

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第6703479号
(P6703479)

(45) 発行日 令和2年6月3日(2020.6.3)

(24) 登録日 令和2年5月12日(2020.5.12)

(51) Int.Cl.		F I	
G 1 6 H 50/30	(2018.01)	G 1 6 H	50/30
C 1 2 Q 1/68	(2018.01)	C 1 2 Q	1/68
C 1 2 N 15/00	(2006.01)	C 1 2 N	15/00
C 1 2 M 1/00	(2006.01)	C 1 2 M	1/00

請求項の数 5 (全 26 頁)

(21) 出願番号	特願2016-532548 (P2016-532548)	(73) 特許権者	500586875
(86) (22) 出願日	平成26年12月11日 (2014.12.11)		フィリップ モリス プロダクツ エス アー
(65) 公表番号	特表2017-505932 (P2017-505932A)		スイス国 2000 ヌーシャテル ケ ジャンルノー 3
(43) 公表日	平成29年2月23日 (2017.2.23)	(74) 代理人	100078282
(86) 国際出願番号	PCT/EP2014/077473		弁理士 山本 秀策
(87) 国際公開番号	W02015/091225	(74) 代理人	100113413
(87) 国際公開日	平成27年6月25日 (2015.6.25)		弁理士 森下 夏樹
審査請求日	平成29年12月11日 (2017.12.11)	(74) 代理人	100181674
(31) 優先権主張番号	61/916,443		弁理士 飯田 貴敏
(32) 優先日	平成25年12月16日 (2013.12.16)	(74) 代理人	100181641
(33) 優先権主張国・地域又は機関	米国 (US)		弁理士 石川 大輔
前置審査		(74) 代理人	230113332
			弁護士 山本 健策

最終頁に続く

(54) 【発明の名称】 個人の喫煙ステータスを予測するためのシステムおよび方法

(57) 【特許請求の範囲】

【請求項1】

被験者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、受け取る回路によって、前記サンプルに関連付けられたデータセットを受け取ることであって、前記データセットが L R R N 3、C D K N 1 C、P A L L D、S A S H 1、R G L 1、および T N F R S F 1 7 を含む遺伝子シグネチャに対する定量的な発現データを含むことと、

プロセッサによって、前記遺伝子シグネチャに対する前記定量的な発現データに対して分類子を適用することにより、前記遺伝子シグネチャに対する前記定量的な発現データに基づいてスコアを発生することであって、前記分類子が、遺伝子発現レベルの線形結合に基づいており、前記分類子が、Matthewsの相関係数(MCC)を演算することにより線形判別解析(LDA)モデルの特性を見積もることによって、訓練されており、前記MCCは、1つ以上の訓練データセットの前記遺伝子シグネチャに対する定量的な発現データと、喫煙者ステータスとに基づいて演算され、前記スコアが前記被験者の予測される喫煙ステータスを示すことと、を含む、コンピュータ実装された方法。

【請求項2】

前記遺伝子シグネチャに対する前記定量的な発現データが、I G J、R R M 2、S E R P I N G 1、F U C A 1、および I D 3 を更に含む、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】

前記遺伝子シグネチャにおける LRRN3、CDKN1C、PALLD、SASH1、RGL1、および TNFRSF17 の各々に対してフォールド変化値を演算することを更に含む、請求項 1 ~ 2 のいずれか 1 項に記載のコンピュータ実装された方法。

【請求項 4】

各演算されたフォールド変化値が所定の閾値を超えることを決定することを更に含む、請求項 3 に記載のコンピュータ実装された方法。

【請求項 5】

コンピュータ可読指示を含むコンピュータプログラムであって、少なくとも 1 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した時、請求項 1 ~ 4 のいずれか 1 項に記載の前記方法の 1 つ以上の工程を前記プロセッサに実施させる、コンピュータプログラム。

10

【発明の詳細な説明】

【背景技術】

【0001】

関連出願の相互参照

本出願は、合衆国法典第 35 編米国特許法第 119 条の下で、2013 年 12 月 16 日出願の「Systems and Methods for Predicting a Smoking Status of an Individual」と題する米国仮特許出願第 61/916,443 号の優先権を主張するものであり、その全体が参照により本明細書に組み込まれる。

20

【0002】

全ゲノム発現レベルを測定し、そして様々な条件に対する生物学的洞察を得るために、全ゲノムマイクロアレイを実用的な手段として使用する。このアプローチは、活性物質への曝露に対する身体の応答を評価し、そして結果として得られる表現型を予測するためにも使用される。組織学的異常が見られない時でさえ、煙への曝露に応答した喫煙者の太い気道の細胞のトランスクリプトームの分子変化を検出することができる。この観察は、様々な物質の曝露に対する生物系の応答を評価するためにトランスクリプトームデータを使用することができる可能性があることを示す。

【0003】

数多くの製品リスク評価研究では、所望の原発部位（気道などの）からのサンプルの取得は侵襲的であり、不便でもある。別の方法としての末梢血サンプリングは、侵襲性が最低限であり、一般的な母集団で広く使用される。従って、代理の組織として機能する末梢血を信頼して使用することができるバイオマーカーを見出し、かつ確立することには利益がある。

30

【0004】

以前に、ケース母集団と対照母集団との間で差次的に発現される遺伝子の特定に焦点を当てた分子バイオマーカーの発見が試みられた。最近の方法により、ますます予想性の高い新しいケースに遭遇することができ、これによって診断の強化、予後の改善、個人向け医療の進化がもたらされる。しかしながら、臨床用途のためのロバストでかつ多目的な演算方法論の開発は、まだ困難なままである。喫煙に関連する疾病については、末梢血サンプル内の診断シグネチャが特定されてきた。少なくとも 2 つの研究は、差次的に発現される遺伝子が、対照被験者または非悪性肺疾患を有する被験者から初期の非小細胞肺癌を有する被験者を判別することができることを示した (Rotunno, M., Hu, N., Su, H., Wang, C., Goldstein, A.M., Bergen, A.W., Consonni, D., Pesatori, A.C., Bertazzi, P.A., Wacholder, S., et al. (2011). A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. Cancer Prev Res (Phila) 4, 1599-1608; Showe, M.K., Vachani, A., Kossenkov, A.V., Yousef, M., N

40

50

ichols, C., Nikonova, E. V., Chang, C., Kucharczuk, J., Tran, B., Wakeam, E., et al. (2009). Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. *Cancer Res* 69, 9202 - 9210).

【先行技術文献】

【非特許文献】

【0005】

10

【非特許文献1】Rotunno, M., Hu, N., Su, H., Wang, C., Goldstein, A. M., Bergen, A. W., Consonni, D., Pesatori, A. C., Bertazzi, P. A., Wacholder, S., et al. (2011)

【非特許文献2】A gene expression signature from peripheral whole blood for stage I lung adenocarcinoma. *Cancer Prev Res (Phila)* 4, 1599 - 1608; Showe, M. K., Vachani, A., Kossenkov, A. V., Yousef, M., Nichols, C., Nikonova, E. V., Chang, C., Kucharczuk, J., Tran, B., Wakeam, E., et al. (2009)

20

【非特許文献3】Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. *Cancer Res* 69, 9202 - 9210)

【発明の概要】

【課題を解決するための手段】

【0006】

個人の喫煙者ステータスを予測するために使用することができ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。本明細書に記述される遺伝子シグネチャは、現在喫煙している被験者を、喫煙したことがない被験者または喫煙をやめた被験者と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する能力を有する。

30

【0007】

特定の態様では、本開示のシステムおよび方法は、被験者から取得したサンプルを評価するためのコンピュータ化した方法を提供する。コンピュータ化した方法は、回路を受け取ることによって、サンプルに関連付けられたデータセットを受け取ることを含み、このデータセットは、LRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17に対する定量的な発現データを含む。プロセッサは、受け取ったデータセットに基づいて、予測される被験者の喫煙ステータスを示すスコアを発生する。予測される喫煙ステータスは、現在喫煙している者として、または現在喫煙していない者として被験者を分類する場合がある。

40

【0008】

特定の実施では、データセットは、IGJ、RRM2、SERPING1、FUCA1、およびID3に対する定量的な発現データを更に含む。特定の実施では、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内の定量的な発現データに基づいて決定される。

【0009】

特定の実施では、この方法は、LRRN3、CDKN1C、PALLD、SASH1、

50

RGL1、およびTNFRSF17の各々に対してフォールド変化値を演算し、少なくとも1つの基準を満たす各フォールド変化値を決定することを更に含む。基準は、少なくとも2つの独立した母集団データセットに対してそれぞれの演算されたフォールド変化値が各々所定の閾値を超えることを必要とする場合がある。

【0010】

特定の態様では、本開示のシステムおよび方法は、被験者から取得したサンプルを評価するためのコンピュータ化した方法を提供する。1つの装置は、試験サンプル内のLRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17を含む遺伝子シグネチャ内の前記遺伝子の発現レベルを検出するための手段を含む。この装置は、発現レベルを喫煙者ステータスの分類と関連するための手段、および被験者の喫煙者ステータスの予測として喫煙者ステータスの分類を出力するための手段も含む。

10

【0011】

特定の態様では、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。このキットは、試験サンプル内のLRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17を含む遺伝子シグネチャ内の遺伝子の発現レベルを検出する一組の試薬と、個人の喫煙者ステータスを予測するためのキットを使用するための説明書とを含む。

【0012】

特定の態様では、本開示のシステムおよび方法は、喫煙製品の代替品の個人に対する効果を評価するためのキットを提供する。このキットは、試験サンプル内のLRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17を含む遺伝子シグネチャ内の遺伝子の発現レベルを検出する一組の試薬と、代替品の個人に対する効果を評価するための前記キットを使用するための説明書とを含む。喫煙製品の代替品は、加熱たばこ製品(HTP)であってもよく、そして代替品の個人に対する効果は、その個人を非喫煙者として分類することであってもよい。

