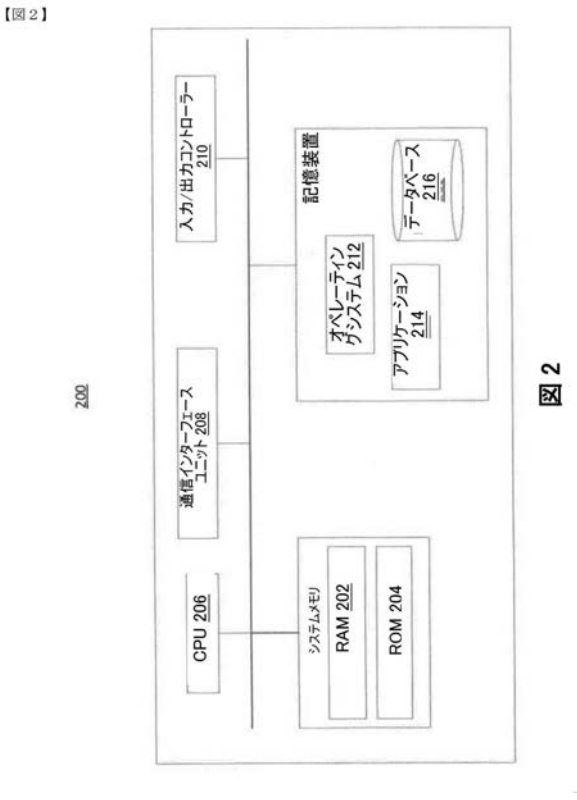

【図2】
200
CPU 206
システムメモリ
RAM 202
ROM 204
通信インターフェースユニット 208
入力/出力コントローラー 210
記憶装置
オペレーティングシステム 212
アプリケーション 214
データベース 216

