

(19)日本国特許庁(JP)

(12)公表特許公報(A)

(11)公表番号

特表2023-511368

(P2023-511368A)

(43)公表日 令和5年3月17日(2023.3.17)

(51)国際特許分類		F I		テーマコード (参考)	
C 1 2 Q	1/6883(2018.01)	C 1 2 Q	1/6883	Z Z N A	4 B 0 6 3
C 1 2 Q	1/6869(2018.01)	C 1 2 Q	1/6869	Z	
C 1 2 Q	1/6886(2018.01)	C 1 2 Q	1/6886	Z	
C 1 2 Q	1/686(2018.01)	C 1 2 Q	1/686	Z	
C 1 2 Q	1/6851(2018.01)	C 1 2 Q	1/6851	Z	
		審査請求 未請求 予備審査請求 未請求		(全60頁)	

(21)出願番号	特願2022-544274(P2022-544274)	(71)出願人	521032771
(86)(22)出願日	令和3年1月22日(2021.1.22)		ゲートハウス バイオ インコーポレイテッド
(85)翻訳文提出日	令和4年9月16日(2022.9.16)		アメリカ合衆国 0 2 4 9 4 マサチューセッツ ニーダム リンドバーグ アベニュー 2 0 2
(86)国際出願番号	PCT/US2021/014755	(74)代理人	110000796
(87)国際公開番号	WO2021/150990		弁理士法人三枝国際特許事務所
(87)国際公開日	令和3年7月29日(2021.7.29)	(72)発明者	サルズマン デビッド ダブリュー.
(31)優先権主張番号	62/964,412		アメリカ合衆国 0 2 4 9 4 マサチューセッツ ニーダム リンドバーグ アベニュー 2 0 2 シー/オー ゲートハウス バイオ インコーポレイテッド
(32)優先日	令和2年1月22日(2020.1.22)	(72)発明者	サルズマン アラン ピー.
(33)優先権主張国・地域又は機関	米国(US)		アメリカ合衆国 0 2 4 9 4 マサチューセッツ ニーダム リンドバーグ アベニュー 2 0 2 シー/オー ゲートハウス バイオ インコーポレイテッド
(81)指定国・地域	AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,最終頁に続く		最終頁に続く

(54)【発明の名称】 低分子RNA疾患分類器

(57)【要約】

本開示は、1つ以上の異なる生物学的状態または1つ以上の疾患サブタイプについて対象を評価するための疾患分類器を構築するための方法を提供する。本発明は、発見試料セットの配列データから候補低分子RNA (sRNA) 配列を特定することを含む。発見試料セットにわたる候補sRNA配列(各々個別に入手されたもの)の存在または存在量は、目的とする生物学的状態を予測し(例えば、他の異なる生物学的状態または非疾患対照に対して)、これらの候補sRNA配列が本開示の実施形態に従ってさらにフィルタリングまたは選択される。その後、機械学習技法が適用されて、複数疾患分類器を含む疾患分類器を構築及びトレーニングする。トレーニングされた分類器を使用して、新たな試料を分類する、例えば、疾患について患者を評価することができる。

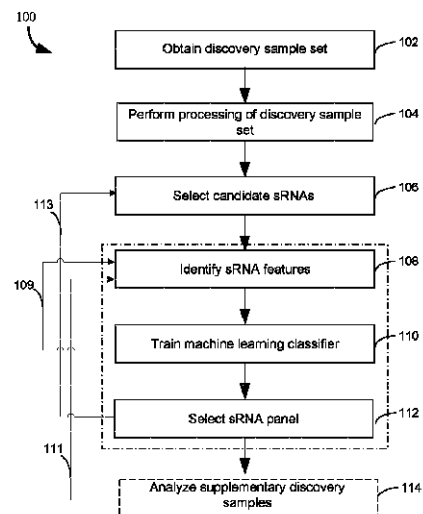


FIG. 1

【特許請求の範囲】

【請求項 1】

1 つ以上の生物学的状態について対象を評価するための分類器を作製するための方法であって、

発見試料セットにわたる sRNA 配列の存在もしくは不在または存在量を含む sRNA 配列データを提供することであって、前記発見試料セットが 1 つ以上の生物学的状態の存在または不在を表す、前記提供することと、

候補 sRNA 配列であって、それらの存在もしくは不在または存在量が生物学的状態の存在または不在と相関している、前記候補 sRNA 配列を選択することと、

前記候補 sRNA 配列から、前記 1 つ以上の生物学的状態について試料を評価するための特徴を含む分類器をトレーニングすることと、を含む、前記方法。

10

【請求項 2】

前記発見試料が 2 つ以上の生物学的状態に対して陽性または陰性であるとラベル付けされる、請求項 1 に記載の方法。

【請求項 3】

前記 sRNA 配列データが、sRNA 配列リードから 5' 及び 3' 配列決定アダプターをトリミングすることによって、かつ参照配列または遺伝子座に基づいて sRNA 配列バリエーションを統合することなく処理される、請求項 1 に記載の方法。

【請求項 4】

候補 sRNA 配列が、それらの存在もしくは不在または存在量が生物学的状態と相関する程度に基づいて選択される、請求項 3 に記載の方法。

20

【請求項 5】

少なくとも 1 つの候補 sRNA 配列が、生物学的状態に対して陽性である複数の発見試料に存在し、すべての非疾患試料または異なる生物学的状態でラベル付けされたすべての試料に不在である、請求項 4 に記載の方法。

【請求項 6】

候補 sRNA 配列であって、それらの存在または存在量によって生物学的状態の存在または不在を個別に予測する、前記候補 sRNA 配列が選択される、請求項 4 に記載の方法。

【請求項 7】

候補 sRNA 配列であって、それらの存在または存在量が生物学的状態の存在または不在を予測し、かつ少なくとも 0.01 の p 値を有する、前記候補 sRNA 配列が選択される、請求項 6 に記載の方法。

30

【請求項 8】

少なくとも 1 つの候補 sRNA 配列であって、それらの存在または存在量が生物学的状態の不在の存在を予測し、少なくとも 0.0001 の p 値を有する、前記少なくとも 1 つの候補 sRNA 配列が選択される、請求項 7 に記載の方法。

【請求項 9】

少なくとも 1 つの候補 sRNA 配列であって、それらの存在または存在量が生物学的状態の存在または不在を予測し、かつ少なくとも 0.000001 の p 値を有する、前記少なくとも 1 つの候補 sRNA 配列が選択される、請求項 7 に記載の方法。

40

【請求項 10】

少なくとも 1 つの候補 sRNA 配列であって、それらの存在または存在量が生物学的状態の存在または不在を予測し、かつ少なくとも 0.00000001 の p 値を有する、前記少なくとも 1 つの候補 sRNA 配列が選択される、請求項 7 に記載の方法。

【請求項 11】

少なくとも 1 つの候補 sRNA 配列であって、それらの存在または存在量が生物学的状態の存在または不在を予測し、かつ少なくとも 0.0000000001 の p 値を有する、前記少なくとも 1 つの候補 sRNA 配列が選択される、請求項 7 に記載の方法。

【請求項 12】

50

少なくとも2つの生物学的状態の存在または不在を個別に予測する候補 sRNA 配列が選択される、請求項7に記載の方法。

【請求項13】

前記発見試料セットが少なくとも2つの別個の研究から得られ、前記選択された候補 sRNA 配列の各々が各研究からの少なくとも1つの試料に存在した、請求項1に記載の方法。

【請求項14】

前記別個の研究が異なる場所での生体試料の収集を含む、請求項13に記載の方法。

【請求項15】

前記別個の研究が異なる場所での核酸または sRNA の抽出をさらに含む、請求項14に記載の方法。 10

【請求項16】

前記別個の研究が異なる場所での sRNA 配列決定をさらに含む、請求項15に記載の方法。

【請求項17】

前記発見試料セットが生物学的状態のステージ、グレード、または重症度についてさらにラベル付けされ、リードカウントがかかるステージ、グレード、または重症度と関連する候補 sRNA 配列が選択される、請求項1～16のいずれか1項に記載の方法。

【請求項18】

前記 sRNA 配列が、sRNA 配列のレベルを内在性 sRNA 対照及び／またはスパイクイン対照に対して正規化するために、前記対照を使用して sRNA 配列決定によって決定されたものである、請求項17に記載の方法。 20

【請求項19】

複数の試料由来の RNA が配列決定のためにプールされ、異なる試料由来の配列が試料特定タグ配列を含む、請求項18に記載の方法。

【請求項20】

候補 sRNA 配列が100万リードあたり少なくとも0.1のトリミングされたリードの平均リードカウントを有する、請求項19に記載の方法。

【請求項21】

候補 sRNA 配列が、生物学的状態において増加した配列多様性を有する sRNA ファミリーを特定し、かつ前記 sRNA ファミリー内の sRNA 配列を候補 sRNA 配列として選択することによって選択される、及び／またはエクソソーム中での存在に関連する配列特徴を有する候補 sRNA 配列が選択される、請求項1に記載の方法。 30

【請求項22】

前記発見試料セットが、少なくとも3つの生物学的状態または少なくとも5つの生物学的状態の存在及び不在を表す、請求項1～21のいずれか1項に記載の方法。

【請求項23】

前記発見試料セットが、少なくとも10の生物学的状態の存在及び不在を表す、請求項22に記載の方法。

【請求項24】

前記分類器が、sRNA 配列のパネルの存在もしくは不在または存在量に基づいて試料を分類するようにトレーニングされ、前記パネルが、1クラスあたり約4～約200個の sRNA 配列、または1クラスあたり約4～約100個の sRNA 配列、または1クラスあたり約4～約50個の sRNA 配列を含む、請求項1～23のいずれか1項に記載の方法。 40

【請求項25】

前記発見試料セットが、固体組織試料、生体液試料、または培養細胞を含む、請求項1～24のいずれか1項に記載の方法。

【請求項26】

前記発見試料セットが、血液、血清、血漿、脳脊髄液、尿、または唾液である、請求項 50

25に記載の方法。

【請求項27】

前記発見試料セットが固体組織生検である、請求項25に記載の方法。

【請求項28】

前記発見試料セットが、前記少なくとも2つの生物学的状態に対して陽性である少なくとも10個の試料を含む少なくとも100個の試料を含む、請求項1～27のいずれか1項に記載の方法。

【請求項29】

前記発見試料が、少なくとも25個の非疾患または健常対照を含む、請求項28に記載の方法。

10

【請求項30】

前記分類器が、パラメトリック／ノンパラメトリック距離測定法、ロジスティック回帰、サポートベクトルマシン、決定木、ランダムフォレスト、ニューラルネットワーク、プロビット回帰、フィッシャー線形判別、単純ベイズ分類器、パーセプトロン、二次分類器、カーネル推定、k近傍法、学習ベクトル量子化、及び主成分分析などの教師あり、教師なし、半教師あり機械学習モデルのうちの1つ以上を使用してトレーニングされる、請求項1～29のいずれか1項に記載の方法。

【請求項31】

前記分類器が、線形サポートベクトルマシンを使用してトレーニングされる、請求項30に記載の方法。

20

【請求項32】

追加の発見試料からのsRNA配列データが、分類器特徴を減少させるために評価される、請求項31に記載の方法。

【請求項33】

前記生物学的状態が、中枢神経系の状態である、請求項1～32のいずれか1項に記載の方法。

【請求項34】

少なくとも2つの生物学的状態が、認知症の症状を伴う神経変性疾患である、請求項33に記載の方法。

【請求項35】

少なくとも2つの生物学的状態が、アルツハイマー病、パーキンソン病、ハンチントン病、軽度認知障害、進行性核上性麻痺、前頭側頭型認知症、レビー小体型認知症、及び血管性認知症から選択される、請求項33に記載の方法。

30

【請求項36】

少なくとも2つの生物学的状態が、運動制御喪失の症状を伴う神経変性疾患である、請求項33に記載の方法。

【請求項37】

少なくとも2つの生物学的状態が、アルツハイマー病、パーキンソン病、ハンチントン病、多発性硬化症、筋萎縮性側索硬化症、及び脊髄性筋萎縮症から選択される、請求項36に記載の方法。

40

【請求項38】

少なくとも2つの生物学的状態が、任意選択で多発性硬化症、視神経炎、横断性脊髄炎、及び視神経脊髄炎を含む、脱髄疾患である、請求項33に記載の方法。

【請求項39】

1つ以上の生物学的状態が、アルツハイマー病、パーキンソン病、ハンチントン病、多発性硬化症、筋萎縮性側索硬化症、及び脊髄性筋萎縮症から選択され、トレーニング試料が、疾患ステージ、疾患重症度、薬物応答性、または疾患進行の経過についてラベル付けされる、請求項1～32のいずれか1項に記載の方法。

【請求項40】

前記生物学的状態が、異なる組織または細胞起源のがんである、請求項1～32のい

50

れか 1 項に記載の方法。

【請求項 4 1】

前記生物学的状態が、薬物感受性及び薬物抵抗性のがんを含む、請求項 4 0 に記載の方法。

【請求項 4 2】

前記対象由来の前記生体試料が、腫瘍またはがん細胞生検である、請求項 4 0 または 4 1 に記載の方法。

【請求項 4 3】

前記生物学的状態が、任意選択で全身性エリテマトーデス（SLE）、強皮症、自己免疫性血管炎、糖尿病（1 型または 2 型）、グレーブス病、アディソン病、シェーグレン症候群、甲状腺炎、リウマチ性関節炎、重症筋無力症、多発性硬化症、線維筋痛症、乾癬、クローン病、潰瘍性大腸炎、憩室症、セリアック病、及び臓器線維症疾患のうちの 1 つ以上を含む、炎症性疾患または免疫疾患である、請求項 1 ～ 3 2 のいずれか 1 項に記載の方法。

10

【請求項 4 4】

前記生体試料が、血液、血清、または血漿である、請求項 4 3 に記載の方法。

【請求項 4 5】

前記生物学的状態が、任意選択で急性事象のリスクに対する層別化を含む、心血管疾患である、請求項 1 ～ 3 2 のいずれか 1 項に記載の方法。

【請求項 4 6】

前記心血管疾患が、冠動脈疾患（CAD）、心筋梗塞、脳卒中、うっ血性心不全、高血圧性心疾患、心筋症、心臓不整脈、先天性心疾患、心臓弁膜症、心臓炎、大動脈瘤、末梢動脈疾患、及び静脈血栓症のうちの 1 つ以上を含む、請求項 4 5 に記載の方法。

20

【請求項 4 7】

少なくとも 2 つの生物学的状態が疾患サブタイプである、請求項 1 ～ 3 2 のいずれか 1 項に記載の方法。

【請求項 4 8】

前記試料セットが複雑な疾患の疾患サブタイプについてラベル付けされず、疾患サブタイプ分類器が教師なし機械学習モデルを使用してトレーニングされるか、または前記試料セットが複雑な疾患の疾患サブタイプについて部分的にのみラベル付けされ、疾患サブタイプ分類器が半教師あり機械学習モデルを使用してトレーニングされる、請求項 4 7 に記載の方法。

30

【請求項 4 9】

前記パネルにおける sRNA が標的遺伝子または経路にマッピングされて、前記疾患サブタイプの創薬可能な標的または治療的介入を特定する、請求項 4 8 に記載の方法。

【請求項 5 0】

1 つ以上の生物学的状態について対象を評価するための方法であって、前記対象の生体試料を提供し、sRNA パネルにおける sRNA の存在もしくは不在または存在量を決定すること、請求項 1 ～ 4 9 のいずれか 1 項に従って調製された疾患分類器を使用して 1 つ以上の生物学的状態間で前記対象の状態を分類すること、を含む、前記方法。

40

【請求項 5 1】

前記試料中の sRNA の存在もしくは不在または存在量が、定量的 PCR アッセイによって決定される、請求項 5 0 に記載の方法。

【請求項 5 2】

前記試料中の sRNA の存在もしくは不在または存在量が、任意選択で sRNA 標的捕捉を用いる sRNA 配列決定によって決定される、請求項 5 0 に記載の方法。

【請求項 5 3】

前記疾患分類器が、少なくとも 3 つの生物学的状態または少なくとも 5 つの生物学的状態間で試料を分類する、請求項 5 0 ～ 5 2 のいずれか 1 項に記載の方法。

50

【請求項 5 4】

前記疾患分類器が、少なくとも 10 の生物学的状態間で分類する、請求項 5 3 に記載の方法。

【請求項 5 5】

前記パネルが、約 4 ～約 200 個の sRNA、または約 4 ～約 100 個の sRNA、または約 4 ～約 50 個の sRNA を含む、請求項 5 0 ～5 4 のいずれか 1 項に記載の方法。

【請求項 5 6】

前記生体試料が、固体組織試料、生体液試料、または培養細胞のうちの 1 つ以上を含む、請求項 5 5 に記載の方法。

【請求項 5 7】

前記生体試料が、血液、血清、血漿、脳脊髄液、尿、または唾液である、請求項 5 6 に記載の方法。

【請求項 5 8】

前記対象の生体試料が固体組織生検である、請求項 5 6 に記載の方法。

【請求項 5 9】

前記分類器が、中枢神経系の生物学的状態を表す発見セットを使用してトレーニングされる、請求項 5 7 に記載の方法。

【請求項 6 0】

前記対象が中枢神経系疾患と一致する症状を呈する、請求項 5 9 に記載の方法。

【請求項 6 1】

前記対象が認知症の症状を有する、請求項 6 0 に記載の方法。

【請求項 6 2】

前記対象が運動制御喪失の症状を有する、請求項 6 0 に記載の方法。

【請求項 6 3】

前記対象が、アルツハイマー病、パーキンソン病、ハンチントン病、軽度認知障害、進行性核上性麻痺、前頭側頭型認知症、レビー小体型認知症、血管性認知症、多発性硬化症、筋萎縮性側索硬化症、及び脊髄性筋萎縮症のうちの 1 つ以上を有するまたは有しないと分類される、請求項 6 1 または 6 2 に記載の方法。

【請求項 6 4】

前記対象が、任意選択で多発性硬化症、視神経炎、横断性脊髄炎、及び視神経脊髄炎のうちの 1 つ以上を含む、脱髄疾患を有するまたは有しないと分類される、請求項 6 0 に記載の方法。

【請求項 6 5】

前記対象が、アルツハイマー病、パーキンソン病、ハンチントン病、多発性硬化症、筋萎縮性側索硬化症、及び脊髄性筋萎縮症のうちの 1 つ以上を有すると診断または決定される、前記対象が、疾患ステージ、疾患重症度、薬物応答性、または疾患進行の経過について分類される、請求項 6 0 に記載の方法。