20

【0013】

特定の態様では、本開示のシステムおよび方法は、被験者から取得したサンプルを評価するための方法を提供する。この方法は、回路を受け取ることによって、サンプルに関連付けられたデータセットを受け取ることを含む。このデータセットは、LRRN3、CDKN1C、PALLD、SASH1、RGL1、TNFRSF17、IGJ、RRM2、SERPING1、FUCA1、およびID3から成る群から選択される少なくとも5つのマーカーに対する定量的な発現データを含む。この方法は、受け取ったデータセットに基づいてプロセッサによって、予測される被験者の喫煙ステータスを示すスコアを発生することを更に含む。被験者の予測される喫煙ステータスは、現在喫煙している者として、または現在喫煙していない者として被験者を分類する場合がある。

30

【0014】

このスコアは、このデータセットに適用される分類スキームの結果である場合があり、分類スキームはデータセット内の定量的な発現データに基づいて決定される。この方法は、LRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17の各々に対してフォールド変化値を演算し、少なくとも1つの基準を満たす各フォールド変化値を決定することを更に含む場合がある。基準は、少なくとも2つの独立した母集団データセットに対してそれぞれの演算されたフォールド変化値が各々所定の閾値を超えることを必要とする場合がある。

40

【0015】

特定の態様では、本開示のシステムおよび方法は、被験者から取得したサンプルを評価するための装置を提供する。この装置は、LRRN3、CDKN1C、PALLD、SASH1、RGL1、TNFRSF17、IGJ、RRM2、SERPING1、FUCA1、およびID3から成る群から選択される少なくとも5つのマーカーを含む遺伝子シグネチャ内の遺伝子の発現レベルを検出するための手段を含む。この装置は、発現レベルを喫煙者ステータスの分類と関連するための手段、および被験者の喫煙者ステータスの予測

50

として喫煙者ステータスの分類を出力するための手段を更に含む。

【0016】

特定の態様では、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。このキットは、試験サンプル中のLRRN3、CDKN1C、PALLD、SASH1、RGL1、TNFRSF17、IGJ、RRM2、SERPING1、FUCA1、およびID3から成る群から選択される少なくとも5つのマーカーを含む遺伝子シグネチャ内の遺伝子の発現レベルを検出する一組の試薬と、個人の喫煙者ステータスを予測するための前記キットを使用するための説明書とを備える。

【0017】

特定の態様では、本開示のシステムおよび方法は、喫煙製品の代替品の個人に対する効果を評価するためのキットを提供する。このキットは、試験サンプル中のLRRN3、CDKN1C、PALLD、SASH1、RGL1、TNFRSF17、IGJ、RRM2、SERPING1、FUCA1、およびID3から成る群から選択される少なくとも5つのマーカーを含む遺伝子シグネチャ内の遺伝子の発現レベルを検出する一組の試薬と、代替品の個人に対する効果を評価するための前記キットを使用するための説明書とを含む。喫煙製品に対する代替品は、HTPであってもよく、そして代替品の個人に対する効果は、その個人を非喫煙者として分類することであってもよい。

本発明は、例えば、以下を提供する。

(項目1)

被験者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、
受け取る回路によって前記サンプルに関連付けられたデータセットを受け取ることであ
って、前記データセットがLRRN3、CDKN1C、PALLD、SASH1、RGL
1、およびTNFRSF17に対する定量的な発現データを含むことと、
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発生することであ
って、前記スコアが前記被験者の予測される喫煙ステータスを示すことと、を含む、コン
ピュータ実装された方法。

(項目2)

前記データセットが、IGJ、RRM2、SERPING1、FUCA1、およびID
3に対する定量的な発現データを更に含む、項目1に記載のコンピュータ実装された方法
。

(項目3)

前記スコアが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って、前記分類ス
キームが前記データセット内の前記定量的な発現データに基づいて決定される、項目1ま
たは2のいずれか1項に記載のコンピュータ実装された方法。

(項目4)

LRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF
17の各々に対してフォールド変化値を演算することを更に含む、項目1～3のいずれか
1項に記載のコンピュータ実装された方法。

(項目5)

各演算されたフォールド変化値のそれぞれが少なくとも2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1つの基準を、各フォールド変
化値が満たすことを決定することを更に含む、項目4に記載のコンピュータ実装された方
法。

(項目6)

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、
試験サンプル内のLRRN3、CDKN1C、PALLD、SASH1、RGL1、お
よびTNFRSF17を含む遺伝子シグネチャ内の前記遺伝子の発現レベルを検出する一
組の試薬と、
前記個人の喫煙者ステータスを予測するための前記キットを使用するための説明書と、を
備えるキット。

10

20

30

40

50

(項目7)

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前記キットが使用される、項目6に記載のキット。

(項目8)

前記喫煙製品の前記代替品が加熱式たばこ製品である、項目7に記載のキット。

(項目9)

前記個人による前記加熱式たばこ製品の使用開始後0～5日の間にLRRN3発現の減少が検出される、項目8に記載のキット。

(項目10)

前記代替品の前記個人に対する前記効果が、前記個人を非喫煙者として分類するものである、項目7～9のいずれか1項に記載のキット。

10

(項目11)

前記遺伝子シグネチャが、IGJ、RRM2、SERPING1、FUCA1、およびID3のうちの少なくとも1つを更に含む、項目6～10のいずれか1項に記載のキット

。

(項目12)

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、試験サンプル中のLRRN3、CDKN1C、PALLD、SASH1、RGL1、TNFRSF17、IGJ、RRM2、SERPING1、FUCA1、およびID3から成る群から選択される少なくとも5つのマーカーを含む遺伝子シグネチャ内の前記遺伝子の発現レベルを検出する一組の試薬と、

20

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

(項目13)

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キットであって、前記喫煙製品の前記代替品が加熱式たばこ製品である、項目12に記載のキット。

(項目14)

前記個人による前記加熱式たばこ製品の使用開始後0～5日の間にLRRN3発現の減少が検出される、項目13に記載のキット。

(項目15)

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であって、少なくとも1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した時、項目1～5のいずれか1項に記載の前記方法の1つ以上の工程を前記プロセッサに実施させる、コンピュータプログラム製品。

30

【図面の簡単な説明】**【0018】**

本開示の更なる特徴、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部分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。

40

【0019】

【図1】 遺伝子の集合を特定し、遺伝子の集合に基づいて分類モデルを取得するためのプロセスのフローチャートである。

【図2】 被験者から取得したサンプルを評価するためのプロセスのフローチャートである。

【図3】 本明細書に記載されるコンピュータ化したシステムのいずれかの中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。

【図4】 サンプルのデータセットで差次的に発現される遺伝子に対するボルケーノプロットである。

50

【図5】異なる研究に対する分類スキームを示す様々なボックスプロットである。

【発明を実施するための形態】

【0020】

(詳細な説明)

個人の喫煙者ステータスを予測するために使用することができ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ための演算システムおよび方法を本明細書に記述する。特に、本明細書に記述する遺伝子シグネチャは、現在喫煙している被験者を喫煙したことがない被験者または喫煙をやめた被験者特別する能力を有する。

【0021】

本明細書で使用する場合、「ロバスト」な遺伝子シグネチャは、研究、研究期間、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特性を維持するものを指す。ロバストシグネチャは、大きい個人差を含む母集団データの集合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、シグネチャの特性の過度の楽観的な報告を避けるためにも適切に認証されるべきである。

【0022】

本開示の1つの目的は、個人の喫煙者ステータスを正確に予測することができる遺伝子シグネチャを取得することである。遺伝子シグネチャの特性を見積もるために、本明細書に予測の結果を表に示す。この表では、予測されるステータスを横列に示し、実際のステータスを縦列に示す。以下に示す表1は、予測結果を表示するための1つのやり方の実施例である。この表の第1の横列は、現在喫煙している者の実際の数および現在喫煙していない者の実際の数を示し、これらの者のサンプルは現在喫煙している者と関連付けられることが予測されており、この表の第2の横列は、実際に現在喫煙している者の数および実際に現在喫煙していない者の数を示し、これらの者のサンプルは現在喫煙していない者と関連付けられることが予測されている。

【表1】

	現在喫煙している者	現在喫煙していない者
(予想) 現在喫煙している者	真陽性	偽陽性
(予想) 現在喫煙していない者	偽陰性	真陰性

完全な予測子は、すべての現在喫煙している者を現在喫煙している者として正確に予測し（真陽性が100%となり、偽陰性が0%となる）、そしてすべての現在喫煙していない者を現在喫煙していない者として正確に予測する（真陰性が100%となり、偽陽性が0%となる）。本明細書に記述するように、個人は、喫煙ステータス（例えば、現在喫煙している者、現在喫煙していない者、以前喫煙していた者、喫煙したことがない者、等々）に従って分類されるが、一般に、当業者であれば本明細書に記述されるシステムおよび方法が任意の分類スキームに適用可能であることを理解するであろう。

【0023】

予測子の強さを見積もるために、予測結果表内の値に基づく様々な評価基準が使用される場合がある。本明細書では、現在喫煙している者の集合の中から現在喫煙している者として正確に分類される個人の割合である1つの評価基準が「感受性」として参照される。換言すれば、感受性評価基準は、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、すなわちと $TP / (TP + FN)$ に等しい。感受性値が1であるということは、現在喫煙している者についての完全な分類を示す。本明細書では、現在喫煙していない者の集合の中から現在喫煙していない者として正確に分類される個人の割合である1つの別の評価基準が「特異性」として参照される。換言すれば、特異性評価基準は、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、すなわちと $TN / (TN + FP)$ に等しい。特異性値が1であるということは、現在喫煙していない者についての完全な分類を示す。強い

10

20

30

40

50

予測子であると考えられるためには、感受性と特異性との両方の値が高いことが望ましい。本明細書では、予測子の特性を見積もるために感受性評価基準および特異性評価基準が使用されるが、一般には、陽性試験の予測値 ($TP / (TP + FP)$) または陰性試験の予測値 ($TN / (TN + FN)$) などの任意の他の評価基準も本開示の範囲を逸脱することなく使用される場合がある。