図 2

【図３】

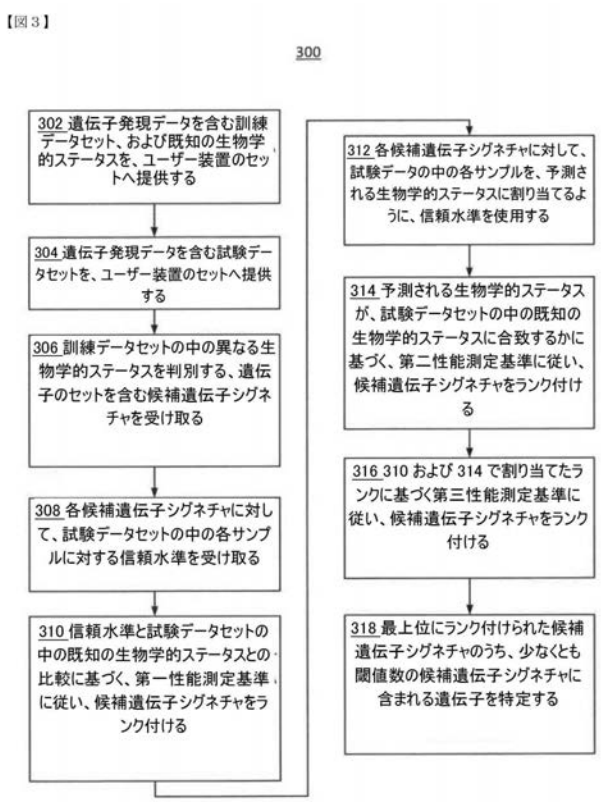


図 3

【図４】

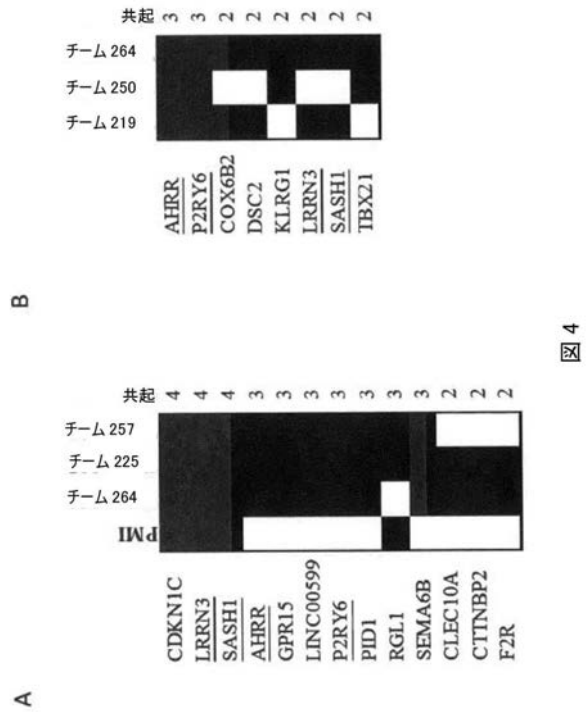


図 4

【図５】

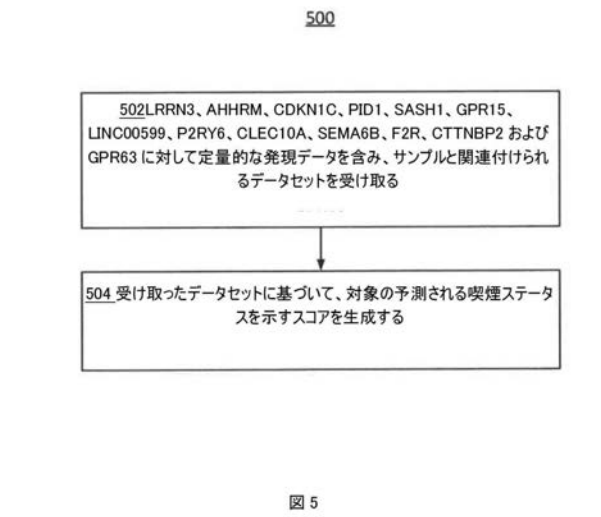