【請求項 6 6】

前記対象が、がんのリスクがあるか、がんを有する疑いがあるか、またはがんを有すると診断される、請求項 5 0 ～5 8 のいずれか 1 項に記載の方法。

【請求項 6 7】

前記対象ががんを有し、前記試料が、薬物感受性、薬物抵抗性、及び組織起源から選択される 1 つ以上について分類される、請求項 6 6 に記載の方法。

【請求項 6 8】

前記対象由来の前記生体試料が、腫瘍またはがん細胞生検である、請求項 6 7 に記載の方法。

【請求項 6 9】

前記対象が炎症性疾患または免疫疾患の症状を呈する、請求項 5 0 ～5 8 のいずれか 1 項に記載の方法。

【請求項 7 0】

10

20

30

40

50

前記対象の試料が、全身性エリテマトーデス（SLE）、強皮症、自己免疫性血管炎、糖尿病（1型または2型）、グレーブス病、アディソン病、シェーグレン症候群、甲状腺炎、リウマチ性関節炎、重症筋無力症、多発性硬化症、線維筋痛症、乾癬、特発性肺線維症、クローン病、潰瘍性大腸炎、憩室症、及びセリアック病のうちの1つ以上の存在または不在について分類される、請求項69に記載の方法。

【請求項71】

前記生体試料が、血液、血清、または血漿である、請求項69または70に記載の方法。

【請求項72】

前記疾患状態が、任意選択で急性事象のリスクに対する層別化を含む、心血管疾患である、請求項50～58のいずれか1項に記載の方法。

10

【請求項73】

前記心血管疾患が、冠動脈疾患（CAD）、心筋梗塞、脳卒中、うっ血性心不全、高血圧性心疾患、心筋症、心臓不整脈、先天性心疾患、心臓弁膜症、心臓炎、大動脈瘤、末梢動脈疾患、及び静脈血栓症のうちの1つ以上を含む、請求項72に記載の方法。

【請求項74】

前記対象が、複雑な疾患の疾患サブタイプについて分類される、請求項50～73のいずれか1項に記載の方法。

【発明の詳細な説明】

【技術分野】

20

【0001】

優先権

本出願は、参照により全体が本明細書に組み込まれる2020年1月22日に出願された特許仮出願第62／964,412号の利益を主張するものである。

【背景技術】

【0002】

複雑な疾患における診断法及び治療法の進歩は、限られた成功しか収めていない。メンデル型遺伝病とは対照的に、複雑な疾患は、多くの場合、単一遺伝子変異によって引き起こされない表現型として定義される。複雑な疾患は、多数の遺伝的事象によって引き起こされる可能性があり、これは、罹患した個体によって異なる場合があり、環境要因からの多大な寄与を含み得る。複雑な疾患の研究に対する従来のアプローチは、類似した表現型を有する患者を特定し、関連研究を使用して表現型に関する共通の原因遺伝的事象を特定しようと試みている。これらのアプローチは、例えば、表現型に関連する一塩基多型（SNP）などの遺伝子変異を特定することによって、DNAレベルで機能する。この古典的なアプローチは、限られた成功しか収めておらず、多くの高価な治験が有効性を示すことができず、これは、部分的には、基礎疾患が十分に特徴付けされていないまたは理解されていないままであるか、または確立されたまたは認識された疾患特徴付けとは異質のままであるためである。Jameson LJ et al., Precision Medicine – Personalized, Problematic and Promising, NEJM 372:2229–2234 (2015)、Lyman GH, et al., Biomarker Tests for Molecularly Targeted Therapies – Laying the Foundation and Fulfilling the Dream, J. Clin. Oncol. 34 (17):2061–2066 (2016)を参照されたい。

30

40

【0003】

複雑な疾患をサブタイプ化するためのアプローチを含む、疾患を分類するための新たなアプローチが必要とされている。複雑な疾患を分類またはサブタイプ化するための正確な分子アプローチは、診断及び療法における大きなブレイクスルーにつながり、次世代の患者ケアにつながる可能性がある。本発明は、これら及び他の目的を満たす。

【発明の概要】

50

【0004】

本開示は、1つ以上の異なる生物学的状態または1つ以上の疾患サブタイプについて対象を評価するための疾患分類器を構築するための方法を提供する。本発明は、発見試料セットの配列データから候補低分子RNA (sRNA) 配列を特定することを含む。発見試料セットにわたる候補sRNA配列 (各々個別に入手されたもの) の存在または存在量は、目的とする生物学的状態を予測する (例えば、他の異なる生物学的状態または非疾患対照に対して) か、または疾患進行もしくは治療に対する反応を予測し、これらの候補sRNA配列が本開示の実施形態に従ってさらにフィルタリングまたは選択される。その後、機械学習技法が適用されて、疾患分類器、マルチクラス疾患分類器、及び異なる病状または疾患状態の分類器を含む分類器を構築及びトレーニングする。トレーニングされた分類器を使用して、新たな試料を分類する、例えば、疾患について患者を評価する、または治療的治療法に応答する疾患患者群を予測することができる。

10

【0005】

いくつかの実施形態では、疾患分類器は、マルチクラス予測器である。例えば、マルチクラス予測器は、類似の臨床症状 (例えば、認知症、運動障害など) とともに現れる可能性のある状態、及び／または類似の病理学的注釈 (例えば、疾患ステージ、線維症、炎症など) を有する状態などの目的とする生物学的状態を区別し得る。発見セットにわたる候補sRNA配列、具体的には、それらのバイナリプロファイル (存在もしくは不在) または存在量レベルプロファイルは、本明細書により完全に記載されるように、様々な機械学習モデルを使用して疾患分類器を構築するために使用される。疾患分類器は、分子検出アッセイを使用して、または他の実施形態では、sRNA配列決定を使用して、1つ以上の疾患状態の存在について対象をスクリーニングまたは評価するために使用することができる。

20

【0006】

いくつかの実施形態では、発見セットにおける候補sRNA配列の存在もしくは不在または存在量は、疾患サブタイプを特定または分類するために使用される。疾患サブタイプには、表現型が類似しているが、生物学的経路の異種の調節不全または異種のsRNA生物発生に起因し得る疾患が含まれる。異種のサブタイプは、治療的介入に異なる応答を示し得る。さらに、予測sRNA配列を標的遺伝子及びそれらの生物学的経路にマッピングすることにより、疾患サブタイプの異なる創薬可能な標的及び治療レジメンを解明することができる。疾患サブタイプ分類器は、患者を適切な治療レジメンと一致させるために個別化された医療用途で使用される。疾患サブタイプ分類器はさらに、試験薬の作用機序に依拠して患者募集を調整するために臨床試験設計で使用される。

30

【0007】

様々な実施形態では、本発明は、1つ以上の生物学的状態について対象を評価するための分類器を生成するための方法を提供する。本方法は、発見試料セットにわたって存在する異なるsRNA配列の編集を含むsRNA配列データを提供することと、候補sRNA配列であって、それらの存在もしくは不在または存在量 (例えば、発現レベル) が目的とする生物学的状態の存在、不在、ステージ、または他の特徴と相関する、候補sRNA配列を選択することとを含む。これらの異なるsRNA変種 (例えば、isomiR) は、参照配列または遺伝子座に基づいて統合されず、それ故に、miRNAを分析するための従来のアプローチとは異なる。発見試料セットは、概して、目的とする1つ以上の生物学的状態の存在または不在を表す試料を含み、非疾患対照をさらに含み得る。その後、分類器は、臨床表現型または病理学的ラベルを含む試料メタデータとともに、様々な機械学習モデルを使用して、例えば、トレーニングセットにわたる候補sRNA配列の存在もしくは不在、またはいくつかの実施形態では存在量を使用してトレーニングされる。この態様による分類器は、生物学的状態の存在及び／または不在について対象の試料を評価するためのsRNA特徴を含む。

40

【0008】

様々な実施形態では、発見セット試料は、目的とする1つ以上の生物学的状態に対して

50

陽性または陰性であるとラベル付けされる。かかる実施形態では、本発明は、教師あり機械学習モデルを使用して試料を分類するための sRNA パネル及び特徴を特定することを含む。これらの実施形態では、本発明は、疾患の初期段階を含む同様の症状または病理を呈し得る生物学的状態を正確に分類するための分類器を提供する。例としては、とりわけ、認知症または振戦を呈する CNS 障害及び胃腸炎を呈する障害が挙げられる。いくつかの異なる疾患状態にわたって共有され得る他の疾患表現型が本明細書の他の場所で提供される。

【0009】

さらに他の実施形態では、発見セット試料は、複雑な疾患及び非疾患対照の試料を表す。例えば、複雑な疾患は、発見セットでラベル付けされていない1つ以上の疾患サブタイプを含み得る。いくつかの実施形態では、本明細書に記載の方法は、かかる疾患サブタイプを潜在的に初めて特定する。これらの実施形態では、本発明は、教師なしまたは半教師あり機械学習を使用して、かかる疾患サブタイプの存在または不在について試料を分類するための sRNA 特徴を特定する。したがって、試料をラベル付けするために代理マーカが利用できない場合、または病理医の評価が異なる疾患サブタイプを区別するのに不十分な場合であっても、本発明による候補 sRNA 配列の存在もしくは不在または相対的存在量は、試料を分類するための驚くほど効果的な手段を提供する。いくつかの実施形態では、本明細書に記載の本発明は、さもなければ病理学的に類似しているとみなされる発見試料セットからこれらの疾患サブタイプを特定及び分類するために使用される。

【0010】

機械学習を改善するために、トレーニングセット内のおよそ1億個の異なる配列であり得る異なる sRNA 配列が、事前選択基準を使用して、数千個の候補 sRNA にフィルタリングされる。候補 sRNA 配列は、それらの存在、不在、または存在量が目的とする生物学的状態の存在または不在と相関する程度に基づいて選択することができる。いくつかの実施形態では、少なくとも1つの候補 sRNA 配列は、目的とする生物学的状態に対して陽性である発見試料（例えば、トレーニングセット）にのみ存在し、他のすべての発見試料には不在である。いくつかの実施形態では、少なくとも1つの候補 sRNA 配列は、目的とする生物学的状態（例えば、非疾患対照または他の生物学的状態クラス）に対して陰性である発見試料（例えば、トレーニングセット）にのみ存在し、目的とする生物学的状態に対して陽性としてラベル付けされたすべての試料には不在である。様々な実施形態では、候補 sRNA 配列であって、それらの存在または不在によってトレーニングセットにおける目的とする生物学的状態を個別に予測する、候補 sRNA 配列が選択される。すなわち、候補 sRNA は、トレーニングセット及び／または非疾患対照で表される他の生物学的状態に対して少なくとも1つの生物学的状態の存在または不在を決定するそれらの予測力について個別に選択された配列を含む。いくつかの実施形態では、候補 sRNA 配列は、それらの存在量（例えば、過剰または不足）が目的とする生物学的状態の存在または不在と相関する程度に基づいて、配列データから選択される。

【0011】

いくつかの実施形態では、発見試料セットは、目的とする1つ以上の生物学的状態のステージ、グレード、または他の特徴についてさらにラベル付けされる。これらの実施形態では、候補 sRNA であって、それらのリードカウントが、例えば、疾患ステージまたはグレードなどの疾患活動性と相関する、候補 sRNA が選択され得る。例えば、疾患ステージまたはグレードが進行すると、より高いまたはより低いリードカウントを示す候補 sRNA 配列が選択され得る。すなわち、平均リードカウントは、疾患の後期段階で、またはより高い疾患活動性ととも増加または減少する。あるいは、疾患ステージが低下すると（例えば、処理群において）、処理された対象においてより低いまたはより高いリードカウントを示す候補 sRNA 配列を選択することができる。

【0012】

様々な実施形態では、生物学的状態において増加した配列多様性を有する sRNA ファミリー（例えば、同じシード配列を有する miRNA）が特定される。これらの sRNA

10

20

30

40

50

ファミリー内の sRNA アイソフォームは、分類のための候補 sRNA 配列として選択される。例えば、いくつかの実施形態では、配列変異が、疾患状態において増加する、及び／または疾患状態の重症度とともに増加する、及び／または変異が治療レジメンに応答して正規化するか、または改善され得る sRNA ファミリーが特定され得る。いくつかの実施形態では、機械学習のための sRNA の事前選択は、同じシード配列を有する isomiR の選択に重きが置かれているか、またはエクソソーム内での存在（例えば、3' 非翻訳型ヌクレオチドの存在）に関連する変異を有する isomiR などの他の sRNA 特性に重きが置かれている。

【0013】

sRNA 特徴が選択された後、1つ以上の機械学習アプローチを使用して、機械学習分類器をトレーニングすることができる。いくつかの実施形態では、分類器は、候補 sRNA のパネルの存在もしくは不在または存在量に基づいて、試験セットの試料を分類するように構成されている。パネルのサイズは、関連するクラスの数に依存する。例えば、パネルは、1～約50,000個の sRNA 配列を含み得る。いくつかの実施形態では、パネルは、約4～約200個の sRNA 配列を含む。いくつかの実施形態では、パネルの最大サイズを選択することができる（例えば、約100の sRNA）。いくつかの実施形態では、分類器は、例えば、サポートベクトルマシンアルゴリズム、決定木アルゴリズム、教師なしクラスタリングアルゴリズム、教師ありクラスタリングアルゴリズム、ロジスティック回帰アルゴリズム、混合モデル、隠れマルコフモデル、またはニューラルネットワークアルゴリズムに基づく。

【0014】

トレーニングされた機械学習分類器は、対象由来の生体試料中のパネルにおける sRNA マーカーの存在もしくは不在または存在量を検出し、かつ分類器を適用することによって、疾患状態または疾患サブタイプ（生物学的状態）についての独立した対象の評価に使用することができる。生体試料は、試験される各クラスに関して計算される対応する確率または別の尺度を有する2つ以上のクラスに割り当てることができる。いくつかの事例では、ある特定の閾値を超える関連付けられた確率値を有する割り当てのみが、分類器によって提供され得る。さらに、いくつかの実施形態では、治療推奨またはレジメンは、対象の生体試料の分類の結果に基づいて生成することができる。

【0015】

他の態様では、本発明は、1つ以上の疾患状態または疾患サブタイプについて対象を評価するための方法を提供する。様々な実施形態では、本方法は、対象の生体試料を提供し、sRNA パネルにおける sRNA の存在もしくは不在または存在量を決定することを含む。その後、この sRNA プロファイルが、本開示により調製された疾患分類器を使用して1つ以上の疾患状態または疾患サブタイプ間で対象の状態を分類するために使用される。患者の状態または疾患サブタイプが特定された場合、患者を、疾患状態に適切な治療レジメンとマッチング（すなわち、それを投与）することができる、及び／または臨床試験に組み入れるまたは除外することができる。例えば、いくつかの実施形態では、患者は、調節不全または異常経路を標的とし、かつクラスタ分析に使用されるパネルにおける1つ以上の sRNA（例えば、miRNA）によって標的とされる経路に対応する療法を投与される。

【0016】

様々な実施形態では、対象の試料中の sRNA の存在もしくは不在または存在量は、定量的 PCR アッセイなどの分子診断アッセイによって決定される。例えば、sRNA 配列の検出は、定量的もしくは定性的 PCR、例えば、リアルタイム PCR を含む、プローブの逆転写、増幅、及び／またはハイブリダイゼーションを用いることができる様々な検出プラットフォームのうちの1つに移行する。PCR 検出フォーマットは、いくつかの実施形態では、かつ任意選択で、蛍光標識プローブに関連して、RT-PCR 用のステムループプライマーを用いることができる。

【0017】

10

20

30

40

50

さらに他の実施形態では、対象の試料に存在する sRNA は、本明細書の他の場所に記載の sRNA 配列決定及びアダプタートリミングによって決定または定量化される。sRNA 配列決定は、当該技術分野で既知の標的捕捉（標的濃縮配列決定）を含み得る。

【0018】

本発明の他の態様及び実施形態は、以下の発明を実施するための形態から明白となるであろう。

【図面の簡単な説明】

【0019】

【図1】いくつかの実施形態による、分類器を生成する方法を説明するフローチャートである。

【図2】いくつかの実施形態による、図1の方法を使用して生成された分類器を適用する方法を説明するフローチャートである。

【図3】A～Dは、高度に正確なマルチクラス疾患予測：対照（A）、クローン病（B）、潰瘍性大腸炎（C）、及び憩室症（D）を説明する、様々なIBDクラス及び対照のROC/AUC曲線を示す。

【図4】正確なマルチクラス疾患予測のそれらの真の参照同一性に対する割合を示しているヒートマップを示す。クラスは、クローン病、対照（CTR）、憩室症、及び潰瘍性大腸炎である。

【図5】スパイクイン低分子RNAを使用した正規化の例を説明する。

【図6A】教師あり機械学習と教師なし機械学習との組み合わせを使用して複雑な疾患をサブタイプ化するための方法を説明する。

【図6B】いくつかの実施形態による教師なし機械学習のステップを図6Bに図式的に示す。

【図7】sRNAの事前選択中に共通のシード領域を有するmiRNAバリエーションが凝集したときの分類器性能の向上を示す。

【発明を実施するための形態】

【0020】

本開示は、1つ以上の異なる生物学的状態または1つ以上の疾患サブタイプ（集合的に「生物学的状態」または「疾患状態」と称されることもある）について対象を評価するための疾患分類器を構築するための方法を提供する。本発明は、発見試料セットの配列データから候補低分子RNA（sRNA）配列を特定することを含む。発見試料セット（またはトレーニングセット）にわたる候補sRNA配列（各々個別に入手されたもの）の存在または存在量は、目的とする生物学的状態を予測し（例えば、他の異なる生物学的状態または非疾患対照に対して）、これらの候補sRNA配列が本開示の実施形態に従ってさらにフィルタリングまたは選択される。その後、機械学習技法が適用されて、複数疾患分類器及び疾患サブタイプ分類器を含む疾患分類器を構築及びトレーニングする。トレーニングされた分類器を使用して、新たな試料を分類する、例えば、疾患について患者を評価することができる。

【0021】

いくつかの実施形態では、疾患分類器は、マルチクラス予測器である。例えば、マルチクラス予測器は、典型的には類似の臨床症状（例えば、認知症、運動障害など）とともに現れるか、またはそれを呈する状態などの目的とする生物学的状態を区別し得る。発見セットにわたる候補sRNA配列、具体的には、それらのバイナリプロファイル（存在もしくは不在）または発現レベルプロファイルは、本明細書により完全に記載されるように、様々な機械学習モデルを使用して疾患分類器を構築するために使用される。疾患分類器は、分子検出アッセイを使用して、または他の実施形態では、sRNA配列決定を使用して、1つ以上の疾患状態の存在について対象を評価するために使用することができる。

【0022】

いくつかの実施形態では、sRNAパネルは、疾患サブタイプを特定または分類するために使用される。疾患サブタイプには、表現型が類似しているが、生物学的経路の異なる

10

20

30

40

50

異常もしくは調節不全、または異種の sRNA 生物発生に起因し得る疾患が含まれる。異種のサブタイプは、治療的介入に異なる応答を示し得る。さらに、予測 sRNA 配列を標的遺伝子及びそれらの生物学的経路にマッピングすることにより、疾患サブタイプの異なる創薬可能な標的及び治療レジメンを解明することができる。疾患サブタイプ分類器は、患者を適切な治療法または治療レジメンと一致させるために個別化された医療用途で利用される。疾患サブタイプ分類器はさらに、治療薬の作用機序に応じて患者募集を調整するために臨床試験設計で利用される。