【0024】

本明細書に記述されるシステムおよび方法は、先ず異なる訓練データセットからその発現レベルが高いフォールド変化を呈する遺伝子を特定することによって予測モデルを構築する。次いで、独立したデータセットを用いて特定された遺伝子の集合が認証される。認証の後、既知の喫煙者ステータスを持つ被験者の血液トランスクリプトームを見積もり、1つの喫煙者ステータスを有する個人の特定された遺伝子集合からの発現レベルを、別の喫煙者ステータスを有する個人と比較することで、遺伝子集合が試験された。結果として得られる、順調に認証されかつ試験された遺伝子の集合は、本明細書では「遺伝子シグネチャ」と称される。

10

【0025】

遺伝子シグネチャは、個人を特定の予測される喫煙者ステータス群へと正確に分類するために使用することができる。更に、個人の喫煙者ステータスを正確に予測することができることで、遺伝子シグネチャは、HTPを使用する個人と従来の紙巻たばこを喫煙する個人の結果を比較することによって様々なHTPの使用を検出することができるようになる。遺伝子シグネチャは、喫煙行動に関する適合性が要求される状況で使用されてもよい。一例では、個人の予測される喫煙者ステータス（遺伝子シグネチャによって決定される）は、個人がHTPに切り替えた後、生物学的に変化するかどうか、またはいつ生物学的に変化するかを特定するために、HTPに対する臨床試験で使用されてもよい。一般に、遺伝子シグネチャは、紙巻たばこ喫煙、禁煙、またはHTPへの切り替えをモニターする、任意の健康に関係する研究に使用される場合がある。

20

【0026】

一例では、いくつかの公的に入手可能な遺伝子発現データセットから、現在喫煙している者および非喫煙者または以前喫煙していた者からの血液サンプルをプロファイリングしたいくつかのデータを取得した。様々な独立した研究からのフォールド変化の高い遺伝子に基づいて遺伝子を予め選択することは有利である。そうすることによってシグネチャのロバスト性を異なる研究にわたって強化し、かつ単一のデータセットによって予測モデルにバイアスがないようにするためである。COPDに対する新規なバイオマーカーを発見することを目的とした臨床的な研究から誘導された独立したデータセットによって、認証が遂行される。更に、別の臨床研究から、従来の紙巻たばこ（たばこを燃焼する）からHTP（たばこを燃焼しない、本明細書ではたばこ加熱システム（THS）2.1と称する）に連続5日間切り替えた喫煙者の血液トランスクリプトームが見積もられ、従来の紙巻たばこを継続的に喫煙している喫煙者の血液トランスクリプトームと比較された。本明細書に記述されるシグネチャは、独立したデータセットを使用したその特性によって実証されるように、現在喫煙している者と現在喫煙していない者との分類を顕著に良好に行う。更に、5日間のTHS 2.1への切り替えの影響は、血液トランスクリプトーム内で検出可能であり、THS 2.1に切り替えた被験者は現在喫煙していない者に分類された。これは、本明細書の遺伝子シグネチャ並びにシステムおよび方法が、喫煙者ステータスを決定するためだけでなく、紙巻たばこ喫煙の短期的な影響を見積もる上でも有用である場合があることを示唆する。

30

40

【0027】

限られた数の遺伝子に基づくシグネチャを使用することは、コストおよび作業負荷を低減するという点で全トランスクリプトームを使用することと比較して有利である。解析は、最終的には定量的逆転写酵素ポリメラーゼ連鎖反応（qRT-PCR）測定に基づくからである。qRT-PCRを使用するための装置への投資および試薬代などのランニングコストは、マイクロアレイの使用よりも好ましいものである。

50

【0028】

－実施例では、遺伝子シグネチャを特定するために、最初の工程で異なる訓練データセットが取得される。具体的には、本明細書では2つの訓練データセット、すなわちBLD-SMK-01およびQASMCが使用される。しかしながら、一般に、本開示の範囲を逸脱することなく、任意の数の任意の組み合わせの訓練データセットが使用されてもよい。

【0029】

BLD-SMK-01用にPAXgene血液DNAキット(Qiagen)を使用して収集した血液サンプルを、保存レポジトリ(BioServe Biotechnologies Ltd, (米国メリーランド州、Beltsville、20705))から取得した。サンプリングのとき、被験者の年齢は、23歳～65歳であった。病歴のない被験者および処方薬を服用している被験者は除外された。現在喫煙している者は、少なくとも1日につき10本の紙巻たばこを少なくとも3年間喫煙していた。以前喫煙していた者は、少なくともサンプリングの2年前に喫煙をやめていて、少なくとも1日につき10本の紙巻たばこを少なくとも3年間喫煙していた。現在喫煙している者と非喫煙者とは、年齢および性別を一致させた。現在喫煙している者から合計で31個の血液サンプルを取得し、喫煙したことがない者から30個の血液サンプルを取得し、そして以前喫煙していた者から30個の血液サンプルを取得した。

【0030】

Queen Ann Street Medical Center(QASMC)の臨床研究からも血液サンプルを取得した。これは、The Heart and Lung Centre(英国、London)においてGood Clinical Practice(GCP)に従って行われ、そしてこれは識別子、NCT01780298でClinicalTrials.govに登録されている。QASMCの研究は、バイオマーカーまたはバイオマーカーのパネルを特定することを目的とするものだが、これらのバイオマーカーはCOPDを患う被験者(現在喫煙している者、10パック年の喫煙履歴のある、GOLD Stage 1または2)と一致させた非喫煙被験者の3つの対照群(喫煙したことがない者、以前喫煙していた者、および現在喫煙している者)との間の区別を可能にすることになる。60人の被験者からのサンプルを4つの群の各々で取得した(合計被験者240人)。40～70歳の男性および女性の被験者が含まれた。すべての被験者は、研究で採用されたCOPD被験者と人種、性別、および年齢(5年以内)を一致させた。血液サンプルをAROS Applied Biotechnology AS(デンマーク、Aarhus)に送付し、ここで血液サンプルを更に処理し、次いで以下に記述するように、Affymetrix Human Genome U133 Plus 2.0 GeneChipsと交配した。

【0031】

製造業者の説明書に従ってPAXgene Blood miRNA Kit(catalog number, 763134, Qiagen)を使用して、全RNA(microRNAを含む)を分離した。RNAサンプルの濃度および純度は、UV分光光度計(NanoDrop ND1000, Thermo Fisher Scientific(米国マサチューセッツ州Waltham))を使用して、230、260、および280 nmにおける吸光度を測定することによって決定された。RNAの完全性は、Agilent 2100 Bioanalyzerを使用して更に確認された。上記6つのRNA完全性番号(RIN)を有するRNAのみが更なる解析のために処理された。

【0032】

RNAの調製およびAffymetrix交配。50 ngのRNAから、NuGEN(商標)Ovation(商標)全血試薬およびNuGEN(商標)Ovation(商標)RNA Amplification System V2を使用して転写の3'終止部を標的としたAffymetrixプローブセットを調製した。cDNAの量はNanodrop 1000または8000、分光光度計(Thermo Fisher Sc

10

20

30

40

50

ientific)、またはSpectraMax 384 Plus (Molecular Devices)を用いて測定された。cDNAの質は、断片化していないcDNAのサイズをAgilent 2100 Bioanalyzer使用して評価することによって決定された。最終的な断片化およびピオチン化した生成物のサイズ分布も、エレクトロフェログラムを使用してモニターされた。cDNAをラベリングした後、製造業者のガイドラインに従って、断片をGeneChip Human Genome U133 Plus 2.0 Arrayに交配した。標的の調製のサンプルは、Affymetrix遺伝子発現マイクロアレイのために完全にランダム化された。

【0033】

Taqman qRT-PCRアッセイ。逆転写反応を、iScript (商標) cDNA Synthesis Kit (カタログ番号170-8890、Bio-Rad (米国カリフォルニア州Hercules))を使用して、製造業者の説明書に従って500 ngの開始RNAを用いて遂行した。次いで、このcDNAを希釈して厳密に10 ng/ μ Lにした。市販のヒトユニバーサルRNA (UHR) 参照 (Cat # 740000、Agilent Technologies (米国カリフォルニア州Santa Clara))を、複数の実験および機器にわたって信頼性のあるデータの比較を行うために較正用としてサンプルに加えた。Taqmanアッセイ中で使用されたプローブはエクソンをスパンニングし、そしてデータの正規化工程のために5つのハウスキーピング遺伝子 (B2M、GAPDH、FARP1、A4GALT、GINS2) が選択された。qPCR工程は、Taqman (登録商標) アッセイおよびTaqMan (登録商標) Fast Advanced Master Mix (cat: 444963) を使用して実施された。簡潔に述べると、cDNAを希釈して、384ウェルプレートで、ウェルにつき1.25 ngで適用できるようにした。並行して、各Taqmanアッセイ用にマスターミックス (Taqmanアッセイ試薬およびTaqman Advanced Mixのミックス) を調製した。最終的な反応容積は10 μ Lであった。ViiA7機器 (Life Technologies) を使用してqPCRを実行し、そして結果を解析するために自動ベースラインおよびデフォルト C_t 閾値設定を適用した。ユニバーサルヒト参照 (UHR) サンプルを追加すると、各々の遺伝子用に (サブトラクションによって) C_t 値がUHR C_t 値に関して正規化され、次いでGAPDHハウスキーピング遺伝子値 (いわゆる C_t 値をもたらず) に関して正規化された。

【0034】

TaqmanプライマーをLife Technologies (米国カリフォルニア州) から取得した。下記の表2にqRT-PCRを遂行するために使用されるプライマー配列を一覧表示する。