図 5

【図６】

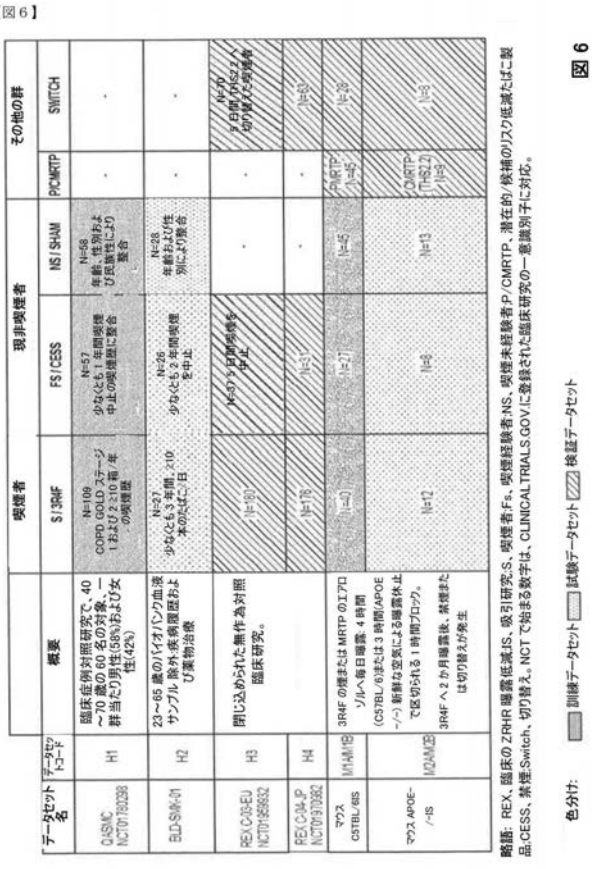
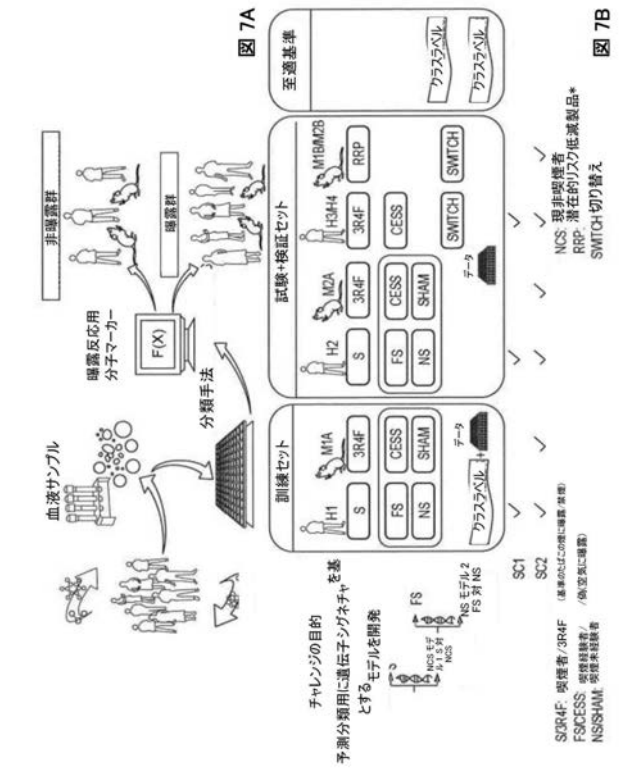
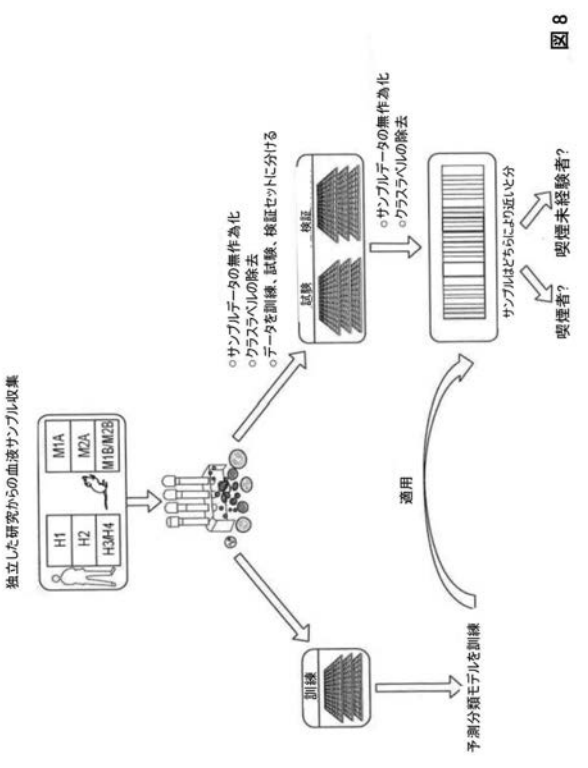


