

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号
特表2025-509329
(P2025-509329A)

(43)公表日 令和7年4月11日(2025.4.11)

(51)国際特許分類		F I		テーマコード (参考)	
C 1 2 Q	1/6886(2018.01)	C 1 2 Q	1/6886	Z	4 B 0 2 9
C 1 2 Q	1/6869(2018.01)	C 1 2 Q	1/6869	Z	4 B 0 6 3
C 1 2 Q	1/689(2018.01)	C 1 2 Q	1/689	Z	
C 1 2 Q	1/686(2018.01)	C 1 2 Q	1/686	Z	
C 1 2 Q	1/6895(2018.01)	C 1 2 Q	1/6895	Z	
		審査請求	未請求	予備審査請求	未請求 (全64頁) 最終頁に続く

(21)出願番号	特願2024-553534(P2024-553534)	(71)出願人	523174457
(86)(22)出願日	令和5年3月9日(2023.3.9)		マイクロノーマ, インク.
(85)翻訳文提出日	令和6年10月24日(2024.10.24)		アメリカ合衆国 9 2 1 2 1 カリフォル
(86)国際出願番号	PCT/US2023/064065		ニア州 サンディエゴ フェリス・スクエ
(87)国際公開番号	WO2023/173034		ア 6 3 4 2
(87)国際公開日	令和5年9月14日(2023.9.14)	(74)代理人	110003797
(31)優先権主張番号	63/318,479		弁理士法人清原国際特許事務所
(32)優先日	令和4年3月10日(2022.3.10)	(72)発明者	フラッチオ, セレーナ
(33)優先権主張国・地域又は機関			アメリカ合衆国 9 2 1 2 1 カリフォル
	米国(US)		ニア州 サンディエゴ フェリス・スクエ
(81)指定国・地域	AP(BW,CV,GH,GM,KE,LR,LS,MW,MZ		ア 6 3 4 2
	,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),	(72)発明者	ワンドロ, ステファン
	EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(		アメリカ合衆国 9 2 1 2 1 カリフォル
	AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,		ニア州 サンディエゴ フェリス・スクエ
	FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV		ア 6 3 4 2
	最終頁に続く		最終頁に続く

(54)【発明の名称】 標的化された微生物アンプリコン配列決定からの疾患分類子

(57)【要約】
本明細書の他箇所に記載の、1つ以上の疾患を診断するマルチモーダルな方法及びシステムが本明細書で提供される。

【選択図】 図 1

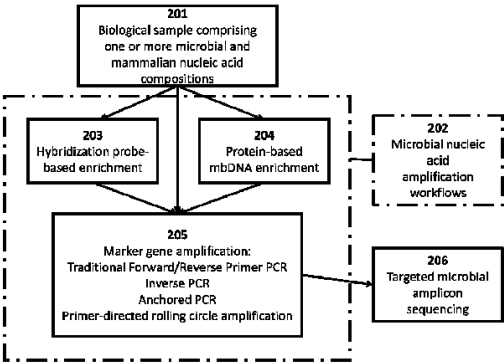


FIG. 1

【特許請求の範囲】**【請求項 1】**

1つ以上の対象のがん及び非がんの健康状態を区別するためのゲノム特徴セットを生成する方法であって、前記方法が、

(a) 1つ以上の対象の1つ以上の核酸分子及び対応する健康状態を提供することと、

(b) 前記1つ以上の核酸分子の1つ以上の非哺乳類核酸分子の1つ以上のゲノム特徴を増幅し、それによって、増幅された1つ以上のゲノム特徴を生成することと、

(c) 前記増幅された1つ以上のゲノム特徴を配列決定して、1つ以上の非哺乳類配列決定リードを生成することと、

(d) 前記1つ以上の非哺乳類配列決定リードの前記1つ以上のゲノム特徴の存在量と、前記1つ以上の対象の前記健康状態と、を組み合わせることによって、がん及び非がんの健康状態を区別するように構成された特徴セットを生成することと、を含む、方法。

10

【請求項 2】

前記ゲノム特徴が、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 1 に記載の方法。

【請求項 3】

前記微生物系統発生マーカー遺伝子が、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 2 に記載の方法。

【請求項 4】

前記微生物系統発生マーカー遺伝子が、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 2 に記載の方法。

20

【請求項 5】

前記細菌マーカー遺伝子が、リボソームRNA遺伝子5S、リボソームRNA遺伝子16S、リボソームRNA遺伝子23S、細菌ハウスキーピング遺伝子dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf、又はそれらの任意の組み合わせを含む、請求項 3 に記載の方法。

【請求項 6】

前記真菌マーカー遺伝子が、リボソームRNA遺伝子18S、リボソームRNA遺伝子5.8S、リボソームRNA遺伝子28S、内部転写されたスペーサー領域1及び2、又はそれらの任意の組み合わせを含む、請求項 4 に記載の方法。

30

【請求項 7】

微生物系統発生マーカー遺伝子が、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含む、請求項 2 に記載の方法。

【請求項 8】

増幅が、ポリメラーゼ連鎖反応又はその派生物を実施することを含む、請求項 1 に記載の方法。

【請求項 9】

前記その派生物が、逆PCR、アンカーPCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む、請求項 8 に記載の方法。

40

【請求項 10】

前記ポリメラーゼ連鎖反応が、1つ以上のゲノム特徴の増幅を防止するように構成された、ブロックプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む、請求項 8 に記載の方法。

【請求項 11】

前記1つ以上のゲノム特徴が、ミトコンドリアDNAゲノム特徴を含む、請求項 10 に記載の方法。

【請求項 12】

50

前記 1 つ以上の核酸分子を富化することを含む、請求項 1 に記載の方法。

【請求項 1 3】

前記 1 つ以上の核酸分子が、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸分子を含む、請求項 1 2 に記載の方法。

【請求項 1 4】

前記 1 つ以上の哺乳類核酸分子が、DNA、RNA、無細胞 DNA、無細胞 RNA、エクソソーム DNA、エクソソーム RNA、又はそれらの任意の組み合わせを含む、請求項 1 3 に記載の方法。

【請求項 1 5】

前記 1 つ以上の非哺乳類核酸分子が、微生物 DNA、微生物 RNA、微生物無細胞 DNA、微生物無細胞 RNA、又はそれらの任意の組み合わせを含む、請求項 1 3 に記載の方法。

【請求項 1 6】

前記富化が、

(a) 前記 1 つ以上の哺乳類核酸分子及び前記非哺乳類核酸分子をハイブリダイゼーションプローブと組み合わせることであって、前記ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせることと、

(b) 前記ハイブリダイゼーションプローブ、前記 1 つ以上の哺乳類核酸分子、及び前記 1 つ以上の非哺乳類核酸を、前記 1 つ以上の非哺乳類核酸分子の標的核酸特徴と前記ハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュベーションすることと、

(c) 非結合のハイブリダイゼーションプローブ及び前記 1 つ以上の非哺乳類核酸分子に結合したハイブリダイゼーションされたプローブを分離することと、

(d) 前記 1 つ以上の非哺乳類核酸分子に結合した前記ハイブリダイゼーションされたプローブを洗浄し、それによって、1 つ以上の富化された非哺乳類核酸分子を生成することと、を含む、請求項 1 3 に記載の方法。

【請求項 1 7】

洗浄が、非特異的に会合された核酸及び他の反応構成要素を除去するように構成される、請求項 1 6 に記載の方法。

【請求項 1 8】

前記 1 つ以上の核酸分子の富化が、非哺乳類 DNA 富化を含む、請求項 1 3 に記載の方法。

【請求項 1 9】

前記非哺乳類 DNA 富化が、

(a) 前記 1 つ以上の哺乳類核酸分子及び前記 1 つ以上の非哺乳類核酸分子を 1 つ以上の組換え C X X C ドメインタンパク質と組み合わせて、タンパク質-DNA 結合反応を形成することと、

(b) 前記組換え C X X C ドメインタンパク質と、前記 1 つ以上の哺乳類核酸分子又は前記 1 つ以上の非哺乳類核酸分子の非メチル化 CpG モチーフとの間の相互作用を促進する条件下で、前記タンパク質-DNA 結合反応をインキュベーションすることと、

(c) 非結合の組換え C X X C ドメインタンパク質及び前記非メチル化 CpG 核酸断片に結合した組換え C X X C ドメインタンパク質を前記タンパク質-DNA 結合反応の残りから分離することと、

(d) 前記非メチル化 CpG 核酸断片に結合した前記組換え C X X C ドメインタンパク質を洗浄し、それによって、増幅のための 1 つ以上の富化された核酸を生成することと、を含む、請求項 1 8 に記載の方法。

【請求項 2 0】

洗浄が、非特異的に会合された核酸分子及び前記タンパク質-DNA 結合反応構成要素の残りを除去するように構成される、請求項 1 9 に記載の方法。

【請求項 2 1】

前記 1 つ以上の核酸分子が、前記 1 つ以上の対象の 1 つ以上の生体試料に由来する、請求項 1 に記載の方法。

【請求項 2 2】

前記 1 つ以上の生体試料が、組織、液体、又はそれらの任意の組み合わせの生検試料を含む、請求項 2 1 に記載の方法。

【請求項 2 3】

前記 1 つ以上の対象が、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む、請求項 1 に記載の方法。

【請求項 2 4】

前記 1 つ以上の非哺乳類配列決定リードをフィルタリングすることを含む、請求項 1 に記載の方法。

【請求項 2 5】

フィルタリングが、前記 1 つ以上の非哺乳類配列決定リードをフィルタリングして、1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードを作製することを含む、請求項 2 4 に記載の方法。

【請求項 2 6】

フィルタリングが、1 つ以上の微生物参照データベースに対して前記 1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードをマッピングして、前記 1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む、請求項 2 5 に記載の方法。

【請求項 2 7】

前記 1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードを浄化すること (decontaminating) を含む、請求項 2 6 に記載の方法。

【請求項 2 8】

浄化が、インシリコ浄化を含む、請求項 2 7 に記載の方法。

【請求項 2 9】

浄化が、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する、請求項 2 7 に記載の方法。

【請求項 3 0】

マッピングが、Q I I M E 2 又はその他のサポートされたバージョンで実施される、請求項 2 6 に記載の方法。

【請求項 3 1】

前記 1 つ以上の微生物参照データベースが、細菌 1 6 S r R N A のデータベース G r e e n g e n e s 、細菌、真菌、及び古細菌 r R N A のデータベース S I L V A 、真核核リボソーム I T S 領域のデータベース U N I T E 、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む、請求項 2 6 に記載の方法。

【請求項 3 2】

前記 1 つ以上の非哺乳類配列決定リードの前記 1 つ以上のゲノム特徴の存在量が、微生物機能性遺伝子、生化学的経路、又はそれらの任意の組み合わせの存在量を含む、請求項 1 に記載の方法。

【請求項 3 3】

前記浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測し、それによって 1 つ以上の機能的存在量を生成することを含む、請求項 2 9 に記載の方法。

【請求項 3 4】

前記メタゲノム機能的含有量の予測が、P I C R U S t 2 によって実施される、請求項 3 3 に記載の方法。

【請求項 3 5】

前記液体生検試料が、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、

10

20

30

40

50

又はそれらの任意の組み合わせを含む、請求項 22 に記載の方法。

【請求項 36】

前記がんが、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む、請求項 1 に記載の方法。

【請求項 37】

前記肺がんが、非小細胞肺がんを含む、請求項 36 に記載の方法。

【請求項 38】

前記非がんの健康状態が、健康、疾患、又はそれらの任意の組み合わせを含む、請求項 1 に記載の方法。

【請求項 39】

前記疾患の状態が、肺疾患を含み、前記肺疾患が、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、サルコイドーシス、又はそれらの任意の組み合わせを含む、請求項 38 に記載の方法。

【請求項 40】

前記がんが、Ⅰ、Ⅱ、又はⅢ期のがんを含む、請求項 1 に記載の方法。

【請求項 41】

訓練された予測モデルを生成することを含み、前記訓練された予測モデルが、前記 1 つ以上の対象の前記特徴セット及び前記健康状態で訓練される、請求項 1 に記載の方法。

【請求項 42】

前記訓練された予測モデルが、機械学習モデル、1 つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む、請求項 41 に記載の方法。

【請求項 43】

前記訓練された予測モデルが、正則化された機械学習モデルを含む、請求項 41 に記載の方法。

【請求項 44】

前記機械学習モデルが、機械学習分類子を含む、請求項 42 に記載の方法。

【請求項 45】

前記機械学習モデルが、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k 平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせを含む、請求項 42 に記載の方法。

【請求項 46】

1 つ以上の対象のがんを診断する方法であって、前記方法が、

(a) 1 つ以上の対象の 1 つ以上の核酸分子を提供することと、

(b) 前記 1 つ以上の核酸分子の 1 つ以上の非哺乳類核酸分子の 1 つ以上のゲノム特徴を増幅し、それによって、増幅された 1 つ以上のゲノム特徴を生成することと、

(c) 前記増幅された 1 つ以上のゲノム特徴を配列決定して、1 つ以上の非哺乳類配列決定リードを生成することと、

(d) 少なくとも、前記 1 つ以上のゲノム特徴を、訓練された予測モデルへの入力として提供する結果として、前記 1 つ以上の対象のがん又は非がんの健康状態の診断を出力することと、を含む、方法。

【請求項 47】

前記非哺乳類核酸分子が、微生物核酸を含む、請求項 46 に記載の方法。

【請求項 48】

前記 1 つ以上の対象が、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む、請求項 46 に記載の方法。

【請求項 49】

前記 1 つ以上の核酸分子が、DNA、RNA、無細胞 DNA、無細胞 RNA、エクソソーム DNA、エクソソーム RNA、無細胞微生物 DNA、無細胞微生物 RNA、又はそれらの任意の組み合わせの総集団を含む、請求項 46 に記載の方法。

10

20

30

40

50

【請求項 50】

前記 1 つ以上のゲノム特徴が、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 46 に記載の方法。

【請求項 51】

前記微生物系統発生マーカー遺伝子が、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 50 に記載の方法。

【請求項 52】

前記微生物系統発生マーカー遺伝子が、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 50 に記載の方法。

【請求項 53】

前記細菌マーカー遺伝子が、リボソーム RNA 遺伝子 5 S、リボソーム RNA 遺伝子 16 S、リボソーム RNA 遺伝子 23 S、細菌ハウスキーピング遺伝子 *dnaG*、*frr*、*infC*、*nusA*、*pgk*、*pyrG*、*rplA*、*rplB*、*rplC*、*rplD*、*rplE*、*rplF*、*rplK*、*rplL*、*rplM*、*rplN*、*rplP*、*rplS*、*rplT*、*rpmA*、*rpoB*、*rpsB*、*rpsC*、*rpsE*、*rpsI*、*rpsJ*、*rpsK*、*rpsM*、*rpsS*、*smpB*、*tsf*、又はそれらの任意の組み合わせを含む、請求項 51 に記載の方法。

10

【請求項 54】

前記真菌マーカー遺伝子が、リボソーム RNA 遺伝子 18 S、リボソーム RNA 遺伝子 5.8 S、リボソーム RNA 遺伝子 28 S、内部転写されたスパーサー領域 1 及び 2、又はそれらの任意の組み合わせを含み得る、請求項 52 に記載の方法。

20

【請求項 55】

前記微生物系統発生マーカー遺伝子が、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含む、請求項 50 に記載の方法。

【請求項 56】

増幅が、ポリメラーゼ連鎖反応又はその派生物を実施することを含む、請求項 46 に記載の方法。

【請求項 57】

前記その派生物が、逆 PCR、アンカー PCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む、請求項 56 に記載の方法。

30

【請求項 58】

前記ポリメラーゼ連鎖反応が、1 つ以上のゲノム特徴の増幅を防止するように構成された、ブロックプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む、請求項 56 に記載の方法。

【請求項 59】

前記 1 つ以上のゲノム特徴が、ミトコンドリア DNA ゲノム特徴を含む、請求項 58 に記載の方法。

【請求項 60】

前記 1 つ以上の核酸分子を富化することを含む、請求項 46 に記載の方法。

【請求項 61】

前記 1 つ以上の核酸分子が、1 つ以上の哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸分子を含む、請求項 60 に記載の方法。

40

【請求項 62】

富化が、

(a) 前記 1 つ以上の哺乳類核酸分子及び前記 1 つ以上の非哺乳類核酸分子をハイブリダイゼーションプローブと組み合わせることであって、前記ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせることと、

(b) 前記ハイブリダイゼーションプローブ並びに前記 1 つ以上の哺乳類核酸分子及び前記 1 つ以上の非哺乳類核酸分子を、前記 1 つ以上の非哺乳類核酸分子の標的核酸特徴と前記ハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュ

50

ベーションすることと、

(c) 非結合のハイブリダイゼーションプローブ及び前記1つ以上の非哺乳類核酸分子に結合したハイブリダイゼーションされたプローブを分離することと、

(d) 前記1つ以上の非哺乳類核酸分子に結合した前記ハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸分子を生成することと、を含む、請求項61に記載の方法。

【請求項63】

洗浄が、非特異的に会合された核酸分子及び他の反応構成要素を除去するように構成される、請求項62に記載の方法。

【請求項64】

前記1つ以上の核酸分子の富化が、非哺乳類DNA富化を含む、請求項61に記載の方法。

【請求項65】

前記非哺乳類DNA富化が、

(a) 前記1つ以上の哺乳類核酸分子及び前記1つ以上の非哺乳類核酸分子を1つ以上の組換えCXXCDメインタンパク質と組み合わせて、タンパク質-DNA結合反応を形成すること、

(b) 前記組換えCXXCDメインタンパク質と、前記1つ以上の哺乳類核酸分子又は前記1つ以上の非哺乳類核酸の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、前記タンパク質-DNA結合反応をインキュベーションすること、

(c) 非結合の組換えCXXCDメインタンパク質及び前記非メチル化CpG核酸分子断片に結合した組換えCXXCDメインタンパク質を前記タンパク質-DNA結合反応の残りから分離すること、

(d) 前記非メチル化CpG核酸分子断片に結合した前記組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸を生成すること、を含む、請求項64に記載の方法。

【請求項66】

前記組換えCXXCDメインタンパク質が、組換えジンクフィンガーCXXCDメイン含有タンパク質KDM2A、KDM2A、KDM2B、FBXL19、CFP1、DNMT1、MLL1、MLL2、MDB1、TET1、TET3、IDAX、CXXC5、CGBP、それらに由来する組換えCXXCDメイン、又はそれらの任意の組み合わせを含む、請求項65に記載の方法。

【請求項67】

洗浄が、非特異的に会合された核酸分子及び前記タンパク質-DNA結合反応構成要素の前記残りを除去するように構成される、請求項65に記載の方法。

【請求項68】

前記1つ以上の核酸分子が、前記1つ以上の対象の1つ以上の生体試料に由来する、請求項46に記載の方法。

【請求項69】

前記1つ以上の生体試料が、組織、液体、又はそれらの任意の組み合わせの生検試料を含む、請求項68に記載の方法。

【請求項70】

前記1つ以上の哺乳類核酸分子が、DNA、RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、又はそれらの任意の組み合わせを含む、請求項61に記載の方法。

【請求項71】

前記1つ以上の非哺乳類核酸分子が、微生物DNA、微生物RNA、微生物無細胞DNA、微生物無細胞RNA、又はそれらの任意の組み合わせを含む、請求項61に記載の方法。

【請求項72】

10

20

30

40

50

前記 1 つ以上の非哺乳類配列決定リードをフィルタリングすることを含む、請求項 6 1 に記載の方法。

【請求項 7 3】

フィルタリングが、前記 1 つ以上の非哺乳類配列決定リードをフィルタリングして、1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードを作製することを含む、請求項 7 2 に記載の方法。

【請求項 7 4】

フィルタリングが、1 つ以上の微生物参照データベースに対して前記 1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードをマッピングして、前記 1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む、請求項 7 3 に記載の方法。

【請求項 7 5】

前記 1 つ以上のミトコンドリア DNA 枯渇非哺乳類配列決定リードを浄化することを含む、請求項 7 4 に記載の方法。

【請求項 7 6】

浄化が、インシリコ浄化を含む、請求項 7 5 に記載の方法。

【請求項 7 7】

浄化が、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する、請求項 7 5 に記載の方法。

【請求項 7 8】

マッピングが、Q I I M E 2 又はその他のサポートされたバージョンで実施される、請求項 7 4 に記載の方法。

【請求項 7 9】

前記 1 つ以上の微生物参照データベースが、前記細菌 1 6 S r R N A のデータベース G r e e n g e n e s、前記細菌、真菌、及び古細菌 r R N A のデータベース S I L V A、前記真核核リボソーム I T S 領域のデータベース U N I T E、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む、請求項 7 4 に記載の方法。

【請求項 8 0】

前記 1 つ以上のゲノム特徴が、前記 1 つ以上の非哺乳類配列決定リードの微生物機能性遺伝子、生化学的経路の存在量、又はそれらの任意の組み合わせの存在量を含む、請求項 4 6 に記載の方法。

【請求項 8 1】

前記浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測し、それによって 1 つ以上の機能的存在量を生成することを含む、請求項 7 7 に記載の方法。

【請求項 8 2】

前記メタゲノム機能的含有量の予測が、P I C R U S t 2 によって実施される、請求項 8 1 に記載の方法。

【請求項 8 3】

前記液体生検試料が、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む、請求項 6 9 に記載の方法。

【請求項 8 4】

前記がんの健康状態が、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む、請求項 4 6 に記載の方法。

【請求項 8 5】

前記肺がんが、非小細胞肺がんを含む、請求項 8 4 に記載の方法。

【請求項 8 6】

前記非がんの状態が、健康、疾患、又はそれらの任意の組み合わせの非がんの状態を含む、請求項 4 6 に記載の方法。

10

20

30

40

50

【請求項 87】

前記疾患が、肺疾患を含み、前記肺疾患が、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、又はそれらの任意の組み合わせを含む、請求項 86 に記載の方法。

【請求項 88】

前記がんの健康状態が、Ⅰ、Ⅱ、又はⅢ期のがんを含む、請求項 46 に記載の方法。

【請求項 89】

前記訓練された予測モデルが、1つ以上の対象の特徴セット及び健康状態で訓練される、請求項 46 に記載の方法。

10

【請求項 90】

前記訓練された予測モデルが、機械学習モデル、1つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む、請求項 46 に記載の方法。

【請求項 91】

前記訓練された予測モデルが、正則化された機械学習モデルを含む、請求項 46 に記載の方法。

【請求項 92】

前記機械学習モデルが、機械学習分類子を含む、請求項 90 に記載の方法。

【請求項 93】

前記機械学習モデルが、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k 平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む、請求項 90 に記載の方法。

20

【請求項 94】

1つ以上の対象のがん性又は非がん性の健康状態を診断するためのシステムであって、前記システムが、

(a) 1つ以上のプロセッサと、

(b) 前記 1つ以上のプロセッサに、

(i) 前記 1つ以上の対象の生体試料の 1つ以上の対象の 1つ以上の核酸分子配列決定リードを受け取ることであって、前記 1つ以上の核酸分子配列決定リードが、1つ以上の非哺乳類核酸分子の増幅された 1つ以上のゲノム特徴の配列を含む、受け取ること、及び

30

(ii) 少なくとも、前記 1つ以上の非哺乳類核酸配列決定リードの 1つ以上のゲノム特徴を、訓練された予測モデルへの入力として提供する結果として、前記 1つ以上の対象のがん性又は非がん性の健康状態の診断を出力することを行わせるように構成されたソフトウェアを含む、非一時的なコンピュータ可読記憶媒体と、を含む、システム。