10

20

30

【表2 - 1】

アッセイID	可用性	カタログ番号	アッセイタイプ	遺伝子記号	遺伝子名
Hs00539582_s1	在庫あり	4331182	遺伝子	LRRN3	hCG1643830 C e l e r a アノテーション;ロイシン富化繰返しニューロナル3
Hs03045080_m1	在庫あり	4331182	遺伝子	TNFRSF17	hCG14623 C e l e r a アノテーション;腫瘍壊死因子レセプタースーパーファミリー;メンバー17
Hs00376160_m1	在庫あり	4331182	遺伝子	IGJ	hCG17003 C e l e r a アノテーション;免疫グロブリンJポリペプチド;免疫グロブリン α 及び μ ポリペプチド用リンカータンパク質
Hs00323932_m1	在庫あり	4331182	遺伝子	SASH1	hCG16768 C e l e r a アノテーション;SAMおよびSH3ドメイン含有1
Hs00357247_g1	在庫あり	4331182	遺伝子	RRM2	hCG23833 C e l e r a アノテーション;リボヌクレオチド還元酵素M2
Hs00363100_m1	在庫あり	4331182	遺伝子	PALLD	パラジン;細胞骨格関連タンパク質;hCG2026123 C e l e r a アノテーション
Hs00954037_g1	在庫あり	4331182	遺伝子	ID3	hCG1982882 C e l e r a アノテーション;DNA結合阻害物質3;優性負のらせんループ-らせんタンパク質
Hs00163781_m1	在庫あり	4331182	遺伝子	Hs00163781_m1	セルピンペプチダーゼ阻害物質;クレードG (C1阻害物質);メンバー1;hCG39766 C e l e r a アノテーション
Hs00175938_m1	在庫あり	4331182	遺伝子	CDKN1C	サイクリン依存性キナーゼ阻害物質1C (p57;Kip2);hCG1782992 C e l e r a アノテーション
Hs00609173_m1	在庫あり	4331182	GE	FUCA1	hCG1739246 C e l e r a アノテーション;フコシダーゼ; α -L-1;組織
Hs99999907_m1	在庫あり	4331182	GE	B2M	hCG1786707 C e l e r a アノテーション; β -2-ミクログロブリン
Hs02758991_g1	在庫あり	4331182	GE	GAPDH	グリセルアルデヒド3リン酸デヒドロゲナーゼ;hCG2005673 C e l e r a アノテーション

10

20

30

40

【表 2 - 2】

Hs00195010_m1	在庫あり	433 118 2	GE	FARP1	hCG1811328 Celer a アノテーション; FERM; ρ GEF (ARHGEF) およびプレクストリン ドメインタンパク質1 (軟骨細胞由 来)
Hs00213726_m1	在庫あり	433 118 2	GE	A4G ALT	α 1; 4-ガラクトシルトランスフェ ラーゼ; hCG1640515 Ce l e r aアノテーション
Hs00211479_m1	在庫あり	433 118 2	GE	GIN S2	GINS錯体サブユニット2 (Psf 2ホモログ); hCG15657 C e l e r aアノテーション
Hs00248508_m1	在庫あり	433 118 2	GE	RGL 1	hCG2025089 Celer a アノテーション; ralグアニンヌク レオチド解離刺激物質様1

10

【0035】

マイクロアレイ解析 - データ品質の確認および正規化。チップスキャン上のアーチファクトを検出するためのチップ画像の調査後、標準的な品質管理パイプラインを通してデータを処理した。簡潔に述べると、*affy*パッケージのReadAffy機能 (Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). *affy - analysis of Affymetrix GeneChip data at the probe level*. *Bioinformatics* 20, 307 - 315) を使用して、マイクロアレイ解析ツールのBioconductor suite (Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). *Bioconductor: open software development for computational biology and bioinformatics*. *Genome Biol* 5, R80) から (R統計的環境用 (R Development Core Team (2007) . *R: A Language and Environment for Statistical Computing*) に入手可能)、生データファイルを読み込んだ。RNA分解プロット (*affy*パッケージのAffyRNAdeg機能)、[09:42:29] 正規化したスケーリングしていない標準誤差プロット、相対ログ発現プロット (*affyPLM*パッケージ (Brettschneider, J., Collins, F., and Bolstad, B. M. (2008). *Quality Assessment for Short Oligonucleotide Microarray Data*. *Technometrics* 50, 241 - 264))、および相対ログ発現値の平均を発生しかつ試験することによって品質が管理された。更に、空間的影響が存在していないことを確認するために、擬似画像 (プローブレベルモデルの残像) の肉眼による確認を行った。品質管理確認上で閾値の集合以下に下げられたアレイは、更なる解析から除外した。

20

30

40

【0036】

母集団レベル解析 (すなわち、平均フォールド変化の研究) のために、データを引き続きGC-ロバストマイクロアレイ解析 (GC-RMA) を使用して正規化した。品質管理確認に合格したすべてのアレイからマイクロアレイ発現値 (Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). *Exploration, normalization, and summaries of high density oligonucleotide*

50

array probe level data. *Biostatistics* 4, 249 - 264) を発生するために、バックグラウンド補正および分位点正規化を使用した。個人シグネチャ予測モデルとしては、このデータは MAS5 (Affymetrix, I. (2002). *Statistical algorithms description document. Technical paper*) で正規化された。

【0037】

統計的モデリング - 母集団レベル解析。各々の比較については、全体的な線形モデルは、中程度の t 統計量に基づいて、発現アレイ上の各プローブ集合に対して生の p 値を発生するように適合した。多数の遺伝子が見積もられるときに起こる多重検定効果に対する補正のためにベンジャミニ - ホッホバーグの偽発見率 (FDR) 法を使用した。

10

【0038】

統計的モデリング - 個別サンプル予測モデリング。予測モデルのロバスト性を達成するために、血液 (GSE15289) および PBMC (GSE42057) からの独立した遺伝子発現データセットを National Center for Biotechnology Information Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/gds/?term=GEO>) から取得し、処理した。NOWAC 研究 (GSE15289) からのデータセット (Dumeaux, V., Olsen, K.S., Nuel, G., Paulsen, R.H., Borresen-Dale, A.-L., and Lund, E. (2010a). *Deciphering normal blood gene expression variation - The NOWAC postgenome study. PLoS genetics* 6, e1000873) は、211 人の喫煙したことがない者と 74 人の現在喫煙している者を含む、285 人の 48 ~ 63 歳の閉経後の女性からの全血サンプルを含んだ。Bahr らのデータセット (GSE42057) (Bahr, T.M., Hughes, G.J., Armstrong, M., Reisdorph, R., Coldren, C.D., Edwards, M.G., Schnell, C., Kedl, R., LaFlamme, D.J., and Reisdorph, N. (2013). *Peripheral Blood Mononuclear Cell Gene Expression in Chronic Obstructive Pulmonary Disease. American journal of respiratory cell and molecular biology*) は、36 人の現在喫煙している者 (そのうち 22 が COPD を患い、14 人が健常者) と 100 人の以前喫煙していた者 (そのうち 72 人が COPD を患い、28 人が健常者) とから収集した末梢血単核細胞 (PBMC) サンプルから誘導された。すべての被験者は、非ヒスパニック系の白人であった。

20

30

【0039】

GSE15289 および GSE42057 データセットの被験者からサンプリングされたデータを、喫煙者のサンプルと喫煙したことがない (または以前喫煙していた) 者のサンプルとの間で平均発現の大きい変化を示した遺伝子を各データセットで特定するために使用した。 L_1 と L_2 を M (ここでは、 $M = 1000$ だが、一般に M は任意の値とすることができる) 個の 2 つの独立したデータセット (GSE15289 および GSE42057) からのフォールド変化が最も高い遺伝子の集合とした。リスト L_1 を取得するために、データセット GSE15289 を喫煙者ステータス (現在喫煙している者および喫煙したことがない者) に従ってソーティングし、そして各群について平均遺伝子発現レベルを取得した。現在喫煙している者の群と喫煙したことがない者の群との間の平均遺伝子発現レベルの差異は、本明細書ではフォールド変化と称され、そしてフォールド変化が最も高い M 遺伝子は集合 L_1 に含まれる。リスト L_2 は、現在喫煙している者と以前喫煙していた者について同様に取得した。

40

【0040】

図 1 は、遺伝子の集合を特定し、そして遺伝子の集合に基づいて分類モデルを取得する

50

ためのプロセス100のフローチャートである。特に、プロセス100は、カウンターパラメータNを初期化して1にする工程(工程102)と、Matthewsの相関係数(MCC(N))を演算することによって線形判別解析(LDA)モデルの特性を見積もる工程(工程104)と、カウンターパラメータが最大カウンター値Mと等しいかどうかを決定する工程(決定ブロック106)とを含む。NがM未満である場合、プロセス100は工程108に進みNをインクリメントし、そして次の係数MCC(N)を演算することによってLDAモデルの特性を見積もるために工程104に戻る。NがMに達すると(決定ブロック106)、最大MCC値の結果として得られるNの値(N_{MAX})が見積もられ(工程110)、そして2つの遺伝子集合L₁[1:N]とL₂[1:N]との積集合としてコア遺伝子リストが画定される(工程112)。コア遺伝子リストが特定された後、コア遺伝子リストに基づいてLDAモデルが演算される(工程114)。

10

【0041】

工程102では、カウンターパラメータNは初期化されて1になる。カウンターパラメータNは、1から最大値Mまで変化し、そして判断ブロック106でNがMに達するまで、工程108でインクリメントされる。

【0042】

工程104では、LDAモデルの特性は、係数MCC(N)を演算することによって見積もられる。特に、集合L₁内のN個の最大フォールド変化と集合L₂内のN個の最大フォールド変化との積集合である、L₁[1:N] L₂[1:N]を5分割交差検証(1000回)を使用して、LDAモデルの特性が見積もられる場合がある。LDAモデルは、MCC(N)を演算することによって見積もられる。MCC評価基準は、すべての真陽性/偽陽性と陰性との比を組み合わせ、そしてひいては単一の値である妥当な評価基準を提供する。MCCは、複合特性スコアとして使用されてもよい特性評価基準である。MCCは、-1と+1との間の値であり、本質的に既知の2項分類と予測される2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。

20

【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、一般に、LDAモデルの特性を評価するために、特性評価基準の集合に基づいて複合特性評価基準を発生するための任意の好適な技法が使用されてもよい。+1のMCC値は、モデルが完全な予測を取得することを示し、0のMCC値は、ランダムと何ら変わらないモデル予測が遂行することを示し、そして-1のMCC値は完全に不適格なモデル予測を示す。MCCは、分類分けの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、という利点を有する。対照的に、曲線下面積(AUC)の演算については、分類子機能は数値的なスコアの提供を必要とする。しかしながら、一般に、TP、FP、TN、およびFNを説明するいずれかの評価基準が、本開示に従って使用されてもよい。