【0023】

様々な実施形態では、本発明は、1つ以上の生物学的状態について対象を評価するための分類器を生成するための方法を提供する。本方法は、発見試料セット（例えば、トレーニングセット）にわたって存在する異なる sRNA 配列の編集を含む sRNA 配列データを提供することと、候補 sRNA 配列であって、それらの存在もしくは不在または存在量が目的とする生物学的状態の存在、不在、ステージ、または他の特徴と相関する、候補 sRNA 配列を選択することを含む。発見試料セットは、概して、目的とする1つ以上の生物学的状態の存在または不在を表す試料を含み、非疾患対照をさらに含み得る。本発明の実施形態による候補 sRNA 配列を減少させた後（以下に記載されるように）、分類器は、生物学的状態ラベルを含む試料メタデータとともに、様々な機械学習モデルを使用して、例えば、トレーニングセットにわたる候補 sRNA 配列の存在もしくは不在、またはいくつかの実施形態では存在量を使用してトレーニングされる。この態様による分類器は、生物学的状態の存在及び／または不在について対象の試料を評価するための sRNA 特徴を含む。

【0024】

図1は、いくつかの実施形態による、分類器を生成する方法100を概略的に説明する。方法100は、少なくとも部分的に、いくつかの実施態様では1つ以上の中央処理装置 CPU（プロセッサとも称される）、1つ以上のグラフィカル処理装置、1つ以上のネットワークインターフェース、ユーザインターフェース、非永続的メモリ、永続的メモリ、及びこれらの構成要素を相互接続するための1つ以上の通信バスを含む好適なシステムで実行することができる。1つ以上の通信バスは、任意選択で、システム構成要素間の通信を相互接続及び制御する回路（チップセットと呼ばれることもある）を含む。非永続的メモリは、典型的には、DRAM、SRAM、DDR RAM、ROM、EEPROM、フラッシュメモリなどの高速ランダムアクセスメモリを含み、永続的メモリは、典型的には、CD-ROM、デジタル多用途ディスク（DVD）もしくは他の光学記憶装置、磁気カセット、磁気テープ、磁気ディスク記憶装置もしくは他の磁気記憶装置、磁気ディスク記憶装置、光ディスク記憶装置、フラッシュメモリデバイス、または他の不揮発性ソリッドステート記憶装置を含む。

【0025】

永続的メモリは、任意選択で、CPUから遠隔に位置する1つ以上の記憶装置を含む。永続的メモリ、及び非永続的メモリ内の不揮発性メモリデバイスは、非一時的コンピュータ可読記憶媒体を含む。いくつかの実施態様では、非永続的メモリまたはあるいは非一時的コンピュータ可読記憶媒体は、方法100を実装するために使用されるプログラム、モジュール、及びデータ構造を記憶する（場合によっては永続的メモリと併せて）。これらのプログラム、モジュール、及びデータ構造は、任意選択のオペレーティングシステム（様々な基本システムサービスを処理し、かつハードウェア依存タスクを実行するための手順を含む）、システムを他のデバイスまたは通信ネットワークに接続するための任意選択のネットワーク通信モジュール（または命令）、及び他のモジュールを含むことができる。例えば、1つ以上のトレーニングデータセットは、システムのメモリに記憶することができる。これらのモジュール、データ、またはプログラム（例えば、命令セット）は、別個のソフトウェアプログラム、手順、データセット、またはモジュールとして実装される必要はなく、それ故に、これらのモジュール及びデータの様々なサブセットは、様々な実装において組み合わせられるか、またはさもなければ再配置され得る。

【0026】

図1のブロック102では、発見試料セットを取得することができる。発見試料セットは、患者試料に一致するsRNA配列データを提供する任意の1つ以上の研究を含む、任意の好適な供給源から得ることができる。発見試料セットは、概して、目的とする1つ以上の生物学的状態の存在または不在を表す試料を含み得、非疾患対照をさらに含み得る。

【0027】

本明細書で使用される場合、「発見セット」または「発見試料セット」は、目的とする1つ以上の生物学的状態を表す試料セットを含み、様々な実施形態では、目的とする生物学的状態のうちのいずれかを表さない対照（非疾患対照）も含む。いくつかの実施形態では、発見試料は、共通の組織に由来し、目的とする生物学的状態は、共通の表現型または病理を有する。目的とする生物学的状態を定義し得る例示的な表現型または病理は、とりわけ、がん性悪性腫瘍、悪性腫瘍浸潤、認知症、認知試験スコア、 β -アミロイドタンパク質堆積、タウタングル、運動制御または振戦、神経変性、脱髄、不安、抑うつ、または双極性障害、頭痛または疲労、不眠症、慢性組織炎症、血管炎、血管透過性、過敏性腸症候群（腹痛、下痢、便秘、疲労、及び／または体重減少を含み得る）、筋肉または関節痛または疲労、胃腸透過性、筋萎縮、自己免疫、組織線維症、身体、精神、または社会性発達障害、リソソーム蓄積異常、グリコーゲン蓄積、無制御細胞増殖、細胞または組織壊死またはアポトーシス、脂肪肝または肝炎、慢性腎疾患、好中球増加症または好中球減少症、骨再形成異常（異常な骨形成または骨吸収を含む）、インスリン抵抗性、高血圧または低血圧、血管収縮、病理学的血管新生またはリンパ管形成、高コレステロール血症、代謝性疾患または肥満、冠動脈疾患、うっ血性心不全、薬物反応または薬物中毒のうちの1つ以上を含み得るが、これらに限定されない。いくつかの実施形態では、発見セットは、本明細書にさらに記載されるように、候補sRNA及び機械学習を選択するためにトレーニングセット及び試験セットにランダムに分けられる。

【0028】

いくつかの実施形態では、発見セットは、目的とする生物学的状態を表し、かつ異種の治療的介入を受けるか、または治療的介入に対する異種の応答を有する患者から得られた試料を含む。かかる実施形態では、試料は、特定の治療的介入、及び／または治療的介入の有効性もしくは毒性についてラベル付けされ得る。

【0029】

様々な実施形態では、発見試料セットにおける試料は、少なくとも2つの生物学的状態、または少なくとも3つの生物学的状態、または少なくとも5つの生物学的状態の存在及び不在を表し（例えば、それらについてラベル付けされ）、これらは、共通の表現型または病理を共有する。いくつかの実施形態では、発見試料セットは、少なくとも4つ、少なくとも5つ、少なくとも7つ、または少なくとも10の生物学的状態の存在及び不在を表す。いくつかの実施形態では、発見試料は、共通の表現型または病理を共有する3～10または3～5の生物学的状態の存在及び不在を表す。

【0030】

いくつかの実施形態では、発見試料セットは、2つ以上の異なる疾患サブタイプを有する疑いのある少なくとも1つの生物学的状態を表す。本明細書で使用される場合、「疾患サブタイプ」とは、類似の疾患症状とともに現れるが、異なるsRNA生物発生、異種のもしくは区別可能な生物学的経路異常もしくは調節不全の生物学的経路を含み得る、及び／または異なる治療様式を必要とし得る生物学的状態の集合を意味する。本開示によれば、理論に拘束されることを意図するものではないが、多くの複雑な疾患が実際にはsRNA生物発生の分析に基づいて有意義に区別することができる疾患の異質の集合であると考えられている。いくつかの実施形態では、本発明は、別の点で病理学的に類似しているとみなされる発見試料セットからこれらの疾患サブタイプを特定する。

【0031】

様々な実施形態では、発見試料セットは、固体組織試料、生体液試料、または培養細胞を含む。例えば、生体液試料は、血液、血清、血漿、脳脊髄液、尿、または唾液であり得

10

20

30

40

50

る。いくつかの実施形態では、発見試料セットは、固体組織生検（例えば、疾患組織のもの）または解剖試料である。いくつかの実施形態では、発見セットは、がん細胞培養物を含み、これらのがん細胞培養物は、いくつかの実施形態では、一次培養物または不死化細胞株であり得る。

【0032】

様々な実施形態では、発見試料セット（またはトレーニングセット）は、目的とする生物学的状態の各々に対して陽性である少なくとも50個の試料、または少なくとも100個の試料、例えば、少なくとも10個の試料または少なくとも20個の試料または少なくとも50個の試料を含む。いくつかの実施形態では、発見試料セットは、少なくとも25個の非疾患または健常対照、または少なくとも50個の非疾患または健常対照、または少なくとも100個の非疾患または健常対照を含む。

10

【0033】

発見セットは、単一の研究から供給される必要はなく、いくつかの実施形態では、分析前変数、例えば、核酸の抽出、sRNAライブラリの調製、及び次世代配列決定を制御するために、発見セットが別個の研究から調達されることが好ましい。「別個の研究」という用語は、異なる場所（例えば、別個の施設）での生体試料の収集、または異なる場所での核酸もしくはsRNAの抽出、及び任意選択で少なくとも1つの他の場所からの異なる核酸もしくはsRNA抽出プロトコルまたは試薬の使用、ならびに異なる場所でのsRNA配列決定ライブラリ調製及び／または配列決定、及び任意選択で少なくとも1つの他の場所からの異なるsRNA配列決定ライブラリ調製及び／または配列決定プロトコルの使用のうちの1つ以上を必要とする。いくつかの実施形態では、別個の研究は、異なる地理（例えば、少なくとも2つの異なる国または大陸）での組織の調達もしくは処理及び／または配列決定を含む。これらの実施形態では、別個の調達、処理、または配列決定は、研究プロトコルの追加の多様性を提供し、患者の遺伝的または民族的差異も提供し得る。いくつかの実施形態では、本明細書に記載されるように、特徴減少のために追加の発見試料が続いて用いられる。

20

【0034】

様々な実施形態では、発見セット試料は、目的とする1つ以上の生物学的状態に対して陽性または陰性であるとラベル付けされる。かかる実施形態では、本発明は、本発明は、教師あり機械学習モデルを使用して試料を分類するためのsRNA特徴を特定することを含む。これらの実施形態では、本発明は、疾患の初期段階を含む同様の症状を呈し得る生物学的状態を正確に分類するための分類器を提供する。例としては、とりわけ、認知症または振戦を呈するCNS障害、胃腸炎症を呈する障害、臓器または組織の炎症または線維症（例えば、特発性肺線維症）を呈する障害、腫瘍形成または細胞悪性腫瘍を特徴とする障害が挙げられる。いくつかの異なる疾患状態にわたって共有され得る他の疾患表現型が本明細書の他の場所で提供される。

30

【0035】

さらに他の実施形態では、発見セット試料は、少なくとも1つの複雑な疾患及び非疾患対照の試料を表す。例えば、複雑な疾患は、発見セットでラベル付けされていないか、または部分的にのみラベル付けされている1つ以上の疾患サブタイプを含み得る。いくつかの実施形態では、本明細書に記載の方法は、疾患サブタイプを潜在的に初めて特定する。これらの実施形態では、本発明は、教師なしまたは半教師あり機械学習を使用して、かかる疾患サブタイプの存在または不在について試料を分類するためのsRNA特徴を特定する。したがって、試料をラベル付けするために代理マーカーが利用できない場合、または病理医の評価が異なる疾患サブタイプを区別するのに不十分な場合であっても、本発明の実施形態による教師あり機械学習によって特定されるパネルにおけるsRNA配列の存在もしくは不在または相対的存在量は、複雑な疾患の試料をサブタイプ化するための驚くほど効果的な手段を提供する。いくつかの実施形態では、本明細書に記載の本発明は、さもないれば病理学的に類似しているとみなされる発見試料セットからこれらの疾患サブタイプを特定及び分類するために使用される。

40

50

【0036】

図1に戻って参照すると、いくつかの実施形態では、ブロック104に示されるように、発見試料セットにおけるsRNA配列決定データが処理され、これはアダプタートリミングを伴う。いくつかの実施形態では、アダプタートリミングは、例えば、全内容が参照により本明細書に組み込まれるPCT/US2018/014856に記載されるように行うことができる。

【0037】

本開示のいくつかの実施形態では、発見試料セットのsRNA配列データが提供される。sRNA配列データは、存在する5'及び3'変異を特定するために、sRNA配列リードから5'及び3'配列決定アダプターをトリミングすることによって処理される。これらの異なる変異は、miRNAを分析するための従来のアプローチである参照配列または遺伝子座に基づいて統合されない。したがって、発見セットからのsRNA配列データは、発見試料にわたる各試料における異なるsRNA配列（すなわち、アイソフォーム）の編集を伴う。

【0038】

sRNAの5'末端及び3'末端での変異を特定するために、例えば、好適な計算モジュール（例えば、ソフトウェアプログラム）を使用して、ユーザ定義の配列決定アダプターを生じるsRNA配列リードからトリミングしてもよい。アダプターは、配列決定プラットフォームに基づいて、ユーザによって定義される。アダプター配列を除去することにより、sRNAアイソフォームを特定し、試料中で定量化することができる。例えば、いくつかの実施形態では、ソフトウェアプログラムは、ユーザ定義の3'アダプターに対応する正規表現を検索し、それらを生じるsRNA配列リードから削除する。

【0039】

いくつかの実施形態では、ユーザ定義の3'アダプターの正規表現は、いくつかの「ワイルドカード」を含む。ワイルドカードは、以下の4つのデオキシリボ核酸：（A）アデニン、（T）チミン、（G）グアニン、または（C）シトシンのうちのいずれか1つであると定義される。しかしながら、ユーザ指定の3'アダプター配列の5'末端の第1のヌクレオチドは改変されておらず（例えば、挿入もしくは欠失とみなされないか、または別様にワイルドカード変化を受けやすいとみなされない）、それ故に、sRNAの3'末端ヌクレオチドが3'アダプターの5'末端ヌクレオチドにライゲーションされる接合部でsRNA配列を保持する。ユーザ指定の3'アダプターの5'末端ヌクレオチドが、ユーザが指定したものと一致しない場合、3'アダプター配列はトリミングされないが、必要に応じて、独立して検証することができる。いくつかの実施形態では、（トリミング後に）少なくとも17ヌクレオチド長を有するsRNAが分析に考慮される。いくつかの実施形態では、約75以下のヌクレオチド長以下、または約50以下のヌクレオチド、または約43以下のヌクレオチド長を有するsRNAが分析に考慮される。

【0040】

いくつかの実施形態では、異なるsRNA配列の存在もしくは不在または存在量が決定される。かかる実施形態では、sRNA配列は、1つ以上の内因性sRNA対照または外因性（すなわち、「スパイクイン」）sRNA対照に対して正規化され得る。いくつかの実施形態では、スパイクインは、（1）合成オリゴヌクレオチド、（2）合成オリゴヌクレオチドの等モルプール、または（3）増加濃度で混合された合成オリゴヌクレオチドのプールであり得る。各実施形態では、スパイクインは、5'及び3'アダプターライゲーション前に試料に添加される。上記の事例の各々では、オリゴヌクレオチドが5'ホスフェート及び3'ヒドロキシルで合成されて、内因性sRNAを模倣する。

【0041】

いくつかの実施形態では、実施例2（図5）により詳細に記載されるように、5'ホスフェート及び3'ヒドロキシルで合成されるある特定の数の外因性オリゴヌクレオチドのプールは、様々な濃度で組み合わせられ、5'及び3'アダプターライゲーション前に各試料に添加され得る。

【0042】

sRNA配列決定は、低分子RNA種、例えば、マイクロRNA (miRNA)、Piwi相互作用RNA (piRNA)、低分子干渉RNA (siRNA)、ヴォールトRNA (vtRNA)、核小体低分子RNA (snRNA)、トランスファーRNA由来の低分子RNA (tsRNA)、リボソームRNA由来の低分子RNA断片 (rsRNA)、低分子rRNA由来のRNA (srRNA)、及び核内低分子RNA (U-RNA)を濃縮し、配列決定する。例えば、sRNA配列決定データを提供する際に、入力材料が低分子RNAについて濃縮され得る。配列ライブラリ構築は、用いられるハイスループット配列決定プラットフォームに応じていくつかのプロセスまたは市販のキットのうちのいずれかを使用して、sRNA濃縮材料を用いて行われる。概して、sRNA配列決定ライブラリ調製は、試料からの全RNAの単離、サイズ分画、配列決定アダプターのライゲーション、逆転写及びPCR増幅、ならびにDNA配列決定を含む。

10

【0043】

より具体的には、いくつかの実施形態では、所与の試料において、すべてのRNA (すなわち、全RNA) が抽出され、単離される。低分子RNAは、サイズ分画によって、例えば、変性ポリアクリルアミドゲル上に単離されたRNAを泳動させることによって、または様々な市販のキットのうちのいずれかを使用することによって単離される。その後、ライゲーションステップは、逆転写中及びPCR増幅中にプライマー結合部位として作用する低分子RNAの両端にアダプターを付加する。例えば、事前にアデニル化された一本鎖DNAの3'アダプター、続いて5'アダプターは、T4 RNAリガーゼ2切断 (T4 Rnl2tr K227Q) などのライゲーション酵素を使用して、低分子RNAにライゲーションされる。これらのアダプターは、異なる5'及び3'末端化学を有するRNA分解産物ではなく、生物学的に処理された低分子RNA (例えば、マイクロRNA) の特徴である5'ホスフェート及び3'ヒドロキシル基を有する低分子RNAを捕捉するように設計されている。その後、sRNAライブラリが逆転写され、PCRによって増幅される。このステップは、アダプターライゲーションRNAを、配列決定反応の鑄型であるcDNAクローンに変換する。固有のヌクレオチドインデックス配列で設計されたプライマーをこのステップで使用して、IDタグ (すなわち、バーコード) を作成して、ライブラリプール及びマルチプレックス配列決定を容易にすることもできる。

20

【0044】

とりわけ、パイロシーケンシング (例えば、454 Life Sciences)、ポリメラーゼベースの合成による配列 (例えば、Illumina)、またはライゲーションによる配列決定 (例えば、ABI Solid Sequencingプラットフォーム) などの任意の次世代配列決定プラットフォームを含む、任意のDNA配列決定プラットフォームを用いることができる。

30

【0045】

図1に戻って参照すると、ブロック106では、候補sRNAがブロック104で処理されたsRNAから選択され得る。いくつかの実施形態では、候補sRNAは、miRNAアイソフォーム、トランスファーRNA由来の断片、及びリボソームRNA由来の断片のうちの1つ以上に限定される。いくつかの実施形態では、これらのmiRNA種、tRNA種、及びrRNA種は、sRNA配列からフィルタリングされ、候補選択のために使用される。いくつかの実施形態では、1つ以上のsRNAは、isomiRである。「isomiR」とは、参照miRNA配列 (例えば、miRBaseによって使用される) に関して変異を有する配列を指す。miRBaseでは、各miRNAは、miRNA前駆体、及び1つまたは2つの成熟miRNA (-5p及び-3p) に関連する。ディープ配列決定は、miRNA生合成における大きな変異を検出し、これは、同じmiRNA前駆体から多くの異なる配列が検出され得ることを意味する。SRNAの6つの主な変異:

40

- (1) 5' 改変 (5' 末端ヌクレオチドが参照sRNA配列の上流または下流にある)、
- (2) 3' 改変 (3' 末端ヌクレオチドが参照sRNA配列の上流または下流にある)、
- (3) 5' ヌクレオチド付加 (ヌクレオチドが参照sRNAの5' 末端に酵素的に付加される)