【請求項 95】

前記非哺乳類核酸分子が、微生物核酸を含む、請求項 94 に記載のシステム。

【請求項 96】

前記生体試料が、組織、液体、又はそれらの任意の組み合わせの生検試料を含む、請求項 94 に記載のシステム。

40

【請求項 97】

前記 1つ以上の対象が、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む、請求項 94 に記載のシステム。

【請求項 98】

前記 1つ以上の核酸分子配列決定リードが、DNA、RNA、無細胞 DNA、無細胞 RNA、エクソソーム DNA、エクソソーム RNA、無細胞微生物 DNA、無細胞微生物 RNA、又はそれらの任意の組み合わせの配列決定リードを含む、請求項 94 に記載システム。

【請求項 99】

50

前記 1 つ以上のゲノム特徴が、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 94 に記載のシステム。

【請求項 100】

前記微生物系統発生マーカー遺伝子が、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 99 に記載のシステム。

【請求項 101】

前記微生物系統発生マーカー遺伝子が、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 99 に記載のシステム。

【請求項 102】

前記細菌マーカー遺伝子が、リボソーム RNA 遺伝子を含む、請求項 100 に記載のシステム。

【請求項 103】

前記リボソーム RNA 遺伝子が、5 S、16 S、23 S、又はそれらの任意の組み合わせのリボソーム RNA 遺伝子を含む、請求項 102 に記載のシステム。

【請求項 104】

前記細菌マーカー遺伝子が、リボソーム RNA 遺伝子 5 S、リボソーム RNA 遺伝子 16 S、リボソーム RNA 遺伝子 23 S、細菌ハウスキーピング遺伝子 dna G、frr、inf C、nus A、pgk、pyr G、rpl A、rpl B、rpl C、rpl D、rpl E、rpl F、rpl K、rpl L、rpl M、rpl N、rpl P、rpl S、rpl T、rpm A、rpo B、rps B、rps C、rps E、rps I、rps J、rps K、rps M、rps S、smp B、tsf、又はそれらの任意の組み合わせを含む、請求項 100 に記載のシステム。

【請求項 105】

前記真菌マーカー遺伝子が、リボソーム RNA 遺伝子 18 S、リボソーム RNA 遺伝子 5.8 S、リボソーム RNA 遺伝子 28 S、内部転写されたスペーサー領域 1 及び 2、又はそれらの任意の組み合わせを含む、請求項 101 に記載のシステム。

【請求項 106】

前記 1 つ以上の非哺乳類核酸分子の前記増幅された 1 つ以上のゲノム特徴が、ポリメラーゼ連鎖反応又はその派生物によって増幅される、請求項 94 に記載のシステム。

【請求項 107】

前記その派生物が、逆 PCR、アンカー PCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む、請求項 106 に記載のシステム。

【請求項 108】

前記ポリメラーゼ連鎖反応が、1 つ以上のゲノム特徴の増幅を防止するように構成された、ブロックングプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む、請求項 106 に記載のシステム。

【請求項 109】

前記 1 つ以上のゲノム特徴が、ミトコンドリア DNA ゲノム特徴を含む、請求項 108 に記載のシステム。

【請求項 110】

前記 1 つ以上の核酸分子配列決定リードが、1 つ以上の富化された核酸分子の配列決定リードを含む、請求項 94 に記載のシステム。

【請求項 111】

前記 1 つ以上の核酸分子配列決定リードの前記 1 つ以上の核酸分子が、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸分子配列決定リードを含む、請求項 110 に記載のシステム。

【請求項 112】

前記 1 つ以上の富化された核酸分子が、

(a) 前記 1 つ以上の核酸分子配列決定リードの 1 つ以上の核酸分子をハイブリダイゼーションプローブと組み合わせることであって、前記ハイブリダイゼーションプローブが

10

20

30

40

50

、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせること、

(b) 前記ハイブリダイゼーションプローブ及び前記1つ以上の核酸分子を、前記1つ以上の核酸分子の標的核酸特徴と前記ハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュベーションすること、

(c) 非結合のハイブリダイゼーションプローブ及び1つ以上の非哺乳類核酸分子に結合したハイブリダイゼーションされたプローブを分離すること、並びに

(d) 前記1つ以上の非哺乳類核酸分子に結合した前記ハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸分子を生成することによって生成される、請求項110に記載のシステム。

【請求項113】

洗浄が、非特異的に会合された核酸及び他の反応構成要素を除去するように構成される、請求項112に記載のシステム。

【請求項114】

前記1つ以上の富化された核酸分子が、非哺乳類DNA富化によって生成される、請求項110に記載のシステム。

【請求項115】

前記非哺乳類DNA富化が、

(a) 前記1つ以上の核酸分子を1つ以上の組換えCXXCDメインタンパク質と組み合わせて、タンパク質-DNA結合反応を形成すること、

(b) 前記組換えCXXCDメインタンパク質と、前記1つ以上の核酸分子の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、前記タンパク質-DNA結合反応をインキュベーションすること、

(c) 非結合の組換えCXXCDメインタンパク質及び前記非メチル化CpG核酸分子断片に結合した組換えCXXCDメインタンパク質を前記タンパク質-DNA結合反応の残りから分離すること、

(d) 前記非メチル化CpG核酸断片に結合した前記組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸分子を生成することを含む、請求項114に記載のシステム。

【請求項116】

洗浄が、非特異的に会合された核酸分子及び前記タンパク質-DNA結合反応構成要素の前記残りを除去するように構成される、請求項115に記載のシステム。

【請求項117】

前記ソフトウェアが、前記1つ以上のプロセッサを、前記1つ以上の核酸分子配列決定リードをフィルタリングするように構成する、請求項94に記載のシステム。

【請求項118】

前記フィルタリングが、前記1つ以上の核酸分子配列決定リードをフィルタリングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを作製することを含む、請求項117に記載のシステム。

【請求項119】

フィルタリングが、1つ以上の微生物参照データベースに対して前記1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードをマッピングして、前記1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む、請求項118に記載のシステム。

【請求項120】

前記ソフトウェアが、前記1つ以上のプロセッサを、前記1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを浄化するように構成する、請求項119に記載のシステム。

【請求項121】

浄化が、インシリコ浄化を含む、請求項120に記載のシステム。

【請求項122】

10

20

30

40

50

前記浄化が、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する、請求項 1 2 0 に記載のシステム。

【請求項 1 2 3】

マッピングが、Q I I M E 2 又はその他のサポートされたバージョンで実施される、請求項 1 1 9 に記載のシステム。

【請求項 1 2 4】

前記 1 つ以上の微生物参照データベースが、細菌 1 6 S r R N A のデータベース G r e e n g e n e s 、細菌、真菌、及び古細菌 r R N A のデータベース S I L V A 、真核核リボソーム I T S 領域のデータベース U N I T E 、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む、請求項 1 1 9 に記載のシステム。

10

【請求項 1 2 5】

前記増幅された 1 つ以上のゲノム特徴が、1 つ以上の非哺乳類配列決定リードの微生物機能性遺伝子、生化学的経路の存在量、又はそれらの任意の組み合わせの存在量を含む、請求項 1 1 1 に記載のシステム。

【請求項 1 2 6】

前記ソフトウェアが、前記 1 つ以上のプロセッサを、前記浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測するように構成し、それによって、1 つ以上の機能的存在量を生成する、請求項 1 2 2 に記載のシステム。

20

【請求項 1 2 7】

前記メタゲノム機能的含有量の予測が、P I C R U S t 2 によって実施される、請求項 1 2 6 に記載のシステム。

【請求項 1 2 8】

前記生体試料が、組織、液体、又はそれらの任意の組み合わせの生検試料を含む、請求項 9 4 に記載のシステム。

【請求項 1 2 9】

前記液体生検試料が、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む、請求項 1 2 8 に記載のシステム。

【請求項 1 3 0】

前記がん性の健康状態が、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む、請求項 9 4 に記載のシステム。

30

【請求項 1 3 1】

前記肺がんが、非小細胞肺がんを含む、請求項 1 3 0 に記載のシステム。

【請求項 1 3 2】

前記非がん性の健康状態が、健康、疾患、又はそれらの任意の組み合わせの非がんの状態を含む、請求項 9 4 に記載のシステム。

【請求項 1 3 3】

前記疾患の状態が、肺疾患を含み、前記肺疾患が、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、サルコイドーシス、又はそれらの任意の組み合わせを含む、請求項 1 3 2 に記載のシステム。

40

【請求項 1 3 4】

前記がん性の健康状態が、I、II、又はIII期のがんを含む、請求項 9 4 に記載のシステム。

【請求項 1 3 5】

前記訓練された予測モデルが、機械学習モデル、1 つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む、請求項 9 4 に記載のシステム。

【請求項 1 3 6】

前記訓練された予測モデルが、正則化された機械学習モデルを含む、請求項 9 4 に記載

50

のシステム。

【請求項 1 3 7】

前記機械学習モデルが、機械学習分類子を含む、請求項 1 3 5 に記載のシステム。

【請求項 1 3 8】

前記機械学習モデルが、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k 平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む、請求項 1 3 5 に記載のシステム。

【請求項 1 3 9】

前記がん性の健康状態又は前記非がん性の健康状態が、がん若しくは疾患のカテゴリ、組織特異的位置、又はそれらの任意の組み合わせを含む、請求項 9 4 に記載のシステム。

10

【請求項 1 4 0】

前記がん性の健康状態が、1 つ以上のタイプのがん、1 つ以上のサブタイプのがん、がんの病期、がんの予後、又はそれらの任意の組み合わせを含む、請求項 9 4 に記載のシステム。

【請求項 1 4 1】

前記訓練された予測モデルが、前記 1 つ以上の対象のがん療法応答を予測するために使用される、請求項 9 4 に記載のシステム。

【請求項 1 4 2】

前記訓練された予測モデルが、前記 1 つ以上の対象の疾患又はがんに対する最適な療法を選択するために利用される、請求項 9 4 に記載のシステム。

20

【請求項 1 4 3】

前記訓練された予測モデルが、療法に対する前記 1 つ以上の対象の 1 つ以上のがん又は 1 つ以上の疾患の応答の経過を縦断的にモデル化し、次いで、治療レジメンを調整するために利用される、請求項 9 4 に記載のシステム。

【請求項 1 4 4】

前記がん性の健康状態が、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、肺腺がん、肺扁平上皮がん、リンパ腫瘍びまん性大細胞型 B 細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む、請求項 9 4 に記載のシステム。

30

【請求項 1 4 5】

前記訓練された予測モデルが、前記 1 つ以上の非哺乳類核酸分子の汚染しない非哺乳類特徴の配列を選択的に保持しながら、前記 1 つ以上の核酸分子配列決定リードから汚染する非哺乳類特徴を除去するように構成される、請求項 9 4 に記載のシステム。

【請求項 1 4 6】

1 つ以上の対象のがんのタイプを区別するための特徴セットを生成する方法であって、前記方法が、

40

(a) 1 つ以上の対象の 1 つ以上の核酸分子及び対応する健康状態を提供することと、

(b) 前記 1 つ以上の核酸分子の 1 つ以上の非哺乳類核酸分子の 1 つ以上のゲノム特徴を増幅し、それによって、増幅された 1 つ以上のゲノム特徴を生成することと、

(c) 前記増幅された 1 つ以上のゲノム特徴を配列決定して、1 つ以上の非哺乳類配列決定リードを生成することと、

(d) 前記 1 つ以上の非哺乳類配列決定リードの前記 1 つ以上のゲノム特徴の存在量と、前記 1 つ以上の対象の前記健康状態と、を組み合わせることによって、がんのタイプを区別するように構成された特徴セットを生成することと、を含む、方法。

【請求項 1 4 7】

前記 1 つ以上のゲノム特徴が、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子

50

断片を含む、請求項 1 4 6 に記載の方法。

【請求項 1 4 8】

前記微生物系統発生マーカー遺伝子が、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 1 4 7 に記載の方法。

【請求項 1 4 9】

前記微生物系統発生マーカー遺伝子が、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む、請求項 1 4 7 に記載の方法。

【請求項 1 5 0】

前記細菌マーカー遺伝子が、リボソーム RNA 遺伝子 5 S、リボソーム RNA 遺伝子 16 S、リボソーム RNA 遺伝子 23 S、細菌ハウスキーピング遺伝子 dna G、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf、又はそれらの任意の組み合わせを含む、請求項 1 4 8 に記載の方法。

【請求項 1 5 1】

前記真菌マーカー遺伝子が、リボソーム RNA 遺伝子 18 S、リボソーム RNA 遺伝子 5.8 S、リボソーム RNA 遺伝子 28 S、並びに内部転写されたスペーサー領域 1 及び 2 のうちの 1 つ以上を含み得る、請求項 1 4 9 に記載の方法。

【請求項 1 5 2】

微生物系統発生マーカー遺伝子が、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含む、請求項 1 4 7 に記載の方法。

【請求項 1 5 3】

増幅が、ポリメラーゼ連鎖反応又はその派生物を実施することを含む、請求項 1 4 6 に記載の方法。

【請求項 1 5 4】

前記その派生物が、逆 PCR、アンカー PCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む、請求項 1 5 3 に記載の方法。

【請求項 1 5 5】

前記ポリメラーゼ連鎖反応が、1 つ以上のゲノム特徴の増幅を防止するように構成された、ブロックプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む、請求項 1 5 3 に記載の方法。

【請求項 1 5 6】

前記 1 つ以上のゲノム特徴が、ミトコンドリア DNA ゲノム特徴を含む、請求項 1 5 5 に記載の方法。

【請求項 1 5 7】

前記 1 つ以上の核酸分子を富化することを含む、請求項 1 4 6 に記載の方法。

【請求項 1 5 8】

前記 1 つ以上の核酸分子が、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸分子を含む、請求項 1 5 7 に記載の方法。

【請求項 1 5 9】

富化が、
(a) 前記 1 つ以上の哺乳類核酸分子及び前記 1 つ以上の非哺乳類核酸分子をハイブリダイゼーションプローブと組み合わせることであって、前記ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせることと、
(b) 前記ハイブリダイゼーションプローブ並びに前記 1 つ以上の哺乳類核酸分子及び前記 1 つ以上の非哺乳類核酸分子を、前記 1 つ以上の非哺乳類核酸分子の標的核酸特徴と前記ハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュベーションすることと、
(c) 非結合のハイブリダイゼーションプローブ及び前記 1 つ以上の非哺乳類核酸分子

10

20

30

40

50

に結合したハイブリダイゼーションされたプローブを分離することと、

(d) 前記1つ以上の非哺乳類核酸分子に結合した前記ハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸分子を生成することと、を含む、請求項158に記載の方法。

【請求項160】

洗浄が、非特異的に会合された核酸分子及び他の反応構成要素を除去するように構成される、請求項159に記載の方法。

【請求項161】

前記1つ以上の核酸分子の富化が、非哺乳類DNA富化を含む、請求項158に記載の方法。

【請求項162】

前記非哺乳類DNA富化が、

(a) 前記1つ以上の哺乳類核酸分子及び前記1つ以上の非哺乳類核酸分子を1つ以上の組換えCXXCDメインタンパク質と組み合わせて、タンパク質-DNA結合反応を形成すること、

(b) 前記組換えCXXCDメインタンパク質と、前記1つ以上の哺乳類核酸分子又は前記1つ以上の非哺乳類核酸分子の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、前記タンパク質-DNA結合反応をインキュベーションすること、

(c) 非結合の組換えCXXCDメインタンパク質及び前記非メチル化CpG核酸分子断片に結合した組換えCXXCDメインタンパク質を前記タンパク質-DNA結合反応の残りから分離すること、

(d) 前記非メチル化CpG核酸分子断片に結合した前記組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸分子を生成すること、を含む、請求項161に記載の方法。

【請求項163】

洗浄が、非特異的に会合された核酸分子及び前記タンパク質-DNA結合反応構成要素の残りを除去するように構成される、請求項162に記載の方法。

【請求項164】

前記1つ以上の核酸が、前記1つ以上の対象の1つ以上の生体試料に由来する、請求項146に記載の方法。

【請求項165】

前記1つ以上の生体試料が、組織、液体、又はそれらの任意の組み合わせの生検試料を含む、請求項164に記載の方法。

【請求項166】

前記1つ以上の対象が、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む、請求項146に記載の方法。

【請求項167】

前記1つ以上の哺乳類核酸分子が、DNA、RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、又はそれらの任意の組み合わせの核酸分子を含む、請求項158に記載の方法。

【請求項168】

前記1つ以上の非哺乳類核酸分子が、微生物DNA、微生物RNA、微生物無細胞DNA、微生物無細胞RNA、又はそれらの任意の組み合わせを含む、請求項158に記載の方法。

【請求項169】

前記1つ以上の非哺乳類配列決定リードをフィルタリングすることを含む、請求項146に記載の方法。

【請求項170】

フィルタリングが、前記1つ以上の非哺乳類配列決定リードをフィルタリングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを作製することを含む、請求

10

20

30

40

50

項 1 6 9 に記載の方法。

【請求項 1 7 1】

フィルタリングが、1つ以上の微生物参照データベースに対して前記1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードをマッピングして、前記1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む、請求項 1 7 0 に記載の方法。

【請求項 1 7 2】

前記1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを浄化することを含む、請求項 1 7 1 に記載の方法。

【請求項 1 7 3】

浄化が、インシリコ浄化を含む、請求項 1 7 2 に記載の方法。

【請求項 1 7 4】

浄化が、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する、請求項 1 7 2 に記載の方法。

【請求項 1 7 5】

マッピングが、Q I I M E 2 又はその他のサポートされたバージョンで実施される、請求項 1 7 1 に記載の方法。

【請求項 1 7 6】

前記1つ以上の微生物参照データベースが、細菌 1 6 S r R N A のデータベース G r e e n g e n e s 、細菌、真菌、及び古細菌 r R N A のデータベース S I L V A 、真核核リボソーム I T S 領域のデータベース U N I T E 、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む、請求項 1 7 1 に記載の方法。

【請求項 1 7 7】

前記1つ以上の非哺乳類配列決定リードの前記1つ以上のゲノム特徴の存在量が、微生物機能性遺伝子、生化学的経路、又はそれらの任意の組み合わせの存在量を含む、請求項 1 4 6 に記載の方法。

【請求項 1 7 8】

前記浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測し、それによって1つ以上の機能的存在量を生成することを含む、請求項 1 7 4 に記載の方法。

【請求項 1 7 9】

前記メタゲノム機能的含有量の予測が、P I C R U S t 2 によって実施される、請求項 1 7 8 に記載の方法。

【請求項 1 8 0】

前記液体生検試料が、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む、請求項 1 6 5 に記載の方法。

【請求項 1 8 1】

前記がんのタイプが、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む、請求項 1 4 6 に記載の方法。

【請求項 1 8 2】

前記肺がんが、非小細胞肺がんを含む、請求項 1 8 1 に記載の方法。

【請求項 1 8 3】

前記がんのタイプが、I、II、又はIII期のがんを含む、請求項 1 4 6 に記載の方法。

【請求項 1 8 4】

訓練された予測モデルを生成することを含み、前記訓練された予測モデルが、前記1つ以上の対象の前記特徴セット及び前記健康状態で訓練される、請求項 1 4 6 に記載の方法。

【請求項 1 8 5】

10

20

30

40

50

前記訓練された予測モデルが、機械学習モデル、1つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む、請求項184に記載の方法。

【請求項186】

前記訓練された予測モデルが、正則化された機械学習モデルを含む、請求項184に記載の方法。

【請求項187】

前記機械学習モデルが、機械学習分類子を含む、請求項185に記載の方法。

【請求項188】

前記機械学習モデルが、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む、請求項185に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

相互参照

本出願は、2022年3月10日に提出された米国仮出願第63／318,479号の利益を主張し、これは、全ての目的のために参照により本明細書に組み込まれる。

【発明の概要】

【0002】

本開示の態様は、1つ以上の対象のがん及び非がんの健康状態を区別するための特徴セットを生成する方法を提供する。いくつかの実施形態では、方法は、1つ以上の微生物ゲノム特徴の標的化されたアンプリコン配列決定に基づく。いくつかの実施形態では、方法は、(a) 1つ以上の対象の1つ以上の核酸及び対応する健康状態を提供するステップと、(b) 当該1つ以上の核酸の1つ以上の非哺乳類核酸の1つ以上のゲノム特徴を増幅し、それによって、増幅された1つ以上のゲノム特徴を生成するステップと、(c) 当該増幅された1つ以上のゲノム特徴を配列決定して、1つ以上の非哺乳類配列決定リードを生成するステップと、(d) 当該1つ以上の非哺乳類配列決定リードの当該1つ以上のゲノム特徴の存在量と、当該1つ以上の対象の当該健康状態と、を組み合わせることによって、がん及び非がんの健康状態を区別するように構成された特徴セットを生成するステップと、を含む。いくつかの実施形態では、ゲノム特徴は、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、細菌マーカー遺伝子は、リボソームRNA遺伝子5S、リボソームRNA遺伝子16S、リボソームRNA遺伝子23S、細菌ハウスキーピング遺伝子dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、真菌マーカー遺伝子は、リボソームRNA遺伝子18S、リボソームRNA遺伝子5.8S、リボソームRNA遺伝子28S、内部転写されたスペーサー領域1及び2、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含む。いくつかの実施形態では、増幅は、ポリメラーゼ連鎖反応又はその派生物を実施することを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応の派生物は、逆PCR、アンカーPCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応は、1つ以上のゲノム特徴の増幅を防止するように構成された、ブロッキングプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む。いくつか

10

20

30

40

50

の実施形態では、1つ以上のゲノム特徴は、ミトコンドリアDNAゲノム特徴を含む。いくつかの実施形態では、ブロッキングプライマーは、ミトコンドリアDNAゲノム特徴の増幅を阻害する。いくつかの実施形態では、方法は、1つ以上の核酸を富化することを更に含む。いくつかの実施形態では、1つ以上の核酸は、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸を含む。いくつかの実施形態では、核酸富化は、(a) 1つ以上の哺乳類及び非哺乳類核酸をハイブリダイゼーションプローブと組み合わせるステップであって、ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせるステップと、(b) ハイブリダイゼーションプローブ及び1つ以上の哺乳類及び非哺乳類核酸を、標的核酸特徴と当該ハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュベーションするステップと、(c) 非結合のハイブリダイゼーションプローブ及び非哺乳類核酸に結合したハイブリダイゼーションされたプローブを分離するステップと、(d) 非哺乳類核酸に結合した当該ハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸を生成するステップと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及び他の反応構成要素を除去するように構成される。いくつかの実施形態では、1つ以上の核酸の富化は、非哺乳類DNA富化を含む。いくつかの実施形態では、非哺乳類DNA富化は、(a) 1つ以上の哺乳類及び非哺乳類核酸を1つ以上の組換えCXXCDメインタンパク質と組み合わせ、タンパク質-DNA結合反応を形成するステップと、(b) 組換えCXXCDメインタンパク質と、1つ以上の哺乳類又は非哺乳類核酸の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、タンパク質-DNA結合反応をインキュベーションするステップと、(c) 非結合の組換えCXXCDメインタンパク質及び非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質をタンパク質-DNA結合反応の残りから分離するステップと、(d) 非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸を生成するステップと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及びタンパク質-DNA結合反応構成要素の残りを除去するように構成される。いくつかの実施形態では、1つ以上の核酸は、当該1つ以上の対象の1つ以上の生体試料に由来する。いくつかの実施形態では、1つ以上の生体試料は、組織、液体、又はそれらの任意の組み合わせの生検試料を含む。いくつかの実施形態では、液体生検試料は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の対象は、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む。いくつかの実施形態では、哺乳類及び非哺乳類核酸は、DNA、RNA、微生物無細胞DNA、微生物無細胞RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、又はそれらの任意の組み合わせの核酸を含む。いくつかの実施形態では、方法は、1つ以上の非哺乳類配列決定リードをフィルタリングすることを含む。いくつかの実施形態では、フィルタリングは、1つ以上の非哺乳類配列決定リードをフィルタリングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを作製することを含む。いくつかの実施形態では、フィルタリングは、1つ以上の微生物参照データベースに対して1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードをマッピングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む。いくつかの実施形態では、方法は、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを浄化すること (decontaminating) を含む。いくつかの実施形態では、浄化 (decontamination) は、インシリコ浄化を含む。いくつかの実施形態では、浄化は、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する。いくつかの実施形態では、非哺乳類配列決定リードマッピングは、QIIME2又はその他のサポートされたバージョンで実施される。いくつかの実施形態では、1つ以上の微生物参照データベースは、細菌16S rRNAのデータベースGreengenes、細菌、真菌、及び古細菌rRNAのデータベースSILVA、真核核リボソームI