図 6

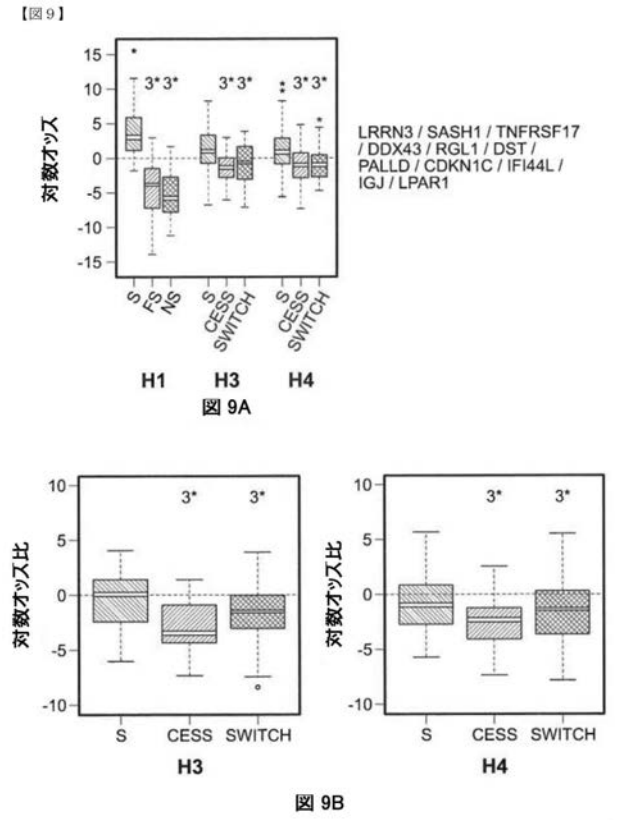
【図７】



【図８】



【図９】



【図１０】

【図10】

SC2

チーム	AUPR	MCC	平均ランク
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

チーム	AUPR	MCC	平均ランク
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１１Ａ】

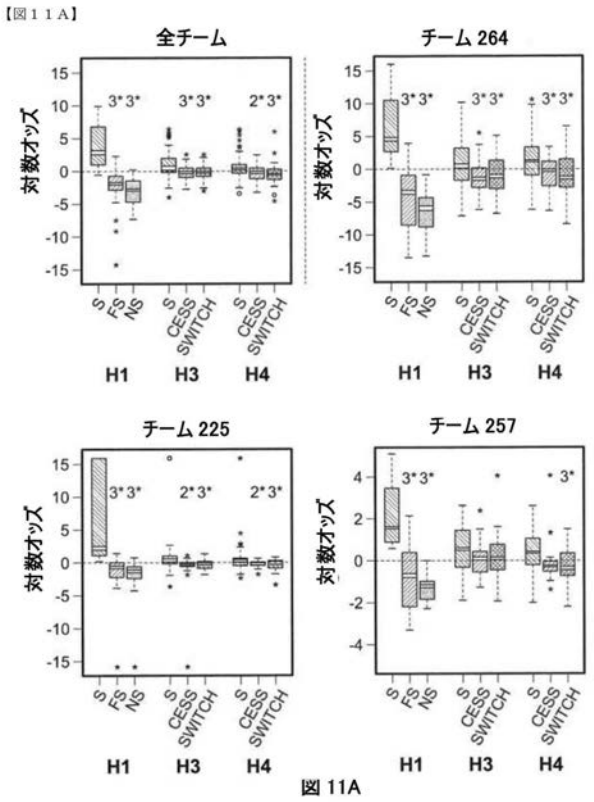


図 11A

【図１１Ｂ】

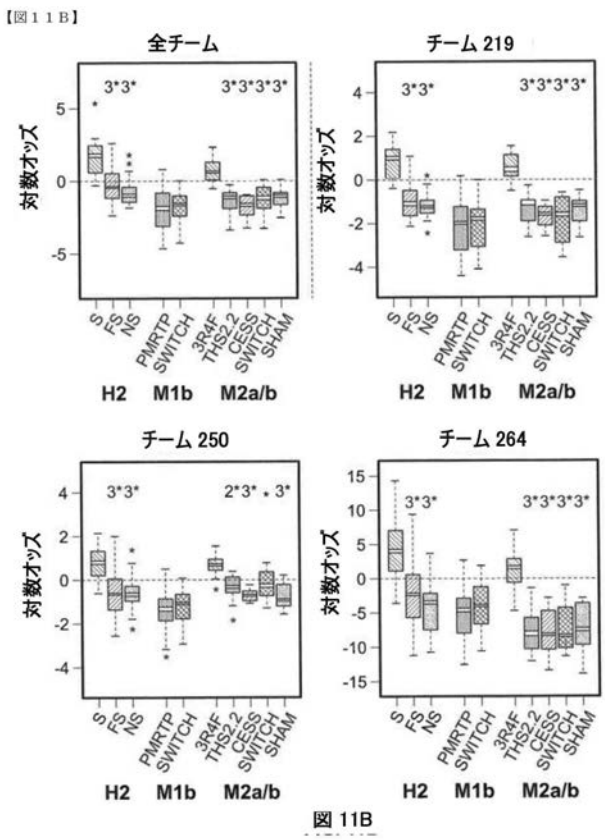


図 11B

【図１２】

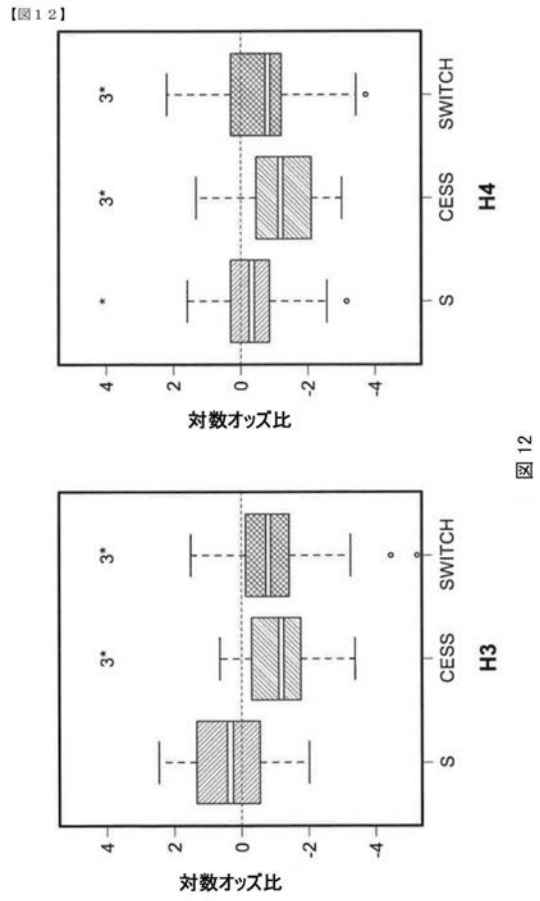


図 12

【図１３】

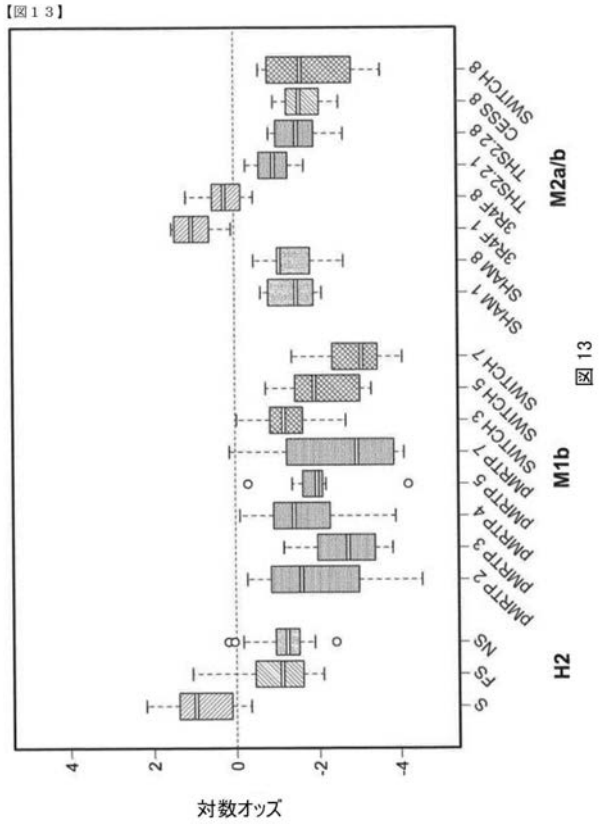


図 13

【図１４Ａ】

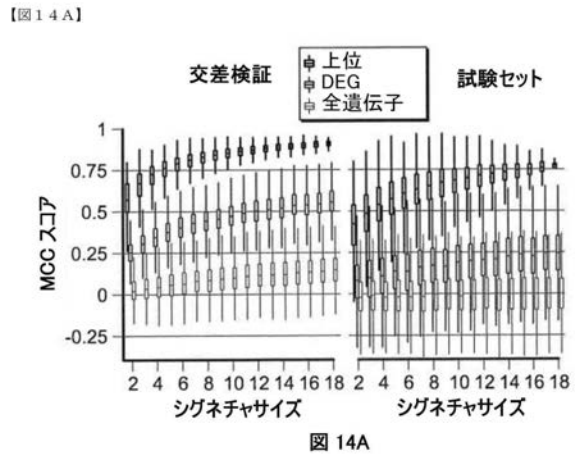


図 14A

【図１４Ｂ】