30

【0043】

MCCを演算するためには、まず分類の集合を選択しなければならない。喫煙したことがない者、以前喫煙していた者、および現在喫煙している者からBLD-SMK-01データセットを取る。図4A、図4B、および図4Cは、BLK-SMK-01サンプルで差次的に発現される遺伝子に対するボルケーノプロットを示す。各々のボルケーノプロットは、-log₁₀(調整したP値)に対して見積もったlog₂(フォールド変化)を示す。中程度のt統計に基づいてP値を演算し、そしてベンジャミニ-ホッホバーグ法によって調整した。特に、図4Aは現在喫煙している者と非喫煙者との間で遺伝子発現プロファイルと比較し、図4Bは現在喫煙している者と以前喫煙していた者との間で遺伝子発現プロファイルと比較し、そして図4Cは以前喫煙していた者と喫煙したことがない者との間で遺伝子発現プロファイルと比較する。図4Cに示すボルケーノプロットは喫煙した

40

50

ことがない者と以前喫煙していた者との間の差次的な遺伝子発現を示さない（すなわち、図4Cでは傾向が見られない）が、図4Aおよび図4Bは、現在喫煙している者と喫煙したことがない者との間（図4A）、および現在喫煙している者と以前喫煙していた者との間（図4B）に数多くの差次的な遺伝子発現の変化が見られることを示す。

【0044】

従って、BLD-SMK-01サンプルの母集団レベルのトランスクリプトーム解析は、全血では喫煙したことがない者と以前喫煙していた者との間に差次的な遺伝子発現の変化がないことを示し、従って血液トランスクリプトームに基づいて以前喫煙していた者と喫煙したことがない者とを区別するのは極めて困難であることになる。これとは逆に、現在喫煙している者と喫煙したことがない者および以前喫煙していた者との間にはそれぞれ数多くの差次的に発現される遺伝子がある（図4Aおよび図4B）。喫煙したことがない者の母集団と以前喫煙していた者の母集団との間には差が見られないので、工程104でモデルを見積もるためには2つの分類、すなわち現在喫煙している者および現在喫煙していない者のみで使用された。

10

【0045】

特に、工程104では、遺伝子の集合 $L_1 [1 : N]$ 、 $L_2 [1 : N]$ は、2つの独立したデータセットGSE15289およびGSE42057からのフォールド変化が最も大きいNの積集合に対応する。 $L_1 [1 : N]$ 、 $L_2 [1 : N]$ のいずれかに基づく各々の予測モデルは、LDAモデルの結果が独立したデータセットに対して一般化可能かどうかを評価するために交差検証される。一実施例では、 $L_1 [1 : N]$ 遺伝子集合に対して5分割交差検証の一事例を遂行するために、 $L_1 [1 : N]$ 集合は、ランダムに5つの部分集合A、B、C、D、およびEに分けられた。LDA技法を使用して分類子を訓練するために、4つの部分集合（A、B、C、およびD）が使用され、そして他の4つの部分集合上で訓練された分類子をテストするために第5の部分集合（E）が使用された。他の部分集合（A、B、C、およびD）の各々を、その他の4つの部分集合に対して訓練された分類子をテストするためのテスト部分集合として用いて、この訓練およびテストプロセスが更に4回繰り返された。

20

【0046】

一般に、LDA技法の基準は、特徴を分類yの中に記述する入力ベクトルxを分類することである。分類は、観察された特徴の線形結合である機能に基づく。線形結合の係数は、訓練部分集合に基づいて見積もられる。具体的には、LDA技法を使用して分類子を訓練するために、4つの訓練部分集合からのデータでの遺伝子発現レベルの線形結合が特定される。本明細書では、線形結合は分類子と称され、そして予測される喫煙者ステータスと予測される非喫煙者ステータスとの間に境界を画定する。分類子は、テスト部分集合内で各々の個人に対して予測されるステータスを取得するために使用される。このプロセスは、更に4回繰り返され、これによって5つの部分集合の各々が1回はテスト部分集合として処理される。5つの部分集合の各々は1回テスト部分集合となった後、5分割交差検証のうちの1つの事例が完了し、そして訓練データの観察値（ $L_1 [1 : N]$ 、 $L_2 [1 : N]$ 集合の特徴を有する）を、5つの新しい部分集合A'、B'、C'、D'、およびE'に分割し、5分割交差検証の第2の事例を開始する。

30

40

【0047】

本明細書に記述される実施例は、5分割交差検証の100個の事例の結果であるが、当業者であれば、一般に、本開示の範囲を逸脱することなく任意の数のk分割交差検証の事例を使用してもよいことを理解するであろう。更に、本明細書に記述される実施例は、遺伝子発現レベルの線形結合に基づいて分類子を形成するLDA技法の結果である。しかしながら、当業者であれば、一般に、分類子を形成するために、 $R \times N$ で次元マニフォールドを形成する場合がある遺伝子発現レベルの任意の関数（二次関数、多項式関数、対数関数、または任意の他の好適な関数などの）を使用して、分類子を画定してもよいことを理解するであろう。

【0048】

50

工程 110 では、 N が最大数 M に達した後、 MCC の M 値の集合が考慮され、そして $N_{max} = \arg \max_N (MCC(N))$ として MCC 最大値に対応する N の値が見積もられる。図 1 に示すように、 N_{max} を見積もる工程は、 MCC のすべての M 値が演算された後遂行される。しかしながら、当業者であれば、一般に、代替的に次の値 $MCC(N+1)$ を見積もる前に工程 104 で演算された $MCC(N)$ の値をいずれかの所定の閾値と比較する場合があることを理解するであろう。この場合、 MCC の値が所定の閾値を超えることが判明した場合、プロセス 100 は直ちに工程 110 に進み、残りの値 $N = N_{max} + 1$ を M に対して考慮することなく、 N_{max} の値に現在の N の値を割り当てる場合がある。

【0049】

10

工程 112 では、シグネチャに対するコア遺伝子リストは、 $L_1 [1 : N_{max}]$ $L_2 [1 : N_{max}]$ で表される積集合、すなわち $L_1 [1 : N_{max}]$ と $L_2 [1 : N_{max}]$ との両方の中にある遺伝子の集合によって画定される。この例で記述されるように、 L_1 および L_2 の 2 つのデータセットのみが使用される。しかしながら、当業者であれば、一般に、 MCC 値を演算するため、および遺伝子シグネチャを画定する遺伝子のコア集合を特定するために任意の数のデータセットを使用してもよいことを当業者は理解するであろう。特に、 m データセットの積集合、または対になった積集合の和集合が使用されてもよい。

【0050】

工程 114 では、 LDA モデルを演算するために工程 112 で決定されたコア遺伝子リストが使用される。特に、コア遺伝子リストに基づいて演算した LDA モデルを、5 分割交差検証または任意の数の n 分割交差検証を 100 回遂行することによって演算してもよい。

20

【0051】

一例では、工程 102 ~ 114 に関連して記述される統計的モデリング方法論を適用することによって、以下の 6 つの遺伝子、すなわち $LRRN3$ 、 $SASH1$ 、 $PALLD$ 、 $RGL1$ 、 $TNFRSF17$ 、および $CDKN1C$ を含むコア遺伝子シグネチャが特定される。このモデルの 5 分割交差検証 (100 回) MCC は、現在喫煙している者から取得したサンプルを喫煙したことがない者に対して分類する時に 0.77 (感受性スコア (Se)) が 0.91 であり、かつ特異性スコア (Sp) が 0.85) である。方法論の設計によって、シグネチャ内のコア遺伝子は、 $NOWAC (GSE15289)$ および $Bahr$ ら ($GSE42057$) の両方の研究での高分割変化遺伝子の中にあり、そしてこれらの 2 つの GSE 研究 ($Se = 0.73$ 、 $Sp = 0.81$) の間で共通の 77 個の遺伝子すべてに基づいて LDA モデルの特性について予測が改善した。6 つの遺伝子 $LRRN3$ 、 $SASH1$ 、 $PALLD$ 、 $RGL1$ 、 $TNFRSF17$ 、および $CDKN1C$ が、本明細書ではコア遺伝子シグネチャとして参照されているとしても、6 つの遺伝子のうちの 3 つ、4 つ、または 5 つなどの任意の組み合わせなどのように、6 つの遺伝子の任意の組み合わせがコア遺伝子シグネチャとして使用されてもよいことを、当業者であれば理解するであろう。

30

【0052】

40

一部の実施形態では、シグネチャ内の遺伝子は、高い特異性スコアおよび感受性スコアに関連付けられたコア集合に含まれていない追加的な遺伝子を含む、遺伝子の拡張集合を含むように広げられた。特に、フォールド変化の高い遺伝子の各々のリストを個々に活用することによって取得された予測的なモデルを研究する時、 IGJ 、 $RRM2$ 、 $ID3$ 、 $SERPING1$ 、および $FUCA1$ は、特異性および感受性の高いシグネチャ内の潜在的な候補として繰返し特定される。これらの 5 つの遺伝子も、 $NOWAC$ (現在喫煙している者対喫煙したことがない者) および $Bahr$ ら (現在喫煙している者対以前喫煙していた者) の両方の研究の血液トランスクリプトーム中のフォールド変化の高い遺伝子の中にあり、かつコア遺伝子シグネチャを拡張シグネチャへと拡張するために使用される。拡張シグネチャ ($LRRN3$ 、 $SASH1$ 、 $PALLD$ 、 $RGL1$ 、 $TNFRSF17$ 、 C

50

DKN1C、IGJ、RRM2、ID3、SERPING1、およびFUCA1)に基づくモデルの交差検証MCCは、現在喫煙している者対喫煙したことがない者を分類すると、0.73 (Se = 0.88、Sp = 0.84)であった。11個の遺伝子LRRN3、SASH1、PALLD、RGL1、TNFRSF17、CDKN1C、IGJ、RRM2、ID3、SERPING1、およびFUCA1のすべてが本明細書で拡張遺伝子シグネチャとして引用されているが、11個の遺伝子のうちの5つ、6つ、7つ、8つ、9つ、または10個などの任意の組み合わせなどのように、11個の遺伝子の任意の組み合わせがコア遺伝子シグネチャとして使用されてもよいことを、当業者であれば理解するであろう。更に、この組み合わせは、コア遺伝子シグネチャ内の6つの遺伝子のうちの3つ、4つ、または5つの組み合わせ、および拡張遺伝子シグネチャ内の追加的な遺伝子内の5つの遺伝子のうちの2つ、3つ、または4つを含んでもよい。