10

20

30

40

50

T S 領域のデータベース U N I T E、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の非哺乳類配列決定リードの1つ以上のゲノム特徴の存在量が、微生物機能性遺伝子、生化学的経路、又はそれらの任意の組み合わせの存在量を含む。いくつかの実施形態では、方法は、浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測し、それによって1つ以上の機能的存在量を生成することを含む。いくつかの実施形態では、メタゲノム機能的含有量の予測は、P I C R U S t 2によって実施される。いくつかの実施形態では、がんは、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む。いくつかの実施形態では、肺がんは、非小細胞肺がんを含む。いくつかの実施形態では、がんは、I、II、又はIII期のがんを含む。いくつかの実施形態では、非がんの状態は、健康、疾患、又はそれらの任意の組み合わせの非がんの状態を含む。いくつかの実施形態では、疾患の状態は、肺疾患を含み、肺疾患が、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、サルコイドーシス、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、訓練された予測モデルを生成することを含み、訓練された予測モデルが、1つ以上の対象の特徴セット及び健康状態で訓練される場合である。いくつかの実施形態では、訓練された予測モデルは、機械学習モデル、1つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、正則化された機械学習モデルを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む。

10

20

30

40

50

【0003】

本明細書に開示される態様は、1つ以上の対象のがん又は非がんの健康状態を診断するために訓練された予測モデルの出力を使用する方法を提供する。いくつかの実施形態では、方法は、(a) 1つ以上の対象の1つ以上の核酸を提供するステップと、(b) 1つ以上の非哺乳類核酸の1つ以上のゲノム特徴を増幅し、それによって、増幅された1つ以上のゲノム特徴を生成するステップと、(c) 増幅された1つ以上のゲノム特徴を配列決定して、1つ以上の非哺乳類配列決定リードを生成するステップと、(d) 少なくとも、1つ以上のゲノム特徴を訓練された予測モデルへの入力として提供する結果として、1つ以上の対象のがん又は非がんの健康状態の診断を出力するステップと、を含む。いくつかの実施形態では、非哺乳類核酸は、微生物核酸を含む。いくつかの実施形態では、1つ以上の核酸は、1つ以上の対象の1つ以上の生体試料に由来する。いくつかの実施形態では、1つ以上の生体試料は、組織、液体、又はそれらの任意の組み合わせの生検試料を含む。いくつかの場合では、液体生検試料は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の対象は、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む。いくつかの実施形態では、1つ以上の核酸分子は、DNA、RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、無細胞微生物DNA、無細胞微生物RNA、又はそれらの任意の組み合わせの総集団を含む。いくつかの実施形態では、1つ以上のゲノム特徴は、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、細菌マーカー遺伝子は、リボソームRNA遺伝子5S、リボソームRNA遺伝子16S、リボソームRNA遺伝子23S、細菌ハウスキーピング遺伝子dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rp

s K、r p s M、r p s S、s m p B、t s f、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、真菌マーカー遺伝子は、リボソームRNA遺伝子18S、リボソームRNA遺伝子5.8S、リボソームRNA遺伝子28S、内部転写されたスペーサー領域1及び2、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含む。いくつかの実施形態では、増幅は、ポリメラーゼ連鎖反応又はその派生物を実施することを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応の派生物は、逆PCR、アンカーPCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応は、1つ以上のゲノム特徴の増幅を防止するように構成された、ブロッキングプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上のゲノム特徴は、ミトコンドリアDNAゲノム特徴を含む。いくつかの実施形態では、ブロッキングプライマーは、ミトコンドリアDNAゲノム特徴の増幅を阻害する。いくつかの実施形態では、方法は、1つ以上の核酸を富化することを含む。いくつかの実施形態では、1つ以上の核酸は、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸を含む。いくつかの実施形態では、哺乳類及び非哺乳類核酸は、DNA、RNA、微生物無細胞DNA、微生物無細胞RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、又はそれらの任意の組み合わせの核酸を含む。いくつかの実施形態では、核酸富化は、(a) 当該1つ以上の哺乳類及び非哺乳類核酸をハイブリダイゼーションプローブと組み合わせるステップであって、ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせるステップと、(b) ハイブリダイゼーションプローブ及び1つ以上の哺乳類及び非哺乳類核酸を、標的核酸特徴と当該ハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュベーションするステップと、(c) 非結合のハイブリダイゼーションプローブ及び非哺乳類核酸に結合したハイブリダイゼーションされたプローブを分離するステップと、(d) 非哺乳類核酸に結合したハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸を生成するステップと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及び他の反応構成要素を除去するように構成される。いくつかの実施形態では、当該1つ以上の核酸の富化は、非哺乳類DNA富化を含む。いくつかの実施形態では、非哺乳類DNA富化は、(a) 1つ以上の哺乳類及び非哺乳類核酸を1つ以上の組換えCXXCDメインタンパク質と組み合わせ、タンパク質-DNA結合反応を形成するステップと、(b) 組換えCXXCDメインタンパク質と、1つ以上の哺乳類又は非哺乳類核酸非メチル化CpGモチーフとの間の相互作用を促進する条件下で、タンパク質-DNA結合反応をインキュベーションするステップと、(c) 非結合の組換えCXXCDメインタンパク質及び非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質をタンパク質-DNA結合反応の残りから分離するステップと、(d) 非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸を生成するステップと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及びタンパク質-DNA結合反応構成要素の当該残りを除去するように構成される。いくつかの実施形態では、組換えCXXCDメインタンパク質は、組換えジンクフィンガーCXXCDメイン含有タンパク質KDM2A、KDM2A、KDM2B、FBXL19、CFP1、DNMT1、MLL1、MLL2、MDB1、TET1、TET3、IDAX、CXXC5、CGBP、それらに由来する組換えCXXCDメイン、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の非哺乳類配列決定をフィルタリングすることを含む。いくつかの実施形態では、フィルタリングは、1つ以上の非哺乳類配列決定リードをフィルタリングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを作製することを含む。いくつかの実施形態では、フィルタリングは、1つ以上の微生物参照データベースに対して1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードをマッピングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物

10

20

30

40

50

分類学的同一性を決定することを含む。いくつかの実施形態では、方法は、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを浄化することを含む。いくつかの実施形態では、浄化は、インシリコ浄化を含む。いくつかの実施形態では、浄化は、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する。いくつかの実施形態では、非哺乳類配列決定リードマッピングは、QIIME2又はその他のサポートされたバージョンで実施される。いくつかの実施形態では、1つ以上の微生物参照データベースは、細菌16S rRNAのデータベースGreengenes、細菌、真菌、及び古細菌rRNAのデータベースSILVA、真核核リボソームITS領域のデータベースUNITE、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上のゲノム特徴は、1つ以上の非哺乳類配列決定リードの微生物機能性遺伝子、生化学的経路の存在量、又はそれらの任意の組み合わせの存在量を含む。いくつかの実施形態では、方法は、浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測し、それによって、1つ以上の機能的存在量を生成することを含む。いくつかの実施形態では、メタゲノム機能的含有量は、PICRUSt2によって実施される。いくつかの実施形態では、がんの健康状態は、肺、乳房、卵巣、胃腸、頭頸部、肝臓、膵臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む。いくつかの実施形態では、肺がんは、非小細胞肺がんを含む。いくつかの実施形態では、がんは、I、II、又はIII期のがんを含む。いくつかの実施形態では、非がんの状態は、健康、疾患、又はそれらの任意の組み合わせの非がんの状態を含む。いくつかの実施形態では、疾患の状態は、肺疾患を含み、肺疾患が、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、サルコイドーシス、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、1つ以上の対象の特徴セット及び健康状態で訓練される。いくつかの実施形態では、訓練された予測モデルは、機械学習モデル、1つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、正則化された機械学習モデルを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む。

10

20

30

【0004】

本明細書に開示される態様は、1つ以上の対象のがん性又は非がん性の健康状態を診断するためのシステムを提供する。いくつかの実施形態では、システムは、(a)プロセッサと、(b)プロセッサに、(i)1つ以上の対象の生体試料の1つ以上の対象の1つ以上の核酸配列決定リードを受け取ることであって、1つ以上の核酸配列決定リードが、1つ以上の非哺乳類核酸の増幅された1つ以上のゲノム特徴を含む、受け取ることで、及び、(ii)少なくとも、1つ以上の非哺乳類核酸配列決定リードの1つ以上のゲノム特徴を、訓練された予測モデルへの入力として提供する結果として、1つ以上の対象のがん性又は非がん性の健康状態の診断を出力することを行わせるように構成されたソフトウェアを含む、非一時的なコンピュータ可読記憶媒体と、を含む。いくつかの実施形態では、非哺乳類核酸は、微生物核酸を含み得る。いくつかの実施形態では、1つ以上の生体試料は、組織、液体、又はそれらの任意の組み合わせの生検試料を含む。いくつかの実施形態では、1つ以上の対象は、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含み得る。いくつかの実施形態では、液体生検試料は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の核酸は、DNA、RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、無細胞微生物DNA、無細胞微生物RNA、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ゲノム特徴は、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの実施形態では、微生物系

40

50

統発生マーカー遺伝子は、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの実施形態では、細菌マーカー遺伝子は、リボソームRNA遺伝子を含む。いくつかの実施形態では、リボソームRNA遺伝子は、5S、16S、23S、又はそれらの任意の組み合わせのリボソームRNA遺伝子を含む。いくつかの実施形態では、細菌マーカー遺伝子は、リボソームRNA遺伝子5S、リボソームRNA遺伝子16S、リボソームRNA遺伝子23S、細菌ハウスキーピング遺伝子dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、真菌マーカー遺伝子は、リボソームRNA遺伝子18S、リボソームRNA遺伝子5.8S、リボソームRNA遺伝子28S、内部転写されたスペーサー領域1及び2、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含み得る。いくつかの実施形態では、1つ以上の非哺乳類核酸の増幅された1つ以上のゲノム特徴は、ポリメラーゼ連鎖反応又はその派生物によって増幅される。いくつかの実施形態では、ポリメラーゼ連鎖反応の派生物は、逆PCR、アンカーPCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応は、1つ以上のゲノム特徴の増幅を防止するように構成された、ブロックプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上のゲノム特徴は、ミトコンドリアDNAゲノム特徴を含む。いくつかの実施形態では、ブロックプライマーは、ミトコンドリアDNAゲノム特徴の増幅を阻害する。いくつかの実施形態では、1つ以上の核酸配列決定リードは、1つ以上の富化された核酸の配列決定リードを含む。いくつかの実施形態では、1つ以上の核酸は、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸を含み得る。いくつかの実施形態では、1つ以上の富化された核酸は、(a) 1つ以上の哺乳類及び非哺乳類核酸をハイブリダイゼーションプローブと組み合わせるステップであって、ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせるステップ、(b) ハイブリダイゼーションプローブ及び1つ以上の哺乳類及び非哺乳類核酸を、標的核酸特徴とハイブリダイゼーションプローブとの間の核酸塩基対合を促進する条件下でインキュベーションするステップ、(c) 非結合のハイブリダイゼーションプローブ及び非哺乳類核酸に結合したハイブリダイゼーションされたプローブを分離するステップ、並びに(d) 非哺乳類核酸に結合したハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸を生成するステップによって生成される。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及び他の反応構成要素を除去するように構成される。いくつかの実施形態では、1つ以上の富化された核酸は、非哺乳類DNA富化によって生成される。いくつかの実施形態では、非哺乳類富化は、(a) 1つ以上の哺乳類及び非哺乳類核酸を1つ以上の組換えCXXCDメインタンパク質と組み合わせ、タンパク質-DNA結合反応を形成することと、(b) 組換えCXXCDメインタンパク質と、1つ以上の哺乳類又は非哺乳類核酸の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、タンパク質-DNA結合反応をインキュベーションすることと、(c) 非結合の組換えCXXCDメインタンパク質及び非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質をタンパク質-DNA結合反応の残りから分離することと、(d) 非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸を生成することと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及びタンパク質-DNA結合反応構成要素の残りを除去するように構成される。いくつかの実施形態では、組換えCXXCDメインタンパク質は、組換えジンクフィンガーCXXCDメイン含有タンパク質KDM2A、K

10

20

30

40

50

DM2A、KDM2B、FBXL19、CFP1、DNMT1、MLL1、MLL2、MDB1、TET1、TET3、IDAX、CXXC5、CGBP、それらに由来する組換えCXXCドメイン、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ソフトウェアは、プロセッサを、1つ以上の核酸配列決定リードをフィルタリングするように構成する。いくつかの実施形態では、フィルタリングは、1つ以上の配列決定リードをフィルタリングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを作製することを含む。いくつかの実施形態では、フィルタリングは、1つ以上の微生物参照データベースに対して1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードをマッピングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む。いくつかの実施形態では、ソフトウェアは、プロセッサを、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを浄化するように構成する。いくつかの実施形態では、浄化は、インシリコ浄化を含む。いくつかの実施形態では、浄化は、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する。いくつかの実施形態では、マッピングは、QIIME2又はその他のサポートされたバージョンで実施される。いくつかの実施形態では、1つ以上の微生物参照データベースは、細菌16S rRNAのデータベースGreengenes、細菌、真菌、及び古細菌rRNAのデータベースSILVA、真核核リボソームITS領域のデータベースUNITE、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、増幅された1つ以上のゲノム特徴は、1つ以上の非哺乳類配列決定リードの微生物機能性遺伝子、生化学的経路の存在量、又はそれらの任意の組み合わせの存在量を含む。いくつかの実施形態では、メタゲノム機能的含有量の予測は、浄化された微生物分類学的割り当てに対して実施され、それによって、1つ以上の機能的存在量を生成する。いくつかの実施形態では、ソフトウェアは、浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測するようにプロセッサを構成し、それによって1つ以上の機能的存在量を生成する。いくつかの実施形態では、メタゲノム機能的含有量の予測は、PICRUSt2によって実施される。いくつかの実施形態では、がん性の健康状態は、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む。いくつかの実施形態では、肺がんは、非小細胞肺がんを含む。いくつかの実施形態では、がん性の状態は、I、II、又はIII期のがんを含む。いくつかの実施形態では、非がん性の健康状態は、健康、疾患、又はそれらの任意の組み合わせの非がん性の状態を含む。いくつかの実施形態では、疾患の状態は、肺疾患を含み得、肺疾患が、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、サルコイドーシス、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、1つ以上の対象の1つ以上のゲノム特徴セット及び当該健康状態で訓練される。いくつかの実施形態では、訓練された予測モデルは、機械学習モデル、1つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、正則化された機械学習モデルを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む。いくつかの実施形態では、がん性の健康状態は、1つ以上のタイプのがん、1つ以上のサブタイプのがん、がんの病期、がんの予後、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がん性又は非がん性の健康状態は、がん若しくは疾患のカテゴリ、組織特異的位置、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、1つ以上の対象のがん療法応答を予測するために使用される。いくつかの実施形態では、訓練された予測モデルは、1つ以上の対象に対する最適な療法を選択するために利用される。いくつかの実施形態では、訓練された予測モデルは、療法に対する1つ以上の対象の1つ以上のがんの応答の経過を縦断的にモデル化し、次の

10

20

30

40

50

で、治療レジメンを調整するために利用される。いくつかの実施形態では、がん性の健康状態は、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、肺腺がん、肺扁平上皮がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含み得る。いくつかの実施形態では、訓練された予測モデルは、他の汚染しない非哺乳類特徴を選択的に保持しながら、汚染する非哺乳類特徴を除

10

去するように構成され得る。

【0005】

本明細書に開示される態様は、1つ以上の対象のがんのタイプを区別するための特徴セットを生成する方法であって、方法が、(a) 1つ以上の対象の1つ以上の核酸及び対応する健康状態を提供することと、(b) 1つ以上の核酸の1つ以上の非哺乳類核酸の1つ以上のゲノム特徴を増幅し、それによって、増幅された1つ以上のゲノム特徴を生成することと、(c) 増幅された1つ以上のゲノム特徴を配列決定して、1つ以上の非哺乳類配列決定リードを生成することと、(d) 1つ以上の非哺乳類配列決定リードの1つ以上のゲノム特徴の存在量と、当該1つ以上の対象の健康状態と、を組み合わせることによって、がんのタイプを区別するように構成された特徴セットを生成することと、を含む、方法を提供する。いくつかの実施形態では、ゲノム特徴は、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含む。いくつかの実施形態では、細菌マーカー遺伝子は、リボソームRNA遺伝子5S、リボソームRNA遺伝子16S、リボソームRNA遺伝子23S、細菌ハウスキーピング遺伝子dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、真菌マーカー遺伝子は、リボソームRNA遺伝子18S、リボソームRNA遺伝子5.8S、リボソームRNA遺伝子28S、内部転写されたスパーサー領域1及び2、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、微生物系統発生マーカー遺伝子は、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含む。いくつかの実施形態では、増幅は、ポリメラーゼ連鎖反応又はその派生物を実施することを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応の派生物は、逆PCR、アンカーPCR、プライマー指向ローリングサークル増幅、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ポリメラーゼ連鎖反応は、1つ以上のゲノム特徴の増幅を防止するように構成された、ブロックングプライマー、マーカー遺伝子プライマー、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上のゲノム特徴は、ミトコンドリアDNAゲノム特徴を含む。いくつかの実施形態では、ブロックングプライマーは、ミトコンドリアDNAゲノム特徴の増幅を阻害する。いくつかの実施形態では、方法は、1つ以上の核酸を富化することを更に含む。いくつかの実施形態では、1つ以上の核酸は、哺乳類、非哺乳類、又はそれらの任意の組み合わせの核酸を含む。いくつかの実施形態では、核酸富化は、(a) 1つ以上の哺乳類及び非哺乳類核酸をハイブリダイゼーションプローブと組み合わせるステップであって、ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に対する核酸配列相補性を含む、組み合わせるステップと、(b) ハイブリダイゼーションプローブ及び1つ以上の哺乳類及び非哺乳類核酸を、標的核酸特徴と当該ハイブリダイゼーションプローブと

20

30

40

50

の間の核酸塩基対合を促進する条件下でインキュベーションするステップと、(c) 非結合のハイブリダイゼーションプローブ及び非哺乳類核酸に結合したハイブリダイゼーションされたプローブを分離するステップと、(d) 非哺乳類核酸に結合した当該ハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸を生成するステップと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及び他の反応構成要素を除去するように構成される。いくつかの実施形態では、1つ以上の核酸の富化は、非哺乳類DNA富化を含む。いくつかの実施形態では、非哺乳類DNA富化は、(a) 1つ以上の哺乳類及び非哺乳類核酸を1つ以上の組換えCXXCドメインタンパク質と組み合わせて、タンパク質-DNA結合反応を形成するステップと、(b) 組換えCXXCドメインタンパク質と、1つ以上の哺乳類又は非哺乳類核酸の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、タンパク質-DNA結合反応をインキュベーションするステップと、(c) 非結合の組換えCXXCドメインタンパク質及び非メチル化CpG核酸断片に結合した組換えCXXCドメインタンパク質をタンパク質-DNA結合反応の残りから分離するステップと、(d) 非メチル化CpG核酸断片に結合した組換えCXXCドメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸を生成するステップと、を含む。いくつかの実施形態では、洗浄は、非特異的に会合された核酸及びタンパク質-DNA結合反応構成要素の残りを除去するように構成される。いくつかの実施形態では、1つ以上の核酸は、当該1つ以上の対象の1つ以上の生体試料に由来する。いくつかの実施形態では、1つ以上の生体試料は、組織、液体、又はそれらの任意の組み合わせの生検試料を含む。いくつかの実施形態では、液体生検試料は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の対象は、ヒト、非ヒト哺乳動物、又はそれらの任意の組み合わせの対象を含む。いくつかの実施形態では、哺乳類及び非哺乳類核酸は、DNA、RNA、微生物無細胞DNA、微生物無細胞RNA、無細胞DNA、無細胞RNA、エクソソームDNA、エクソソームRNA、又はそれらの任意の組み合わせの核酸を含む。いくつかの実施形態では、方法は、1つ以上の非哺乳類配列決定リードをフィルタリングすることを含む。いくつかの実施形態では、フィルタリングは、1つ以上の非哺乳類配列決定リードをフィルタリングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを作製することを含む。いくつかの実施形態では、フィルタリングは、1つ以上の微生物参照データベースに対して1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードをマッピングして、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含む。いくつかの実施形態では、方法は、1つ以上のミトコンドリアDNA枯渇非哺乳類配列決定リードを浄化することを含む。いくつかの実施形態では、浄化は、インシリコ浄化を含む。いくつかの実施形態では、浄化は、非内因性微生物配列決定リードを除去するように構成され、それによって、浄化された微生物分類学的割り当て及び配列決定リードの関連する量を生成する。いくつかの実施形態では、非哺乳類配列決定リードマッピングは、QIIME 2又はその他のサポートされたバージョンで実施される。いくつかの実施形態では、1つ以上の微生物参照データベースは、細菌16S rRNAのデータベースGreengenes、細菌、真菌、及び古細菌rRNAのデータベースSILVA、真核核リボソームITS領域のデータベースUNITe、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の非哺乳類配列決定リードの1つ以上のゲノム特徴の存在量が、微生物機能性遺伝子、生化学的経路、又はそれらの任意の組み合わせの存在量を含む。いくつかの実施形態では、方法は、浄化された微生物分類学的割り当てのメタゲノム機能的含有量を予測し、それによって1つ以上の機能的存在量を生成することを含む。いくつかの実施形態では、メタゲノム機能的含有量の予測は、PICRUST 2によって実施される。いくつかの実施形態では、がんは、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含む。いくつかの実施形態では、肺がんは、非小細胞肺がんを含む。いくつかの実施形態では、がんは、I、II、又はI

10

20

30

40

50

ⅠⅠ期のがんを含む。いくつかの実施形態では、方法は、訓練された予測モデルを生成することを含み、訓練された予測モデルが、1つ以上の対象の特徴セット及び健康状態で訓練される場合である。いくつかの実施形態では、訓練された予測モデルは、機械学習モデル、1つ以上の機械学習モデル、機械学習モデルのアンサンブル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、訓練された予測モデルは、正則化された機械学習モデルを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、勾配ブースティング機械、ニューラルネットワーク、サポートベクター機械、k平均、分類木、ランダムフォレスト、回帰、又はそれらの任意の組み合わせの機械学習モデルを含む。