50

、(4) 3'ヌクレオチド付加(ヌクレオチドが参照sRNAの3'末端に酵素的に付加される)、(5)ヌクレオチド置換(ヌクレオチドがDNAバリエーション(例えば、一塩基多型、挿入、または欠失)に起因して改変される)、(6)ヌクレオチド編集(ヌクレオチドがmiRNA前駆体または成熟miRNAまたは他のsRNA中の1つ以上のヌクレオチド塩基の酵素的改変に起因して改変される)が存在する。いくつかの実施形態では、isomiRの包含は、5'及び3'バリエーションに限定されるが、置換または「スワップ」は限定されない。いくつかの実施形態では、遺伝子間マッピングmiRNAは、候補sRNA選択プロセスにおいて許可されない。

【0046】

いくつかの実施形態では、1つ以上の候補sRNAバリエーションは、スワップなしのトランスファーRNA由来の断片である。いくつかの実施形態では、1つ以上の候補sRNAバリエーションは、スワップなしのリボソームRNA由来の断片である。

【0047】

様々な実施形態によれば、図1のブロック106では、発見セットからのsRNA配列データを使用して、機械学習のための候補sRNA配列を選択する。機械学習を改善するために、発見セット内のおよそ1億個の異なる配列であり得る異なるsRNA配列を、事前選択基準を使用して、数千個の候補sRNAにフィルタリングする。例えば、いくつかの実施形態では、約100,000個以下のsRNA配列が機械学習分析のために選択されるか、または約50,000個以下のsRNA配列、もしくは約10,000個以下のsRNA配列、もしくは約5,000個以下のsRNA配列、もしくは約2,000個以下のsRNA配列が、機械学習モデルを使用して疾患分類器をトレーニングするために選択される。様々な実施形態では、少なくとも約1000個、または少なくとも約2000個、または少なくとも約5000個、または少なくとも約10,000個の候補sRNAが、教師あり機械学習のために事前選択される。いくつかの実施形態では、約2,500〜約60,000個のsRNA配列が、疾患分類器をトレーニングするために事前選択される。

【0048】

図1のブロック106では、いくつかの実施形態では、発見セットからのsRNA配列データが処理された後、候補sRNA配列がsRNA配列データから選択される。候補sRNA配列は、それらの存在、不在、または存在量が、例えば、発見セットに存在する他の状態または非疾患対照と比較して、目的とする生物学的状態の存在または不在と相関する程度に基づいて選択することができる。いくつかの実施形態では、少なくとも1つの候補sRNA配列は、目的とする生物学的状態に対して陽性であり、かつすべての他の発見試料に不在である発見試料(例えば、トレーニングセットの)にのみ存在する。いくつかの実施形態では、目的とする生物学的状態に対して陽性または陰性であり、かつすべての他の発見試料に不在である試料にのみ存在する少なくとも5つ、または少なくとも10個、または少なくとも20個の候補sRNA配列が選択される。いくつかの実施形態では、sRNAは、定義された頻度閾値で疾患試料に存在する(かつ少なくとも1つの他のクラス(例えば、健常対照または他の生物学的状態)のすべての他の試料に不在である)ものについてフィルタリングされる。例えば、sRNAは、目的とする生物学的状態に対して陽性である試料の少なくとも約5%、または少なくとも約10%、または少なくとも約15%、または少なくとも約20%、または少なくとも約25%に存在するものについてフィルタリングされ得る。加えて、sRNA配列は、定義された頻度閾値で対照試料に存在する(かつ少なくとも1つの生物学的状態クラスのすべての試料に不在である)ものについてフィルタリングすることができる。例えば、sRNAは、健常(非疾患)対照である試料の少なくとも約5%、または少なくとも約10%、または少なくとも約15%、または少なくとも約20%、または少なくとも約25%に存在するものについてフィルタリングされ得る。1つのクラスの試料に存在するが、少なくとも1つの他のクラスのすべての試料に不在であると特定されるsRNAマーカーは、本明細書では「バイナリ」マーカーと称されることがある。

10

20

30

40

50

【0049】

様々な実施形態では、候補 sRNA 配列であって、それらの存在または不在によって発見セット、特にトレーニング群における試料セットにおける目的とする生物学的状態を個別に予測する、候補 sRNA 配列が選択される。例えば、候補 sRNA 配列であって、それらの存在または不在が目的とする生物学的状態を予測し、かつトレーニング群において少なくとも 0.01 の p 値を有する、候補 sRNA 配列を選択することができる。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列（例えば、少なくとも 2、3、4、または 5 つの候補 sRNA 配列）であって、その存在または不在が目的とする生物学的状態を予測し、かつトレーニング群において少なくとも 0.0001 の p 値を有する、少なくとも 1 つの候補 sRNA 配列が選択される。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列（例えば、少なくとも 2、3、4、または 5 つの候補 sRNA 配列）であって、その存在または不在が目的とする生物学的状態を予測し、かつトレーニング群において少なくとも 0.000001 の p 値を有する、少なくとも 1 つの候補 sRNA 配列が選択される。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列（例えば、少なくとも 2、3、4、または 5 つの候補 sRNA 配列）であって、その存在または不在が目的とする生物学的状態を予測し、かつトレーニング群において少なくとも 0.00000001 の p 値を有する、少なくとも 1 つの候補 sRNA 配列が選択される。様々な実施形態では、かかる候補 sRNA 配列は、目的とする各生物学的状態に対して選択される。すなわち、候補 sRNA は、発見セット及び／または非疾患対照で表される他の生物学的状態に対して少なくとも 1 つの生物学的状態の存在または不在を決定するそれらの予測力について個別に選択された配列を含む。

【0050】

いくつかの実施形態では、事前選択は、少なくとも部分的に、トレーニング群における候補 sRNA の頻度閾値を選択することによって実施される。すなわち、候補 sRNA は、（トレーニング群において）特定のクラスでは最小頻度で存在しなければならないが、少なくとも 1 つの他のクラスでは指定された周波数閾値未満で存在しなければならない。例えば、候補 sRNA は、（トレーニング群において）特定のクラスで試料の少なくとも約 50%、または特定のクラスで試料の少なくとも約 40%、または特定のクラスで試料の少なくとも約 25%、または特定のクラスで試料の少なくとも約 20%、または特定のクラスで試料の少なくとも約 15%、または特定のクラスで試料の少なくとも約 10%、または特定のクラスで試料の少なくとも約 5% に存在し得る。いくつかの実施形態では、候補 sRNA は、そのクラスで表される各々の独立した研究のためのこの閾値要件を満たす。かかる候補 sRNA に関して、これらは、トレーニング群において少なくとも 1 つの他のクラスで閾値未満、例えば、少なくとも 1 つの他のクラスで試料の約 15% 未満、または少なくとも 1 つの他のクラスで試料の約 10% 未満、または少なくとも 1 つの他のクラスで試料の約 5% 未満で存在する。いくつかの実施形態では、候補 sRNA は、トレーニング群において少なくとも 1 つの他のクラスのすべての試料に不在である。

【0051】

いくつかの実施形態では、候補 sRNA 配列は、例えば、発見セットに存在する他の状態または非疾患対照と比較して、それらの存在量が目的とする生物学的状態の存在または不在と相関する程度に基づいて、配列データから選択される。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列は、目的とする生物学的状態の存在または不在を示す存在量レベルを有する（例えば、存在量は、ある特定の閾値を上回るか、または下回る）。いくつかの実施形態では、疾患試料と非疾患試料との間の相対存在量の差は、少なくとも約 5 倍、または少なくとも約 10 倍、または少なくとも約 100 倍、または少なくとも約 1000 倍、または少なくとも約 10,000 倍である。少なくとも 2 つのクラス間の

存在量の差に基づいて選択される sRNA マーカーは、本明細書では「差次的に発現された」マーカーと称されることがある。

【0052】

いくつかの実施形態では、候補 sRNA 配列であって、それらの存在量に基づいて目的とする生物学的状態の存在または不在を個別に予測する、候補 sRNA 配列が選択される。例えば、候補 sRNA 配列であって、それらの存在量が目的とする生物学的状態の存在または不在を予測し、かつトレーニング群において少なくとも 0.01 の p 値を有する、候補 sRNA 配列を選択することができる。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列（例えば、少なくとも 2、3、4、または 5 つの候補 sRNA 配列）であって、その存在量が目的とする生物学的状態の存在または不在を予測し、かつトレーニング群において少なくとも 0.0001 の p 値を有する、少なくとも 1 つの候補 sRNA 配列が選択される。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列（例えば、少なくとも 2、3、4、または 5 つの候補 sRNA 配列）であって、その存在量が目的とする生物学的状態の存在または不在を予測し、かつトレーニング群において少なくとも 0.000001 の p 値を有する、少なくとも 1 つの候補 sRNA 配列が選択される。いくつかの実施形態では、少なくとも 1 つの候補 sRNA 配列（例えば、少なくとも 2、3、4、または 5 つの候補 sRNA 配列）であって、その存在量が目的とする生物学的状態の存在または不在を予測し、かつトレーニング群において少なくとも 0.00000001 の p 値を有する、少なくとも 1 つの候補 sRNA 配列が選択される。様々な実施形態では、かかる候補 sRNA 配列は、目的とする各生物学的状態に対して選択される。すなわち、候補 sRNA は、トレーニング群において発見セット及び／または非疾患対照で表される他の生物学的状態に対して少なくとも 1 つの生物学的状態の存在を決定するそれらの予測力について個別に選択された配列を含む。

【0053】

いくつかの実施形態では、存在量が増加した sRNA の事前選択は、少なくとも部分的に、候補 sRNA の頻度閾値を選択することによって実施される。すなわち、候補 sRNA は、トレーニング群において少なくとも 1 つの他のクラスの試料で観察される相対的存在量レベル（例えば、平均または中央値）と比較して、特定のクラスでは最小頻度で存在量が有意に高いか、または低くなければならない。例えば、候補 sRNA は、（少なくとも 1 つの他のクラスで観察される sRNA の相対的存在量と比較して）トレーニング群において特定のクラスで試料の少なくとも約 50 %、または特定のクラスで試料の少なくとも約 40 %、または特定のクラスで試料の少なくとも約 25 %、または特定のクラスで試料の少なくとも約 20 %、または特定のクラスで試料の少なくとも約 15 %、または特定のクラスで試料の少なくとも約 10 %、または特定のクラスで試料の少なくとも 5 % で相対的存在量が有意に高い場合があるか、または低い場合がある。いくつかの実施形態では、候補 sRNA は、トレーニング群においてそのクラスで表される各々の独立した研究のためのこの閾値要件を満たす。かかる候補 sRNA に関して、相対存在量の変化は、トレーニング群において少なくとも 1 つの他のクラスで閾値未満、例えば、少なくとも 1 つの他のクラスで試料の約 15 % 未満、または少なくとも 1 つの他のクラスで試料の約 10 % 未満、または少なくとも 1 つの他のクラスで試料の約 5 % 未満で観察される。いくつかの実施形態では、候補 sRNA は、トレーニング群において少なくとも 1 つの他のクラスのいずれの試料でも観察されない特定のクラスの試料における相対的存在量の統計的に有意な変化を有する。

【0054】

候補 sRNA の数は、例えば、線形またはロジスティック回帰モデルを使用してさらに減少させることができる。

【0055】

いくつかの実施形態では、発見試料セットは、目的とする生物学的状態のステージ、グレード、または他の特徴についてさらにラベル付けされる。これらの実施形態では、候補 sRNA であって、それらのリードカウントが、例えば、疾患ステージまたはグレードなどの疾患活動性と（例えば、直接）相関する、候補 sRNA が選択され得る。例えば、疾患ステージまたはグレードが進行すると、より高いリードカウントを示す候補 sRNA 配列が選択され得る。すなわち、平均リードカウントは、疾患の後期段階で、またはより高い疾患活動性ととも増加する。あるいは、疾患重症度が低下すると（例えば、処理群において）、処理された対象においてより低いリードカウントを示す候補 sRNA 配列を選択することができる。したがって、いくつかの実施形態では、少なくとも 1、2、3、4、または 5 つの候補 sRNA 配列であって、その存在または存在量が発見セットにおける試料によって表される生物学的状態を予測し、かつその読み取り数がかかる試料における疾患ステージまたは疾患グレードと相関する、候補 sRNA 配列が選択される。候補 sRNA 配列を選択するために平均リードカウントが望ましい場合、sRNA 配列は、例えば、以下の実施例 2 に記載されるように、内因性 sRNA 及び／またはスパイクイン正規化対照のうちの 1 つ以上を使用して決定することができる。

10

【0056】

様々な実施形態では、目的とする生物学的状態において増加した配列多様性を有する sRNA ファミリーが特定される。これらの sRNA ファミリー内の sRNA 配列が候補 sRNA 配列として選択される。例えば、いくつかの実施形態では、配列変異が、疾患状態において増加する、及び／または疾患状態の重症度とともに増加する、及び／または変異が治療レジメンに応答して正規化するか、または改善され得る sRNA ファミリーが特定され得る。例えば、sRNA 事前選択は、生物学的に関連する配列特徴に基づいて sRNA アイソフォーム（isomiR など）を「ファミリー」にグループ化することを含むことができる。いくつかの実施形態では、配列特徴は、一般に注釈付き sRNA の 5' 末端から 2～8 のヌクレオチドを含む miRNA 「シード配列」である。いくつかの実施形態では、配列特徴は、一塩基多型またはインデルである。これらの sRNA ファミリーは、5' 末端及び 3' 末端での変異について評価される。例えば、変異は、テンプレート化及び／または非テンプレート化ヌクレオチド付加、または 5' 及び／または 3' トリミングを含む 5' 及び／または 3' 変異を含み得、これは、疾患の存在または疾患活動と相関し得る。これらの全ファミリーまたはファミリー内の予測バリエーションは、機械学習の候補として選択することができる。いくつかの実施形態では、これらのファミリーは、目的とする生物学的状態が固有である少なくとも 1 つの sRNA 配列を含む。

20

30

【0057】

いくつかの実施形態では、線形またはロジスティック回帰モデルは、共通のシード配列を有する sRNA アイソフォーム（isomiR）、またはエクソソーム内での存在に関連する特性を有する sRNA（3' 非テンプレート化ヌクレオチド付加、例えば、U 付加など）について重み付けされる。いくつかの実施形態では、共通のシード領域を有する miRNA が候補 sRNA 減少中に（例えば、事前選択フィルタを使用して）凝集する。

【0058】

他のパラメータを使用して、候補 sRNA 配列の選択を補助することができる。例えば、発見試料セットは、本明細書の他の場所に記載されるように、少なくとも 2 つの別個の研究から調達することができ、いくつかの実施形態では、少なくとも 2 つの異なる機関、国、または大陸からの調達を含む。これらの実施形態では、選択された候補 sRNA 配列は各々、各研究からの少なくとも 1 つの試料に存在し（または各研究における頻度閾値を上回り）、それにより、その配列が研究アーチファクトである可能性を低下させる。別個の研究は、異なる場所での生体試料の収集、または異なる場所での核酸もしくは sRNA の抽出、または異なる場所での配列決定ライブラリ調製及び／または配列決定を含み得る。いくつかの実施形態では、異なる研究は、異なる核酸もしくは sRNA 抽出プロトコル、または異なる配列決定ライブラリ調製プロトコル及び／または配列決定プロトコルを用

40

50

いる。

【0059】

様々な実施形態では、sRNA配列は、発見セットにおける閾値平均リードカウントに基づいて事前選択される。例えば、選択されたsRNA配列は、100万リードあたり少なくとも0.1のトリミングされたリードの平均リードカウントを有し得る。いくつかの実施形態では、指定されたフロアを上回り、かつ指定されたシーリングを下回るリードカウントを有するsRNA配列が選択される。いくつかの実施形態では、配列決定深度は、生物学的マトリックスに基づくスライディングスケールである。例えば、固体組織試料は、1試料あたり50,000億〜150,000億リードで配列決定され得、脳脊髄液、血清、及び血漿試料は、1試料あたり150,000億〜350,000億リードで配列決定され得、PAXgene（全血）試料は、1試料あたり350,000億〜550,000億リードで配列決定され得る。より高い深度で配列決定することにより、本方法は、sRNAが組織から出て末梢に入る際のsRNAの希釈を考慮する。

10

【0060】

様々な実施形態では、候補sRNA配列は、ヒトゲノムにマッピングするそれらの能力に基づいて選択される。

【0061】

図1に戻って参照すると、候補sRNAがトレーニングセットから選択されると（機械学習のための所望の数の候補sRNAへの減少を含む）、sRNA特徴が分類器をトレーニングするために特定され得る（ブロック108）。様々な特徴選択または抽出アプローチを使用して、機械学習分類器に適切な特徴を選択することができる。いくつかの実施形態では、特徴は、処理されたデータの形態、例えば、ブロック106で選択されたsRNAのポリヌクレオチド配列（これらは、例えば、アダプタートリミングによって以前に処理されたものである）であり得る。さらに、いくつかの実施形態では、多次元データポイントである特徴を生成することができる。計算負荷を減少させるために、かかる特徴の次元は、例えば、当該技術分野で既知の統計的特徴選択または特徴抽出手順、例えば、主成分分析、非負行列因数分解、特徴ランキングのためのROC曲線、カーネルPCA、グラフベースのカーネルPCA、UMAP、線形判別分析、一般化判別分析を使用して減少させることができる。同様に、いくつかの実施形態では、機械学習技法、例えば、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、オートエンコーダ、サポートベクトルマシン、ベイズネットワーク、または遺伝的アルゴリズムが、多次元データポイントの次元の数を減少させるために使用される。

20

30

【0062】

いくつかの実施形態では、ブロック110を参照すると、sRNA特徴が選択された後、1つ以上の機械学習アプローチを使用して、機械学習分類器をトレーニングすることができる。いくつかの実施形態では、分類器は、（候補sRNAからの）sRNA配列のパネルの存在もしくは不在または存在量に基づいて、試料を分類するように構成されている。いくつかの実施形態では、所望のパネルサイズを選択することができる。一般に、パネルのサイズは、より多くの疾患クラスが存在する場合、より大きい可能性がある。例えば、いくつかの実施形態では、パネルは、約1〜約50,000個のsRNA配列、例えば、1クラスあたり約1〜約200個のsRNA配列、または1クラスあたり約4〜約100個のsRNA配列、または1クラスあたり約4〜約50個のsRNA配列を含む。いくつかの実施形態では、パネルは、1クラスあたり約10〜約100個のsRNA配列、または1クラスあたり約10〜約50個のsRNA配列、または1クラスあたり約10〜約40個のsRNA配列、または1クラスあたり約10〜約30個のsRNA配列を含む。いくつかの実施形態では、パネルは、1クラスあたり約50〜約150個のsRNA配列、または約50〜約100個のsRNA配列を含む。いくつかの実施形態では、総パネルが1〜約500個のsRNA配列、または1〜約200個のsRNA配列、または約4〜約100個のsRNA配列、または約4〜約50個のsRNA配列、または約10〜約100個のsRNA配列、または約10〜約50個のsRNA配列、または約10〜約40