10

【0053】

工程114で演算されたLDAモデルの結果を、BLD-SMK-01単独(すなわち、2つの公表されているデータセットGSE15289およびGSE42057を使用しない)からわずかなシグネチャを学習する時に取得したモデルの予測交差検証結果と比較した。喫煙者対非喫煙者を予測する上でこのモデルの5分割交差検証特性の結果としてSp = 0.96およびSe = 0.93が得られ、これはコアシグネチャおよび拡張シグネチャに基づくモデルの特性よりわずかに高い。本明細書に記述される方法論を用いて誘導される予測モデルの交差検証の特異性および感受性(Sp = 0.88、Se = 0.84)が、結果として独立したデータセット(Sp = 0.96、Se = 0.93)を使用せずに取得したモデルよりわずかに低い特性をもたらすものの、本明細書で誘導される予測モデルは、より広い範囲の用途に関連付けられるために有利である。特に、本開示の方法に従って誘導される予測モデルは、工程116と関連して詳細に記述されるようにモデルが認証された時に、ロバストである。

20

【0054】

工程116では、工程114で演算されたLDAモデルが認証される。LDAモデルの認証は、BLD-SMK-01研究からの以前喫煙していた者の群、およびQASMC研究からの血液データセットを使用することによって遂行される。QASMCトランスクリプトームのサンプルを品質確認した後、52人のCOPD罹患患者、58人の現在喫煙している者、58人の以前喫煙していた者、および59人の喫煙したことがない者のCELファイルが予測のために利用可能だった。コアシグネチャおよび拡張シグネチャの予測特性を見積もるために、QASMCサンプルを2つの群、すなわち、現在喫煙している者(COPD罹患患者および健常者)と以前喫煙していた者および喫煙したことがない者を含む現在喫煙していない者とへ層化した。これらの群により、COPDステータスに関してシグネチャのロバスト性を見積もることができた。各々の中心データセットは、コア遺伝子シグネチャまたは拡張シグネチャを基に構築されたモデルを使用して予測された。

30

【0055】

表3は、独立したデータセット上で様々なシグネチャに対してLDAモデルを使用して予測結果を示す。表3の書式は表1の書式に従い、予測される分類を異なる横列に示し、かつ実際の分類を異なる縦列に示す。特に、表3に示す予測結果は、コア遺伝子シグネチャ(最初の3つの横列)、拡張遺伝子シグネチャ(中間の3つの横列)、BLD-SMK-01サンプル単独から誘導されたシグネチャ(最後から2番目の横列)、およびBeinekeら(Beineke, P., Fitch, K., Tao, H., Elashoff, M.R., Rosenberg, S., Kraus, W.E., およびWingrove, J.A. (2012) A whole blood gene expression-based signature for smoking status. BMC medical genomics 5, 58.)に記述される遺伝子の集合に基づくシグネチャ(一番下の横列)に対するものを含む。表3に示すように、コアシグネチャと拡張シグネチャとの両方は、BLD-SMK-01サンプル単独およびBeinekeによって特定されたシグネチャから誘導されるシグネチャより高い感受性スコアおよび特

40

50

異性スコアをもたらす。

【表 3】

		BLD-SMK-01	QASMC	
		以前喫煙していた者	現在喫煙している者	現在喫煙していない者
コア	現在喫煙している者	3	99	12
	現在喫煙していない者	23	11	105
	真のレート	$S_p=0.88$	$S_e=0.90$	$S_p=0.90$
拡張	現在喫煙している者	4	100	12
	現在喫煙していない者	22	10	105
	真のレート	$S_p=0.85$	$S_e=0.91$	$S_p=0.90$
その他	BLD-SMK-01	$S_p=0.73$	$S_e=0.81$	$S_p=0.77$
	Beineke	$S_p=0.73$	$S_e=0.87$	$S_p=0.79$

QASMC 研究に対するシグネチャの分類特性は、COPD ステータスにかかわらずモデルが口バストであることを確認した (コアシグネチャに対して $S_e = 0.9$ 、 $S_p = 0.9$ 、拡張シグネチャに対して $S_e = 0.91$ 、 $S_p = 0.90$)。

【0056】

更に、図 5 A、図 5 B、図 5 D、および図 5 E は、異なる研究に対する分類スキームを示す様々なボックスプロットを示す。特に、図 5 A および図 5 B は、BLD-SMK-01 研究および QASMC 研究それぞれに対して、LDA モデルから現在喫煙している者に分類されるサンプルの事後確率のボックスプロットをプロットする。図 5 D および図 5 E は、BLD-SMK-01 研究および QASMC 研究それぞれに対して、線形判別関数からの予測スコアのボックスプロットをプロットする。特に、負のスコアを有するサンプルは、現在喫煙している者として分類され、そして正のスコアを有するサンプルは、現在喫煙していない者として分類される。

【0057】

性別および年齢などの追加的な共変動の影響も検討された。BLD-SMK-01 および QASMC 研究は、性別および年齢に関してはバランスの取れたものであった。年齢または性別と喫煙ステータスとの間の統計的関連性は、統計的なカイ二乗検定 (BLD-SMK-01 に対して χ^2 (性別、喫煙ステータス) P 値 = 1、および QASMC に対して χ^2 (性別、喫煙ステータス) P 値 = 0.9)、ならびに統計的 t 検定 (BLD-SMK-01 に対して t 検定 (年齢対喫煙ステータス) P 値 = 0.8、および QASMC に対して t 検定 (年齢対喫煙ステータス) P 値 = 0.46) によって示されるように、存在しなかった。

【0058】

更に、シグネチャ内の各遺伝子は、BLD-SMK-01 での性別および年齢との関連性、および ANOVA P 値について、性別の影響をあまり示さない PALLD 遺伝子を除くいずれの遺伝子についても 0.05 を下回ることがないことが試験された。以前に特定された遺伝子シグネチャには性別および/または年齢の影響が見られ、そしてかかる因子に対して調整が必要であることが決定された。Beineke ら、2012。特に、喫煙者は、平均して喫煙したことがない者または以前喫煙していた者より年齢が高く、BLD-SMK-01 研究では喫煙ステータスと統計的関連性がなかったため、この共変動は予測子に含まれなかったため、年齢は公開されている 2 つのデータセット (GSE15289 および GSE42057) では重要な共変動であった。しかしながら、特異性および感受性のスコアによって画定されたより良好な特性に加えて、本明細書に記述される遺伝子

シグネチャは、一般的に性別または年齢と相関しない。このことは、これらの因子に対する調整が不要で、これによって演算プロセスを単純化するという点で、本明細書に記述されるコアシグネチャおよび拡張シグネチャが既知の遺伝子シグネチャに優る利点を提供することを示唆する。

【 0 0 5 9 】

発見したシグネチャを q R T - P C R をベースにした曝露バイオマーカーへと変換できるかどうかを決定するために、20個の無作為に選択したサンプル（10人の現在喫煙している者および10人の喫煙したことがない者）の部分集合を、拡張シグネチャ内の遺伝子の発現レベルを測定するために q R T - P C R の対象とした。L D A モデルは、正規化した q R T - P C R データ上で拡張シグネチャ内の遺伝子に基づいて訓練され、10分割交差検証（1000回、サンプルサイズが小さいため10分割が選択された）によって評価され、0.85の特異性および0.96の感受性がもたらされた（表4）。同一のものをコアシグネチャに適用した場合は、0.62の特異性および0.80のより低い感受性が得られた（表4）。

【表4】

		現在喫煙している者	現在喫煙していない者
コア	現在喫煙している者	7.9	3.39
	現在喫煙していない者	2.1	5.61
	真のレート	Se=0.80	Sp=0.62
拡張	現在喫煙している者	9.61	1.36
	現在喫煙していない者	0.39	7.64
	真のレート	Se=0.96	Sp=0.85

【 0 0 6 0 】

本開示の1つの目標は、コア遺伝子シグネチャおよび拡張遺伝子シグネチャを適用して、このシグネチャを使用して加熱式たばこ製品（HTP）への切換えの影響を検出することができるかどうかを決定することである。この目標を容易にするために、R E X - E X - 0 1 研究からデータを取得した。R E X - E X - 0 1 研究は、両方の性別の23～65歳の42人の健常喫煙者を採用した非盲検、無作為化、対照、並行2群間研究であった。この研究は従来の紙巻たばこの喫煙者を最近連続して5日間以上HTP（本明細書ではたばこ加熱システム2.1（THS2.1）と称する）に切り換えた喫煙者と比較するために実施した。この研究は医薬品の臨床試験の実施基準（Good Clinical Practices、GCP）に従って実施され、また識別番号NCT01780714でClinicalTrials.govに登録した。血液サンプルをPAXgene採血管の中に保存し、そしてAROS Applied Biotechnology AS（デンマーク、Aarhus）に送付し、そこで更に処理され、かつAffymetrix Human Genome U133 Plus 2.0 GeneChipで交配された。

【 0 0 6 1 】

この本明細書で特定された遺伝子シグネチャが、臨床試験で曝露-応答を評価するための感応性がありかつ非侵襲的なツールを提供するかどうかを試験する目的で、5日後の全血トランスクリプトームでHTPへの切換えを検出することができるかどうかを決定するために、このシグネチャをTHS2.1データに適用した。この研究の仮説は、THS2.1に切り替えた喫煙者の全血トランスクリプトームが、現在喫煙している者よりも以前喫煙していた者の全血トランスクリプトームに類似していることである。5日間切り換えている特定のHTP使用者の遺伝子発現プロファイルを特徴付ける（例えば、R E X - E X - 0 1 研究データからシグネチャを抽出することによって）よりも、より長期間の切り替えパターンの指示計として機能することができる、トランスクリプトームに基づく曝露