【0006】

本明細書で提供される本開示の態様は、対象の疾患を決定する方法であって、対象の生体試料、電子カルテ情報、及び1つ以上の放射線画像を受け取ることと、生体試料から単離された1つ以上の核酸分子を配列決定し、それによって、1つ以上の核酸分子配列決定リードを生成することと、予測モデルに、入力として、対象の1つ以上の核酸分子配列決定リード、電子カルテ情報、及び1つ以上の放射線画像に由来するデータが提供される場合、予測モデルの出力として対象の疾患を決定することと、を含む、方法を記載する。いくつかの実施形態では、方法は、対象の生体試料から1つ以上のタンパク質バイオマーカーを特定することを更に含む。いくつかの実施形態では、予測モデルには、対象の生体試料からの1つ以上のタンパク質バイオマーカーが提供される。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、がん抗原15-3、がん抗原19-9、がん抗原125、インターロイキン-8、プロラクチン、サイトケラチン19断片(CYFRA21-1)、MMP-9、sTNFRII、MMP-7、レジスチン、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFRII、又はそれらの組み合わせを含む。いくつかの実施形態では、疾患は、がん又は非がん性疾患性を含む。いくつかの実施形態では、生体試料は、液体生検、組織生検、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の放射線画像は、X線、コンピュータ断層撮影(CT)、低線量コンピュータ断層撮影、磁気共鳴画像法(MRI)、超音波、陽電子放出断層撮影、蛍光透視、血管造影、又はそれらの任意の組み合わせの画像を含む。いくつかの実施形態では、がんは、直径が3センチメートル未満の腫瘍塊を含む。いくつかの実施形態では、配列決定は、アンプリコンに基づく16S rRNA配列決定を含む。いくつかの実施形態では、アンプリコンに基づく16S rRNA配列決定は、1つ以上の核酸分子のV6領域を配列決定する。いくつかの実施形態では、1つ以上の核酸分子は、哺乳類RNA、哺乳類DNA、哺乳類無細胞DNA、哺乳類無細胞RNA、哺乳類エクソソームDNA、哺乳類エクソソームRNA、非ヒトRNA、非ヒトDNA、非ヒト無細胞DNA、非ヒト無細胞RNA、非ヒトエクソソームDNA、非ヒトエクソソームRNA、循環腫瘍DNA、循環腫瘍RNA、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、液体生検は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、肺腺がん(LUAD)、肺扁平上皮がん(LUSC)、小細胞肺がん(SCLC)、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の放射線画像の1つ以上の特徴を計算することを更に含み、1つ以上の放射線画像の1つ以

10

20

30

40

50

上の特徴は、予測モデルへの入力として提供される。いくつかの実施形態では、1つ以上の特徴は、ブロックがん確率スコア、病変の直径、病変の棘形成、病変の固体性、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の核酸配列決定リードをゲノムデータベースにマッピング又はアラインメントして、1つ以上の核酸配列決定リードの1つ以上のヒト、非ヒト、又はそれらの組み合わせの特徴を決定することを更に含む。いくつかの実施形態では、ゲノムデータベースは、ヒトゲノムデータベースを含む。いくつかの実施形態では、予測モデルは、機械学習モデルを含む。いくつかの実施形態では、予測モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク、ロジスティック回帰、ランダムフォレスト、夕食ベクター機械、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、スタックされた機械学習モデル、1つ以上の機械学習モデル、アンサンブル機械学習モデル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ抜き検証で訓練される。いくつかの実施形態では、予測モデルは、がんの病期、がんの解剖学的起源、又はそれらの組み合わせを決定するように構成される。いくつかの実施形態では、がんの病期は、I期、II期、III期、又はIV期である。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードを浄化して、1つ以上の浄化された配列決定リードを作製することを更に含む。いくつかの実施形態では、浄化は、インシリコ浄化、実験的対照浄化、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約91%、少なくとも約92%、少なくとも約93%、少なくとも約94%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の精度で疾患を決定する。いくつかの実施形態では、配列決定は、ショットガンメタゲノム配列決定、次世代配列決定、ロングリード配列決定、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードの1つ以上の特徴を決定することを更に含む。いくつかの実施形態では、1つ以上の核酸分子の1つ以上の特徴は、非微生物分類学的存在量、哺乳類ゲノム座標、注釈付きゲノム遺伝子座、哺乳類機能性遺伝子及び／又は生化学的経路存在量、又はそれらの任意の組み合わせの特徴、並びに当該1つ以上の特徴に関連する多数の配列決定リードを含む。いくつかの実施形態では、予測モデルは、対象のがん及び非がん性疾患を区別するように構成される。いくつかの実施形態では、マッピング又はアラインメントは、Deblur、Bowtie2、Kraken、又はそれらの任意の組み合わせで完了される。
【0007】

本明細書で提供される本開示の別の態様は、1つ以上の対象の生体試料、電子カルテ情報、1つ以上の放射線画像に由来するデータ、及び対応する疾患を受け取ることと、生体試料から単離された1つ以上の核酸分子を配列決定し、それによって、1つ以上の核酸分子配列決定リードを生成することと、1つ以上の対象の疾患に対応する1つ以上の核酸分子配列決定リード、電子カルテ情報、及び1つ以上の放射線画像に由来するデータの1つ以上の特徴を特定することと、を含む、方法を記載する。いくつかの実施形態では、特定は、1つ以上の配列決定リードをゲノムデータベースにアラインメントすることを含む。いくつかの実施形態では、方法は、予測モデルを、1つ以上の対象の核酸分子配列決定リード、電子カルテ情報、及び1つ以上の放射線画像に由来するデータ、並びに対応する疾患の1つ以上の特徴で訓練することを更に含む。いくつかの実施形態では、疾患は、がん又は非がん性疾患を含む。いくつかの実施形態では、方法は、対象の生体試料の1つ以上のタンパク質バイオマーカーの1つ以上の特徴を特定することを更に含む。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、がん抗原15-3、がん抗原19-9、がん抗原125、インターロイキン-8、プロラクチン、サイトケラチン19断片(CYFRA21-1)、MMP-9、sTNFR II、MMP-7、レジスチン、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、M

10

20

30

40

50

I F、P D G F、A B / b b、R A N T E S、S A A、T N F R I I、又はそれらの組み合わせを含む。いくつかの実施形態では、生体試料は、液体生検、組織生検、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の放射線画像は、X線、コンピュータ断層撮影（C T）、低線量コンピュータ断層撮影、磁気共鳴画像法（M R I）、超音波、陽電子放出断層撮影、蛍光透視、血管造影、又はそれらの任意の組み合わせの画像を含む。いくつかの実施形態では、がんは、直径が3センチメートル未満の腫瘍塊を含む。いくつかの実施形態では、配列決定は、アンプリコンに基づく16S rRNA配列決定を含む。いくつかの実施形態では、アンプリコンに基づく16S rRNA配列決定は、1つ以上の核酸分子のV6領域を配列決定する。いくつかの実施形態では、1つ以上の核酸分子は、哺乳類RNA、哺乳類DNA、哺乳類無細胞DNA、哺乳類無細胞RNA、哺乳類エクソソームDNA、哺乳類エクソソームRNA、非ヒトRNA、非ヒトDNA、非ヒト無細胞DNA、非ヒト無細胞RNA、非ヒトエクソソームDNA、非ヒトエクソソームRNA、循環腫瘍DNA、循環腫瘍RNA、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、液体生検は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、肺腺がん（L U A D）、肺扁平上皮がん（L U S C）、小細胞肺がん（S C L C）、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巢漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の放射線画像特徴は、ブロックがん確率スコア、病変の直径、病変の棘形成、病変の固体性、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の核酸配列決定リードをゲノムデータベースにマッピング又はアラインメントして、1つ以上の核酸配列決定リードの1つ以上のヒト、非ヒト、又はそれらの組み合わせの特徴を決定することを更に含む。いくつかの実施形態では、ゲノムデータベースは、ヒトゲノムデータベースを含む。いくつかの実施形態では、予測モデルは、機械学習モデルを含む。いくつかの実施形態では、予測モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク、ロジスティック回帰、ランダムフォレスト、夕食ベクター機械、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、スタックされた機械学習モデル、1つ以上の機械学習モデル、アンサンブル機械学習モデル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ抜き検証で訓練される。いくつかの実施形態では、予測モデルは、がんの病期、がんの解剖学的起源、又はそれらの組み合わせを決定するように構成される。いくつかの実施形態では、がんの病期は、I期、II期、又はIII期、又はIV期である。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードを浄化して、1つ以上の浄化された配列決定リードを作製することを更に含む。いくつかの実施形態では、浄化は、インシリコ浄化、実験的対照浄化、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約91%、少なくとも約92%、少なくとも約93%、少なくとも約94%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の精度で疾患を決定する。いくつかの実施形態では、配列決定は、ショットガン配列決定、次世代配列決定、ロングリード配列決定、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の核酸分子配列決定リードの1つ以上の特徴を決定することを更に含む方法。いくつかの実施形態では、1つ以上の核酸分子の1つ以上の特徴は、非微生物分類学的存在量、哺乳

10

20

30

40

50

乳類ゲノム座標、注釈付きゲノム遺伝子座、哺乳類機能性遺伝子及び／又は生化学的経路存在量、又はそれらの任意の組み合わせの特徴、並びに当該1つ以上の特徴に関連する多数の配列決定リードを含む。いくつかの実施形態では、予測モデルは、対象のがん及び非がん性疾患を区別するように構成される。いくつかの実施形態では、マッピング又はアラインメントは、Deblur、Bowtie2、Kraken、又はそれらの任意の組み合わせで完了される。

【0008】

本明細書で提供される本開示の別の態様は、対象の疾患を決定するように構成されたコンピュータシステムであって、(a) 1つ以上のプロセッサと、(b) ソフトウェアを含む非一時的なコンピュータ可読記憶媒体であって、ソフトウェアが、実行の結果として、コンピュータシステムの1つ以上のプロセッサに、(i) 対象の生体試料、電子カルテ情報、及び1つ以上の画像の1つ以上の配列決定リードを受け取ること、並びに(ii) 予測モデルに、入力として、対象の1つ以上の核酸分子配列決定リード、電子カルテ情報、及び1つ以上の放射線画像に由来するデータが提供される場合、予測モデルの出力として対象の疾患を決定することを行わせる実行可能命令を含む、非一時的なコンピュータ可読記憶媒体と、を含む、コンピュータシステムを記載する。いくつかの実施形態では、疾患は、がん又は非がん性疾患を含む。いくつかの実施形態では、生体試料は、組織生検、液体生検、又はそれらの組み合わせを含む。いくつかの実施形態では、実行可能命令は、対象の生体試料から1つ以上のタンパク質バイオマーカーを受け取ることを含む。いくつかの実施形態では、予測モデルには、対象の生体試料からの1つ以上のタンパク質バイオマーカーが提供される。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ以上の対象の核酸分子配列決定リード、電子カルテ情報、及び1つ以上の放射線画像に由来するデータ、並びに対応する疾患の1つ以上の特徴で訓練される。いくつかの実施形態では、実行可能命令は、対象の生体試料の1つ以上のタンパク質バイオマーカーの1つ以上の特徴を特定することを含む。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、がん抗原15-3、がん抗原19-9、がん抗原125、インターロイキン-8、プロラクチン、サイトケラチン19断片(CYFRA21-1)、MMP-9、sTNFR1I、MMP-7、レジスチン、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFR1I、又はそれらの組み合わせを含む。いくつかの実施形態では、1つ以上の放射線画像は、X線、コンピュータ断層撮影(CT)、低線量コンピュータ断層撮影、磁気共鳴画像法(MRI)、超音波、陽電子放出断層撮影、蛍光透視、血管造影、又はそれらの任意の組み合わせの画像を含む。いくつかの実施形態では、がんは、直径が3センチメートル未満の腫瘍塊を含む。いくつかの実施形態では、1つ以上の核酸分子配列決定リードは、1つ以上のアンプリコンに基づく16SrRNA配列決定リードを含む。いくつかの実施形態では、アンプリコンに基づく16SrRNA配列決定リードは、1つ以上の核酸分子のV6領域の配列決定リードを含む。いくつかの実施形態では、1つ以上の核酸分子配列決定リードは、哺乳類RNA、哺乳類DNA、哺乳類無細胞DNA、哺乳類無細胞RNA、哺乳類エクソソームDNA、哺乳類エクソソームRNA、非ヒトRNA、非ヒトDNA、非ヒト無細胞DNA、非ヒト無細胞RNA、非ヒトエクソソームDNA、非ヒトエクソソームRNA、循環腫瘍DNA、循環腫瘍RNA、又はそれらの任意の組み合わせの配列決定リードを含む。いくつかの実施形態では、液体生検は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、肺腺がん(LUAD)、肺扁平上皮がん(LUSC)、小細胞肺がん(SCLC)、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、

10

20

30

40

50

腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の放射線画像特徴は、ブロックがん確率スコア、病変の直径、病変の棘形成、病変の固体性、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の核酸配列決定リードをゲノムデータベースにマッピング又はアラインメントして、1つ以上の核酸配列決定リードの1つ以上のヒト、非ヒト、又はそれらの組み合わせの特徴を決定することを更に含む実行可能命令。いくつかの実施形態では、ゲノムデータベースは、ヒトゲノムデータベースを含む。いくつかの実施形態では、予測モデルは、機械学習モデルを含む。いくつかの実施形態では、予測モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク、ロジスティック回帰、ランダムフォレスト、タ食ベクター機械、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、スタックされた機械学習モデル、1つ以上の機械学習モデル、アンサンブル機械学習モデル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ抜き検証で訓練される。いくつかの実施形態では、予測モデルは、がんの病期、がんの解剖学的起源、又はそれらの組み合わせを決定するように構成される。いくつかの実施形態では、がんの病期は、I期、II期、III期、又はIV期である。いくつかの実施形態では、実行可能命令は、1つ以上の核酸分子配列決定リードを浄化して、1つ以上の浄化された配列決定リードを作製することを更に含む。いくつかの実施形態では、浄化は、インシリコ浄化、実験的対照浄化、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約91%、少なくとも約92%、少なくとも約93%、少なくとも約94%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の精度で疾患を決定する。いくつかの実施形態では、1つ以上の配列決定リードは、ショットガン配列決定、次世代配列決定、ロングリード配列決定、又はそれらの任意の組み合わせによって生成される。いくつかの実施形態では、実行可能命令は、1つ以上の核酸分子配列決定リードの1つ以上の特徴を決定することを更に含む。いくつかの実施形態では、1つ以上の核酸分子の1つ以上の特徴は、非微生物分類学的存在量、哺乳類ゲノム座標、注釈付きゲノム遺伝子座、哺乳類機能性遺伝子及び／又は生化学的経路存在量、又はそれらの任意の組み合わせの特徴、並びに当該1つ以上の特徴に関連する多数の配列決定リードを含む。いくつかの実施形態では、予測モデルは、対象のがん及び非がん性疾患を区別するように構成される。いくつかの実施形態では、マッピング又はアラインメントは、Deblur、Bowtie2、Kraken、又はそれらの任意の組み合わせで完了される。

10

20

30

40

50

【0009】

本明細書で提供される本開示の別の態様は、対象の疾患を決定する方法であって、対象から生体試料を受け取ることと、生体試料の1つ以上の核酸分子を配列決定し、それによって、1つ以上の核酸分子配列決定リードを生成することと、予測モデルに、入力として、対象の1つ以上の核酸分子配列決定リードが提供される場合、予測モデルの出力として対象の疾患を決定することと、を含み、予測モデルが、1つ以上の対象の1つ以上の液体生体試料及び1つ以上の組織生体試料の1つ以上の核酸分子配列決定リード、並びに対応する疾患で訓練される、方法を記載する。いくつかの実施形態では、疾患は、がん、非がん性疾患性、又はそれらの組み合わせを含む。いくつかの実施形態では、方法は、対象の生体試料から1つ以上のタンパク質バイオマーカーを特定することを更に含む。いくつかの実施形態では、予測モデルには、対象の生体試料からの1つ以上のタンパク質バイオマーカーが提供される。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、がん抗原15-3、がん抗原19-9、がん抗原

125、インターロイキン-8、プロラクチン、サイトケラチン19断片(CYFRA21-1)、MMP-9、sTNFRII、MMP-7、レジスチン、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFRII、又はそれらの組み合わせを含む。いくつかの実施形態では、がんは、直径が3センチメートルミリメートル未満の腫瘍塊を含む。いくつかの実施形態では、配列決定は、アンプリコンに基づく16S rRNA配列決定を含む。いくつかの実施形態では、アンプリコンに基づく16S rRNA配列決定は、1つ以上の核酸分子のV6領域を配列決定する。いくつかの実施形態では、1つ以上の核酸分子は、哺乳類RNA、哺乳類DNA、哺乳類無細胞DNA、哺乳類無細胞RNA、哺乳類エクソソームDNA、哺乳類エクソソームRNA、非ヒトRNA、非ヒトDNA、非ヒト無細胞DNA、非ヒト無細胞RNA、非ヒトエクソソームDNA、非ヒトエクソソームRNA、循環腫瘍DNA、循環腫瘍RNA、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、液体生検は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、肺腺がん(LUAD)、肺扁平上皮がん(LUSC)、小細胞肺がん(SCLC)、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の核酸配列決定リードをゲノムデータベースにマッピング又はアラインメントして、予測モデルへの入力として提供される1つ以上の核酸配列決定リードの1つ以上のヒト、非ヒト、又はそれらの組み合わせの特徴を決定することを更に含む。いくつかの実施形態では、ゲノムデータベースは、ヒトゲノムデータベースを含む。いくつかの実施形態では、予測モデルは、機械学習モデルを含む。いくつかの実施形態では、予測モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク、ロジスティック回帰、ランダムフォレスト、タ食ベクター機械、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、スタックされた機械学習モデル、1つ以上の機械学習モデル、アンサンブル機械学習モデル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ抜き検証で訓練される。いくつかの実施形態では、予測モデルは、がんの病期、がんの解剖学的起源、又はそれらの組み合わせを決定するように構成される。いくつかの実施形態では、がんの病期は、I期、II期、III期、又はIV期である。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードを浄化して、1つ以上の浄化された核酸分子配列決定リードを作製することを更に含み、1つ以上の浄化された核酸分子が入力として予測モデルに提供される。いくつかの実施形態では、浄化は、インシリコ浄化、実験的対照浄化、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約91%、少なくとも約92%、少なくとも約93%、少なくとも約94%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の精度で疾患を決定する。いくつかの実施形態では、配列決定は、ショットガン配列決定、次世代配列決定、ロングリード配列決定、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードの1つ以上の特徴を決定することを更に含む。いくつかの実施形態では、1つ以上の核酸分子の1つ以上の特徴は、非微生物分類学的存在量、哺乳類ゲノム座標、注釈付きゲノム遺伝子座、哺乳類機能性遺伝子及び／又は生化学的経路存在

10

20

30

40

50

量、又はそれらの任意の組み合わせの特徴、並びに当該1つ以上の特徴に関連する多数の配列決定リードを含む。いくつかの実施形態では、予測モデルは、対象のがん及び非がん性疾患を区別するように構成される。いくつかの実施形態では、マッピング又はアラインメントは、Deblur、PICRUST2、Bowtie2、Kraken、又はそれらの任意の組み合わせで完了される。

【0010】

本明細書で提供される本開示の別の態様は、1つ以上の非ヒトゲノム特徴を特定する方法であって、1つ以上の対象の1つ以上の液体生体試料、1つ以上の組織生体試料、及び対応する疾患を受け取ることと、1つ以上の液体生体試料及び1つ以上の組織生体試料の1つ以上の核酸分子を配列決定し、それによって、1つ以上の配列決定リードを生成することと、1つ以上の配列決定リードから1つ以上の対象の疾患に対応する1つ以上の非ヒトゲノム特徴を特定することと、を含む、方法を記載する。いくつかの実施形態では、特定は、1つ以上の配列決定リードをゲノムデータベースにアラインメント又はマッピングして、1つ以上の核酸配列決定リードの1つ以上のヒト、非ヒト、又はそれらの組み合わせの特徴を決定することを含む。いくつかの実施形態では、方法は、1つ以上の対象の1つ以上の非ヒトゲノム特徴及び対応する疾患で予測モデルを訓練することを更に含む。いくつかの実施形態では、疾患は、がん又は非がん性疾患を含む。いくつかの実施形態では、方法は、1つ以上の液体生体試料、1つ以上の組織生体試料、又はそれらの組み合わせの1つ以上のタンパク質バイオマーカーの1つ以上の特徴を特定することを更に含む。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、がん抗原15-3、がん抗原19-9、がん抗原125、インターロイキン-8、プロラクチン、サイトケラチン19断片(CYFRA21-1)、MMP-9、sTNFRII、MMP-7、レジスチン、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFRII、又はそれらの組み合わせを含む。いくつかの実施形態では、がんは、直径が3センチメートル未満の腫瘍塊を含む。いくつかの実施形態では、配列決定は、アンプリコンに基づく16S rRNA配列決定を含む。いくつかの実施形態では、アンプリコンに基づく16S rRNA配列決定は、1つ以上の核酸分子のV6領域を配列決定する。いくつかの実施形態では、1つ以上の核酸分子は、哺乳類RNA、哺乳類DNA、哺乳類無細胞DNA、哺乳類無細胞RNA、哺乳類エクソソームDNA、哺乳類エクソソームRNA、非ヒトRNA、非ヒトDNA、非ヒト無細胞DNA、非ヒト無細胞RNA、非ヒトエクソソームDNA、非ヒトエクソソームRNA、循環腫瘍DNA、循環腫瘍RNA、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、液体生体試料は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、肺腺がん(LUAD)、肺扁平上皮がん(LUSC)、小細胞肺がん(SCLC)、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、ゲノムデータベースは、ヒトゲノムデータベースを含む。いくつかの実施形態では、予測モデルは、機械学習モデルを含む。いくつかの実施形態では、予測モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク、ロジスティック回帰、ランダムフォレスト、夕食ベクター機械、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、スタックされた機械学習モデル、1つ以上の機械学習

10

20

30

40

50

モデル、アンサンブル機械学習モデル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ抜き検証で訓練される。いくつかの実施形態では、予測モデルは、がんの病期、がんの解剖学的起源、又はそれらの組み合わせを決定するように構成される。いくつかの実施形態では、がんの病期は、Ⅰ期、Ⅱ期、Ⅲ期、又はⅣ期である。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードを浄化して、1つ以上の浄化された配列決定リードを作製することを更に含む。いくつかの実施形態では、浄化は、インシリコ浄化、実験的対照浄化、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約91%、少なくとも約92%、少なくとも約93%、少なくとも約94%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の精度で疾患を決定する。いくつかの実施形態では、配列決定は、ショットガン配列決定、次世代配列決定、ロングリード配列決定、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、方法は、1つ以上の核酸分子配列決定リードの1つ以上の特徴を決定することを更に含む。いくつかの実施形態では、1つ以上の核酸分子の1つ以上の特徴は、非微生物分類学的存在量、哺乳類ゲノム座標、注釈付きゲノム遺伝子座、哺乳類機能性遺伝子及び／又は生化学的経路存在量、又はそれらの任意の組み合わせの特徴、並びに当該1つ以上の特徴に関連する多数の配列決定リードを含む。いくつかの実施形態では、予測モデルは、対象のがん及び非がん性疾患を区別するように構成される。いくつかの実施形態では、マッピング又はアラインメントは、Deblur、PICRUST2、Bowtie2、Kraken、又はそれらの任意の組み合わせで完了される。