40

50

個の sRNA 配列、または約 10～約 30 個の sRNA 配列、または約 50～約 150 個の sRNA 配列、または約 50～約 100 個の sRNA 配列である最小パネルまたは減少パネルが選択される。いくつかの実施形態では、パネルは、約 100 個以下の sRNA 配列、または 96 個以下の sRNA 配列、または 75 個以下の sRNA 配列、または 50 個以下の sRNA 配列を含む。

【0063】

いくつかの実施形態では、分類器は、サポートベクトルマシンアルゴリズム、決定木アルゴリズム、教師なしクラスタリングアルゴリズム、教師ありクラスタリングアルゴリズム、ロジスティック回帰アルゴリズム、混合モデル、隠れマルコフモデル、またはニューラルネットワークアルゴリズムに基づく。

10

【0064】

様々な実施形態では、分類器は、例えば、パラメトリック／ノンパラメトリック距離測定法、ロジスティック回帰、サポートベクトルマシン、決定木、ランダムフォレスト、ニューラルネットワーク、プロビット回帰、フィッシャー線形判別、単純ベイズ分類器、パーセプトロン、二次分類器、カーネル推定、k 近傍法、学習ベクトル量子化、及び PCA などの教師あり、教師なし、半教師あり機械学習モデルのうちの 1 つ以上を使用してトレーニングされる。例えば、いくつかの実施形態では、分類器は、少なくとも線形サポートベクトルマシンを使用してトレーニングされる。

【0065】

いくつかの実施形態では、分類器は、教師なしクラスタリングモデルである。いくつかの実施形態では、分類器は、教師ありクラスタリングモデルである。クラスタリングは、参照により全体が本明細書に組み込まれる、Duda and Hart, Pattern Classification and Scene Analysis, 1973, John Wiley & Sons, Inc., New York (以下、「Duda 1973」) の 211～256 頁に記載されている。クラスタリング問題は、データセット内の自然なグループ化を見つけ出すことを含む。自然なグループ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（または相違性）を測定する方法が決定される。この測定基準（例えば、類似性尺度）は、あるクラスタ内の試料が、それらが他のクラスタ内の試料よりも互いに類似していることを確実にするために使用される。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が決定される。クラスタリング調査を開始するために、距離関数を定義し、トレーニングセット内のすべての試料対間の距離の行列を計算することができる。距離が良好な類似性尺度である場合、同じクラスタ内の参照エンティティ間の距離は、異なるクラスタ内の参照エンティティ間の距離よりも有意に短くなる。しかしながら、クラスタリングは距離測定基準の使用を必要としない。例えば、非計量的類似性関数 $s(x, x')$ を使用して、2 つのベクトル x 及び x' を比較することができる。従来、 $s(x, x')$ は、 x と x' がどこことなく「類似」している場合に値が大きい対称関数である。

20

30

【0066】

データセット内のポイント間の「類似性」または「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、データの任意のパーティション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パーティションは、データをクラスタリングするために使用される。クラスタリング技術についてのさらなる情報は、各々参照により本明細書に組み込まれる、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 で見つけることができる。本開示で使用する可以使用する特定

40

50

の例示的なクラスタリング技術には、階層的クラスタリング（最近傍アルゴリズム、最遠方（*farthest-neighbor*）アルゴリズム、平均連結アルゴリズム、重心アルゴリズム、または二乗和アルゴリズムを使用した凝集型クラスタリング）、*k* 平均クラスタリング、ファジー *k* 平均クラスタリングアルゴリズム、及びジャーヴィス・パトリッククラスタリングが含まれるが、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、トレーニングセットがクラスタリングされるときにどのクラスが形成されるべきであるかの先入観が課されていない教師なしクラスタリングを含む。いくつかの実施形態では、教師なしクラスタリングを使用して疾患サブタイプを特定することができ、これにより、有意義なパターンを *sRNA* データ内で発見し、研究及び臨床用途で利用することができるようになる。

10

【0067】

いくつかの実施形態では、分類器は、全体が参照により本明細書に組み込まれる、*Agresti, An Introduction to Categorical Data Analysis, 1996, John Wiley & Sons, Inc., New York, Chapter 8* に記載の多カテゴリロジットモデルなどの回帰モデルである。いくつかの実施形態では、分類器は、*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使用する。

【0068】

いくつかの実施形態では、分類器は、メタゲノムリードを扱うためにローゼンらによって開発されたツールなどのナイーブベイズアルゴリズムである（*Bioinformatics 27 (1): 127-129, 2011* を参照されたい）。いくつかの実施形態では、分類器は、*Kamvar et al., Front Genetics 6: 208 doi: 10.3389/fgene.2015.00208, 2015*）に記載のノンパラメトリック法などの最近傍アルゴリズムである。いくつかの実施形態では、分類器は、*McLachlan et al., Bioinformatics 18 (3): 413-422, 2002* に記載のものなどの混合モデルである。いくつかの実施形態では、特に時間的成分を含む実施形態では、分類器は、*Schliep et al., 2003, Bioinformatics 19 (1): i255-i263* に記載のものなどの隠れマルコフモデルである。

20

30

【0069】

主成分分析（*PCA*）アルゴリズムは、参照により本明細書に組み込まれる、*Jolliffe, 1986, Principal Component Analysis, Springer, New York* に記載されている。*PCA* は、参照により本明細書に組み込まれる、*Draghici, 2003, Data Analysis Tools for DNA Microarrays, Chapman & Hall / CRC* にも記載されている。主成分（*PC*）には相関関係がなく、*k* 番目の *PC* が *PC* の中で *k* 番目に大きい分散を有するように順序付けられる。*k* 番目の *PC* は、1 番目の *k-1 PC* に直交するようにデータポイントの投影の変動を最大化する方向として解釈することができる。最初のいくつかの *PC* は、トレーニングセットの変動のほとんどを捕捉する。対照的に、最後のいくつかの *PC* は、多くの場合、トレーニングセット内の残りの「ノイズ」のみを捕捉すると想定される。

40

【0070】

SVM アルゴリズムは、各々参照により全体が本明細書に組み込まれる、*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*

50

Learning Theory, ACM Press, Pittsburgh, Pa., pp. 142-152、Vapnik, 1998, Statistical Learning Theory, Wiley, New York、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、及びFurey et al., 2000, Bioinformatics 16, 906-914に記載されている。分類のために使用される場合、SVMは、ラベル付けされたデータから最大限に離れた超平面でバイナリラベル付けデータトレーニングセット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」技法と組み合わせて機能することができる。特徴空間におけるSVMによって見つけれられる超平面は、入力空間における非線形決定境界に対応する。

【0071】

いくつかの実施形態では、特徴の選択及び機械学習分類器のトレーニング（それぞれ、図1のブロック108及び110）は、図1の矢印109によって概略的に示されるように、分類器が適切な特徴の相互検証及び選択のために使用されるように同じ処理の一部であり得る。トレーニングされた機械学習分類器を使用して、図1のブロック112に示されるように、sRNAパネルを選択することができる。機械学習分類器のトレーニング及びsRNAパネルの選択が同じプロセスの一部であり得ることを理解されたい。また、sRNAパネルに含まれるsRNAのリストは、図1の矢印113によって概略的に示されるように、反復的に調整することができる。

【0072】

いくつかの実施形態では、再び図1のブロック110を参照して、機械学習分類器をトレーニングするために、試料の10%~90%がトレーニングセットにランダムに分けられる。事前選択を使用して、例えば、0.1~100の最小TRPM（100万あたりのトリミングされたリード）を有するトレーニングセットから2,400~60,000個の低分子RNA特徴を選択する。sRNA特徴セットは、回帰モデルを使用して、1クラスあたり1~1,000個のsRNA特徴に減少させることができる。最終sRNA特徴セットは、51%~100%信頼区間の閾値で線形回帰またはサポートベクトルマシンを使用して、試料の残りの10%~90%で試験して、試料を分類するために使用される。精度は、真陽性率、偽陽性率、真陰性率、及び偽陰性率、全精度、ならびに曲線下面積を計算するために標準の受信者操作特性を使用して計算される。「ROC」または「ROC曲線」という用語は、受信者動作特性曲線を指す。ROC曲線は、バイナリ分類器システムの性能のグラフィック表現であり得る。任意の所与の方法の場合、ROC曲線は、様々な閾値設定で特異度に対する感度をプロットすることによって生成され得る。さらに、3つのパラメータ（例えば、感度、特異度、及び閾値設定）のうちの少なくとも1つが提供されると、ROC曲線は、任意の未知パラメータの値または期待値を決定することができる。未知パラメータは、ROC曲線に適合した曲線を使用して決定することができる。例えば、試料中のsRNAのパネルの存在／不在または存在量が提供されると、試験の期待感度及び／または特異度を決定することができる。「AUC」または「ROC-AUC」という用語は、受信者動作特性曲線下の面積を指すことができる。この測定基準は、方法の感度及び特異度の両方を考慮して、方法の診断的有用性の尺度を提供することができる。ROC-AUCは、0.5~1.0の範囲であり得、0.5に近い値は、方法が限定された診断的有用性（例えば、より低い感度及び／または特異度）を有することを示すことができ、1.0に近い値は、方法がより高い診断的有用性（例えば、より高い感度及び／または特異度）を有することを示す。例えば、参照により全体が本明細書に組み込まれる

、Pepe et al., 2004, “Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker,” Am. J. Epidemiol 159 (9): 882–890を参照されたい。診断的有用性を特徴付けるための追加のアプローチには、尤度関数、オッズ比、情報理論、予測値、較正（適合度を含む）、及び再分類測定の使用が含まれる。アプローチの例は、例えば、参照により全体が本明細書に組み込まれる、Cook, “Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction,” Circulation 2007, 115: 928–935に要約されている。本開示の実施形態では、分類器は、バイナリ分類器であり得る（すなわち、例えば、状態を表す2つのクラスを分類することができる）か、または3、4、5、もしくはそれ以上の生物学的状態を分類し得る。いくつかの実施形態では、分類器は、少なくとも3、少なくとも5、少なくとも10、少なくとも15、少なくとも20、少なくとも25、少なくとも30、または少なくとも35の生物学的状態を分類することができる。

10

20

30

40

50

【0073】

いくつかの実施形態では、図1のブロック114に示されるように、機械学習分類器をトレーニングした後、追加の発見試料を評価して、パネルにおける分類器特徴の数またはsRNAの数を減少させることができる（図1の矢印111を参照のこと）。例えば、追加の試料の分類に関する分類器特徴の値を使用して、個々の特徴を重み付けすることができるか、または特徴セットを減少させることができる。いくつかの実施形態では、少なくとも100個のsRNA配列が発見試料に基づいて元の特徴セットに含まれ、この特徴セットは、追加の試料からのsRNA配列データを使用して、75未満、または50未満、または20未満に減少される。様々な実施形態では、追加の発見試料を使用して、sRNAパネルが、いくつかの実施形態では、概して、少なくとも10%、または少なくとも25%、または少なくとも50%減少する。様々な実施形態では、追加の発見試料は、発見セットに関して、異なる収集基準を有する試料、例えば、異なる場所での生体試料の収集、または異なる場所での核酸もしくはsRNAの別個の抽出、または異なる場所での別個のsRNA配列決定ライブラリ調製及び／または配列決定を含む。いくつかの実施形態では、追加の試料は、異なる核酸もしくはsRNA抽出プロトコル、または異なる配列決定ライブラリ調製プロトコル及び／または配列決定プロトコルを用いる。sRNAパネルが選択される前にブロック114（図1）での処理が実行され得ることに留意されたい。

【0074】

トレーニングされた機械学習分類器は、対象由来の生体試料中のパネルにおけるsRNAマーカーの存在もしくは不在または存在量を検出し、かつ分類器を適用することによって、疾患状態についての独立した対象の評価、または疾患サブタイプ（例えば、複雑な疾患のもの）についてのさらなる特定及び評価に使用することができる。図2は、いくつかの実施形態による、疾患もしくは状態または疾患サブタイプについて対象を評価（試験）する方法200の実施形態を説明する。ブロック202では、生体試料を対象（例えば、ヒト）から得ることができる。生体試料は、機械学習分類器をトレーニングするために使用されなかった試料とすることができ、これは、いくつかの実施形態では、試験試料と称することができる。ブロック204では、sRNAデータをsRNAパネル内で検出及び定量化することができ、1つ以上のsRNAパネルにおける生体試料由来のsRNAの存在、不在、または存在量の決定を伴い得る。sRNAは、分子検出アッセイ（定量的もしくは半定量的PCR、または本明細書に記載の他のアプローチなど）を使用して、試料中で検出及び／または定量化され得るか、またはsRNA配列決定及びリードからのアダプター配列のトリミングによって行われ得る。sRNA配列決定は、捕捉RNA配列決定（例えば、捕捉濃縮sRNA配列決定）を伴い得る。sRNAパネルの種類に応じて、いくつかの実施形態では、試料由来のsRNAの存在量が決定される。ブロック206では、トレーニングされた分類器が検出されたsRNAデータに適用されて、図2のブロック2

08を参照して、生体試料をクラスに割り当てることができる。いくつかの実施形態では、生体試料のクラスへの割り当ては、分類器が生体試料をそのクラスに割り当てた（すなわち、生体試料がそのクラスに属すると予測した）信頼性を示すスコアまたは別の尺度と関連付けられ得る。したがって、いくつかの実施形態では、生体試料は、各クラスに関して計算された対応する確率または別の尺度で2つ以上のクラスに割り当てられ得る。いくつかの事例では、ある特定の閾値を超える関連付けられた確率値を有する割り当てのみが、分類器によって提供され得る（例えば、ユーザインターフェース上に示される、ネットワークを介して通信する、及び／または別様にユーザに出力される）。閾値は、様々な方法で、例えば、ユーザ入力に基づいて選択することができる。

【0075】

さらに、いくつかの実施形態では、図2（ブロック210）に示されるように、治療推奨またはレジメンは、対象の生体試料の分類の結果に基づいて生成することができる。

【0076】

様々な生物学的状態に関する分類は、本開示の主題に従って行うことができる。いくつかの実施形態では、分類のための生物学的状態は、中枢神経系の状態である。例えば、いくつかの実施形態では、生物学的状態は、認知症の症状を伴う神経変性疾患である。いくつかの実施形態では、生物学的状態は、アルツハイマー病、パーキンソン病、ハンチントン病、軽度認知障害、進行性核上性麻痺、前頭側頭型認知症、レビー小体型認知症、及び血管性認知症から選択される。これらまたは他の実施形態では、分類のための少なくとも2つの生物学的状態は、運動制御喪失の症状を伴う神経変性疾患である。例えば、いくつかの実施形態では、少なくとも2つの生物学的状態は、アルツハイマー病、進行性核上麻痺、海馬硬化症、レビー小体型認知症、パーキンソン病、ハンチントン病、多発性硬化症、筋萎縮性側索硬化症、及び脊髄性筋萎縮症から選択される。いくつかの実施形態では、分類のための生物学的状態は、多発性硬化症、視神経炎、横断性脊髄炎、及び視神経脊髄炎を含み得る脱髄疾患である。

【0077】

いくつかの実施形態では、発見セットは、疾患ステージ、疾患重症度、薬物応答性、または疾患進行の経過についてラベル付けされる。これらの実施形態は、特に、アルツハイマー病、パーキンソン病、ハンチントン病、多発性硬化症、筋萎縮性側索硬化症、及び脊髄性筋萎縮症などの生物学的状態を評価するために使用される。

【0078】

さらに他の実施形態では、分類のための生物学的状態は、異なる組織または細胞起源のがんである。これらまたは他の実施形態では、発見セットは、薬物感受性または薬物抵抗性についてもラベル付けされえ、これらの特性が対象の試料中で評価されることを可能にする。いくつかの実施形態では、対象由来の生体試料は、腫瘍またはがん細胞生検である。さらに他の実施形態では、生体試料は、血液、血清、または血漿試料である。

【0079】

いくつかの実施形態では、分類のための生物学的状態は、炎症性疾患または免疫疾患である。例示的な炎症性疾患または免疫疾患は、全身性エリテマトーデス（SLE）、強皮症、自己免疫性血管炎、糖尿病（1型または2型）、グレーブス病、アディソン病、シェーグレン症候群、甲状腺炎、リウマチ性関節炎、重症筋無力症、多発性硬化症、線維筋痛症、乾癬、クローン病、潰瘍性大腸炎、憩室症、及びセリアック病のうちの1つ以上を含む。いくつかの実施形態では、発見セットは、組織、血液、血清、血漿、または脳脊髄液などの生体液試料を含む。

【0080】

いくつかの実施形態では、分類のための生物学的状態は、心血管疾患である。いくつかの実施形態では、発見セットは、急性心血管イベントのリスクについてラベル付けされる。かかる実施形態では、疾患分類器は、急性イベントのリスクに対する患者の層別化のための便利なツールを提供する。いくつかの実施形態では、心血管疾患は、冠動脈疾患（CAD）、心筋梗塞、脳卒中、うっ血性心不全、高血圧性心疾患、心筋症、心臓不整脈、先

10

20

30

40

50

天性心疾患、心臓弁膜症、心臓炎、大動脈瘤、末梢動脈疾患、及び静脈血栓症のうちの1つ以上を含む。

【0081】

様々な実施形態では、上述のように、分類器は、例えば、複雑な疾患の疾患サブタイプを特定する。かかる実施形態では、目的とする生物学的状態に関連する全発見試料セット（例えば、非疾患対照を除く）、または目的とする生物学的状態に関連する相当数の試料（例えば、約25%超、または約50%超、または約75%超）が疾患サブタイプについてラベル付けされない。かかる実施形態では、複雑な疾患を分類するために教師あり機械学習を使用して作成されたsRNAパネルが教師なしまたは半教師あり機械学習アプローチで用いられて、疾患サブタイプを特定することができる。これらの実施形態では、sRNAパネルは、クラスタ分析のための強力な手段を提供し、異なるsRNA生物発生パターンを伴う異なる疾患サブタイプを特定する。

10

【0082】

サブタイプ分類器で使用されるsRNA（例えば、miRNA）のパネルを使用して、異なる疾患サブタイプの異なる創薬可能な標的または経路を特定することができる。sRNAをmRNA標的及び経路にマッピングする際に使用される生物学的データベースは、参照により全体が本明細書に組み込まれる、Zou D, et al., Biological Databases for Human Research, Genomics Proteomics Bioinformatics, 13 (2015) 55-63に記載されている。例には、とりわけ、Database of Essential Genes (DEG)、Kyoto Encyclopedia of Genes and Genomes (KEGG)、KEGG Pathways、GeneCards、PolymiRTS (miRNA及びそれらの標的部位の多型)、ChIPBase、miRTarBase、miRWalk、piRNABank、Database of Interacting Protein (DIP)、及びMolecular Interaction Database (MINT)が挙げられる。