応答シグネチャを特定するのが望ましい。これは、現在喫煙している者のサンプルと現在喫煙していない者のサンプルの両方を区別することができる、コア遺伝子シグネチャおよび拡張遺伝子シグネチャを確立することによって達成された。

【0062】

REX-EX-01研究のCELファイルの品質確認を遂行した後、5日目において従来の紙巻たばこ喫煙者およびTHS2.1の使用者に対してそれぞれ16個および18個のファイルが残った。以下の表5は、コア遺伝子シグネチャ(上の3つの横列)と拡張遺伝子シグネチャ(下の3つの横列)に対するREX-EX-01サンプルの予測結果を示す。拡張遺伝子シグネチャについては、従来の紙巻たばこを使用し続けている個人(現在喫煙している者)は主として現在喫煙している者に分類され(69%)、一方でTHS2.1に切り替えた被験者はほとんどが現在喫煙していない者として分類された(89%)。コアシグネチャについては、現在喫煙している者に対する真のレートは同一(69%)で、そしてTHS2.1に切り替えた被験者の78%は現在喫煙していない者として分類された。従って、コア遺伝子シグネチャおよび拡張遺伝子シグネチャの両方が、HTP使用者から取得したサンプルが現在喫煙していない者のサンプルであることを予測する。

【表5】

		現在喫煙している者	THS2.1へ切り換えた者
コア	現在喫煙している者	11	4
	現在喫煙していない者	5	14
	真のレート	Se=0.69	Sp=0.78
拡張	現在喫煙している者	11	2
	現在喫煙していない者	2	16
	真のレート	Se=0.69	Sp=0.89

表5に示す結果は、ニコチンおよびコチニン曝露ではTHS2.1と従来の紙巻たばこの間に顕著な違いがない(データ不表示)という事実にもかかわらず、HTPに切り換えた被験者の血液トランスクリプトームが現在喫煙している者ではなく以前喫煙していた者に類似してくるという当初の仮説と一致する。

【0063】

更に、図5Cは、REX-EX-01データ上のLDAモデルから現在喫煙している者として分類されるサンプルの事後確率のボックスプロットをプロットし、そして図5Fは、REX-EX-01データ上の線形判別関数からの予測スコアのボックスプロットをプロットする。負の予測スコアを有するサンプルは現在喫煙している者として分類され、一方で正の予測スコアは現在喫煙していない者のステータスを示す。

【0064】

単一の遺伝子の測定による遺伝子シグネチャと比較すると、通常のそして病理学的な状況で、遺伝子発現プロファイリングは生物学的プロセスのより完全な全体像を提供する。複数の遺伝子の発現の傾向を総合すると、疾病状態に対する曝露応答から所与の生理学的な状態についてのシグネチャまたは分類子を誘導することも可能である。主として罹患している組織は、正常な状態、曝露された状態、または病的な状態をより正確に示すサンプルを提供するが、多くの場合は組織生検を使用して被験者を分類するのは現実的でない。最低限の侵襲的な技法を使用して血液をサンプリングするのは容易なため、血液を基にしたシグネチャはバイオマーカー発見に対してかなり有望である。この研究では、2組の全血に基づくバイオマーカーが特定された。これらのどちらも、身体の喫煙への応答に対するシグネチャとして機能することができ、従って個人の喫煙ステータスに対する強い予測子として使用することができる。

【0065】

この研究で強く強調されている遺伝子は、LRRN3である。LRRN3の発現は、現在喫煙していない者と比較すると、現在喫煙している者では増加した。REX-EX-01研究では、発現はHTPに切り替えた被験者の血液では0日目と5日目との間で著しく減少し、そして従来の紙巻たばこのままだった被験者の血液では一定のままだった。従って、コアシグネチャと拡張シグネチャとの両方で、従来の紙巻たばこからHTPへの切り換えの効果を測定するためにLRRN3は重要な遺伝子であるように思われる。一実施例では、記述されるように遺伝子シグネチャは、LRRN3のみを含みいかなる他の遺伝子も含まないか、または任意の他の遺伝子と組み合わせてLRRN3を含む。特に、LRRN3を含む遺伝子シグネチャは、切り替え後0日目と5日目との間のLRRN3発現の減少を実証することによって、従来の紙巻たばこの喫煙からHTPの使用への切り換えを検出することができる。

10

【0066】

本明細書に記述されるシステム薬理学アプローチは、現在喫煙している者を現在喫煙していない者と区別することができる1つ以上の口バストな全血に基づく喫煙者遺伝子シグネチャの構築を可能にする。本明細書に記述されるコア遺伝子シグネチャは、6つの遺伝子に基づき、そして拡張遺伝子シグネチャはコア遺伝子シグネチャに加えて追加的な5つの遺伝子に基づく。両方の遺伝子シグネチャは、感受性スコアおよび特異性スコアの両方によって評価されるように、個人の喫煙者ステータスを予測する点で並外れた精度を有する。REX-EX-01研究からのサンプルに適用すると、全血トランスクリプトームデータに基づいて、シグネチャはTHS 2.1を5日間使用した後の被験者を現在喫煙していない者として特定した。従って、最小限の侵襲性サンプリングを使用して曝露応答を評価するために、本明細書に記述されるシグネチャは、感応性がありかつ特異的なツールを提供する。

20

【0067】

図2は、本開示の例示的な実施形態に従って被験者から取得したサンプルを評価するためのプロセス200のフローチャートである。プロセス200は、サンプルと関連付けられたデータセットであって、LRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17に対する定量的な発現データを含むデータセットを受け取る工程(工程202)と、受け取ったデータセットに基づいてスコアを発生する工程であって、このスコアが被験者の予測される喫煙ステータスを示す工程(工程204)とを含む。一部の実施形態では、工程202において受け取ったデータセットは、IGJ、RRM2、SERPING1、FUCA1、およびID3に対する定量的な発現データを更に含む。一部の実施形態では、工程202において受け取ったデータセットは、CLDND1、MUC1、GOPC、およびLEF1のうちの一つ以上に対する定量的な発現データを更に含む。

30

【0068】

工程204で発生したスコアは、この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内の定量的な発現データに基づいて決定される。特に、本明細書に記述される実施例では、個人に対して予測される分類を決定するために、LDAモデル上で訓練された分類子が202で受け取られたデータセットに適用されてもよい。

40

【0069】

本明細書に記述される遺伝子シグネチャは、被験者から取得したサンプルを評価するために、コンピュータ実装された方法で使用されてもよい。特に、サンプルに関連付けられたデータセットが取得されてもよく、そしてデータセットは、コア遺伝子シグネチャのために、LRRN3、CDKN1C、PALLD、SASH1、RGL1、およびTNFRSF17に対する定量的な発現データを含んでもよい。受け取ったデータセットに基づいてスコアを発生してもよく、このスコアは被験者の予測される喫煙ステータスを示す。特に、スコアは、本明細書に記述されるLDAモデルアプローチ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、拡張遺伝子シグネチャに含まれる追加的なマーカ

50

ー I G J、R R M 2、S E R P I N G 1、F U C A 1、および I D 3 に対する定量的な発現データを更に含んでもよい。データセットは、C L D N D 1、M U C 1、G O P C、および L E F 1 のうちの 1 つ以上に対する定量的な発現データを更に含んでもよい。

【 0 0 7 0 】

一部の実施形態では、データセットは、マーカーの集合 L R R N 3、C D K N 1 C、P A L L D、S A S H 1、R G L 1、T N F R S F 1 7、I G J、R R M 2、S E R P I N G 1、F U C A 1、I D 3、C L D N D 1、M U C 1、G O P C、および L E F 1 の任意の数の任意の部分集合を含む。シグネチャに含まれるマーカーには、L R R N 3、C D K N 1 C、P A L L D、S A S H 1、R G L 1、および T N F R S F 1 7 のうち少なくとも 3 つ（または任意の他の好適な数）、I G J、R R M 2、S E R P I N G 1、F U C A 1、および I D 3 のうちの少なくとも 2 つ（または任意の他の好適な数）、および C L D N D 1、M U C 1、G O P C、および L E F 1 のうちの少なくとも 1 つ（または任意の他の好適な数）などの 1 つ以上の基準が適用されてもよい。一般に、本開示の範囲を逸脱することなくこれらのマーカーの組み合わせを使用する任意のシグネチャを使用してもよい。

【 0 0 7 1 】

一部の実施形態では、本明細書に記述されるシグネチャ中の遺伝子は、個人の喫煙者のステータスを予測するキットを組み立てる上で使用される。特に、キットは、試験サンプル内の遺伝子シグネチャ内の遺伝子の発現レベルを検出する一組の試薬と、個人の喫煙者ステータスを予測するためのキットを使用する上での説明書とを備える。このキットは、禁煙または喫煙製品の代替品（H T P など）の個人に対する効果を評価するために使用されてもよい。

【 0 0 7 2 】

図 3 は、図 1 および図 2 に関連して記述されるプロセス、または本明細書に記述される、コア遺伝子シグネチャ、拡張遺伝子シグネチャ、または任意の他の遺伝子シグネチャを保存するプロセスなどの本明細書に記述されるいずれかのプロセスを遂行するための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保存された遺伝子シグネチャは、L R R N 3、C D K N 1 C、P A L L D、S A S H 1、R G L 1、および T N F R S F 1 7 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、L R R N 3、C D K N 1 C、P A L L D、S A S H 1、R G L 1、T N F R S F 1 7、I G J、R R M 2、S E R P I N G 1、F U C A 1、および I D 3 から成る群から選択される少なくとも 5 つ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。