10

20

30

40

50

【0011】

本明細書で提供される本開示の別の態様は、対象の疾患を決定するように構成されたコンピュータシステムであって、(a) 1つ以上のプロセッサと、(b) ソフトウェアを含む非一時的なコンピュータ可読記憶媒体であって、ソフトウェアが、実行の結果として、コンピュータシステムの1つ以上のプロセッサに、(i) 対象の生体試料の1つ以上の配列決定リードを受け取ること、並びに(ii) 予測モデルに、入力として、対象の1つ以上の核酸分子配列決定リードが提供される場合、予測モデルの出力として対象の疾患を決定することを行わせる実行可能命令を含む、非一時的なコンピュータ可読記憶媒体と、を含み、予測モデルが、1つ以上の対象の1つ以上の液体生体試料及び1つ以上の組織生体試料の1つ以上の核酸分子配列決定リード、並びに対応する疾患で訓練される、コンピュータシステムを記載する。いくつかの実施形態では、疾患は、がん又は非がん性疾患を含む。いくつかの実施形態では、疾患は、がん又は非がん性疾患を含む。いくつかの実施形態では、実行可能命令は、対象の生体試料から1つ以上のタンパク質バイオマーカーを受け取ることを含む。いくつかの実施形態では、予測モデルには、対象の生体試料からの1つ以上のタンパク質バイオマーカーが提供される。いくつかの実施形態では、実行可能命令は、対象の生体試料の1つ以上のタンパク質バイオマーカーの1つ以上の特徴を特定することを含む。いくつかの実施形態では、1つ以上のタンパク質バイオマーカーは、がん胎児性抗原、オステオポンチン、がん抗原15-3、がん抗原19-9、がん抗原125、インターロイキン-8、プロラクチン、サイトケラチン19断片(CYFRA21-1)、MMP-9、sTNFRII、MMP-7、レジスチン、MPO、MCP-1、GRO、sVEGFR2、sKDR、sFlk-1、VEGF-A、VEGF-C、VEGF-D、HGF、CRp、MIF、PDGF、AB/bb、RANTES、SAA、TNFRII、又はそれらの組み合わせを含む。いくつかの実施形態では、がんは、直径が3センチメートル未満の腫瘍塊を含む。いくつかの実施形態では、1つ以上の核酸分子配列決定リードは、1つ以上のアンプリコンに基づく16S rRNA配列決定リードを含む。いくつかの実施形態では、アンプリコンに基づく16S rRNA配列決定リードは、1つ以上の核酸分子のV6領域の配列決定リードを含む。いくつかの実施形態では、1つ以上の核酸分子配列決定リードは、哺乳類RNA、哺乳類DNA、哺乳類無細胞DNA、哺乳類無細胞RNA、哺乳類エクソソームDNA、哺乳類エクソソームRNA、非ヒトRN

A、非ヒトDNA、非ヒト無細胞DNA、非ヒト無細胞RNA、非ヒトエクソソームDNA、非ヒトエクソソームRNA、循環腫瘍DNA、循環腫瘍RNA、又はそれらの任意の組み合わせの配列決定リードを含む。いくつかの実施形態では、液体生体試料は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、肺腺がん（LUAD）、肺扁平上皮がん（LUSC）、小細胞肺がん（SCLC）、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、がんは、急性骨髄性白血病、副腎皮質がん、膀胱尿路上皮がん、脳低悪性度神経膠腫、乳房浸潤がん、子宮頸部扁平上皮がん及び子宮頸管腺がん、胆管がん、結腸腺がん、食道がん、多形性神経膠芽腫、頭頸部扁平上皮がん、腎臓嫌色素性細胞、腎臓腎明細胞がん、腎臓腎乳頭細胞がん、肝臓肝細胞がん、リンパ腫瘍びまん性大細胞型B細胞リンパ腫、中皮腫、卵巣漿液性嚢胞腺がん、膵臓腺がん、褐色細胞腫及び傍神経節腫、前立腺腺がん、直腸腺がん、肉腫、皮膚の皮膚黒色腫、胃腺がん、精巣胚細胞腫瘍、胸腺腫、甲状腺がん、子宮がん肉腫、子宮体部子宮内膜がん、ブドウ膜黒色腫、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、1つ以上の核酸配列決定リードをゲノムデータベースにマッピング又はアラインメントして、1つ以上の核酸配列決定リードの1つ以上のヒト、非ヒト、又はそれらの組み合わせの特徴を決定することを更に含む実行可能命令。いくつかの実施形態では、ゲノムデータベースは、ヒトゲノムデータベースを含む。いくつかの実施形態では、予測モデルは、機械学習モデルを含む。いくつかの実施形態では、予測モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク、ロジスティック回帰、ランダムフォレスト、夕食ベクター機械、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、機械学習モデルは、機械学習分類子を含む。いくつかの実施形態では、機械学習モデルは、スタックされた機械学習モデル、1つ以上の機械学習モデル、アンサンブル機械学習モデル、又はそれらの任意の組み合わせを含む。いくつかの実施形態では、予測モデルは、1つ抜き検証で訓練される。いくつかの実施形態では、予測モデルは、がんの病期、がんの解剖学的起源、又はそれらの組み合わせを決定するように構成される。いくつかの実施形態では、がんの病期は、I期、II期、III期、又はIV期である。いくつかの実施形態では、実行可能命令は、1つ以上の核酸分子配列決定リードを浄化して、1つ以上の浄化された配列決定リードを作製することを更に含む。いくつかの実施形態では、浄化は、インシリコ浄化、実験的対照浄化、又はそれらの組み合わせを含む。いくつかの実施形態では、予測モデルは、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約91%、少なくとも約92%、少なくとも約93%、少なくとも約94%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の精度で疾患を決定する。いくつかの実施形態では、1つ以上の配列決定リードは、ショットガン配列決定、次世代配列決定、ロングリード配列決定、又はそれらの任意の組み合わせによって生成される。いくつかの実施形態では、実行可能命令は、1つ以上の核酸分子配列決定リードの1つ以上の特徴を決定することを更に含む。いくつかの実施形態では、1つ以上の核酸分子の1つ以上の特徴は、非微生物分類学的存在量、哺乳類ゲノム座標、注釈付きゲノム遺伝子座、哺乳類機能性遺伝子及び／又は生化学的経路存在量、又はそれらの任意の組み合わせの特徴、並びに当該1つ以上の特徴に関連する多数の配列決定リードを含む。いくつかの実施形態では、予測モデルは、対象のがん及び非がん性疾患を区別するように構成される。いくつかの実施形態では、マッピング又はアラインメントは、Deblur、PICRUST2、Bowtie2、Kraken、又はそれらの任意の組み合わせで完了される。

10

20

30

40

【0012】

参照による援用

本明細書に言及される全ての刊行物、特許、及び特許出願は、各個々の刊行物、特許、又は特許出願が、具体的かつ個別に、参照により組み込まれることが示されているかのように、参照により本明細書に組み込まれる。

【0013】

50

本発明の新規特徴は、添付の特許請求の範囲に詳細に述べられる。本開示の特徴及び利点のより良好な理解が、本開示の原理が使用される例示的な実施形態を記載する以下の詳細な説明、及び添付の図面を参照することによって得られるであろう。

【図面の簡単な説明】

【0014】

【図1】本明細書のいくつかの実施形態に記載の微生物核酸増幅及び／又は富化方法のフロー略図を示す。

【図2】本明細書のいくつかの実施形態に記載の微生物分類計算方法のフロー略図を示す。

【図3】本明細書のいくつかの実施形態に記載の微生物機能注釈付き計算方法のフロー略図を示す。

【図4】健康、がん、及び／又は非がん性の非健康な対象の核酸試料から1つ以上の微生物分類に基づく予測モデル分類子を生成する方法についてのフロー略図を示す。

【図5】健康、がん、及び／又は非がん性の非健康な対象の核酸試料から1つ以上の微生物機能注釈予測モデル分類子を生成する方法についてのフロー略図を示す。

【図6】本明細書で提供される本開示の方法を実行するように構成されたシステムを示す。

【図7A】本明細書のいくつかの実施形態に記載の系統発生的に多様な細菌の16S領域を増幅するために使用される16SリボソームRNA超可変領域及び対応する16Sプライマーを示す。

【図7B】本明細書のいくつかの実施形態に記載の系統発生的に多様な細菌の16S領域を増幅するために使用される16SリボソームRNA超可変領域及び対応する16Sプライマーを示す。

【図8】本明細書のいくつかの実施形態に記載の内部転写(ITS)領域を有する真菌リボソームRNA遺伝子クラスタの概略図を示す。

【図9】本明細書のいくつかの実施形態に記載のV6プライマー対及び当該V6プライマー対で増幅された微生物DNA標準組成物を用いた16SリボソームDNA増幅の実験データを示す。

【図10】ヒトゲノムDNAが存在する及び存在しない、V6プライマー対を用いた微生物16SリボソームDNA増幅の実験データを示す。

【図11】本明細書のいくつかの実施形態に記載のV6プライマー対を用いた16SリボソームDNA増幅の特異度の実験データを示す。

【図12A】本明細書のいくつかの実施形態に記載の16S配列決定ライブラリ調製のためのフロー略図を示す。

【図12B】本明細書のいくつかの実施形態に記載のウェスタン検証のためのフロー略図を示す。

【図13】本明細書のいくつかの実施形態に記載の16S配列決定処理についてのフロー略図を示す。

【図14A】本明細書のいくつかの実施形態に記載の16S配列決定処理を介した様々な点における実験データ配列決定リードカウントを示す。

【図14B】本明細書のいくつかの実施形態に記載の16S配列決定処理を介した様々な点における実験データ配列決定リードカウントを示す。

【図15A】本明細書の他の箇所に記載の1つ以上の対象由来の非小細胞肺癌試料と非がん核酸試料との間の区別における訓練された予測モデルの受信者操作特性曲線を示す。

【図15B】本明細書の他の箇所に記載の1つ以上の対象由来の非小細胞肺癌試料と非がん核酸試料との間の区別における訓練された予測モデルの受信者操作特性曲線を示す。

【発明を実施するための形態】

【0015】

本明細書で提供される開示は、1つ以上の対象及び／又は1つ以上の対象の群の健康状態を区別、分類、及び／又は診断し得る1つ以上の対象の1つ以上の核酸分子特徴を決定

10

20

30

40

50

、識別、分類、及び／又は生成する方法及びシステムを記載する。いくつかの場合では、1つ以上の核酸分子特徴は、対象及び／又は複数の対象の1つ以上の生体試料の1つ以上の核酸分子に由来し、それから得られ、受け取られ、かつ／又は決定され得る。いくつかの場合では、1つ以上の核酸分子は、1つ以上の哺乳類核酸分子、1つ以上の非哺乳類核酸分子、又はそれらの組み合わせを含み得る。いくつかの例では、1つ以上の非哺乳類核酸分子は、細菌、真菌、又はそれらの組み合わせからの1つ以上の核酸分子を含み得る。いくつかの場合では、本明細書の他の箇所に記載の1つ以上の対象の健康状態は、がん性の健康状態、非がん性疾患の健康状態、健康な健康状態、又はそれらの組み合わせを含み得る。いくつかの例では、がん性の健康状態は、がんを有する個体を含み得る。いくつかの場合では、がんは、肺、乳房、卵巣、胃腸、頭頸部、肝臓、脾臓、前立腺、皮膚、又はそれらの任意の組み合わせのがんを含み得る。いくつかの場合では、肺がんは、非小細胞肺がんを含み得る。いくつかの場合では、がん性の健康状態は、がんの病期（例えば、Ⅰ期、Ⅱ期、Ⅲ期など）の診断を含み得る。いくつかの場合では、健康状態は、対象又は複数の対象内のがん及び／又は疾患の空間的位置（すなわち、解剖学的位置）を含み得る。いくつかの場合では、生体試料は、液体生体試料、組織生体試料、又はそれらの組み合わせを含み得る。いくつかの場合では、非がん性疾患の健康状態は、肺疾患を含み得る。いくつかの例では、肺疾患は、カルチノイド、過誤腫、肉芽腫、間質性線維症、肺気腫、気管支炎、慢性閉塞性肺疾患、肺炎、サルコイドーシス、又はそれらの任意の組み合わせを含み得る。いくつかの場合では、液体生体試料は、液体生検を含み得る。いくつかの場合では、液体生検は、血漿、血清、全血、尿、脳脊髄液、唾液、汗、涙、呼気凝縮液、又はそれらの任意の組み合わせを含み得る。いくつかの場合では、組織生体試料は、対象の1つ以上の領域、臓器、及び／又は解剖学的位置（例えば、肺、皮膚、肝臓、脾臓、脳など）の組織生検を含み得る。

10

20

30

40

50

【0016】

いくつかの場合では、1つ以上の核酸分子の増幅、富化、フィルタリング、及び／若しくは浄化、並びに／又は1つ以上の核酸分子の1つ以上の配列決定リードは、対応する1つ以上の富化された、フィルタリングされた、かつ／又は浄化された核酸分子特徴が、少なくとも約80%、少なくとも約85%、少なくとも約86%、少なくとも約87%、少なくとも約88%、少なくとも約89%、少なくとも約90%、少なくとも約92%、少なくとも約94%、少なくとも約96%、少なくとも約98%、又は少なくとも約99%の精度で、1つ以上の対象の健康状態を決定、分類、特定、及び／又は診断する場合、予想より良い結果を提供し得る。

【0017】

いくつかの場合では、図1に見られるように、1つ以上の対象の生体試料は、1つ以上の微生物核酸分子組成物、1つ以上の哺乳類核酸分子組成物、又はそれらの組み合わせ、201を含み得る。いくつかの場合では、1つ以上の微生物核酸分子、1つ以上の哺乳類核酸分子組成物、又はそれらの組み合わせは、微生物核酸富化及び増幅ワークフロー202を介して富化され得る。いくつかの場合では、1つ以上の微生物核酸分子及び1つ以上の哺乳類核酸分子を含む生体試料は、ハイブリダイゼーションプローブ富化203及び／又はタンパク質に基づく微生物DNA富化204によって富化され得る。いくつかの場合では、ハイブリダイゼーションに基づく富化は、1つ以上の哺乳類核酸分子及び1つ以上の非哺乳類核酸分子をハイブリダイゼーションプローブと組み合わせることによって、ハイブリダイゼーションプローブが、非哺乳類ゲノム特徴に相補的な核酸配列を含み得る、組み合わせることと、ハイブリダイゼーションプローブ、1つ以上の哺乳類核酸分子、及び1つ以上の哺乳類上核酸分子を、1つ以上の非哺乳類核酸分子の標的核酸特徴とハイブリダイゼーションプローブとの間の核酸分子塩基対合を促進する条件下でインキュベーションすることと、非結合のハイブリダイゼーションプローブ及び1つ以上の非哺乳類核酸分子に結合したハイブリダイゼーションされたプローブを分離することと、1つ以上の非哺乳類核酸分子に結合したハイブリダイゼーションされたプローブを洗浄し、それによって、1つ以上の富化された非哺乳類核酸分子を生成することと、を含み得る。

【0018】

方法

いくつかの例では、本明細書で提供される本開示は、1つ以上の非哺乳類核酸分子（例えば、非哺乳類DNA）を富化する方法を記載する。いくつかの場合では、1つ以上の非哺乳類核酸分子を富化することは、タンパク質に基づく非哺乳類（例えば、微生物）核酸分子富化204によって富化され得る。いくつかの場合では、非哺乳類DNA富化は、1つ以上の哺乳類核酸分子及び1つ以上の非哺乳類核酸分子を1つ以上の組換えCXXCDメインタンパク質と組み合わせて、タンパク質-DNA結合反応を形成することと、組換えCXXCDメインタンパク質と、1つ以上の哺乳類核酸分子又は1つ以上の非哺乳類核酸分子の非メチル化CpGモチーフとの間の相互作用を促進する条件下で、タンパク質-DNA結合反応をインキュベーションすることと、非結合の組換えCXXCDメインタンパク質及び非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質をタンパク質-DNA結合反応の残りから分離することと、非メチル化CpG核酸断片に結合した組換えCXXCDメインタンパク質を洗浄し、それによって、増幅のための1つ以上の富化された核酸分子を生成することと、を含み得る。

10

【0019】

いくつかの場合では、増幅は、マーカー遺伝子増幅205を含み得る。いくつかの場合では、マーカー遺伝子は、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの例では、微生物系統発生マーカー遺伝子は、細菌、真菌、又はそれらの任意の組み合わせのマーカー遺伝子を含み得る。いくつかの場合では、微生物系統発生マーカー遺伝子は、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの場合では、微生物系統発生マーカー遺伝子は、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの場合では、細菌マーカー遺伝子は、リボソームRNA遺伝子5S、リボソームRNA遺伝子16S、リボソームRNA遺伝子23S、細菌ハウスキーピング遺伝子dnaG、frr、infC、nusA、pgk、pyrG、rplA、rplB、rplC、rplD、rplE、rplF、rplK、rplL、rplM、rplN、rplP、rplS、rplT、rpmA、rpoB、rpsB、rpsC、rpsE、rpsI、rpsJ、rpsK、rpsM、rpsS、smpB、tsf、又はそれらの任意の組み合わせを含み得る。いくつかの例では、図7Aに示されるような細菌リボソームRNA遺伝子は、微生物分類を区別及び／又は分類するために利用され得る超可変領域（V1～V9）を含み得る。いくつかの場合では、図7Bに示されるように、1つ以上のフォワードプライマー及び／又はリバースプライマーを使用して、細菌リボソームRNA遺伝子の16S領域を増幅し、対象及び／又は対象の群の健康状態を区別、決定、及び／又は診断するための特徴として使用され得る系統学的に多様な細菌のセットを区別し得る。

20

30

【0020】

いくつかの例では、真菌マーカー遺伝子は、リボソームRNA遺伝子18S、リボソームRNA遺伝子5.8S、リボソームRNA遺伝子28S、内部転写されたスペーサー領域1及び2、又はそれらの任意の組み合わせを含み得る。いくつかの場合では、図8に示されるように、内部転写されたスペーサー領域1及び2（それぞれ、ITS1及びITS2）は、小リボソームRNA（rRNA）と大リボソームRNA（rRNA）との間に位置し、18S rRNA、5.8S rRNA、及び28S RNAを提示する。いくつかの場合では、本明細書の他の箇所に記載のように、ITS1及び／又はITS2領域の増幅及び配列決定は、対象及び／又は対象の群の生体試料中の1つ以上の真菌の存在を検出及び／又は決定するためのゲノム特徴及び／又は標識を提供する。いくつかの例では、1つ以上の真菌は、対象及び／又は複数の対象の健康状態を区別、分類、及び／又は診断し得る分類学的特徴を提供し得る。いくつかの場合では、ITS1及び／又はITS2領域。いくつかの場合では、増幅は、ポリメラーゼ連鎖反応（PCR）又はその派生物を実施することによって達成及び／又は完了され得る。いくつかの場合では、ポリメラーゼ連鎖反応の派生物は、リバースプライマーPCR、逆PCR、アンカーPCR、プライマー

40

50

指向ローリングサークル増幅、又はそれらの任意の組み合わせを含み得る。いくつかの例では、ポリメラーゼ連鎖反応増幅は、1つ以上のゲノム特徴の増幅を防止するように構成された、ブロッキングプライマー、マーカー遺伝子プライマー、又はそれらの組み合わせを含み得る。いくつかの場合では、1つ以上のゲノム特徴は、ミトコンドリアDNAゲノム特徴を含み得る。

【0021】

いくつかの場合では、図12Aに示されるように、富化及び／又は増幅された1つ以上の核酸分子は、配列ライブラリ調製300を介して配列決定するために調製され得る。いくつかの場合では、方法は、増幅及び／又は富化された1つ以上の核酸を提供すること302と、バーコード化インデックス配列をカップリングすること304と、1つ以上のアダプタ配列をバーコード化インデックス配列にカップリングすること306と、を含み得る。いくつかの場合では、核酸分子のライブラリは、約256bpの塩基対の長さを含み得る。いくつかの場合では、増幅及び／又は富化された1つ以上の核酸分子は、約90bpの長さを含み得る。いくつかの例では、増幅された1つ以上の核酸分子は、V6プライマーで増幅された血漿生体試料の無細胞DNAを含み得る。図12Bに示される例示的な実験のゲル電気泳動の結果に示されるように、富化及び／又は増幅された核酸分子の様々な長さは、ライブラリ調製の前308からライブラリ調製の後310に増加する。

【0022】

哺乳類及び／又は非哺乳類核酸分子の1つ以上の富化及び／又は増幅された核酸分子組成物（複数可）の得られた調製されたライブラリは、次いで、標的化されたアンプリコン配列決定206方法によって配列決定され得、例えば、標的化された微生物アンプリコン配列決定は、それぞれ図2及び3に示すように、微生物分類特徴方法213及び／又は微生物機能特徴方法216において使用され得る。いくつかの場合では、標的化された微生物アンプリコン配列決定は、微生物16Sアンプリコン配列決定を含み得る。いくつかの場合では、1つ以上の富化及び／又は増幅された核酸分子組成物を配列決定することによって生成された1つ以上の配列決定リードは、図13に示されるように、前処理され得る。いくつかの例では、前処理は、fastpを介して富化及び／又は増幅された核酸組成物の1つ以上の配列決定リードを処理して、アダプタ配列を除去し、品質管理を実行して、1つ以上の処理された配列決定リードを生成すること312と、処理された1つ以上の配列決定リードから、垂操作（sub-operational）分類単位を生成して、品質を実施すること314と、垂操作分類単位をゲノムデータベースに対してクエリして、1つ以上の垂操作単位分類を割り当てること316と、を含み得る。いくつかの場合では、品質管理は、約30又は少なくとも約30の平均リード品質を含み得る。いくつかの場合では、ゲノムデータベースは、16S GreenGenes13.8を含み得る。いくつかの場合では、Qime2のsklearn分類子は、垂操作単位分類を割り当てるために使用され得る。いくつかの場合では、垂操作分類単位は、品質スコア、予想されるエラー率、各特有の配列の観察される頻度、又はそれらの組み合わせに基づいて配列のエラープロファイルをモデル化するノイズ除去ツールであるDeblurを使用して生成され得る。いくつかの場合では、標的化された微生物アンプリコン配列決定206は、ショットガン配列決定、次世代配列決定207、合成による配列決定、又はそれらの組み合わせを含み得る。いくつかの場合では、微生物分類特徴213及び／又は微生物機能特徴は、本明細書の他の箇所に記載のように、1つ以上の核酸分子特徴のセットの一部であり得る。いくつかの場合では、微生物分類未来方法213は、富化及び／又は増幅された核酸分子の1つ以上の微生物分類学的割り当て及び関連する微生物存在量を決定し得る。いくつかの例では、微生物機能特徴方法は、富化及び／又は増幅された核酸分子の1つ以上の微生物機能経路を決定し得る。

【0023】

いくつかの場合では、微生物機能特徴方法216（図3）は、例えば、次世代配列決定を使用して、富化及び／又は増幅された核酸分子ライブラリを配列決定して、配列決定リードのセットを生成すること207と、リードのセット配列決定から1つ以上の核酸分子