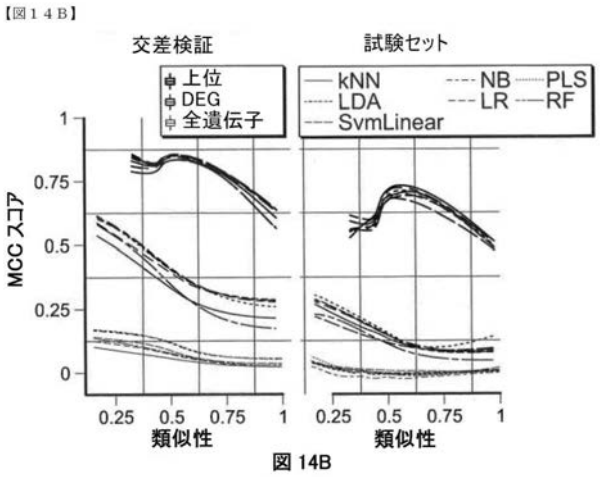


図 14B

【図１４Ｃ】

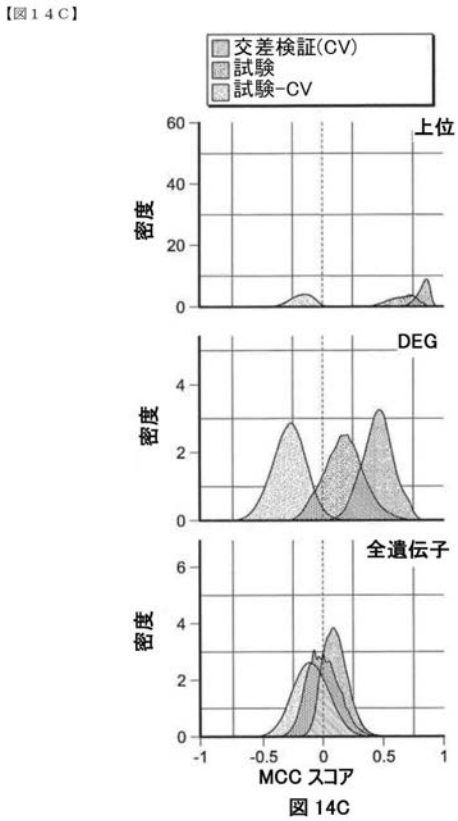


図 14C

【図１５Ａ】

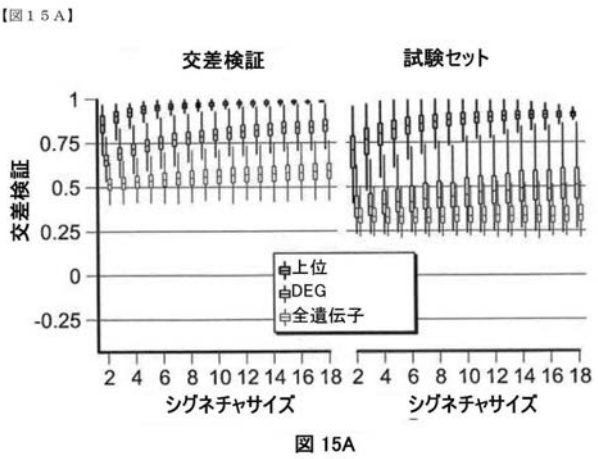


図 15A

【図１５Ｂ】

【図１５Ｂ】

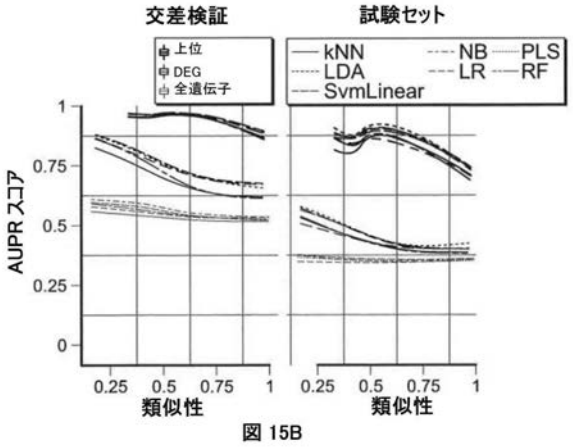


図 15B

【図１５Ｃ】

【図１５Ｃ】

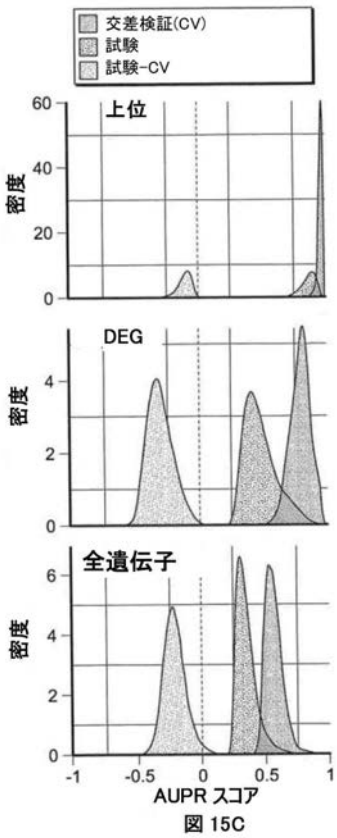


図 15C

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/063073

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/18 G06F19/24 C12Q1/68 G06F19/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, Sequence Search, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2013/032917 A2 (CARDIODX INC [US]; ROSENBERG STEVEN [US]; ELASHOFF MICHAEL REID [US];) 7 March 2013 (2013-03-07) [0030],[0037], [0060]-[0066], [0073]; Table 2, Claim 1 -----	1-24, 46-65
X	PHILIP BEINEKE ET AL: "A whole blood gene expression-based signature for smoking status", BMC MEDICAL GENOMICS, BIOMED CENTRAL LTD, LONDON UK, vol. 5, no. 1, 3 December 2012 (2012-12-03), page 58, XP021137778, ISSN: 1755-8794, DOI: 10.1186/1755-8794-5-58 The whole document, esp. abstract, pgs. 2, 3 ----- -/--	1-24, 46-65