20

【0083】

例えば、sRNAパネルにおける1つ以上のmiRNAバリエーションによって標的とされる遺伝子を伴う生物学的経路を特定することができる。いくつかの実施形態では、生物学的経路は、対応する予測sRNAバリエーションを標的遺伝子にマッピングすることによって、疾患サブタイプ毎に特定される。いくつかの実施形態では、予測isomiRが注釈付きmiRNAにマッピングされ、注釈付けされたmiRNAを使用して、異常なsRNA生物発生によって影響を及ぼされるまたは調節不全にされる潜在的な経路を特定する。Bhattacharya A, et al., PolymiRTS Database 3.0: Linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways, Nucleic Acids Res. 2014; 42: D86-D91を参照されたい。

30

【0084】

図6Aを参照すると、いくつかの実施形態では、本発明は、1つ以上の生物学的状態を分類するために、かつそれらの生物学的状態のうちの少なくとも1つをサブタイプ化するために1つ以上のsRNAパネルを生成する（例えば、複雑な疾患の場合）。特発性肺線維症（IPF）に関して、図6Aの実施形態が実施例3で説明される。

40

【0085】

図6Aにおいて、プロセス（または方法）600は、複数の試料または対応するsRNA配列データ（本明細書に記載のアダプタートリミングされたもの）及び試料メタデータが取得されたときに開始することができる。複数のブートストラップセットが試料から作成され、分析されて、sRNAシグネチャを作成することができる。図6Aのブロック602を参照すると、プロセス600は、試料をトレーニング群と交差検証群または試験群に分けることによってブートストラップセットを作成する。試料は、ランダムに分けるこ

50

とによって、または別の方法で、トレーニング群と試験群に分けることができる。

【0086】

モデルを作成するために、ブロック604で、バイナリまたは差次的に発現された sRNA がトレーニング群で選択され（サブブロック601）、例えば、エラスティックネット（例えば、線形回帰とロジスティック回帰とリッジ回帰との組み合わせ）を使用して（かつ本明細書の他の場所に記載されるように）、候補 sRNA の数を減少させる（サブブロック603）。サポートベクトルマシン（SVM）は、サブブロック605で、減少した sRNA セットを使用してトレーニングされる。図6Aのブロック606を参照すると、SVMが試験群（相互検証群）に対して試験される。図6Aのブロック608を参照して、受信者操作特性（特異度、感度、精度など）がモデル性能を評価するために計算される。

10

【0087】

図6Aに示されるように、ブロック602～608での動作の処理は、ブロック611として集合的に描写される。決定ブロック610では、ブロック611でのステップがN回繰り返されるようにブロック611での処理の回数（繰り返し回数とも称される）がNに達したかが決定される。Nは、事前選択することができるか、ユーザ入力に基づいて設定することができるか、または他の方法で定義することができる。ブロック602～608での処理（ブロック611）がN回繰り返された（「はい」）とブロック610で決定された場合、プロセス600は、受信者動作特性がN回のブートストラップにわたって平均化されるブロック612に進む。

20

【0088】

ブロック614を参照すると、N個のモデルのX%超で選択された sRNA と係数が組み合わせられて、sRNA シグネチャを生成する。いくつかの実施形態では、N個のモデルの25%超で選択された sRNA と係数が組み合わせられて、sRNA シグネチャを生成するが、Xが異なる値であり得ることが認識される。

【0089】

ブロック616を参照すると、任意選択で、生物学的状態（複雑な疾患クラス）の試料に対する sRNA パネル（そのシグネチャにおける sRNA ）の教師なしまたは半教師ありクラスタリングを使用して、異なる疾患サブタイプを特定することができる。ブロック618を参照すると、任意選択で、疾患サブタイプに関連する生物学的経路は、miRNA シード領域及び標的 mRNA の分析によって特定される。これらのステップは、図6Bにさらに図式的に示される。

30

【0090】

ブロック602～608での処理（ブロック611）がN回繰り返されていない（「いいえ」）とブロック610で決定された場合、プロセス600はブロック602に戻り、そこで別のブートストラップセットが作成され、ブロック604、606、及び608での処理が繰り返される。

【0091】

図6Bは、複雑な疾患の試料をサブタイプ化するための本開示の実施形態による sRNA パネルを用いた教師なし学習のプロセス700を図示する。図6Bに示されるように、ブロック704で、プロセス700は、低分子RNA発現値を使用して試料間の距離を計算することを含む。ブロック706では、試料が凝集型または分裂型クラスタリングによってクラスタリングされる。ブロック708では、クラスタラベルが試料に割り当てられる。ブロック710では、任意選択で、クラスタが主成分分析によって検証される。ブロック712では、任意選択で、割り当てられたクラスタラベル上のモデルをトレーニングすることにより、クラスタが教師あり学習（上述のもの）によって検証される。ブロック714では、任意選択で、疾患サブタイプを分類するために使用されるパネルにおける miRNA のシード配列を使用して、標的メッセンジャーRNAが予測される。ブロック710、712、及び714での処理を他の順序で行うことができるため、それらのブロックでの処理の順序がほんの一例として示されていることを理解されたい。

40

50

【0092】

他の態様では、本発明は、1つ以上の疾患状態または疾患サブタイプについて対象を評価するための方法を提供する。様々な実施形態では、本方法は、対象の生体試料を提供し、sRNAパネルにおけるsRNAの存在または不在を決定することを含む。その後、このsRNAプロファイルが、本開示により調製された疾患分類器を使用して1つ以上の疾患状態または疾患サブタイプ間で対象の状態を分類するために使用される。

【0093】

患者の状態または疾患サブタイプが特定された場合、患者を、疾患状態に適切な治療レジメンとマッチング（すなわち、それを投与）することができる、及び／または臨床試験に組み入れるまたは除外することができる。例えば、いくつかの実施形態では、患者は、調節不全または異常経路を標的とし、かつクラスタ分析に使用されるパネルにおける1つ以上のsRNAによって標的とされる経路に対応する療法を投与される。

【0094】

様々な実施形態では、対象の試料中のsRNAの存在もしくは不在またはレベルは、定量的PCRアッセイなどの分子診断アッセイによって決定される。例えば、sRNA配列の検出は、定量的もしくは定性的PCR、例えば、リアルタイムPCRを含む、プローブの逆転写、増幅、及び／またはハイブリダイゼーションを用いることができる様々な検出プラットフォームのうちの1つに移行する。PCR検出フォーマットは、いくつかの実施形態では、かつ任意選択で、蛍光標識されたプローブと関連して、RT-PCRのためのステムループプライマーを用いることができる。

【0095】

一般に、リアルタイムポリメラーゼ連鎖反応（qPCR）は、PCR中、つまり、リアルタイムで、標的DNA分子の増幅をモニタリングする。リアルタイムPCRは、定量的かつ半定量的に使用され得る。リアルタイムPCRでPCR産物を検出するための2つの一般的な方法は、（1）任意の二本鎖DNAにインターカレートする非特異的蛍光色素（例えば、SYBR Green（IまたはII））、及び（2）オリゴヌクレオチドからなる配列特異的DNAプローブであって、オリゴヌクレオチドがその相補的配列（例えば、TAQMAN）とのプローブのハイブリダイゼーション後にのみ検出を可能にする蛍光レポーターで標識されている、配列特異的DNAプローブである。

【0096】

いくつかの実施形態では、アッセイフォーマットは、TAQMANリアルタイムPCRである。TAQMANプローブは、定量的PCRの特異性を増加させるように設計された加水分解プローブである。TAQMANプローブの原理は、相補的標的配列へのハイブリダイゼーション中に、フルオロフォアベースの検出で二重標識プローブを切断するための、Taqポリメラーゼの5'から3'のエキソヌクレアーゼ活性に依存する。TAQMANプローブは、フルオロフォア及びクエンチャで二重標識されており、フルオロフォアがTaqエキソヌクレアーゼ活性によってオリゴヌクレオチドプローブから切断される際に、フルオロフォアシグナルが、検出される（例えば、もはやシグナルはラベルの近接によってクエンチされない）。他の定量的PCR法におけるように、得られた蛍光シグナルは、PCRの指数関数的ステージ中に産物の蓄積の定量的測定を可能にする。TAQMANプローブフォーマットは、検出の高い感度及び特異性を提供する。

【0097】

いくつかの実施形態では、試料に存在するsRNAは、特異的プライマー、例えば、1つ以上のステムループプライマーを使用して、cDNAに変換される。次いで、cDNAの増幅は、例えば、蛍光レポーター分子からのシグナルを検出することによってリアルタイムで定量化され得、シグナル強度は、各増幅サイクルでのDNAのレベルと相関する。

【0098】

あるいは、パネルにおけるsRNA、またはそれらのアンプリコンは、ハイブリダイゼーションによって検出される。例示的なプラットフォームは、表面プラズモン共鳴（SP

10

20

30

40

50

R) 及びマイクロアレイ技術を含む。検出プラットフォームは、簡便な試料処理及び sRNA 検出のために、いくつかの実施形態では、マイクロフルイディクスを使用できる。

【0099】

一般に、試料中の sRNA の存在を決定するための任意の方法が、用いられ得る。かかる方法は、さらに、核酸配列ベースの増幅 (NASBA)、フラップエンドヌクレアーゼベースのアッセイ、同様に、分岐 DNA による直接 RNA キャプチャ (Quantigene (商標))、Hybrid Capture (商標) (Digene)、または nCounter (商標) miRNA 検出 (Nanosttring) を含む。アッセイフォーマットは、miRNA 及び他の sRNA の存在を決定することに加えて、とりわけ、固有のシグナル強度変動の制御も提供できる。かかる制御は、例えば、バックグラウンドシグナル強度及び／または試料処理、及び／またはハイブリダイゼーション効率のための制御、同様に、患者試料中の sRNA を検出するための他の望ましい制御 (例えば、集合的に「正規化制御」と呼ばれる) を含み得る。

10

【0100】

いくつかの実施形態では、アッセイフォーマットは、Invader (商標) アッセイ (Third Wave Technologies) などのフラップエンドヌクレアーゼベースのフォーマットである。インベダー法を使用する場合、標的部位の 3' 領域に特異的な配列を含むインベダープローブと、テンプレートの標的部位の 5' 領域に特異的な配列及び無関係のフラップ配列を含む一次プローブと、が調製される。次いで、クリベースは、これらのプローブ、標的分子、同様に、フラップ配列に相補的な配列と、蛍光色素及びクエンチャの両方で標識された自己相補的配列と、を含む FRET プローブの存在下で、作用することが可能である。一次プローブがテンプレートとハイブリダイズする際に、インベダープローブの 3' 末端は標的部位を貫通し、この構造はクリベースによって切断され、フラップの解離をもたらす。フラップは FRET プローブに結合し、蛍光色素部分は、蛍光の発光をもたらすクリベースによって切断される。

20

【0101】

いくつかの実施形態では、RNA は、検出のための sRNA 処理の前に試料から抽出される。RNA は、例えば、RNA Methodologies, A Laboratory guide for isolation and characterization. 2nd edition, 1998, Robert E. Farrell, Jr., Ed., Academic Press に記載の様々な標準手順を使用して精製され得る。加えて、mirVANA (商標) Paris miRNA 単離キット (Ambion)、miRNeasy (商標) キット (Qiagen)、MagMAX (商標) キット (Life Technologies)、Pure Link (商標) キット (Life Technologies) を含む、低分子量 RNA の単離のための市販の製品と同様に、様々なプロセスがある。例えば、低分子量の RNA は、ガラスファイバーフィルタ上での精製が後に続く有機抽出によって単離され得る。miRNA を単離するための代替方法は、磁気ビーズへのハイブリダイゼーションを含む。代替的には、検出のための miRNA 処理 (例えば、cDNA 合成) は、生体液試料において、すなわち、RNA 抽出ステップを伴わずに実施され得る。

30

40

【0102】

一般に、アッセイは、各アッセイが、注釈付き配列及び／または他の非予測 iso-miR よりも sRNA (例えば、isomiR) に対して少なくとも 80%、または少なくとも 85%、または少なくとも 90%、または少なくとも 95%、または少なくとも 98% 特異的であるように構築され得る。注釈付き配列は、miRBase を参照して決定され得る。例えば、sRNA 予測因子特異的リアルタイム PCR アッセイを調製する際、PCR プライマー及び蛍光プローブは、調製され得、それらの特異性のレベルについて試験され得る。二環式ヌクレオチド (例えば、LNA、cET、及び MOE) または他のヌクレオチド修飾 (塩基修飾を含む) は、検出の感度または特異度を増加させるためにプローブに用いられ得る。

50

【0103】

さらに他の実施形態では、対象の試料に存在する sRNA は、本明細書の他の場所に記載されるように、sRNA 配列決定及びアダプタートリミングによって決定または定量化される。sRNA 配列決定は、捕捉オリゴヌクレオチドプローブを用いて増幅及び／または配列決定のために sRNA 標的を濃縮／捕捉することができる捕捉 RNA 配列決定を用いることができる。WO2011/06967 を参照されたい。

【0104】

本明細書で使用される場合、文脈上他の意味に解すべき場合を除き、「約」という用語は、関連する数値の $\pm 10\%$ を意味する。

【0105】

本発明の他の態様及び実施形態は、以下の実施例によって明らかになるであろう。

【実施例】

【0106】

実施例 1：炎症性腸疾患（IBD）のマルチクラス疾患分類器の構築。

特定の sRNA 分子の存在または不在に基づいて IBD 試料を分類する疾患分類器を構築するために、sRNA パネルを、クローン病、潰瘍性大腸炎、及び憩室症などの、目的の異なる疾患状態を表す様々なトレーニングセットにおける配列データから決定した。

【0107】

試料

すべての試料は、それらのそれぞれの施設内審査委員会（IRB）の承認に従って収集され、無制限の使用について患者の同意を有する。データを、電子診療記録及びカルテ審査から収集した。臨床データは、年齢、性別、人種、民族性、体重、ボディマス指数、喫煙歴、アルコール使用歴、及び家族病歴などの情報を含む。疾患関連データは、診断、炎症性腸疾患（IBD）診断時の年齢、現在及び以前の薬物療法、併存症、大腸全摘及び回腸囊肛門管吻合術（IPAA）時の年齢、同様に、パウチ年齢、イレオストミーの閉鎖からの、またはパウチ手術からの時間（これらの処置を受けている患者から該当する場合）などの情報を含む。

【0108】

生検を、結腸上皮から採取した。手術不能の潰瘍性大腸炎（IUC）、手術可能な潰瘍性大腸炎（OUC）、クローン病（CD）、憩室症（DD）、ポリープ／ポリポーシス（PP）、鋸歯状ポリープ／ポリポーシス（SPP）、結腸癌（CC）、直腸癌（RC）を、臨床的、内視鏡的、組織学的、及び画像研究に従って定義した。さらなる組み入れ基準は、CD 患者についての回腸炎の存在と、内視鏡検査によって見られ、IUC 患者についての組織学によって確認された正常な回腸末端を有することであった。定期的なスクリーニングのために結腸内視鏡検査を必要とし、内視鏡検査及び／または組織学によって非疾患腸組織を有するとして検証された個体を、正常対照としてラベル付けした。

【0109】

すべての生検を、最低 2 名の施設内の IBD 訓練された病理医によって評価し、コンセンサススコア及び診断を、臨床及び業界標準の診断プロトコルに従って提供した。簡単に説明すると、活性炎症特性を、好中球浸潤（0～3）及び潰瘍形成の面積（0～3）に従ってスコアリングし、各試料を、非活性、陰窩炎、陰窩膿瘍、多数の陰窩膿瘍（3 超／高倍率視野）、及び潰瘍形成に分類した。元の Geboes スコア（OGS）または簡略化 Geboes スコア（SGS）を、UC を分類するために使用した。クローン病活性指数（CDAI）及びクローン病内視鏡的重症度指数（CDEIS）を、CD を分類するために使用した。ヒンチェイ分類を、DD を特徴付けるために使用した。大腸がん、ポリープ、及び鋸歯状ポリープを、Multi-Society Task Force on Colorectal Cancer（CRC）の最新の勧告に従って分類した。

【0110】

使用した IBD 試料の概要を、以下に示す。

10

20

30

40

50

【表 1】

表 1

診断	正常	クローン病	潰瘍性大腸炎	憩室症
組織型	結腸上皮	結腸上皮	結腸上皮	結腸上皮
数	64	35	139	20
性別(女性:男性)	26:38	14:21	50:89	6:14
サンプリング時での年齢、年、平均値±SD(範囲)	56.4±13.5(26-82)	36.6±15.8(15-76)	45.5±14.1(32-57)	44.9±10.6(31-69)
IBD 診断時での年齢、年、平均値±SD(範囲)	該当せず	30.4±12.1(18-48)	32.1±11.6(16-51)	26.2±8.7(21-55)
IBD 持続期間、年、平均値±SD(範囲)	該当せず	13.3(3-53)	10.5(3-28)	12.6(25-53)
アシュケナージ起源	5	2	9	1
非アシュケナージ起源	53	31	120	17
混合起源	6	2	10	2
一度も喫煙者でない	56	28	122	19
過去の喫煙者	5	2	10	1
現在の喫煙者	3	5	7	0
ボディマス指数、平均値±SD(範囲)	25.5±2.9(17-30)	27.1±5.3(18-31)	25.8±6.1(15-41)	23.3±5.2(18-40)
IBD の家族歴	2	3	8	1
ステロイド曝露	該当せず	該当せず	110	該当せず
重症度スコア(B1:B2:B3)	該当せず	7:6:8	該当せず	該当せず

10

20

30

40

【0111】

IBDに関連する疾患クラスに対する低分子RNA予測因子を特定するために、低分子RNA配列決定データを、GEOデータベースからダウンロードし、ディスカバリーセットとして使用した。低分子RNA配列決定データを、クローン病（GSE66208）、潰瘍性大腸炎（GSE114591）、憩室症（GSE89667）、及び正常／対照（GSE118504）についてのジオデータベース研究からダウンロードした。

【0112】

データファイルを、Centos用のSRAツールキットv2.8.0を使用して、sra形式からfastq形式に変換し、fastq形式のファイルを、参照によりそ

50

の全体が本明細書に組み込まれる2018年1月23日に出願された米国特許第2018/0258486号及び国際出願第PCT/US2018/014856号に記載されるように処理した。具体的には、すべての、fastqデータファイルを、(Regex)正規表現ベースの検索及びトリムアルゴリズムを使用してアダプター配列をトリミングすることによって処理し、5' TGGAAATTC TCGGGTGCCAGGA A 3' (配列番号1) (最大15ヌクレオチドの3'末端ランケーションを含有する)を、3'アダプター配列を特定するために入力し、Regex検索のための2のレーベンシュタイン距離または5.のハミング距離のパラメータは、ユーザ指定の検索語の1番目のヌクレオチドが、ヌクレオチド挿入、欠失、及び/または交換に関して未変更であるように要求する。