配列（例えば、ミトコンドリアDNA）をフィルタリングし208、それによって1つ以上のミトコンドリアDNA枯渇配列決定リードを生成すること209と、1つ以上のミトコンドリアDNA枯渇配列決定リードの1つ以上の微生物分類学的割り当てを特定すること210と、1つ以上の微生物分類学的割り当てを浄化すること211と、1つ以上の浄化された微生物分類学的配列決定リードの1つ以上の微生物機能特徴を注釈及び／又は特定すること214と、1つ以上の特定された及び／又は注釈された微生物機能特徴の特徴セット215を出力することと、を含み得る。いくつかの場合では、1つ以上の微生物機能特徴215は、本明細書の他の箇所に記載の図5に示されるように、予測モデル（例えば、機械学習分類子）を訓練するために、対象の既知の健康状態（217、218、219）と組み合わせて使用され得る。いくつかの場合では、微生物機能特徴は、メタゲノム機能特徴を含み得る。いくつかの場合では、PICRUST2は、1つ以上の浄化された微生物分類学的配列決定リードの1つ以上のメタゲノム機能特徴を決定及び／又は特定し得る。いくつかの場合では、微生物分類ワークフローは、ミトコンドリアDNA枯渇配列決定リードから微生物分類学的割り当てを決定するためのマッピングを含み得る210。いくつかの場合では、浄化は、インシリコ浄化を含み得る。いくつかの例では、浄化は、1つ以上の非内因性微生物配列決定リードを除去し得、それによって、ミトコンドリアDNA枯渇非哺乳類配列決定リードの1つ以上の微生物分類学的同一性から、1つ以上の浄化された微生物分類学的割り当て及び関連する配列決定リードの量を生成する。

10

【0024】

いくつかの場合では、微生物機能方法213は、例えば、次世代配列決定を使用して、富化及び／又は増幅された核酸分子ライブラリを配列決定して、配列決定リードのセットを生成すること207と、リードのセット配列決定から1つ以上の核酸分子配列（例えば、ミトコンドリアDNA）をフィルタリングし208、それによって1つ以上のミトコンドリアDNA枯渇配列決定リードを生成すること209と、1つ以上のミトコンドリアDNA枯渇配列決定リードの1つ以上の微生物分類学的割り当てを特定すること210と、1つ以上の微生物分類学的割り当てを浄化すること211と、富化及び／又は増幅された核酸分子ライブラリの1つ以上の浄化された微生物分類特徴212を出力することと、を含み得る。いくつかの場合では、1つ以上の微生物分類特徴は、本明細書の他の箇所に記載の図5に示されるように、予測モデル（例えば、機械学習分類子）を訓練するために、対象の既知の健康状態（217、218、219）と組み合わせて使用され得る。いくつかの場合では、微生物分類ワークフローは、ミトコンドリアDNA枯渇配列決定リードから微生物分類学的割り当てを決定するためのマッピングを含み得る210。いくつかの場合では、マッピングは、1つ以上の微生物参照データベースに対して1つ以上のミトコンドリアDNA枯渇核酸分子配列決定リードをマッピングして、ミトコンドリアDNA枯渇非哺乳類配列決定リードの微生物分類学的同一性を決定することを含み得る。いくつかの場合では、マッピングは、QIIME2又はその他のサポートされたバージョンによって実施され得る。いくつかの場合では、1つ以上の微生物参照データベースは、細菌16S rRNAのデータベースGreenGenes、細菌、真菌、及び古細菌rRNAのデータベースSILVA、真核核リボソームITS領域のデータベースUNITE、公的に入手可能で完全な微生物ゲノム配列に由来するカスタムデータベース、又はそれらの任意の組み合わせを含み得る。いくつかの場合では、浄化は、インシリコ浄化を含み得る。いくつかの例では、浄化は、1つ以上の非内因性微生物配列決定リードを除去し得、それによって、ミトコンドリアDNA枯渇非哺乳類配列決定リードの1つ以上の微生物分類学的同一性から、1つ以上の浄化された微生物分類学的割り当て及び関連する配列決定リードの量を生成する。

20

30

40

【0025】

いくつかの場合では、1つ以上の核酸分子特徴は、1つ以上の核酸分子ゲノム特徴を含み得る。いくつかの場合では、ゲノム特徴は、微生物系統発生マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの例では、微生物系統発生マーカー遺伝子は、細菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの例では、微生物系

50

統発生マーカー遺伝子は、真菌マーカー遺伝子又はそのマーカー遺伝子断片を含み得る。いくつかの場合では、1つ以上の核酸分子特徴は、本明細書の他の箇所に記載の1つ以上の非哺乳類核酸分子（例えば、微生物核酸分子）の特徴、特徴のセット及び／又は特徴群を含み得る。

【0026】

いくつかの場合では、1つ以上の核酸分子特徴は、本明細書の他の箇所に記載の、図4及び5に示されるように、1つ以上の予測モデル221（例えば、機械学習分類子）を訓練220するために使用され得る。いくつかの場合では、1つ以上の核酸分子特徴（213、216）は、本明細書の他の箇所に記載の対象又は複数の対象の生体試料の1つ以上の核酸分子の微生物、細菌、真菌、又はそれらの組み合わせの分類及び／又は機能分類及び／又は特徴付けを含み得る。

10

【0027】

いくつかの場合では、予測モデル（例えば、機械学習分類子）は、健康な217、非がん性疾患219、又はがん性218の既知の健康状態を有する対象の生体試料の1つ以上の核酸分子の1つ以上の核酸分子特徴で訓練され得る。いくつかの場合では、予測モデルは、図4に示されるように、1つ以上の微生物分類学的特徴213及び1つ以上の対象の関連する健康状態で訓練220され得る。いくつかの場合では、予測モデルは、図5に示されるように、1つ以上の微生物機能特徴216及び1つ以上の対象の関連する健康状態で訓練され得る。いくつかの場合では、訓練された予測モデル221は、予測モデルの訓練に含まれなかった1つ以上の対象の健康状態を区別、分類、及び／又は診断し得る1つ以上の分類子（222、223、224）を含み得る。いくつかの場合では、1つ以上の分類子は、健康対がんの健康状態分類子222、がん性対非がん性疾患の健康分類子223、非がん性疾患対健康の分類子、又はそれらの任意の組み合わせを含み得る。

20

【0028】

予測モデル

本開示の方法及びシステムは、対象、複数の対象、及び／又は1つ以上の対象の群の健康状態を分類し、診断し、かつ／又は特徴付け得る1つ以上の核酸分子特徴に対して訓練された人工知能、予測モデル、及び／又は機械学習の外部能力を利用又はアクセスし得る。いくつかの場合では、本明細書の他の箇所に記載の1つ以上の核酸分子特徴（例えば、微生物機能特徴、微生物分類学的特徴など）は、1つ以上の対象のがん及び／又は非がん性疾患を予測、分類、及び／又は特定し得る。いくつかの場合では、1つ以上の核酸分子の特徴は、本明細書の他の箇所に記載の1つ以上の予測モデルを訓練するために使用され得る。これらの特徴は、対象、複数の対象、及び／又は1つ以上の対象の群の健康状態、例えば、がん、非がん性疾患、障害、又はそれらの任意の組み合わせを正確に予測し、分類し、かつ／又は特徴付け得る。そのような予測能力を使用して、医療提供者（例えば、医師）は、情報に基づいた正確なリスクに基づいた決定を下すことができ、それによって、がん、非がん性疾患、障害、又はそれらの任意の組み合わせを有する対象に提供されるケア及び監視の質を改善し得る。

30

【0029】

本開示の方法及びシステムは、微生物の存在及び／又は存在量（例えば、特定の属の微生物の存在量、分類、微生物機能経路）を分析し得る。次いで、微生物の存在及び／又は存在量を使用して、1つ以上の対象のがん及び／又は非がん性疾患を予測し得る1つ以上の核酸分子特徴、例えば、非哺乳類核酸分子特徴を決定し得る。いくつかの場合では、本明細書の他の箇所に記載の方法及びシステムは、対象の健康状態、例えば、がん及び／又は非がん性疾患を示す1つ以上の核酸分子特徴を用いて予測モデルを訓練し得る。いくつかの場合では、訓練された予測モデルは、次いで、予測モデルを訓練するために利用された1つ以上の対象とは異なる1つ以上の対象のがん及び／又は非がん性疾患の可能性（例えば、予測）を生成するために使用され得る。訓練された予測モデルは、1つ以上の核酸分子並びに／又は富化、フィルタリング、及び／若しくは増幅された1つ以上の核酸分子から1つ以上の核酸分子特徴を処理して、対象（複数可）ががん、非がん性疾患、又は障

40

50

害を有する可能性を生成するように構成された、機械学習に基づく分類子などの人工知能に基づくモデルを含み得る。モデルは、対象、例えば、がん対象、非がん性疾患を有する対象、疾患を有しない及びがんを有しない対象、がんに対する治療を受けているがん対象、非がん性疾患に対する治療を受ける対象、又はそれらの任意の組み合わせの1つ以上のコホートからの微生物分類学的特徴又は微生物機能経路の存在量を使用して訓練され得る。いくつかの場合では、予測モデルは、予測モデルの訓練データセットの一部ではない1つ以上の対象のがんを治療するための治療予測を提供するように訓練され得る。そのような予測モデルは、患者のハイブリダイゼーション富化された生体試料の1つ以上の微生物の存在及び存在量の入力提供された場合、訓練データセットの一部ではない1つ以上の対象のための治療推奨を出力し得る。

10

【0030】

予測モデルは、1つ以上の予測モデルを含み得る。モデルは、1つ以上の機械学習アルゴリズムを含んでもよい。機械学習アルゴリズムの例としては、サポートベクター機械（SVM）、ナイーブベイズ分類、ランダムフォレスト、ニューラルネットワーク（ディープニューラルネットワーク（DNN）など）、回帰型ニューラルネットワーク（RNN）、ディープRNN、長・短期記憶（LSTM）回帰型ニューラルネットワーク（RNN）、ゲート付き回帰型ユニット（GRU）、勾配ブースティング機械、ランダムフォレスト、又は他の教師あり学習アルゴリズム若しくは教師なし機械学習、統計、線形回帰、k近傍、k平均、決定木、ロジスティック回帰、又はそれらの任意の組み合わせが挙げられ得る。モデルは、分類又は回帰のために使用され得る。モデルは、同様に、複数の予測モデルからなるアンサンブルモデルの推定を伴ってもよく、例えば、勾配ブースティング決定木の構築において、勾配ブースティングなどの技術を利用してもよい。モデルは、1つ以上の核酸分子特徴、対象データ、例えば、対象の医療歴、対象の家族の医療歴、対象のバイタル（例えば、血圧、脈拍、体温、酸素飽和度）、又はそれらの任意の組み合わせを含む1つ以上の訓練データセットを使用して訓練され得る。

20

【0031】

予測モデルは、任意の数の機械学習アルゴリズムを含んでもよい。いくつかの実施形態では、ランダムフォレスト機械学習アルゴリズムは、バギングされた決定木のアンサンブルであってもよい。アンサンブルは、少なくとも、約1、2、3、4、5、10、20、30、40、50、60、70、80、90、100、120、140、160、180、200、250、500、1000又はそれ以上のバギングされた決定木であり得る。アンサンブルは、最大で、約1000、500、250、200、180、160、140、120、100、90、80、70、60、50、40、30、20、10、5、4、3、2又はそれ以下のバギングされた決定木であり得る。アンサンブルは、約1～1000、1～500、1～200、1～100、又は1～10のバギングされた決定木であり得る。

30

【0032】

いくつかの実施形態では、機械学習アルゴリズムは、様々なパラメータを有してもよい。様々なパラメータは、例えば、学習率、ミニバッチサイズ、訓練するエポック数、モーメンタム、学習重み減衰、又はニューラルネットワーク層などであり得る。

40

【0033】

いくつかの実施形態では、学習率は、約0.00001～0.1であり得る。

【0034】

いくつかの実施形態では、ミニバッチサイズは、約16～128であり得る。

【0035】

いくつかの実施形態では、ニューラルネットワークは、ニューラルネットワーク層を含んでもよい。ニューラルネットワークは、少なくとも約2～1000以上のニューラルネットワーク層を有し得る。

【0036】

いくつかの実施形態では、訓練するエポック数は、少なくとも、約1、2、3、4、5

50

、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95、100、150、200、250、500、1000、10000、又はそれ以上であり得る。

【0037】

いくつかの実施形態では、モーメントは、少なくとも、約0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9又はそれ以上であり得る。いくつかの実施形態では、モーメントは、最大で、約0.9、0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1、又はそれ以下であり得る。

【0038】

いくつかの実施形態では、学習重み減衰は、少なくとも、約0.00001、0.0001、0.001、0.002、0.003、0.004、0.005、0.006、0.007、0.008、0.009、0.01、0.02、0.03、0.04、0.05、0.06、0.07、0.08、0.09、0.1、又はそれ以上であり得る。いくつかの実施形態では、学習重み減衰は、最大で、約0.1、0.09、0.08、0.07、0.06、0.05、0.04、0.03、0.02、0.01、0.009、0.008、0.007、0.006、0.005、0.004、0.003、0.002、0.001、0.0001、0.00001、又はそれ以下であり得る。

【0039】

いくつかの実施形態では、機械学習アルゴリズムは、損失関数を使用してもよい。損失関数は、例えば、回帰損失、平均絶対誤差、平均偏り誤差、ヒンジ損失、Adamオプティマイザ、及び／又は交差エントロピーであり得る。

【0040】

いくつかの実施形態では、機械学習アルゴリズムのパラメータは、人間及び／又はコンピュータシステムの助けを借りて調整され得る。

【0041】

いくつかの実施形態では、機械学習アルゴリズムは、特定の特徴を優先してもよい。機械学習アルゴリズムは、がん、非がん性疾患、障害、又はそれらの任意の組み合わせを検出することにより関連し得る特徴を優先してもよい。特徴が、がん、非がん性疾患、及び／又は障害を判定することにおいて別の特徴よりも頻繁に分類される場合、その特徴は、がん、非がん性疾患、及び／又は障害を検出することにより関連し得る。いくつかの場合では、特徴は、重み付けシステムを使用して優先され得る。いくつかの場合では、特徴は、その特徴の出現の頻度及び／又は量に基づいて確率統計上で優先され得る。機械学習アルゴリズムは、人間及び／又はコンピュータシステムの助けを借りて特徴を優先してもよい。

【0042】

いくつかの場合では、機械学習アルゴリズムは、計算コストの低減、処理力の節約、処理時間の節約、信頼性の向上、又はランダムアクセスメモリ使用量の減少などのために、特定の特徴を優先してもよい。

【0043】

訓練データセットは、例えば、一般的ながん、非がん性疾患、又は障害の診断を有する対象の1つ以上のコホートから生成され得る。訓練データセットは、1つ以上の対象の生体試料中に存在する微生物の存在量分類学的割り当て特徴及び／又は生体試料中に存在する微生物の微生物機能経路特徴の形態で、1つ以上の核酸分子特徴を含み得る。特徴は、微生物特徴に対応する1つ以上の対象のがん診断を含み得る。いくつかの場合では、特徴は、患者の年齢、患者の病歴、他の病状、現在又は過去の投薬治療、臨床リスクスコア、及び最後の観察からの時間などの患者情報を含み得る。例えば、所与の時点で所与の患者から収集された特徴セットは、その所与の時点での患者の健康状態又はステータスを示し得るシグネチャとして集合的に機能し得る。

【0044】

標識は、例えば、対象（例えば、患者）における、がん、非がん性疾患、障害、又はそれらの組み合わせの存在、非存在、診断、及び／又は予後などの臨床アウトカムを含んでもよい。臨床アウトカムは、治療有効性（例えば、対象が、がん及び／又は疾患ベースの治療に対する陽性応答者であるか、又は陰性応答者であるか）を含んでもよい。

【0045】

入力特徴は、データをビンに集約することによって、又は代替的にワンホットエンコーディングを使用することによって構造化され得る。入力または、相互相関などの前述の入力から導出された特徴値又はベクトルも含み得る。

【0046】

訓練データセットは、がん、非がん性疾患、障害、又はそれらの任意の組み合わせを示す生体試料の富化及び／又は増幅された核酸分子から特定及び／又は分類された、例えば、1つ以上の微生物分類学的特徴、1つ以上の微生物機能経路、又はそれらの組み合わせの1つ以上の核酸モル特徴の存在及び／又は存在量から構築され得る。

【0047】

モデルは、入力特徴を処理して、1つ以上の分類、1つ以上の予測、又はそれらの組み合わせを含む出力値を生成し得る。例えば、そのような分類又は予測は、がんの存在若しくはがんの非存在のバイナリ分類、非がん性疾患の存在、障害の存在、又は対象のそれらの分類の任意の組み合わせを含んでもよい。いくつかの場合では、1つ以上の予測モデル及び／又は機械学習アルゴリズムは、カテゴリ標識の群（例えば、「がん、非がん疾患及び／又は障害なし」、「明らかながん、非がん疾患及び／又は障害」、並びに「起こり得るがん、非がん疾患及び／又は障害」と、特定のがん、非がん性疾患、及び／又は障害を発症する可能性（例えば、相対的な可能性又は確率）と、がん、非がん疾患及び／又は障害の存在、患者の死亡の可能性についての「リスク因子」、並びに任意の数値予測についての信頼区間を示すスコアとの間で、対象を分類し得る。様々な機械学習技術は、機械学習技術の出力がモデルの後続の層又はサブセクションへの入力特徴としても使用され得るようにカスケードされてもよい。

【0048】

リアルタイムの分類又は予測を生成するためのモデルを（例えば、モデルの重み及び相関を判定することによって）訓練するために、本明細書の他の箇所に記載の訓練データセット及び／又は1つ以上の訓練特徴を使用して、モデルを訓練することができる。そのようなデータセット及び／又は特徴は、統計的に有意な分類又は予測を生成するのに十分な大きさであり得る。例えば、データセットは、1つ以上の対象の生体試料における、真菌、ウイルス、古細菌、細菌、若しくはそれらの任意の組み合わせの微生物の存在及び／又は存在量からの配列決定データに由来する1つ以上の核酸分子特徴を含み得る。

【0049】

データセットは、訓練データセット、開発データセット、及び試験データセットなどのサブセット（例えば、個別又は重複する）に分割され得る。例えば、データセットは、データセットの80%を含む訓練データセット、データセットの10%を含む開発データセット、及びデータセットの10%を含む試験データセットに分割され得る。訓練データセットは、データセットの約10%、約20%、約30%、約40%、約50%、約60%、約70%、約80%、又は約90%を含み得る。開発データセットは、データセットの約10%、約20%、約30%、約40%、約50%、約60%、約70%、約80%、又は約90%を含み得る。試験データセットは、データセットの約10%、約20%、約30%、約40%、約50%、約60%、約70%、約80%、又は約90%を含み得る。いくつかの実施形態では、1つ抜き交差検証が用いられ得る。訓練セット（例えば、訓練データセット）は、試料採取の独立性を確保するために、1つ以上の患者コホートに対応するデータのセットのランダムな試料採取によって選択され得る。代替的に、訓練セット（例えば、訓練データセット）は、試料採取の独立性を確保するために、1つ以上の患者コホートに対応するデータのセットの比例試料採取によって選択され得る。

【0050】

モデル予測の精度を改善し、モデルの過剰適合を低減するために、データセットを拡張して、訓練セット内の試料数を増加させてもよい。例えば、データ拡張は、訓練記録における観測の順序を再配置することを含み得る。欠落した観測を有するデータセットに対応するために、フォワードフィリング、バックフィリング、線形補間、及びマルチタスクガウスプロセスなどの欠落したデータを補完する方法が使用され得る。データセットを、フィルタリング又はバッチ補正して、交絡因子を除去又は軽減し得る。例えば、データベース内では、対象のサブセットが、除外され得る。

【0051】

モデルは、ニューラルネットワーク、畳み込みニューラルネットワーク（CNN）、ディープニューラルネットワーク（DNN）、回帰型ニューラルネットワーク（RNN）、又はディープRNNなどの1つ以上のニューラルネットワークを含んでもよい。回帰型ニューラルネットワークは、長短期記憶（LSTM）ユニット又はゲート付き回帰型ユニット（GRU）であり得るユニットを含み得る。例えば、モデルは、本明細書の他の箇所に記載のように、入力特徴セット、例えば、1つ以上の核酸分子特徴、バイタル測定、対象の医療歴、対象の人口統計、又はそれらの任意の組み合わせを有するニューラルネットワークを含むアルゴリズムアーキテクチャを含み得る。ドロップアウト又は正則化などのニューラルネットワーク手法は、過剰適合を防止するためにモデルの訓練中に使用され得る。ニューラルネットワークは、複数のサブネットワークを含んでもよく、その各々は、異なるタイプの出力情報の分類又は予測を生成するように構成され、これは、ニューラルネットワークの全体的な出力を形成するように組み合わせられ得る。機械学習モデルは、代替的に、ランダムフォレスト、分類及び回帰木、サポートベクター機械、判別分析、回帰技術、並びにそれらのアンサンブル及び勾配ブーストされたバリエーションを含む統計的又は関連アルゴリズムを利用し得る。

【0052】

モデルが、がん、非がん性疾患、障害、又はそれらの組み合わせの分類又は予測を生成する場合、通知（例えば、アラート又はアラーム）が生成され、病院内の対象の治療チームの医師、看護師、又は他のメンバーなどの医療提供者に伝送されてもよい。通知は、自動化電話、ショートメッセージサービス（SMS）、マルチメディアメッセージサービス（MMS）メッセージ、電子メール、及び／又はダッシュボード内のアラートを介して伝送され得る。通知は、がん、非がん性疾患、及び／若しくは障害の予測；予測されるがん、非がん性疾患、及び／若しくは障害の可能性；がん、非がん性疾患、及び／若しくは障害の予想される発症までの時間；可能性若しくは時間の信頼区間、がん、非がん性疾患、及び／若しくは障害に対する推奨される治療経過、又はそれらの情報の任意の組み合わせなどの出力情報を含み得る。

【0053】

モデルの性能を検証するために、異なる性能メトリックが生成されてもよい。例えば、受信者操作特性曲線の下面積（AUROC）を使用して、モデルの診断、予後、スクリーニング、又はそれらの能力の任意の組み合わせを判定し得る。例えば、モデルは、特異度及び感度が調節可能であるように、調整可能な分類閾値を使用してもよく、受信者操作特性曲線（ROC）を使用して、特異度及び感度の異なる値に対応する異なる操作点を特定することができる。

【0054】

データセットが十分に大きくない場合など、いくつかの場合では、相互検証を実施して、異なる訓練及び試験データセットにわたるモデルの堅牢性を評価し得る。

【0055】

感度、特異度、精度、陽性適中率（PPV）、陰性適中率（NPV）、適合率再現率曲線下面積（AUPR）、AUROC、又は類似物などの性能メトリックを計算するために、以下の定義が使用され得る。「偽陽性」は、陽性のアウトカム又は結果が誤って又は早期に生成された（例えば、がん、非がん性疾患及び／又は障害の実際の発症前、又は発症なし）アウトカムを指し得る。「真陽性」は、患者ががん、非がん性疾患及び／又は障害

を有する（例えば、患者ががん、非がん性疾患及び／又は障害の症状を示すか、又は患者の記録ががん、非がん性疾患及び／又は障害を示す）場合に、陽性のアウトカム又は結果が正しく生成されたアウトカムを指し得る。「偽陰性」は、陰性のアウトカム又は結果が生成されたが、患者ががん、非がん性疾患及び／又は障害を有する（例えば、患者ががん、非がん性疾患及び／又は障害の症状を示すか、又は患者の記録ががん、非がん性疾患及び／又は障害を示す）アウトカムを指し得る。「真陰性」は、陰性のアウトカム又は結果が生成された（例えば、がん、非がん性疾患及び／又は障害の実際の発症前、又は発症なし）アウトカムを指し得る。