【 0 0 7 3 】

特定の実装では、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 3 0 0 にわたって実装される場合がある。コンピューティング装置 3 0 0 は、少なくとも 1 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 3 1 0 と、システムメモリと、1 つ以上のデータ保存装置と、を備える。システムメモリは、少なくとも 1 つのランダムアクセスメモリ（R A M 3 0 2）と、少なくとも 1 つの読み取り専用メモリ（R O M 3 0 4）とを含む。これらのすべての要素は、コンピューティング装置 3 0 0 の動作を容易にするために中央処理装置（C P U 3 0 6）と通信する。コンピューティング装置 3 0 0 は、数多くの異なるやり方で構成される場合がある。例えば、コンピューティング装置 3 0 0 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 3 0 0 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いてもよい。コンピューティング装置 3 0 0 は、モデリング動作、スコアリング動作、および集計動作のうちのいくつかまたはすべてを遂行するように構成されていてもよい。図 3 では、コンピューティング装置 3 0 0 は、ネットワークまたはローカルネットワークを介して、他のサーバーまたはシステムにリンクされる。

【 0 0 7 4 】

コンピューティング装置 3 0 0 は、分散したアーキテクチャで構成される場合があり、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

かかるユニットは、主要な処理機能を遂行し、かつ最低でも一般的な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含む。かかる態様では、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308を介して、他のサーバー、クライアントまたはユーザーのコンピュータ、および他の関連した装置との主要通信リンクとして作用する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作用する。様々な通信プロトコルが、システムの一部であってもよい。通信プロトコルとしては、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が挙げられるが、これに限定されない。

【0075】

CPU306は、1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CPU306からの作業負荷を除去するための数値計算コプロセッサなどの1つ以上の補助コプロセッサ、を備える。CPU306は、通信インターフェースユニット308および入力/出力コントローラー310と通信し、CPU306は、これを通して他のサーバー、ユーザー端子、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308および入力/出力コントローラー310は、例えば他のプロセッサ、サーバー、またはクライアント端子などとの同時通信のために複数の通信チャネルを含む場合がある。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、かかる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、実際にはほとんどの時間でデータの交換をやめてもよく、および装置間の通信リンクを確立するために、いくつかの

【0076】

CPU306は、データ保存装置とも通信する。データ保存装置は、磁気、光学、または半導体メモリの適切な組み合わせを含んでもよく、例えばRAM302、ROM304、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。CPU306およびデータ保存装置はそれぞれ、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、または他の類似の無線もしくは有線媒体、あるいはこれら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いてもよい。例えば、CPU306は、通信インターフェースユニット308を介してデータ保存装置に接続されていてもよい。CPU306は、1つ以上の特定の処理機能を遂行するように構成されていてもよい。

【0077】

データ保存装置は、例えば(i)コンピューティング装置300のためのオペレーティングシステム312、(ii)本明細書に記述されるシステムおよび方法に従って、および特にCPU306について詳細に記述されるプロセスに従って、CPU306を導くように適合された、1つ以上のアプリケーション314（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)プログラムによって要求される情報を保存するように利用される場合がある、情報を保存するように適合されたデータベース（複数可）316、を保存してもよい。いくつかの態様では、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保存するデータベースを含む。

【0078】

オペレーティングシステム312およびアプリケーション314は、例えば圧縮された、未コンパイルの、そして暗号化されたフォーマットで保存される場合があり、かつコンピュータプログラムコードを含む場合がある。プログラムの命令は、データ保存装置ではなくコンピュータ可読媒体から（例えばROM304から、またはRAM302からなど）プロセッサの主メモリ内へと読み込まれる場合がある。プログラム内の命令のシーケンスの実行は、本明細書に記述されるプロセスのステップをCPU306に遂行させるが、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

10

20

30

40

50

令と組み合わせて有線回路が使用されてもよい。従って、記述され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。

【0079】

好適なコンピュータプログラムコードは、本明細書に記述されるように、1つ以上の機能を遂行するために提供される場合がある。プログラムは、オペレーティングシステム312、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310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）とインターフェースすることができるようにする「装置ドライバ」などのプログラム要素も含む場合がある。

【0080】

「コンピュータ可読媒体」という用語は本明細書で使用される場合、実行のために、コンピューティング装置300のプロセッサ（または本明細書に記述される装置の任意の他のプロセッサ）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。かかる媒体は、不揮発性媒体、および揮発性媒体を含むがこれに限定されない、数多くの形態をとる場合がある。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、典型的には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一般的な形態としては、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、任意の他の磁気媒体、CD-ROM、DVD、任意の他の光学媒体、パンチカード、紙テープ、任意の他の孔パターン付きの物理的媒体、RAM、PROM、EPROM、もしくはEEPROM（電氣的消去可プログラマブル読み取り専用メモリ）、FLASH-EEPROM、任意の他のメモリチップまたはカートリッジ、あるいはコンピュータが読み取ることができる任意の他の非一時的媒体が挙げられる。

【0081】

実行のために、1つ以上の命令の1つ以上のシーケンスを、CPU306（または、本明細書に記述される、装置の任意の他のプロセッサ）に搬送するために、様々な形態のコンピュータ可読媒体が関与する場合がある。例えば、命令は当初、リモートのコンピュータ（図示せず）の磁気ディスク上に置かれる場合がある。リモートのコンピュータは、命令をそのダイナミックメモリ内にロードし、Ethernet（登録商標）接続、ケーブルライン、またはモデムを使用する電話線を通してさえも、その命令を送る場合がある。コンピューティング装置300に対してローカルの通信装置（例えば、サーバー）は、それぞれの通信ライン上のデータを受け取り、かつプロセッサのためにデータをシステムバス上に位置付ける場合がある。システムバスはデータを主メモリに搬送し、プロセッサは、そこから命令を取得し、かつ実行する。主メモリによって受け取られた命令は、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、メモリ内に保存される場合がある。加えて、命令は通信ポートを介して、電氣的信号、電氣磁氣的信号、または光学的信号として受け取られる場合があり、これはワイヤレス通信または様々なタイプの情報を搬送するデータストリームの例示的な形態である。

【0082】

本明細書で参照した各々の参照は、そのそれぞれの全体が参照により本明細書に組み込まれる。

【0083】

本開示の実装が、特定の実施例を参照して特に示され、かつ記述されてきたが、その中で形態および詳細の様々な変更が、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ように、なされる場合があることが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、従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が含まれることが意図される。

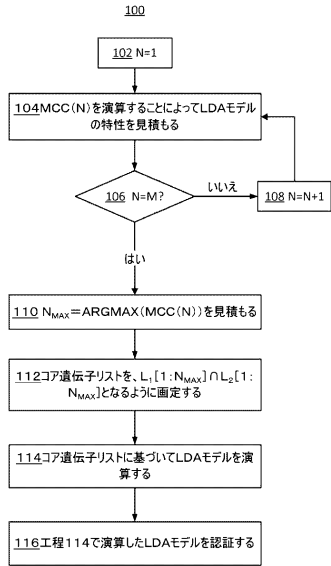
10

20

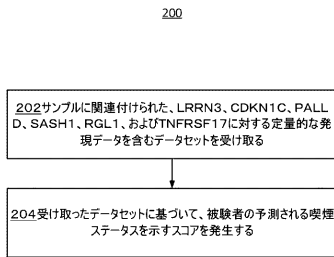
30

40

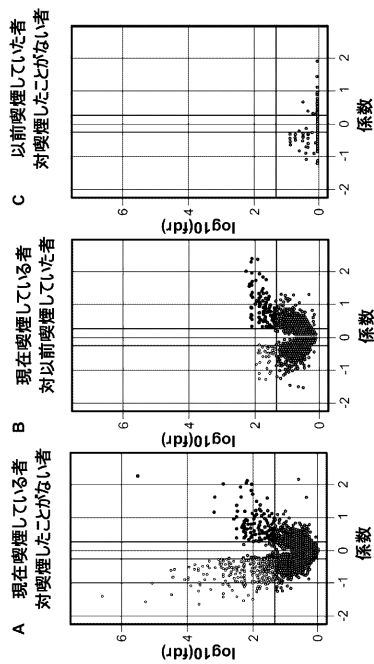
【 図 1 】



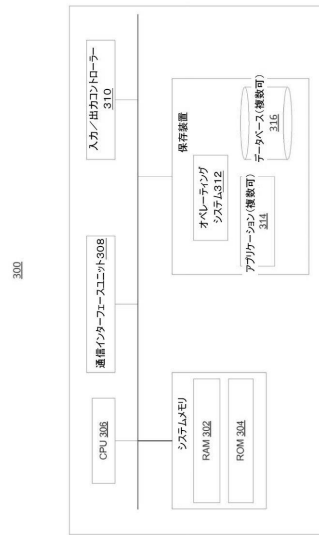
【 図 2 】



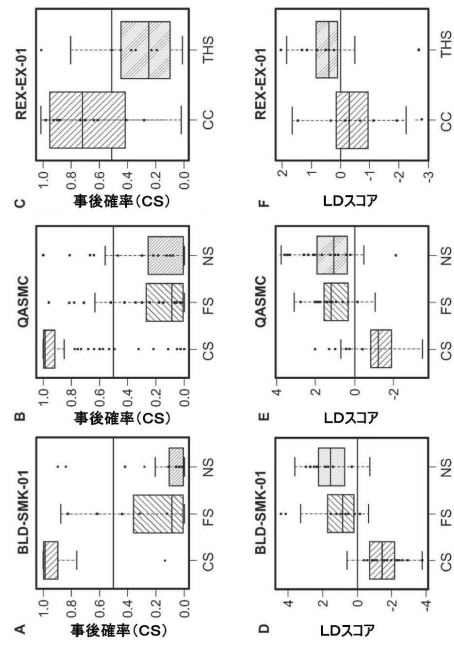
【 図 4 】



【 図 3 】



【 図 5 】



フロントページの続き

(72)発明者 マルティン, フロリアン

スイス国 ツェーハー - 2034 プスー, シュマン ドゥ ロレー 1

(72)発明者 タリッカ, マリヤ

スイス国 ツェーハー - 1350 オルブ, リュ ドゥ ラ トゥルネル 6ベ

審査官 上田 威

(56)参考文献 国際公開第2013/032917(WO, A2)

特表2003-500715(JP, A)

国際公開第2013/067040(WO, A1)

米国特許出願公開第2010/0119474(US, A1)

(58)調査した分野(Int.Cl., DB名)

G16H 10/00 - 80/00

C12M 1/00

C12N 15/00

C12Q 1/68