[X] Further documents are listed in the continuation of Box C. [X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 August 2017	21/11/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Wimmer, Georg

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/063073

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	RICARDO A. VERDUGO ET AL: "Graphical Modeling of Gene Expression in Monocytes Suggests Molecular Mechanisms Explaining Increased Atherosclerosis in Smokers", PLOS ONE, vol. 8, no. 1, 23 January 2013 (2013-01-23), page e50888, XP055188292, DOI: 10.1371/journal.pone.0050888 the whole document, esp. Table 4, pgs. 8, 10 -----	1-24, 46-65
X	"Array-Based Gene Expression Analysis", www.illumina.com , 10 April 2013 (2013-04-10), XP002739403, Retrieved from the Internet: URL:https://web.archive.org/web/20130410192832/http://www.illumina.com/documents/products/datasheets/datasheet_gene_exp_analysis.pdf [retrieved on 2015-05-08] the whole document, esp. pgs. 2, 3 -----	6-14

1

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/EP2017/063073

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-24, 46-65

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

International Application No. PCT/EP2017/063073

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-24, 46-65

Computer-implemented methods and kits for assessing smoker status of an individual

1.1. claims: 1-14(completely); 46-55(partially)

As invention 1,
wherein the set of genes comprises AHHR, CDKN1C, LRRN3, PID1, GPR15, SASH1, CLEC10A, LINC00599, P2RY6, DSC2, F2R, SEMA6B, and TLR5.

1.2. claims: 15-24(completely); 46-55(partially)

As invention 1, but wherein the set of genes comprises LRRN3, AHHR, CDKN1C, PID1, SASH1, GPR15, LINC00599, P2RY6, CLEC10A, SEMA6B, F2R, CTTNBP2, and GPR63.

1.3. claims: 56-65

As invention 1, but wherein
the set of genes comprises AHHR, P2RY6, KLRG1, LRRN3, COX6B2, CTTNBP2, DSC2, F2R, GUCY1B3, MT2, NGFRAP1, REEP6, SASH1, and TBX21.

---

2. claims: 25-45

A computer-implemented method for obtaining a gene signature for predicting biological status, the method comprising: providing, by a computer system including a communications port and at least one computer processor in communication with at least one non-transitory computer readable medium storing at least one electronic database comprising a training data set and a test data set, the training data set over a network to a plurality of user devices, wherein: the training data set includes a set of training samples and the test data set includes a set of test samples, where each training sample and each test sample includes gene expression data, and corresponds to a patient having a known biological status selected from a set of biological statuses; receiving, from the network, candidate gene signatures that are each generated by obtaining a classifier based on the training data set, wherein each candidate gene signature includes a set of genes that are determined to be discriminant between different biological statuses in the training data set;assigning a score to each respective candidate gene signature based on a performance of the respective candidate gene signature in predicting the known biological status of the test samples;identifying a subset of the candidate gene signatures based on the assigned scores;identifying genes that were included in at least a

International Application No. PCT/EP2017/063073

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

threshold number of candidate gene signatures in the subset;
and storing the identified genes as the gene signature

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2017/063073

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2013032917 A2	07-03-2013	AU 2012300375 A1	20-03-2014
		BR 112014004768 A2	13-06-2017
		CA 2846837 A1	07-03-2013
		CN 103890193 A	25-06-2014
		EA 201490533 A1	29-08-2014
		EP 2751290 A2	09-07-2014
		JP 2014531202 A	27-11-2014
		KR 20140051461 A	30-04-2014
		SG 11201400243P A	28-03-2014
		US 2015178462 A1	25-06-2015
		WO 2013032917 A2	07-03-2013

Form PCT/ISA/210 (patent family annex) (April 2005)

フロントページの続き

(81)指定国・地域 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ

(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－レーバン, 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 ブー, ステファニ
スイス国 ツェーハー－２０６８ オトリヴ, ヴェルジェ クロッテュ 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

Ｆターム(参考) 4B063 QA01 QA19 QQ03 QQ52 QR42 QR55 QR84 QS34 QS39 QX02