【0056】

モデルは、診断精度尺度に対応する最小所望値を有するなど、精度又は性能のための特定の所定条件が満たされるまで訓練されてもよい。例えば、診断精度尺度は、対象におけるがん、非がん性疾患及び／又は障害の発生の可能性の予測に対応し得る。別の例として、診断精度尺度は、対象が以前に治療されたがん、非がん性疾患及び／又は障害の悪化又は再発の可能性の予測に対応し得る。診断精度尺度の例としては、がん、非がん性疾患及び／又は障害を検出又は予測する診断精度に対応する感度、特異度、陽性適中率（PPV）、陰性適中率（NPV）、精度、AUPR、及びAUROCが挙げられ得る。

【0057】

例えば、そのような所定条件は、がん、非がん性疾患及び／又は障害を予測する感度が、例えば、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の値を含むことであり得る。

【0058】

別の例として、そのような所定条件は、がん、非がん性疾患及び／又は障害を予測する特異度が、例えば、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の値を含むことであり得る。

【0059】

別の例として、そのような所定条件は、がん、非がん性疾患及び／又は障害を予測する陽性適中率（PPV）が、例えば、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の値を含むことであり得る。

【0060】

別の例として、そのような所定条件は、がん、非がん性疾患及び／又は障害を予測する陰性適中率（NPV）が、例えば、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の値を含むことであり得る。

【0061】

別の例として、そのような所定条件は、がん、非がん性疾患及び／又は障害を予測する受信者操作特性曲線（ROC）の曲線下面積（AUC）（AUROC）が、少なくとも約0.50、少なくとも約0.55、少なくとも約0.60、少なくとも約0.65、少なくとも約0.70、少なくとも約0.75、少なくとも約0.80、少なくとも約0.85、少なくとも約0.90、少なくとも約0.95、少なくとも約0.96、少なくとも約0.97、少なくとも約0.98、又は少なくとも約0.99の値を含むことであり得る。

る。

【0062】

別の例として、そのような所定条件は、がん、非がん性疾患及び／又は障害を予測する適合率－再現率の曲線の下面積（AUPR）が、少なくとも約0.10、少なくとも約0.15、少なくとも約0.20、少なくとも約0.25、少なくとも約0.30、少なくとも約0.35、少なくとも約0.40、少なくとも約0.45、少なくとも約0.50、少なくとも約0.55、少なくとも約0.60、少なくとも約0.65、少なくとも約0.70、少なくとも約0.75、少なくとも約0.80、少なくとも約0.85、少なくとも約0.90、少なくとも約0.95、少なくとも約0.96、少なくとも約0.97、少なくとも約0.98、又は少なくとも約0.99の値を含むことであり得る。

10

【0063】

いくつかの実施形態では、訓練されたモデルは、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の感度で、がん、非がん性疾患及び／又は障害を予測するように訓練又は構成され得る。

【0064】

いくつかの実施形態では、訓練されたモデルは、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の特異度で、がん、非がん性疾患及び／又は障害を予測するように訓練又は構成され得る。

20

【0065】

いくつかの実施形態では、訓練されたモデルは、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の陽性適中率（PPV）で、がん、非がん性疾患及び／又は障害を予測するように訓練又は構成され得る。

30

【0066】

いくつかの実施形態では、訓練されたモデルは、少なくとも約50%、少なくとも約55%、少なくとも約60%、少なくとも約65%、少なくとも約70%、少なくとも約75%、少なくとも約80%、少なくとも約85%、少なくとも約90%、少なくとも約95%、少なくとも約96%、少なくとも約97%、少なくとも約98%、又は少なくとも約99%の陰性適中率（NPV）で、がん、非がん性疾患及び／又は障害を予測するように訓練又は構成され得る。

【0067】

いくつかの実施形態では、訓練されたモデルは、少なくとも約0.50、少なくとも約0.55、少なくとも約0.60、少なくとも約0.65、少なくとも約0.70、少なくとも約0.75、少なくとも約0.80、少なくとも約0.85、少なくとも約0.90、少なくとも約0.95、少なくとも約0.96、少なくとも約0.97、少なくとも約0.98、又は少なくとも約0.99の受信者操作特性曲線（ROC）の曲線下面積（AUC）（AUROC）で、がん、非がん性疾患及び／又は障害を予測するように訓練又は構成され得る。

40

【0068】

いくつかの実施形態では、訓練されたモデルは、少なくとも約0.10、少なくとも約0.15、少なくとも約0.20、少なくとも約0.25、少なくとも約0.30、少なくとも約0.35、少なくとも約0.40、少なくとも約0.45、少なくとも約0.5

50

0、少なくとも約0.55、少なくとも約0.60、少なくとも約0.65、少なくとも約0.70、少なくとも約0.75、少なくとも約0.80、少なくとも約0.85、少なくとも約0.90、少なくとも約0.95、少なくとも約0.96、少なくとも約0.97、少なくとも約0.98、又は少なくとも約0.99の適合率－再現率の曲線の下での面積（AUPR）で、がん、非がん性疾患及び／又は障害を予測するように訓練又は構成され得る。

【0069】

訓練データセットは、訓練対象（例えば、ヒト）から収集されてもよい。各訓練は、それらが生物学的状態と診断されたか、又はがん、非がん性疾患及び／又は障害と診断されていないことを示す診断ステータスを有する。

10

【0070】

いくつかの実施形態では、モデルは、ニューラルネットワーク又は畳み込みニューラルネットワークである。Vincent et al., 2010, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” J Mach Learn Res 11, pp. 3371–3408、Larochelle et al., 2009, “Exploring strategies for training deep neural networks,” J Mach Learn Res 10, pp. 1–40、及びHassoun, 1995, Fundamentals of Artificial Neural Networks, Massachusetts Institute of Technologyを参照されたい。それらの各々は参照により本明細書に組み込まれる。

20

【0071】

いくつかの実施形態では、Lee, T.-W. (1998): Independent component analysis: Theory and applications, Boston, Mass: Kluwer Academic Publishers, ISBN 0-7923-8261-7、及びHyvaerinen, A.; Karhunen, J.; Oja, E. (2001): Independent Component Analysis, New York: Wiley, ISBN 978-0-471-40540-5において記載されるものなど、独立成分分析（ICA）を使用して、データを脱次元化（de-dimensionalize）し、これらは、参照によりその全体が本明細書に組み込まれる。

30

【0072】

いくつかの実施形態では、Jolliffe, I. T. (2002). Principal Component Analysis. Springer Series in Statistics. New York: Springer-Verlag. doi: 10.1007/b98835. ISBN 978-0-387-95442-4に記載されているものなど、主成分分析（PCA）を使用して、データを脱次元化し、これらは、参照によりその全体が本明細書に組み込まれる。

40

【0073】

SVMは、Cristianini and Shawe-Taylor, 2000, “An Introduction to Support Vector Machines,” Cambridge University Press, Cambridge、Boser et al., 1992, “A training algorithm for optimal margin classifiers,” in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, Pittsburgh, Pa., pp. 142–152、Vapnik, 1998, Statistical Learning Theory, Wiley, New York、

50

Mount, 2001, Bioinformatics: sequence and genome analysis, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y.、Duda, Pattern Classification, Second Edition, 2001, John Wiley & Sons, Inc., pp. 259, 262-265、及び Hastie, 2001, The Elements of Statistical Learning, Springer, New York、並びに Furey et al., 2000, Bioinformatics 16, 906-914 に記載されており、これらの各々は、参照によりその全体が本明細書に組み込まれる。分類に使用される場合、SVM は、標識されたデータから、最大限に離れたハイパープレーンを使用して、バイナリ標識されたデータの所与のセットを分離する。線形分離が不可能な場合、SVM は、特徴空間への非線形マッピングを自動的に実現する「カーネル」の技術と組み合わせて機能することができる。特徴空間内の SVM によって見出される超平面は、入力空間内の非線形決定境界に対応する。

10

【0074】

決定木は、参照により本明細書に組み込まれる Duda, 2001, Pattern Classification, John Wiley & Sons, Inc., New York, pp. 395-396 によって概説されている。決定木に基づく方法は、特徴空間を長方形のセットに分割し、各々に（定数のような）モデルを適合させる。いくつかの実施形態では、決定木はランダムフォレスト回帰である。使用され得る1つの特定のアルゴリズムは、分類及び回帰木（CART）である。他の特定の決定木アルゴリズムは、ID3、C4.5、MART、及びランダムフォレストを含むが、これらに限定されない。CART、ID3、及びC4.5は、Duda, 2001, Pattern Classification, John Wiley & Sons, Inc., New York, pp. 396-408 及び pp. 411-412 に記載されており、これは参照により本明細書に組み込まれる。CART、MART、及びC4.5は、Hastie et al., 2001, The Elements of Statistical Learning, Springer-Verlag, New York, Chapter 9 に記載されており、これは、参照によりその全体が本明細書に組み込まれる。ランダムフォレストは、Breiman, 1999, "Random Forests-Random Features," Technical Report 567, Statistics Department, U. C. Berkeley, September 1999 に記載されており、これは参照によりその全体が本明細書に組み込まれる。

20

30

【0075】

クラスタリング（例えば、教師なしクラスタリングモデルアルゴリズム及び教師ありクラスタリングモデルアルゴリズム）は、Duda and Hart, Pattern Classification and Scene Analysis, 1973, John Wiley & Sons, Inc., New York（以降、「Duda 1973」）の211～256ページに記載されており、これは、参照によりその全体が本明細書に組み込まれる。Duda 1973のセクション6.7に記載されるように、クラスタリング問題は、データセットにおいて自然な群化を見出すことの1つとして説明される。自然な群化を特定するために、2つの問題に対処する。第一に、2つの試料間の類似性（又は相違性）を測定する方法を決定する。このメトリック（類似性尺度）を使用して、一方のクラスタ内の試料が他方のクラスタ内の試料よりも互いにより類似していることを確実にする。第二に、類似性尺度を使用してデータをクラスタに分割するための機構を決定する。類似性尺度は、Duda 1973のセクション6.7で考察されており、クラスタリング調査を開始する1つの方法は、距離関数を定義し、訓練セット内の試料の全ての対間の距離の行列を計算することであると記載されている。距離が類似性の良好な尺度である場合、同じクラスタ内の参照実体間の距離は、異なるクラスタ内の参照実体間の距離よりも顕著に小さいであろう。しかしながら、Duda 1973の215ページに

40

50

記載されているように、クラスタリングは距離メトリックの使用を必要としない。例えば、非メトリック類似性関数 $s(x, x')$ を使用して、2つのベクトル x 及び x' を比較し得る。従来、 $s(x, x')$ は、 x 及び x' が何らかの形で「類似」している場合に値が大きい対称関数である。非メトリック類似性関数 $s(x, x')$ の例は、Duda 1973の218ページに示されている。データセット内の点間の「類似性」又は「相違性」を測定するための方法が選択されると、クラスタリングは、データの任意の分割のクラスタリング品質を測定する基準関数を必要とする。基準関数を極端化するデータセットの分割を使用して、データをクラスタリングする。Duda 1973の217ページを参照されたい。基準関数は、Duda 1973のセクション6.8で考察されている。より最近では、Duda et al., Pattern Classification, 2nd edition, John Wiley & Sons, Inc. New Yorkが出版されている。537～563ページは、クラスタリングを詳細に記載している。クラスタリング技術の詳細は、Kaufman and Rousseeuw, 1990, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, N. Y., Everitt, 1993, Cluster analysis (3d ed.), Wiley, New York, N. Y., 及び Backer, 1995, Computer-Assisted Reasoning in Cluster Analysis, Prentice Hall, Upper Saddle River, New Jerseyに見出され得、これらの各々は、参照により本明細書に組み込まれる。本開示において使用され得る特定の例示的なクラスタリング技術としては、階層的クラスタリング（最近傍アルゴリズム、最遠傍アルゴリズム、平均リンクエッジアルゴリズム、重心アルゴリズム、又は二乗和アルゴリズムを使用した凝集クラスタリング）、 k 平均クラスタリング、ファジー k 平均クラスタリング、及びジャービス・パトリック（Jarvis-Patrick）クラスタリングが挙げられるが、これらに限定されない。いくつかの実施形態では、クラスタリングは、教師なしクラスタリングを含み、訓練セットがクラスタリングされる場合、どのクラスタが形成するべきであるかの先入観が課されない。

【0076】

複数カテゴリロジットモデルのものなどの回帰モデルは、Agresti, An Introduction to Categorical Data Analysis, 1996, John Wiley & Sons, Inc., New York, Chapter 8に記載されており、これは、参照によりその全体が本明細書に組み込まれる。いくつかの実施形態では、モデルは、Hastie et al., 2001, The Elements of Statistical Learning, Springer-Verlag, New Yorkに開示されている回帰モデルを利用し、これは、参照によりその全体が本明細書に組み込まれる。いくつかの実施形態では、勾配ブースティングモデルは、例えば、本明細書に記載の分類アルゴリズムに対して使用され、これらの勾配ブースティングモデルは、Boehmke, Bradley; Greenwell, Brandon (2019). "Gradient Boosting". Hands-On Machine Learning with R. Chapman & Hall. pp. 221-245. ISBN 978-1-138-49568-5.に記載されており、これは、参照によりその全体が本明細書に組み込まれる。いくつかの実施形態では、アンサンブルモデル化技術が使用され、これらのアンサンブルモデル化技術は、本明細書の分類モデルの実装に記載され、Zhou Zhihua (2012) Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC. ISBN 978-1-439-83003-1に記載されている（これは、参照によりその全体が本明細書に組み込まれる）。

【0077】

いくつかの実施形態では、機械学習分析は、データ分析を実施するための命令を含む1つ以上のプログラム（例えば、非永続メモリ又は永続メモリに記憶された1つ以上のプロ

グラム)を実行するデバイスによって実施される。いくつかの実施形態では、データ分析は、少なくとも1つのプロセッサ(例えば、処理コア)と、データ分析を実施する命令を含むメモリ(例えば、非永続メモリ又は永続メモリに格納された1つ以上のプログラム)と、を備えるシステムによって実施される。

【0078】

コンピュータシステム

本開示は、本開示の方法を実装するようにプログラムされたコンピュータシステムを提供する。図6は、1つ以上の対象のがん、非がん性疾患、若しくはそれらの任意の組み合わせの健康状態を予測し、本明細書の他の箇所に記載の予測モデルを訓練し、推奨される治療を生成し、又は本明細書の他の箇所に記載のそれらの任意の組み合わせの方法を行うようにプログラムされるか、又はそうでなければそのように構成されるコンピュータシステム600を示す。コンピュータシステム600は、ユーザの電子デバイス、又は電子デバイスに関する遠隔に位置されたコンピュータシステムであり得る。電子デバイスは、モバイル電子デバイスであり得る。

10

【0079】

コンピュータシステム600は、単一コア若しくは複数コアプロセッサ、又は並列処理のための複数のプロセッサであり得る中央処理ユニット(CPU、また、本明細書では「プロセッサ」及び「コンピュータプロセッサ」)606を含む。コンピュータシステム600はまた、メモリ又はメモリ位置604(例えば、ランダムアクセスメモリ、読み取り専用メモリ、フラッシュメモリ)、電子記憶ユニット602(例えば、ハードディスク)、1つ以上の他のシステムと通信するための通信インターフェース608(例えば、ネットワークアダプタ)、並びにキャッシュ、他のメモリ、データ記憶装置、及び／又は電子ディスプレイアダプタなどの周辺デバイス610を含む。メモリ604、記憶ユニット602、インターフェース608、及び周辺デバイス610は、マザーボードなどの通信バス(実線)を介してCPU606と通信している。記憶ユニット602は、データを記憶するためのデータ記憶ユニット(又はデータ保存場所)であり得る。コンピュータシステム600は、通信インターフェース608を用いて、コンピュータネットワーク(「ネットワーク」)612に動作可能にカップリングされ得る。ネットワーク612は、インターネット、インターネット及び／若しくはエクストラネット、又はインターネットと通信しているイントラネット及び／若しくはエクストラネットであり得る。いくつかの場合では、ネットワーク612は、電気通信ネットワーク及び／又はデータネットワークである。ネットワーク612は、クラウドコンピューティングなどの分散コンピューティングを可能にすることができる1つ以上のコンピュータサーバを含み得る。ネットワーク612は、コンピュータシステム600を用いるいくつかの場合では、ピアツーピアネットワークを実装することができ、これは、コンピュータシステム600にカップリングされたデバイスがクライアント又はサーバとして動作することを可能にし得る。

20

30

【0080】

CPU606は、プログラム又はソフトウェアにおいて具現化され得る機械可読命令のシーケンスを実行し得る。命令は、メモリ604などのメモリ位置に記憶され得る。命令は、CPU606に向けられ得、その後、本明細書の他の箇所に記載の本開示の方法を実装するようにCPU606をプログラムするか、又はそうでなければ構成し得る。CPU606によって実施される動作の例としては、フェッチ、デコード、実行、及びライトバックが挙げられ得る。

40

【0081】

CPU606は、集積回路などの回路の一部であり得る。システム600の1つ以上の他の構成要素は、回路に含まれ得る。いくつかの場合では、回路は、用途特異的な集積回路(ASIC)である。

【0082】

記憶ユニット602は、ドライバ、ライブラリ、及び保存されたプログラムなどのファイルを記憶し得る。記憶ユニット602は、ユーザデータ、例えば、ユーザプリファレン

50

ス及びユーザプログラムを保存することができる。いくつかの場合では、コンピュータシステム600は、イントラネット又はインターネットを通してコンピュータシステム600と通信する遠隔サーバ上に配置されるなど、コンピュータシステム600の外部にある1つ以上の追加のデータ記憶ユニットを含むことができる。

【0083】

コンピュータシステム600は、ネットワーク612を介して1つ以上の遠隔コンピュータシステムと通信し得る。例えば、コンピュータシステム600は、ユーザの遠隔コンピュータシステムと通信し得る。遠隔コンピュータシステムの例としては、パーソナルコンピュータ（例えば、ポータブルPC）、スレート若しくはタブレットPC（例えば、Apple（登録商標）iPad、Samsung（登録商標）Galaxy Tab）、電話、スマートフォン（例えば、Apple（登録商標）iPhone、Android対応デバイス、BlackBerry（登録商標））、又はパーソナルデジタルアシスタントが挙げられ得る。ユーザは、ネットワーク612を介してコンピュータシステム600にアクセスし得る。

【0084】

本明細書に記載の方法は、例えば、メモリ604又は電子記憶ユニット602上など、コンピュータシステム600の電子記憶位置上に記憶された機械（例えば、コンピュータプロセッサ）実行可能コードによって実装され得る。機械実行可能コード又は機械可読コードは、ソフトウェアの形態で提供され得る。使用中、コードは、プロセッサ606によって実行され得る。いくつかの場合では、コードは、記憶ユニット602から取得され、プロセッサ606による即時アクセスのためにメモリ604上に記憶され得る。いくつかの状況では、電子記憶ユニット602を除外してもよく、機械実行可能命令は、メモリ604に記憶される。

【0085】

コードは、事前にコンパイルされ、コードを実行するように適合されたプロセッサを有する機械での使用のために構成され得るか、又はランタイム中にコンパイルされ得る。コードは、事前にコンパイルされた又はコンパイルされたままの様式でコードを実行することを可能にするように選択され得るプログラミング言語で提供され得る。

【0086】

いくつかの実施形態では、本明細書の他の箇所に記載のシステムは、1つ以上の対象のがん性又は非がん性の健康状態を診断するためのシステムを含み得る。いくつかの場合では、システムは、（a）1つ以上のプロセッサと、（b）当該1つ以上のプロセッサに、（i）当該1つ以上の対象の生体試料の1つ以上の対象の1つ以上の核酸分子配列決定リードを受け取ることであって、当該1つ以上の核酸分子配列決定リードが、1つ以上の非哺乳類核酸分子の増幅された1つ以上のゲノム特徴の配列を含む、受け取ること、及び、（ii）少なくとも、1つ以上の非哺乳類核酸配列決定リードの1つ以上のゲノム特徴を、訓練された予測モデルへの入力として提供する結果として、1つ以上の対象のがん性又は非がん性の健康状態の診断を出力することを行わせるように構成されたソフトウェアを含む、非一時的なコンピュータ可読記憶媒体と、を含み得る。

【0087】

コンピュータシステム600などの本明細書で提供されるシステム及び方法の態様は、プログラミングで具現化され得る。技術の様々な態様は、典型的には、機械（又はプロセッサ）実行可能コード及び／又は機械可読媒体のタイプで搬送されるか、若しくは具現化される関連データの形態で「製品」又は「製造品」として考えられ得る。機械実行可能コードは、メモリ（例えば、読み取り専用メモリ、ランダムアクセスメモリ、フラッシュメモリ）又はハードディスクなどの電子記憶ユニットに記憶され得る。「記憶」タイプの媒体は、コンピュータ、プロセッサなどの有形メモリのいずれか若しくは全て、又は様々な半導体メモリ、テープドライブ、ディスクドライブなどのそれらの関連するモジュールを含み得、ソフトウェアプログラミングのためにいつでも非一時的な記憶を提供し得る。ソフトウェアの全部又は一部分は、時にはインターネット又は様々な他の通信ネットワーク

10

20

30

40

50

を介して通信され得る。そのような通信は、例えば、一方のコンピュータ又はプロセッサから別のコンピュータ又はプロセッサへ、例えば、管理サーバ又はホストコンピュータからアプリケーションサーバのコンピュータプラットフォームへのソフトウェアのロードを可能にし得る。したがって、ソフトウェア要素を担持し得る別のタイプの媒体は、ローカルデバイス間の物理的インターフェースにわたって、有線及び光地上通信ネットワークを介して、並びに様々なエアリンクにわたって使用されるような、光、電気、及び電磁波を含む。有線又は無線リンク、光リンクなどのそのような波を搬送する物理的要素もまた、ソフトウェアを担持する媒体とみなされ得る。本明細書で使用される場合、非一時的な有形の「記憶」媒体に限定されない限り、コンピュータ又は機械の「読み取り可能媒体」などの用語は、実行のためにプロセッサに命令を提供することに関与する任意の媒体を指す。

10

【0088】

したがって、コンピュータ実行可能コードなどの機械可読媒体は、有形記憶媒体、搬送波媒体又は物理伝送媒体を含むがこれらに限定されない多くの形態をとり得る。不揮発性記憶媒体は、例えば、図面に示されるデータベースなどを実装するために使用され得るなどの任意のコンピュータ（複数可）などの記憶デバイスのうちのいずれかなどの光ディスク又は磁気ディスクを含む。揮発性記憶媒体は、そのようなコンピュータプラットフォームの主要メモリなどの動的メモリを含む。有形伝送媒体は、コンピュータシステム内のバスを含む線を含む、同軸ケーブル、銅線及び光ファイバを含む。搬送波伝送媒体は、電気若しくは電磁信号、又は無線周波数（RF）及び赤外線（IR）データ通信中に生成されたものなどの音響又は光波の形態をとり得る。したがって、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若しくはDVD-ROM、任意の他の光学媒体、パンチカード紙テープ、穴のパターンを有する任意の他の物理記憶媒体、RAM、ROM、PROM及びEPROM、FLASH-EPROM、任意の他のメモリチップ若しくはカートリッジ、データ若しくは命令を輸送する搬送波、そのような搬送波を輸送するケーブル若しくはリンク、又はコンピュータがプログラミングコード若しくはデータを読み取り得る任意の他の媒体が含まれる。これらの形態のコンピュータ可読媒体の多くは、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プロセッサに搬送することに関与し得る。