10

【0113】

マルチクラス分類器を構築するために、試料の60%をトレーニングに使用し、試料の40%を試験に使用して、試料を24の独立したトレーニング群及び試験群にランダムに分けた。事前選択により、1つのクラスに存在し、かつ他の3つのクラスのうちの(少なくとも)1つのすべての試料に不在である最大20,000個のsRNAが選択された。事前選択されたsRNAは、その特定のクラスにおいて25%、及びそのクラス内の各研究において少なくとも25%の最小頻度で存在しなければならなかった。sRNAはまた、試験試料(例えば、すべての試料からトレーニングセットを差し引いたもの)に25%の最小頻度で存在しなければならなかった。エラスティックネットを使用した特徴減少により、sRNAファミリー用のフィルタ(シード配列または非テンプレート3'付加など)を使用せずに、sRNAの数を1クラスあたり126未満に減少させた。0.5の閾値でサポートベクトルマシンを使用して試験を実行した。

20

【0114】

クラスあたりの評価指標

クラスあたりの評価指標を、疾患クラスを特定するために最も重要であるマーカーを特定するために、各クラスに対して決定した。sRNAパネルを、目的の異なる疾患状態を表す様々なトレーニングセットにおける配列データから決定した。疾患クラスの低分子RNA予測因子を含有する特異的バイオマーカーパネルを、以下のように特定した：

- ・対照(健常個体/「正常」個体)：表2(炎症性腸疾患の対照(「正常」個体)に対する結腸上皮組織からのsRNAバイオマーカーのパネルを示す)、
- ・クローン病：表3(クローン病に対する結腸上皮組織からのsRNAバイオマーカーのパネルを示す)、
- ・潰瘍性大腸炎：表4(潰瘍性大腸炎に対する結腸上皮組織からのsRNAバイオマーカーのパネルを示す)、及び
- ・憩室症：表5(憩室症に対する結腸上皮組織からのsRNAバイオマーカーのパネルを示す)。

30

【0115】

教師あり、ノンパラメトリック、ロジスティック回帰機械学習モデルを使用することによって、最終的な選択マーカー計数を、128から100に低減した。分類モデルの性能を評価するために、ROC/AUC曲線を、クラスあたりに特定されたマーカーの各セットについて取得したが、ここでROCは確率曲線であり、AUCは分離可能性の程度または尺度を表す。ROC曲線は、偽陽性率に対して真陽性率でプロットされる。ROC/AUC曲線を、上記のように、様々なIBDクラス及び対照について確立し、これらを図3A、図3B、図3C、及び図3Dに示す。

40

【表 2 - 1】

表 2．正常個体に関連する結腸上皮組織中の特定された sRNA バイオマーカー。

配列番号	マーカー	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thisbl	otherbl
2	GCTGATTGTCACGTTCTGATT	0.61173	0.11392	hsa-mir-5701	(0:0)	(GC:)	(1:T>C)	0.9	2.305	0.767
3	GCCCCTGGCCTATCCTAGA	-0.50514	0.07172	hsa-mir-331-3p	(0:-1)	(:)	()	1	1.473	2.614
4	AGTTCTTCAGTGGCAAGCT	-0.43217	0.12976	hsa-mir-22-5p	(0:-3)	(:)	()	0.7	-0.639	0.822
5	ACCCTGTAGAACCGAATTTGTA	0.23477	0.08481	hsa-mir-10b-5p	(1:-1)	(:A)	()	0.5	3.3	1.212
6	TAGGTAGTTTCTCTGTTGGAT	0.17757	0.0569	hsa-mir-196a-5p	(0:-1)	(:AT)	(11:A>C)	0.8	0.15	-0.592
7	ACCCTGTAGATCTGAAATTTGT	0.16483	0.10074	hsa-mir-10b-5p	(1:-1)	(:)	(10:A>T,12:C>T)	0.3	0.782	-0.34
8	TGAGATGAAGCTGTAGCTC	0.16362	0.03238	hsa-mir-4770	(0:0)	(:C)	(8:C>A,9:A>G)	0.8	0.779	-0.308
9	TACCCTGTAGAACCGAATTGGT	0.15816	0.04547	hsa-mir-10b-5p	(0:-1)	(:)	(19:T>G)	0.7	1.483	-0.398
10	ACCCTGTAGAACCGAATTTGG	0.1312	0.04783	hsa-mir-10a-5p	(1:-2)	(:G)	(10:T>A)	0.5	0.875	-0.605
11	TACAGTCTACAGCCATGGTCG	-0.12465	0.06087	hsa-mir-132-3p	(0:0)	(:)	()	0.6	3.56	4.436
12	AGTTCTTCAGTGGCAAGCTT	-0.11012	0.05699	hsa-mir-22-5p	(0:-2)	(:)	()	0.3	-0.394	1.187
13	TACCCTGTAGAACCGAATTTGG	0.09977	0.03596	hsa-mir-10b-5p	(0:-2)	(:G)	()	0.5	4.121	1.664
14	CAGTGCAATGATGAAAGGGCAT	-0.08933	0.05037	hsa-mir-130a-3p	(0:0)	(:)	(10:T>A,12:A>G)	0.3	0.717	2.623
15	TACCCTGTAGAACCGAATTTA	0.07544	0.04788	hsa-mir-10b-5p	(0:-3)	(:A)	()	0.4	2.698	0.845
16	TACAGTTGTTCAACCAGTTACT	-0.07464	0.05019	hsa-mir-582-5p	(1:0)	(:)	()	0.2	-0.358	0.671

10

20

30

40

50

【表 2 - 2】

配列番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbi	otherlbi
17	ACCTGTAGAACCGAATTTGGG	0.06375	0.06375	hsa-mir-10a-5p	(1:0)	(:)	(10:T>A,20:T>G)	0.1	0.747	-0.188
18	TACCCTGTAGACCGAATTTGT	0.05883	0.03032	hsa-mir-10b-5p	(0:-1)	(:)	(10:A>G)	0.4	1.962	-0.355
19	TGGCAGTGTCTTAGCTGGTT	-0.05794	0.04762	hsa-mir-34a-5p	(0:-2)	(:)	()	0.2	-0.482	1.044
20	ACCTGTAGAACCGAATTTA	0.04848	0.03233	hsa-mir-10a-5p	(1:-3)	(:A)	(10:T>A)	0.2	0.32	-0.63
21	ACCTGTAGAACCGAATTTGT	0.04605	0.04605	hsa-mir-10b-5p	(1:-1)	(:T)	()	0.1	1.076	-0.146
22	TACCCTGTAGATCCGATTTTGT	0.04078	0.01861	hsa-mir-10b-5p	(0:-1)	(:)	(11:A>T,16:A>T)	0.4	1.192	-0.283
23	TACCCTGTAGAACCGAGTTTGT	0.03972	0.03306	hsa-mir-10b-5p	(0:-1)	(:)	(16:A>G)	0.2	2.752	0.399
24	TTCAAGTAATCCAGGATAGGCC T	0.03965	0.03658	hsa-mir-26a-5p	(0:-1)	(:CT)	()	0.2	0.841	-0.548
25	TACCCTGTAGAACCGAATTTAT	0.03939	0.03051	hsa-mir-10b-5p	(0:-1)	(:)	(20:G>A)	0.2	1.886	0.183
26	TACCCTGTAGAACCGGATTTG	0.03714	0.02781	hsa-mir-10b-5p	(0:-2)	(:)	(15:A>G)	0.2	0.166	-0.663
27	TATTGCACCTTGTCCTGGCCTGTA GC	0.03206	0.03206	hsa-mir-92a-3p	(0:2)	(:C)	(22:G>A)	0.1	0.533	-0.546
28	ACCTGTAGATCTGAATTTGTGA	0.02789	0.02789	hsa-mir-10a-5p	(1:0)	(:A)	(12:C>T)	0.1	0.267	-0.681
29	CACTAGATTGTGAGCTCCT	0.02652	0.02652	hsa-mir-28-3p	(0:-3)	(:)	()	0.1	2.028	0.439
30	TACCCTGTAGTACCGAATTTGT	0.02641	0.02641	hsa-mir-10b-5p	(0:-1)	(:)	(10:A>T)	0.1	1.227	-0.21
31	CAGTGCAATGTTAAAGGGCAA	-0.026	0.01733	hsa-mir-130b-3p	(0:-1)	(:A)	(10:A>T,12:G>A)	0.2	-0.212	1.183
32	CTGACCTATGATTTGACAGCC	0.02413	0.01324	hsa-mir-192-5p	(0:0)	(:)	(11:A>T)	0.3	1.746	0.096

10

20

30

40

50

【表 2 - 3】

配列番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbi	otherlbi
33	CTGACCTATGAATTGACAGCCCT	0.02306	0.01562	hsa-mir-192-5p	(0:0)	(:CT)	()	0.2	2.004	0.427
34	CCACTGCCCCAGGTGCTGCTGG	-0.02248	0.02248	hsa-mir-324-3p	(-2:0)	(:)	()	0.1	-0.481	0.945
35	TGAGGTAGTAGGTTGTGTGGGT	0.02215	0.02215	hsa-let-7c-5p	(0:0)	(:)	(16:A>G,2 0:T>G)	0.1	0.975	-0.325
36	ACTGTGCGTGTGACAGCGGCT	-0.02097	0.01562	hsa-mir-210-3p	(-1:-2)	(:)	()	0.2	-0.666	0.215
37	CTGCGCAAGCTACTGCCTTG	-0.0202	0.0202	hsa-let-7i-3p	(0:-2)	(:)	()	0.1	1.199	2.896
38	CACCGTAGAACCGACCTTGCG A	-0.02011	0.01097	hsa-mir-99b-5p	(0:0)	(:A)	()	0.3	3.612	4.648
39	CTGACCTATGTATTGACAGCC	0.01839	0.01249	hsa-mir-192-5p	(0:0)	(:)	(10:A>T)	0.2	2.279	0.663
40	TACCCTGTAGAACCGAATTGTC	0.01577	0.01577	hsa-mir-10b-5p	(0:-2)	(:C)	()	0.1	4.555	1.079
41	TGAGAACTGAATTCATAGGCT GAA	-0.01551	0.01551	hsa-mir-146a-5p	(0:1)	(:AA)	(17:G>A,2 0:T>C)	0.1	-0.359	0.464
42	TGACCTATGAATTGACAGCCAAT T	0.01402	0.01402	hsa-mir-215-5p	(1:3)	(:T)	(18:A>C)	0.1	0.754	-0.46
43	TACCCTGTAGAACCGAATTTGTA	0.01382	0.01382	hsa-mir-10b-5p	(0:-1)	(:A)	()	0.1	5.669	4.122
44	TGAGATGAAGCACTGTAGATC	0.01158	0.01158	hsa-mir-143-3p	(0:0)	(:)	(18:C>A)	0.1	2.526	1.048
45	TACCCTGTAGAACCGAACTTGT	0.0115	0.00939	hsa-mir-10b-5p	(0:-1)	(:)	(17:T>C)	0.2	1.946	0.086
46	CTGACCTATGAAGTACAGGCC	0.01068	0.0088	hsa-mir-192-5p	(0:0)	(:)	(12:T>C)	0.2	2.713	0.568
47	GATTGTCAGTTCCTGATT	0.00994	0.00994	hsa-mir-5701	(2:0)	(G:)	()	0.1	0.926	-0.013
48	TTACAGTCTACAGCCATGGTCG	-0.007	0.007	hsa-mir-132-3p	(0:0)	(:)	(1:A>T)	0.1	-0.541	0.325
49	CATTGCACCTTGCTCGGTCTGAA T	0.00642	0.00642	hsa-mir-25-3p	(0:0)	(:AT)	()	0.1	2.02	0.798

10

20

30

40

50

【表 2 - 4】

配列番号	マーカ－	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
50	TACCCTGTTGAACCGAATTGT	0.00629	0.00629	hsa-mir-10b-5p	(0:-1)	(:)	(8:A>T)	0.1	0.959	-0.227
51	CAAAGTGCTGTTCTGTCAGGTA	-0.00623	0.00623	hsa-mir-93-5p	(0:-1)	(:)	()	0.1	2.94	3.614
52	CTCGCTTCTGGCGCCAAAGCGCCCGGC	-0.00413	0.00413	<該当せず> >	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	-0.552	0.651
53	AACTGGCCCTCAAAGTCCCG	-0.00368	0.00368	hsa-mir-193b-3p	(0:-2)	(:)	()	0.1	0.083	1.702
54	TGAGAACTGAATTCATAGGCA A	-0.00364	0.00364	hsa-mir-146b-5p	(0:-1)	(:AA)	()	0.1	0.256	1.187
55	TGAGGTAGTAGATTGTATAGTT TT	0.00325	0.00325	hsa-let-7a-5p	(0:2)	(:)	(11:G>A)	0.1	0.75	-0.212
56	ACCCTGTAGATCCGAAT	0.00148	0.00148	hsa-mir-10a-5p	(1:-5)	(:)	()	0.1	0.215	-0.459
57	AGGCTGTGATGCTCTCTCTGAGC CCT	0.00039	0.00039	hsa-mir-7974	(0:-1)	(:CT)	()	0.1	0.595	-0.142
58	TAACTGTCTGTGTAAC	0.00027	0.00027	hsa-mir-200a-3p	(0:-5)	(:)	()	0.1	1.631	-0.336
59	TACCCTGTAGATCCGAATTCTGT	0.00024	0.00024	hsa-mir-10b-5p	(0:-1)	(:)	(11:A>T,19:T>C)	0.1	1.832	-0.081

10

20

30

40

50

【表 3 - 1】

表 3. クローン病に関連する結腸上皮組織の特定された sRNA バイオマーカー。

配列番号	マーカー	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
60	CCGCCCCACCGCGCGCGCGC	0.74618	0.16463	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.8	1.72	-0.59
61	CGCTTCTGGCGCCAAGCGCCCGCGC	0.25545	0.08406	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.7	1.39	-0.62
62	AGATTGAGGGTTCTGTCCTTCGTG GTCGCC	0.25408	0.05563	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.8	2.73	-0.37
63	GGCTTGGTCTAGGGGTATGATTCT CGCTTT	0.21881	0.06902	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.7	2.2	-0.46
64	GGCTTTGTCTAGGGGTATGATTCT CGCTT	0.18401	0.12882	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.4	1.34	-0.65
65	CCGCCCCACCGCGCGCGCGCGC T	0.15615	0.09596	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.3	1.5	-0.64
66	CGTACGGAAGACCGGCTCCCGCGC GCCGCT	0.11296	0.05941	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.3	1.26	-0.61
67	GTACGGAAGACCGCGCTCCCGCGCG CCG	0.10944	0.10944	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	1.36	-0.59
68	TGGTCTAGCGGTTAGGATTCCTGG TTTT	0.09687	0.06389	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.3	1.02	-0.66
69	CGCCCCACCGCGCGCGCGCGC	0.09422	0.03815	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.5	1.64	-0.61
70	CCCGCGAGGGGGCCCCGGGCAC	0.07217	0.05546	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.2	1.03	-0.58
71	GCGCGCGCGCCCCCCCCACGCCCC GGGC	0.06871	0.04611	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.2	1.64	-0.67
72	GCTCCCCGTCTCTCCCTCCCTCCCC	0.06762	0.06762	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	1.58	-0.67
73	GCGCAATGAAGGTGAAGGCCCGCGC GC	0.06288	0.03999	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.4	1.03	-0.6
74	ACGCTGCCAGTTGAAGAACTGT	0.05063	0.05063	hsa-mir-22-3p	(0:0)	(:)	(1:A>C)	0.1	0.86	-0.46

10

20

30

40

50

【表 3 - 2】

配列 番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbi	otherlbi
75	GCCCTGGCCTATCTAGAAAA	0.04958	0.03308	hsa-mir-331- 3p	(0:0)	(AA)	()	0.2	0.68	-0.65
76	GCGGGTCCGCGGTGTCGGCGG	0.04831	0.04831	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	0.65	-0.67
77	GGCTGGTCTAGGGGTATGATTCT CGCT	0.04437	0.04437	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	3.5	0.65
78	CCACCTCCCCTGCAACGTCC	0.03994	0.02586	hsa-mir- 1306-5p	(0:-1)	(:)	()	0.4	0.46	-0.6
79	GGTTAGGATTCCTGGTTTT	0.03829	0.03829	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.08	-0.57
80	TCTGGCATGCTAACTAGTTACGG ACCCC	0.03622	0.03622	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	0.84	-0.67
81	CGCGTCCCCCGAAGAGGGGACG GCGGAGC	0.03391	0.03391	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.08	-0.68
82	GCGGAGCGAGCGCACGGGGTCGG CGCGGAC	0.0323	0.0323	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	0.79	-0.52
83	CCCCGCCCCACCCCGCGCGCGCC GCTCGC	0.02563	0.02563	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.3	-0.68
84	CCGTAGGTGAACCTGCGGAAGGAT CATTA	0.02433	0.01963	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.2	2.36	-0.5
85	GGGCTACGCCTGTCTGAGCGTCGC TT	0.02206	0.02206	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	2.74	0.07
86	GCTACGCCTGTCTGAGCGTCGCTT	0.02103	0.02103	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.48	-0.46
87	CCCCACAACCGGCTTGACTAGCT T	0.0204	0.0204	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.43	-0.36
88	CCCTACCCCCCGGCCCGTC	0.01307	0.01307	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.25	-0.56
89	CCCGCCCCACCCCGCGCGCGCGC TCGC	0.01108	0.01108	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.7	-0.59
90	GGGGGTATAGCTCAGTGGTAGAG CGTGCTT	0.01022	0.01022	<該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.12	-0.58

10

20

30

40

50

【表 3 - 3】

(40)

JP 2023-511368 A 2023.3.17

配列 番号	マーカ－	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
91	GTCGGTCGGGCTGGGGCGCGGAAG CGGGGCT	0.00996	0.00996	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	2.53	-0.51
92	TCAGTGGAGAGCATTTGACT	0.00991	0.00991	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	0.54	-0.66
93	CACCCCTAGAACGACCTTGCG	0.0095	0.0095	hsa-mir-99b- 5p	(0:0)	(:)	(5:G>C)	0.1	0.17	-0.66
94	CCTCACCATCCTTCTGCCTGCA	0.00892	0.00892	hsa-mir- 6511a-3p	(0:1)	(:)	()	0.1	0.2	-0.65
95	GTCAGGATGGCCGAGGGGTCT	0.00647	0.00647	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	2.13	0.36
96	TCCCTGGTCTAGTGGTTAGGATT GGCGCG	0.00644	0.00644	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	1.6	-0.27
97	TGAGATGAAGCACTGTAGATC	-0.00555	0.00555	hsa-mir-143- 3p	(0:0)	(:)	(18:C>A)	0.1	-0.07	1.91
98	GGATCGGCCCGCGGGGTGCGC	0.00523	0.00523	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	1.04	-0.68
99	GGAACCTGCGGAAGGATCATT	0.00215	0.00215	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	2.24	-0.33
100	TGAGGTAGTAGTTGTATGGTTG	0.00179	0.00179	hsa-mir-4510	(0:1)	(:)	(5:G>T, 12:A>T)	0.1	0.92	-0.53
101	GTCTAGTGGTTAGGATTCGGCGCT	0.00093	0.00093	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	1.61	-0.38
102	TCCCTGGTCTAGTGGCTAGGATT GGCGCT	0.00085	0.00085	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	0.72	-0.64
103	GGCGCCCCCCCCACGCCCCGGGC	0.0002	0.0002	<該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	0.59	-0.68