20

30

【0089】

コンピュータシステム600は、例えば、予測結果を視覚化するためのディスプレイ、又は予測モデルを訓練するためのインターフェースを提供するためのユーザインターフェース（UI）614を備える電子ディスプレイ616を含むか、又は電子ディスプレイ616と通信することができる。UIの例としては、グラフィカルユーザインターフェース（GUI）及びウェブに基づくユーザインターフェースが挙げられるが、これらに限定されない。

【0090】

本発明の好ましい実施形態が本明細書に示され、記載されているが、そのような実施形態が例としてのみ提供されることは、当業者には明白であろう。ここで、当業者は、本発明から逸脱することなく、多数の変形、変更、及び置換を想起するであろう。本明細書に記載の本発明の実施形態に対する様々な代替物が、本発明を実施する際に用いられ得ることを理解されたい。以下の特許請求の範囲は、本発明の範囲を定義し、これらの特許請求の範囲及びそれらの等価物の範囲内の方法及び構造が、それらによって包含されることが意図される。

40

【0091】

本明細書に記載され、請求される方法のステップは、実施形態による方法又は一連の動作の各々を示しているが、当業者であれば、本明細書に記載の教示に基づいて、多くの変形を認識するであろう。ステップは、異なる順序で完了され得る。ステップは、追加又は省略され得る。ステップのいくつかは、サブステップを含み得る。これらのステップの多

50

くは、有益なほど頻繁に繰り返され得る。

【0092】

本方法又は一連の動作の各々のステップのうちの1つ以上は、本明細書に記載の回路、例えば、フィールドプログラマブルゲートアレイのためのプログラマブルアレイロジックなどのプロセッサ又は論理回路の1つ以上で実施され得る。回路は、本方法又は一連の動作の各々のステップのうちの1つ以上を提供するようにプログラムされ得、プログラムは、例えば、コンピュータ可読メモリに記憶されたプログラム命令、又はプログラマブルアレイロジック又はフィールドプログラマブルゲートアレイなどの論理回路のプログラムされたステップを含み得る。

【0093】

実施例1：16S rDNA V6プライマー増幅効率及び特異度

図9に見られ得るように、16S rDNA V6増幅プライマー、例えば、967F及び1064Rの効率を評価し、ヒトゲノムと微生物標準物質希釈物シリーズとの間で比較した。図9は、様々な組成物：ヒトゲノムDNA318、微生物標準物質希釈物シリーズ322、及び反応中にDNAが存在しない陰性対照320のPCR反応生成物生成の観察されたシグナル強度に対してプロットされたポリメラーゼ連鎖反応（PCR）サイクルを示す。微生物標準物質希釈物シリーズは、790、7,896、78,955、789,554、及び7,895,540の16Sコピー数の5つの標準物質希釈物から構成された。PCRサイクルプロットから、ヒトゲノムDNAのみが存在する場合318、V6プライマーは、微生物標準物質希釈物322と比較して、ヒトgDNAを低効率で増幅することが分かり得る。これらの結果は、V6プライマーが、微生物核酸分子を選択的に増幅し、それによって、宿主哺乳類ミトコンドリアDNAの増幅を介して生成されたノイズ又は汚染物質アンプリコンを低減することを示唆する。

【0094】

図10は、ヒトgDNAの存在下で増幅された16S rDNA V6増幅プライマー（例えば、967F、1064R）の効率を評価するために実施された実験を示す。V6プライマーによるPCR増幅のために、4つの群：白血球からのヒトゲノムDNAにスパイクされた微生物DNA標準物質（V6/wbDNA）（324、326、328、330、332）、微生物DNA標準物質（V6/mbDNA）（338、340、342、344、346）、ヒトgDNAのみ（334）、及び鑄型なし対照（すなわち、陰性対照）（336）を調製した。この反応に使用されたヒトgDNA含有量は、約1,396ゲノム当量に対応する、5ngに固定した。質量ベースで、ヒトgDNAは、各実験群において、微生物DNA標準物質の量の1000倍以上で存在した。5つの微生物標準物質群は、790（324、338）、7,896（326、340）、78,955（328、342）、789,554（330、344）、及び7,895,540（332、346）の16Sコピー数を含む。図10に示される結果から、PCR反応中にヒトgDNAのみが存在する場合に増幅が観測されたにもかかわらず、ヒトgDNAプライマーハイブリダイゼーション事象は、両方のDNA源が混合されている場合に、微生物DNAの特異的増幅を妨げるものではないことが理解され得る。

【0095】

図11は、ヒトDNAの存在下での微生物DNAの増幅の16S rDNA V6プライマー（例えば、967F、1064R）の特異度を評価するために実施された実験を示す。ヒトゲノムDNAの量を変化させた6つの実験群：3ng（410）、0.3ng（402）、0.03ng（408）、0.003ng（414）、0.0003ng（406）、及び0ng（400）を調製し、5pgの微生物DNA標準物質（7,895ゲノム当量）の存在下で増幅した。加えて、鑄型なし対照群（412）も陰性対照として利用した。図37に示されるPCRサイクルプロットから、スパイクされたヒトgDNAの全てのレベルで、微生物gDNAが優先的に増幅されたことが分かり得る。微生物DNAの非存在下では、ヒトDNAのみ対照「hDNA 3ng」曲線は、効率が低下する場合に増幅し、31のCpが得られる。これらの結果は、PCR増幅曲線が重ね合わされるた

10

20

30

40

50

め、V6プライマーが非常に特異的であり、ヒトDNAの顕著な前景によって妨げられないことを示した。

【0096】

実施例2：アンブリコン配列決定ブランクフィルタリング

本明細書に記載のワークフロー及び／又は方法の様々なステップでのアンブリコン配列決定リードカウントを図14A～14Bに示す。図14Bは、図14Aに示される「OTU」及び「OTU__filtered」配列決定リード／試料のズームインビューを示す。実験群には、血漿（500）、ブランク陰性対照（502）、及び規定濃度で混合されたDNA微生物の工業用Zymo市販試料（504）が含まれた。試料当たりのリード数のプロットは、本明細書の他の箇所に記載の核酸分子配列決定リードの様々な点のリード数／試料を示す：「raw__reads」は、品質フィルタ又は分類学的割り当ての前の総リード／試料であり、「qf__reads」は、PCRの重複を除去するための品質フィルタリングステップの後に残存する配列決定リードであり、「OTU」は、Deblur処理を介して特定された垂操作分類単位に対応する試料当たりのリードの数に対応し、「OTUs__filtered」は、DNA抽出ブランク対照（すなわち、「ブランク」）に存在するsOTUを差し引いた後に残存するsOTUに対応する。更なる下流処理のために、データセット全体内の少なくとも10個の存在量を有する特徴を保持した。

10

【0097】

図14A～14Bに示される結果から、ブランク対照から配列決定リードを差し引くことからDNA抽出プロセス自体に存在する汚染を考慮した後、本明細書の他の箇所に記載の、1つ以上の対象の健康状態を分類及び／又は特定するための1つ以上の核酸特徴を決定するために更に処理するための数千の浄化されたリードが存在する。総存在量の75.75%は、ブランクフィルタリング後の血漿試料から除去され、これは、合計30,689個の微生物から12,509個の微生物が除去されたことを意味する（微生物の40.76%が除去された）。したがって、本明細書の他の箇所に記載の方法及びシステムによって検出及び分析されるものは、DNA汚染ではなく、むしろ、対象の生体試料中に存在する核酸分子シグナルである。

20

【0098】

実施例3：機械学習分類子の性能

機械学習分類子を、既知の健康状態標識、すなわち、特定の非がん性疾患標識及び／又は非小細胞肺癌の病期を有する1つ以上の対象の16Sアンブリコン配列（例えば、V6超可変領域）で訓練した。様々な標識された健康状態の対象及び／又は試料の数の分布を図15Aに示す。各群の全ての対象からの血漿を得て、V6 16Sプライマーを用いて増幅し、続いて、本明細書の他の箇所に記載の次世代配列決定を行った。次いで、配列決定リードを浄化して、1つ以上の微生物分類特徴を特定した。微生物分類特徴には、属レベルに落とし込んだ特定された微生物分類の存在量が含まれた。次いで、微生物分類特徴の関連するリードカウントを使用して、5倍の交差検証を用いて3つのランダムフォレスト機械学習モデルを訓練した。3つのランダムフォレスト機械学習モデルは、I期、II期、及びIII期のがんの健康状態を分類及び／又は特徴付けるための分類子を含んだ。分類子及び関連する曲線下面積（AUC）の性能受信機動作特性曲線、すなわち、I期について0.891、II期について0.71、及びIII期のがんについて0.88を、図15Bに示す。

30

40

【0099】

I期（506）、II期（510）、及びIII期（508）のがんの健康状態分類及び／又は特定のための3つの別個の機械学習分類子を、配列決定リードのショットガン配列決定及びウェブオブライフ（WoL）属の相対的存在量を使用して生成した。ショットガン配列決定及びWoL存在量で訓練された機械学習分類子の関連するAUCは、I期がんで0.837、II期がんで0.66、及びIII期がんで0.871であった。

【0100】

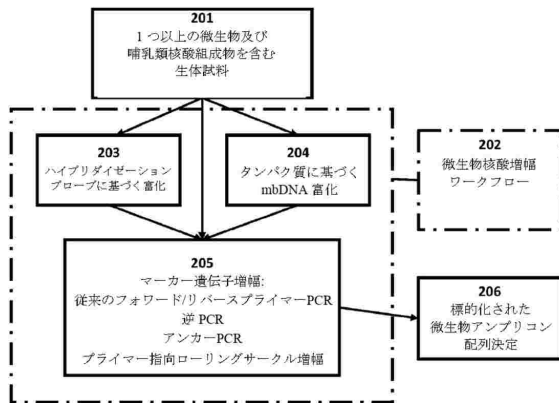
これらの結果から、1つ以上の富化及び／又は増幅された核酸分子を提供するV6 1

50

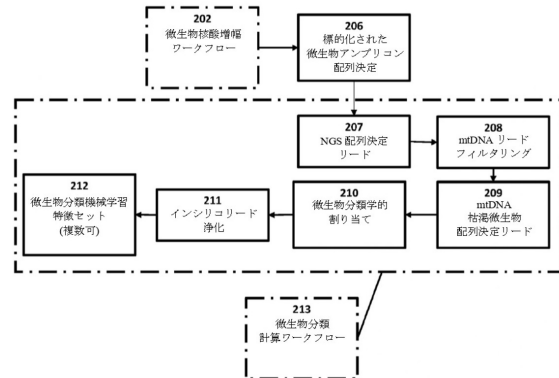
6 S 増幅プライマーを使用して、I 期、II 期、及び III 期のがんの区別において高精度を提供する 1 つ以上の微生物分類学的特徴を開発し得ることが分かり得る。

【図面】

【図 1】

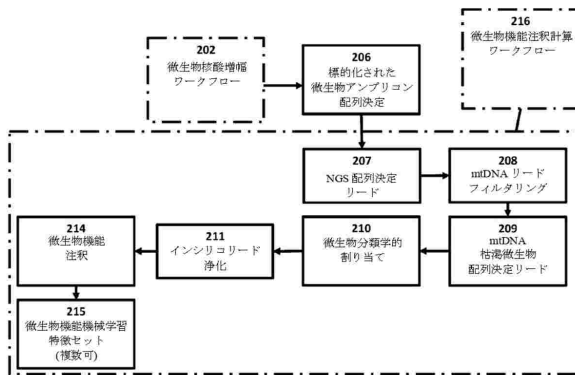


【図 2】

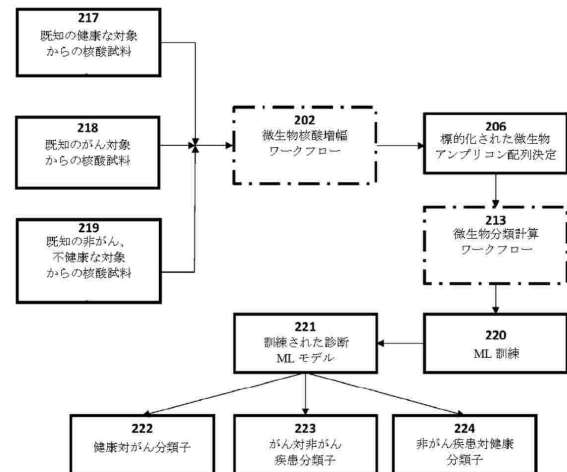


10

【図 3】



【図 4】



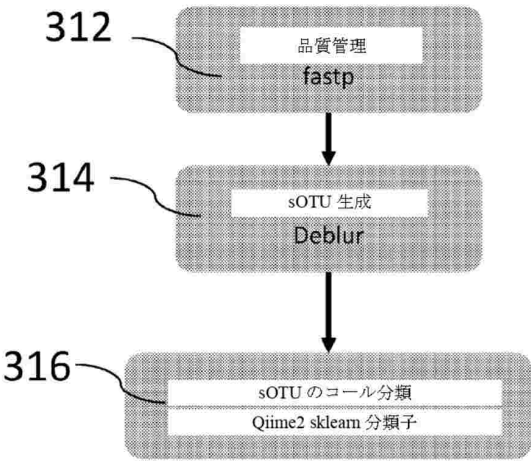
20

30

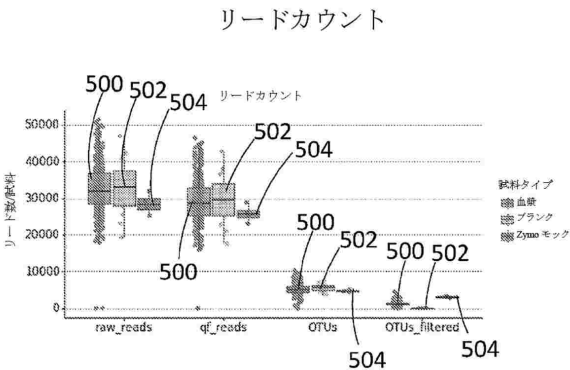
40

50

【図 1 3】



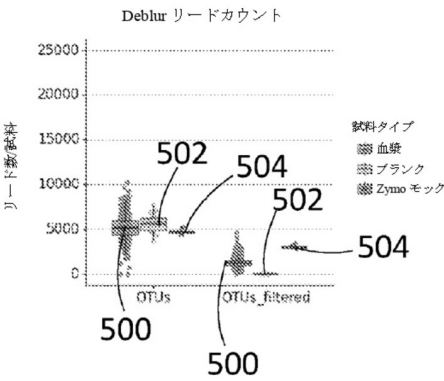
【図 1 4 A】



10

【図 1 4 B】

ブランクフィルタリング



30

【図 1 5 A】

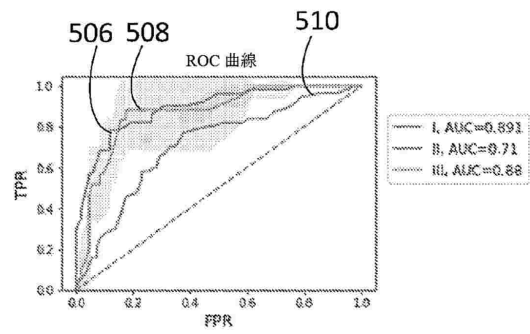
	診断	試料数
	カルチノイド	50
非がん	過誤腫	20
	肉芽腫	16
	間質性線維症	12
	器質化肺炎	6
	気管支拡張症	5
非小細胞肺癌	診断	
	I 期	51
	II 期	57
	III 期	19

20

40

50

【図 15 B】



10

20

30

40

50

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 23/64065

A. CLASSIFICATION OF SUBJECT MATTER

IPC - INV. C12Q 1/6886 (2023.01)

ADD. C12Q 1/6806 (2023.01)

CPC - INV. C12Q 1/6886, C12Q 1/689, G16B 40/20

ADD. C12Q 1/701, G16H 50/20, C12Q 1/701

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)

See Search History document

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

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X — Y	US 2021/0057046 A1 (FREEMOME HOLDINGS, INC.) 25 February 2021 (25.02.2021) abstract para [0006], [0013], [0042], [0043], [0045], [0051], [0066], [0076], [077], [00833], [0086], [0087], [0094], [0096], [0097], [0109], [0110], [0111], [0117], [0120], [0142], [0144], [0177], [0191], [0224], [0241], [0243], [0247], [0248], [0323]	1-15, 21-24, 32, 35-45, 146-149, 150-158, 164-169, 177, 180-188 16-20, 25-31, 33-34, 159-163, 170-176, 178-179
Y	US 2018/0340215 A1 (THE BROAD INSTITUTE, INC.) 29 November 2018 (29.11.2018) abstract, para [0036], [0037], [0039], [0046], [0076], [0077], [0082], [0094]	16-20, 159-163
Y	US 2021/0189459 A1 (ARC BIO, LLC) 24 June 2021 (24.06.2021) abstract, para [0012]	25-31, 33-34, 170-176, 178-179
Y	US 2014/0127687 A1 (ANALYTIK JENA AG) 8 May 2014 (08.05.2014) para [0002]-[0006], [0037]-[0038], [0040], [0042]-[0044]	19-20, 162-163

☐ Further documents are listed in the continuation of Box C.☐ 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

"D" document cited by the applicant in the international application

"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

26 October 2023

Date of mailing of the international search report

NOV 14 2023

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Kari Rodriguez

Telephone No. PCT Helpdesk: 571-272-4300

Form PCT/ISA/210 (second sheet) (July 2022)

10

20

30

40

50

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 23/64065

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 extra 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 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.:
1-45 and 146-188

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 SEARCH REPORT

International application No.

PCT/US 23/64065

Continuation of Box No. III, Observations where unity of invention is lacking:

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I: Claims 1-45 and 146-188, directed to method of generating a genomic feature set for differentiating a cancer, cancer type, or non-cancer health state of subjects by:

- a) providing a subjects' nucleic acids and corresponding health state;
- b) amplifying genomic features of non-mammalian nucleic acid molecules;
- c) sequencing said amplified genomic features to generate non-mammalian sequencing reads;
- d) generating a feature set to differentiate a cancer and non-cancer health state by combining said non-mammalian genomic feature abundances and the health state of the subject(s).

Group II: Claims 48-93, directed to a method of diagnosing a cancer of one or more subjects comprising:

- a) providing nucleic acid molecules from one or more subjects;
- b) amplifying genomic features of non-mammalian nucleic acid molecules;
- c) sequencing said amplified genomic features to generate non-mammalian sequencing reads;
- (d) outputting a diagnosis of a cancer or non-cancer health state of said subject(s) as a result of providing the genomic features as an input to a trained predictive model.

Group III: Claims 94-145, directed to a composition of matter comprising system for diagnosing a cancerous or non-cancerous health state of one or more subjects, the system comprises:

- one or more processors and a non-transitory computer readable storage medium including software configured to cause said one or more processors to
- (i) receive sequencing reads from a subject's nucleic acids comprising amplified genomic features of non-mammalian nucleic acid molecules;
- (ii) output a diagnosis of a cancerous or non-cancerous health state based on providing said reads as an input to a trained predictive model.

The inventions listed as Groups I through III do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Special Technical Features

Group I requires generating a genomic feature set for differentiating a cancer type in a subject by combining said feature set and a corresponding health state in said subject, not required by group II and III.

Group II requires diagnosing a cancer or non-cancer with sequences inputted into a trained predictive model, not required by groups I and III.

Group III requires a system, a composition of matter, comprising processors and non-transitory computer readable storage medium not required by groups I and II.

Common Technical Features

The common technical feature shared by Groups I-III, is genomic features amplified and sequenced from a subject's nucleic acid comprising non-mammalian nucleic acid sequences (microbial sequences, instant application para [0002]) that correlate with a cancer or non-cancer health state. However, this shared technical feature does not represent a contribution over prior art, because the shared technical feature is taught by WO 2014/130162 A1 to Lee et al. (hereinafter 'Lee'). Lee teaches genomic features amplified and sequenced from a subject's nucleic acid comprising microbial nucleic acid sequences that correlate with a cancer or non-cancer health state (abstract "methods of determining that a subject is likely to have cancer are provided. Such methods may include amplifying a microbial DNA sample in a test sample obtained from the subject to determine an amount of microbial DNA in the test sample, wherein the amount of microbial DNA is determined by an amplification or sequencing technique; and determining that the subject is likely to have breast cancer when there is a significantly decreased level of microbial DNA in the test sample when compared to a level of microbial DNA in a control sample").

The common technical feature shared by Groups II-III, is diagnosis of a cancerous or non-cancerous health state based on providing a subject's nucleic acid sequence reads as an input to a trained predictive model. However, this shared technical feature does not represent a contribution over prior art, because the shared technical feature is taught by the article entitled "Assessment of peritoneal microbial features and tumor marker levels as potential diagnostic tools for ovarian cancer" by Miao et al. (hereinafter 'Miao') (PLoS ONE, 9 January 2020, vol 15, no 1, article e0227707, pp 1-19).

***** See Next Extra Sheet to continue *****

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 23/64065

continuation of previous extra sheet:

Miao teaches diagnosis of a cancerous or non-cancerous health state based on providing a subject's nucleic acid sequence reads as an input to a trained predictive model (abstract "Epithelial ovarian cancer (OC) is the most deadly cancer of the female reproductive system. To date, there is no effective screening method for early detection of OC ... We hypothesized that the peritoneal microbial environment was altered in patients with OC and that inclusion of selected peritoneal microbial features with current clinical features into prediction analyses will improve detection accuracy of patients with OC. Blood and peritoneal fluid were collected from consented patients ... and processed for Next Generation Sequencing (NGS) using 16S V4 exon bacterial primers and bioinformatics analyses. We found that patients with OC had a unique peritoneal microbial profile compared to patients with a benign mass. Using ensemble modeling and machine learning pathways, we identified 18 microbial features that were highly specific to OC pathology ... We conclude that OC pathogenesis alters the peritoneal microbial environment and that these unique microbial features are important for accurate diagnosis of OCs", pg. 6, para 2 "The Lasso logistic regression model ... approach is included for comparison due to its simple model structure and easy interpretation", pg. 7, para 5 "The penalty parameter of Lasso is determined through 10-fold cross validation. ...we calculate the classification error for each candidate value through 10-fold cross validation on the N-1 training samples. We then choose the parameter value that yields the smallest classification error, fix the penalty parameter at that value, and fit the model using all N-1 training samples", see Fig. 2, predictive modeling.).

As the technical features were known in the art at the time of the invention, they cannot be considered special technical features that would otherwise unify the groups.

Groups I through III therefore lack unity under PCT Rule 13 because they do not share a same or corresponding special technical feature.

フロントページの続き

(51)国際特許分類

C 1 2 M 1/34 (2006.01)

F I

C 1 2 M

1/34

B

テーマコード (参考)

,MC,ME,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,C
O,CR,CU,CV,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,I
R,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,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,ST,
SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(特許庁注：以下のものは登録商標)

1. ANDROID

(72)発明者 アダムス, エディ

アメリカ合衆国 9 2 1 2 1 カリフォルニア州 サンディエゴ フェリス・スクエア 6 3 4 2

(72)発明者 プーア, グレゴリー ディー.

アメリカ合衆国 9 2 1 2 1 カリフォルニア州 サンディエゴ フェリス・スクエア 6 3 4 2

F ターム (参考) 4B029 AA07 AA23 AA27 BB11 BB20 CC01 DG10 FA02 FA15 GA08

GB01 GB02 GB10 HA10

4B063 QA01 QA13 QA18 QA19 QQ03 QQ42 QQ52 QR08 QR55 QR62

QS25 QS34 QX02