10

20

30

40

50

【表 4 - 1】

表 4. 潰瘍性大腸炎に関連する結腸上皮組織中の特定された sRNA バイオマーカー。

配列番号	マーカー	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
104	TGTCAGTTTGCAAAATACCC CAAG	0.46706	0.1009	hsa-mir-223-3p	(0:2)	(:)	()	0.9	1.892	0.1084
105	CAGCAGCAATTCATGTTTTG AAT	0.29749	0.09883	hsa-mir-424-5p	(0:0)	(:T)	()	0.6	0.578	-0.613
106	GTGGTTGTAGTCGGTGCGA GAATACC	-0.22154	0.09667	<該当せず> (該当せず: 該当せず)			()	0.5	-0.373	1.2368
107	GGATATCATCATATACTGTA AGT	0.1973	0.11602	hsa-mir-144-5p	(0:1)	(:)	()	0.4	2.428	0.8535
108	TAACAGTCTCCAGTCACGG C	0.14329	0.07797	hsa-mir-212-3p	(0:-1)	(:)	()	0.6	1.215	-0.5329
109	TCAGTGCACACAGAACTTT TTT	0.13604	0.06626	hsa-mir-148a-3p	(0:0)	(:T)	(20:G>T)	0.5	0.643	-0.6209
110	CCAGTGGGGCTGCTGTTAT CTGT	-0.13318	0.07284	hsa-mir-194-3p	(0:0)	(:T)	()	0.3	0.857	2.7111
111	GATAAAGTAGAAAGCACTA CT	0.13252	0.06175	hsa-mir-142-5p	(1:0)	(G:)	()	0.4	1.653	-0.6021
112	TAGGTAGTTTCTCTGTTGTG GAT	-0.1183	0.04091	hsa-mir-196a-5p	(0:-1)	(:AT)	(11:A>C)	0.6	-0.676	-0.1724
113	ATGCTTATCAGACTGATGTT GA	0.11425	0.07239	hsa-mir-21-5p	(2:0)	{AT:}	()	0.3	1.241	-0.512
114	TAGTGAATATTGCTTATAG GG	0.10893	0.0759	hsa-mir-454-3p	(0:-1)	(:)	()	0.3	0.82	0.0483
115	CCATAAAGTAGAAAGCAC TACT	0.10582	0.05342	hsa-mir-142-5p	(-2:0)	(:)	()	0.5	1.414	-0.294
116	TACCCATTGCATATCGGAGT T	0.097	0.07557	hsa-mir-660-5p	(0:-1)	(:)	()	0.3	0.876	-0.4505
117	ACTGGACTTGGAGTCAGAA GGAA	-0.09333	0.05017	hsa-mir-378b	(0:3)	(:A)	(13:G>T, 19:A>G)	0.3	2.232	4.1887
118	AAGCAGCAATTCATGTTTTG A	0.09165	0.06219	hsa-mir-424-5p	(1:-1)	(A:)	()	0.2	0.263	-0.6458

10

20

30

40

50

【表 4 - 2】

配列番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbi	otherlbi
119	CTGCAGCACGTAAATATTG GCG	0.0866	0.05794	hsa-mir-16-5p	(2:0)	(CT:)	()	0.2	0.882	-0.5753
120	TGGCAGTGTCTTAGCTGGT T	0.07815	0.06409	hsa-mir-34a-5p	(0:-2)	(:)	()	0.3	1.71	-0.1242
121	ACTGGACTTGGAGTCAGAA GGTT	-0.07752	0.052	hsa-mir-378c	(0:-2)	(:)	(20:A>G, 21:G>T)	0.2	-0.284	1.3769
122	TGAGAACTGAATTCATAG GCTGTAA	0.07149	0.03423	hsa-mir-146b-5p	(0:4)	(:)	(24:G>A)	0.6	2.372	0.6917
123	ACTGGACTTGGAGTCAGAA GGAT	-0.0679	0.04539	hsa-mir-378c	(0:-2)	(:)	(20:A>G, 21:G>A)	0.2	0.289	2.0819
124	TGAGAACTGAATTCATAG GCTGTAAT	0.06566	0.04343	hsa-mir-146b-5p	(0:4)	(:T)	(24:G>A)	0.3	0.687	-0.4488
125	GTTGAGACTCTGAAATCTG ATT	-0.06461	0.05023	hsa-mir-4431	(-2:-7)	(G:GATT)	(3:C>G,1 4:A>A)	0.2	-0.649	0.1771
126	TTAATGCTAATCGTGATAG	0.06346	0.02758	hsa-mir-155-5p	(0:-4)	(:)	()	0.4	2.46	0.3365
127	TGAGAACTGAATTCATAG GAA	0.06095	0.0468	hsa-mir-146a-5p	(0:-2)	(:AA)	(17:G>A)	0.2	1.103	0.1217
128	CTATACGACCTGCTGCCTTT CA	-0.05799	0.05799	hsa-let-7d-3p	(0:-1)	(:A)	()	0.1	0.725	1.845
129	TACCCTGTAGAACC GAATTT GCG	-0.05773	0.04012	hsa-mir-10a-5p	(0:0)	(:)	(11:T>A,2 1:T>C)	0.2	-0.445	0.5034
130	TGGCAGTGTCTTAGCTGGT	0.05695	0.04073	hsa-mir-34a-5p	(0:-3)	(:)	()	0.2	0.721	-0.5822
131	CCAGTGGGGCTGCTGTTAT CT	-0.05534	0.03762	hsa-mir-194-3p	(0:-1)	(:)	()	0.3	1.163	2.2638
132	TTGAGAACTGAATTCATG GGTT	0.05453	0.04544	hsa-mir-146a-5p	(-1:0)	(:)	()	0.2	2.563	0.8429
133	TTACAGTCTACAGCCATGGT CG	0.04999	0.04437	hsa-mir-132-3p	(0:0)	(:)	(1:A>T)	0.2	0.833	-0.4181
134	ACTGGACTTGGAGTCAGAA GGCT	-0.04834	0.0324	hsa-mir-378d	(0:3)	(:)	(19:A>G, 20:A>G)	0.2	5.356	6.5699

10

20

30

40

50

【表 4 - 3】

配列番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thisbl	otherbl
135	TGAGAACTGAATTCATAG GCTGTAG	0.04829	0.0337	hsa-mir-146b-5p	(0:2)	(:AG)	()	0.2	0.761	-0.2346
136	CCCATAAAGTAGAAAGCAC TACA	0.04703	0.03279	hsa-mir-142-5p	(-2:-1)	(:A)	()	0.2	2.327	0.2258
137	TGAGGTAGTAGTTTGCT	0.04637	0.04637	hsa-let-7i-5p	(0:-3)	(:)	()	0.1	3.668	2.5754
138	CGGCGCAAGCTACTGCCTT G	0.04625	0.04625	hsa-let-7i-3p	(0:-2)	(:)	(1:T>G)	0.1	0.127	-0.6692
139	AGTCTTCAGTGGCAAGCT	0.04577	0.04577	hsa-mir-22-5p	(0:-3)	(:)	()	0.1	1.084	-0.0644
140	TCCCCTGTAGAACCGAATTT GT	-0.04267	0.02897	hsa-mir-10b-5p	(0:-1)	(:)	(1:A>C)	0.2	-0.655	0.1801
141	ACTGGACTTGGAGTCAGAA GGCATT	-0.04209	0.02716	hsa-mir-422a	(0:0)	(:ATT)	(9:A>G,1 1:G>A)	0.3	1.615	3.1346
142	AAGTCGGTCTGAGGCCCC TCA	-0.04032	0.03266	hsa-mir-423-3p	(-1:-2)	(:)	()	0.2	0.598	1.7929
143	CCAGTGGGGCTGCTGTAT CTGA	-0.03971	0.03971	hsa-mir-194-3p	(0:0)	(:A)	()	0.1	-0.383	1.5327
144	TGAGGGAGTAGTTTGCT GTTA	0.03743	0.02474	hsa-let-7i-5p	(0:0)	(:A)	(5:T>G)	0.3	0.516	-0.5159
145	AAGAAAGTAGAAAGCACTA CT	0.03726	0.03726	hsa-mir-142-5p	(1:0)	(A:)	(1:T>A)	0.1	0.759	-0.6659
146	CGCTGCCAGTTGAAGAACT GT	0.03671	0.03671	hsa-mir-22-3p	(2:0)	(C:)	()	0.1	1.055	-0.5449
147	GGCTGGTCCGATGGTAGT	-0.03534	0.03534	hsa-mir-6131	(0:-1)	(:)	(8:A>C,1 4:G>T)	0.1	0.079	1.378
148	CTGGGAGAAAGGCTGTTTAC TCT	-0.03467	0.03467	hsa-mir-30c-2-3p	(0:0)	(:)	()	0.1	0.783	1.6525
149	AAGCAATTCTCAAAGGAGC	0.03329	0.01693	hsa-mir-5571-5p	(-3:-5)	(:)	()	0.4	0.38	-0.6931
150	CTCGGCGCCCCCTCGATGCT CT	-0.03132	0.02602	<該当せず> <該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.2	-0.37	0.6322
151	TGCTTGCAGGCCGTCATGC	0.02612	0.01998	hsa-mir-431-5p	(0:-1)	(:)	()	0.2	0.613	-0.6042

10

20

30

40

50

【表 4 - 4】

配列番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
152	CGAATCATTATTGCTGCT	0.02532	0.02532	hsa-mir-15b-3p	(0:-3)	(:)	()	0.1	1.521	-0.0129
153	CAGCAGCAATTCATGTTTGG AAA	0.02138	0.02138	hsa-mir-424-5p	(0:0)	(:A)	()	0.1	0.241	-0.3669
154	ACCAATATTACTGTGCTGCT	0.0205	0.01422	hsa-mir-16-2-3p	(-1:-3)	(:)	()	0.2	3.128	1.1757
155	TTCAAGTAATCCAGGATAG GCTTT	-0.02004	0.02004	hsa-mir-26a-5p	(0:2)	(:)	(22:G>T)	0.1	3.007	4.1471
156	TTGAGAACTGAAATTCATG GGT	0.01968	0.01968	hsa-mir-146a-5p	(-1:-1)	(:)	()	0.1	1.968	0.5389
157	TATTGCACATTACTAAGTTG	0.01865	0.01865	hsa-mir-32-5p	(0:-2)	(:)	()	0.1	3.749	1.603
158	TGACCTATGAAATTGACAGC CTA	-0.01793	0.01793	hsa-mir-215-5p	(1:2)	(:)	(18:A>C, 20:A>T)	0.1	-0.659	0.189
159	ACTGTAAACGCTTTCTGATG	-0.01783	0.01783	hsa-mir-3607-3p	(0:0)	(:)	()	0.1	1.014	1.2253
200	CATTGCACTTGTCTCGGTCT GAAT	-0.01738	0.01738	hsa-mir-25-3p	(0:0)	(:AT)	()	0.1	0.719	1.4522
201	ATAAAGTAGAAAGCACTAC T	0.01695	0.01695	hsa-mir-142-5p	(1:0)	(:)	()	0.1	2.536	0.3764
202	AAGTGCAATGATGAAAGGG CA	0.01537	0.01537	hsa-mir-130a-3p	(1:-1)	(A:)	(9:T>G,1 1:A>T)	0.1	0.631	-0.6633
203	ACCATAAAGTAGAAAGCAC TA	0.01523	0.01523	hsa-mir-142-5p	(-1:-2)	(A:)	()	0.1	1.096	-0.3697
204	CCCCACTGCTAAATTTGACT GGCTTT	-0.01424	0.01424	<該当せず> <該当せず>	(該当せず: 該当せず)	(該当せず: 該当せず)	()	0.1	-0.076	1.0335
205	TGTCAGTTTGTCAAAATACCC CAAGA	0.01423	0.01423	hsa-mir-223-3p	(0:2)	(:A)	()	0.1	0.507	-0.6124
206	TACCCAGTAGAACCGAATTT GT	-0.01326	0.01326	hsa-mir-10b-5p	(0:-1)	(:)	(5:T>A)	0.1	-0.197	0.5859
207	TTTGTTCGTTTCGGCTCGCGT AA	-0.01282	0.01282	hsa-mir-375	(0:0)	(:)	(20:G>A)	0.1	-0.245	1.5709

10

20

30

40

50

【表 4 - 5】

配列 番号	マーカ ー	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
208	ATGCTGCCAGTTGAAGAAC TGTA	0.01218	0.01218	hsa-mir-22- 3p	(0:0)	(:A)	(1:A>T)	0.1	0.462	-0.555
209	TGAGAACACGCTCTGCTCT G	0.01124	0.01124	hsa-mir-589- 5p	(0:-2)	(:)	()	0.1	0.523	-0.2778
210	CTGCCAATTCATAGGTCAC AGT	-0.0098	0.0098	hsa-mir-192- 3p	(0:0)	(:T)	()	0.1	0.349	1.5762
211	TAGCTTATCAGACTGATGTT GAGA	0.00974	0.00974	hsa-mir-21- 5p	(0:0)	(:GA)	()	0.1	0.626	0.2759
212	GTAGCTTATCAGACTGATGT TGACT	0.00953	0.00953	hsa-mir-21- 5p	(-1:2)	(:)	()	0.1	1.628	0.0433
213	TTTGGTCCCCTTCAACCAGC TGA	-0.00945	0.00945	hsa-mir- 133a-3p	(0:0)	(:A)	()	0.1	-0.62	-0.0075
214	TGTAATAGCAACTCCATGTG GAA	-0.00844	0.00844	hsa-mir-194- 5p	(0:1)	(:)	(5:C>T)	0.1	-0.638	0.24
215	GGGACCTATGAATTGACAG AC	0.00774	0.00774	hsa-mir-192- 5p	(2:0)	(GG:)	(17:C>A)	0.1	0.989	-0.4886
216	TAAGGTGCATCTAGTGCAG ATA	0.00772	0.00772	hsa-mir- 18b-5p	(0:-1)	(:)	(19:T>A)	0.1	2.295	0.6414
217	GTA CTGGAAAGTGCACTTG GACGAACA	-0.00721	0.00721	<該当せず> <該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	-0.395	1.7345
218	CCGGGGCTACGCCCTGTCT GAGCGTCGCT	-0.00713	0.00713	<該当せず> <該当せず>	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	1.856	2.7329
219	AAAGCTGGGTTGAGAGGG CGAAA	-0.00655	0.00655	hsa-mir- 320a	(1:2)	(:)	()	0.1	0.172	0.9853
220	CATAAAGTAGAAAGCACTA	0.00604	0.00537	hsa-mir-142- 5p	(0:-2)	(:)	()	0.2	2.95	1.1211
221	TGTCAGTTTGTCAAAATAC	0.00602	0.00602	hsa-mir-223- 3p	(0:-4)	(:)	()	0.1	2.716	-0.261
222	TCCGGTGAGCTCTCGCTGG CC	0.00578	0.00578	hsa-mir- 4792	(-1:1)	(T:)	(9:G>C)	0.1	0.207	-0.4932
223	TATAAAGTAGAAAGCACTA C	0.00555	0.00555	hsa-mir-142- 5p	(1:-1)	(T:)	()	0.1	0.13	-0.6931

10

20

30

40

50

【表 4 - 6】

配列 番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
224	TGCTGCCAGTTGAAGAACT GT	0.00546	0.00546	hsa-mir-22- 3p	(2:0)	(T:)	()	0.1	0.158	-0.6517
225	AGTCGGTCTGAGGGCCCT CAGTTT	-0.00518	0.00518	hsa-mir-423- 3p	(0:2)	(:)	(23:C>T)	0.1	0.091	1.3932
226	TGTCAGTTTGTCAAAATACCC CATG	0.00464	0.00464	hsa-mir-223- 3p	(0:2)	(:)	(22:A>T)	0.1	0.204	-0.642
227	ATCACAGTGGCTAAGTTCC	0.00413	0.00413	hsa-mir-27a- 3p	(1:-2)	(A:)	()	0.1	0.487	-0.6176
228	TGAGAACTGAATTCATAG GCAA	0.0039	0.0039	hsa-mir- 146b-5p	(0:-1)	(:AA)	()	0.1	1.542	0.5058
229	TGGTCTTTGCGGGCGAGA TGA	-0.00383	0.00383	hsa-mir- 193a-5p	(0:0)	(:)	()	0.1	1.582	2.1855
230	TACCCTGTAGAACCAGGATTT G	-0.00313	0.00313	hsa-mir- 10b-5p	(0:-2)	(:)	(15:A>G)	0.1	-0.657	-0.2565
231	TGAGGGAGTAGAATTGTATA GT	0.00301	0.00301	hsa-let-7a- 5p	(0:-1)	(:)	(5:T>G,1 1:G>A)	0.1	1.916	0.1598
232	TACCCTGTTGAACCGAATTT GT	-0.00297	0.00297	hsa-mir- 10b-5p	(0:-1)	(:)	(8:A>T)	0.1	-0.159	0.3187
233	TAAGGTGCATCTAGTCAG AT	0.00245	0.00245	hsa-mir-18a- 5p	(0:-2)	(:)	()	0.1	2.559	0.72
234	GAGAACTGAATTCATAGG CTGT	0.0021	0.0021	hsa-mir- 146b-5p	(1:2)	(:)	()	0.1	0.549	-0.328
235	TAGCAGCACGCAAAATATTG GCG	0.00209	0.00209	hsa-mir-16- 5p	(0:0)	(:)	(10:T>C)	0.1	0.28	-0.5687
236	GGCTCGTTGGTCTAGGGG	-0.0019	0.0019	hsa-mir- 4448	(0:-2)	(:)	(5:C>G)	0.1	-0.534	0.0195
237	CAGCAGCAATTCATGTTTTG	0.00173	0.00173	hsa-mir-424- 5p	(0:-2)	(:)	()	0.1	0.987	-0.0245
238	AACATTCAACGCTGTCGGT G	-0.00169	0.00169	hsa-mir- 181b-5p	(0:-3)	(:)	(8:T>A,9: T>C)	0.1	3.67	3.8391
239	ATGCAGCACGTAAATATTG GCG	0.00169	0.00169	hsa-mir-16- 5p	(2:0)	(AT:)	()	0.1	0.338	-0.6428

10

20

30

40

50

【表 4 - 7】

配列 番号	マーカ－	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbl	otherlbl
240	TCCGACGGCGCTGACCC CCTT	-0.00159	0.00159	< 該当せず >	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	-0.369	0.6898
241	ATTGGTCGTGTTGTAGTC CGTGCGAGAA	-0.00106	0.00106	< 該当せず >	(該当せず; 該当せず)	(該当せず; 該当せず)	()	0.1	-0.405	0.4592
242	TGGCAGTGTCTTAGCTGGT TG	0.001	0.001	hsa-mir-34a- 5p	(0:-1)	(:)	()	0.1	1.828	0.6251
243	TGTCAGTTTGTCAAATA	0.00095	0.00095	hsa-mir-223- 3p	(0:-5)	(:)	()	0.1	0.047	-0.6931
244	ACCCTGAGACCCCTAACTTGT GA	0.00016	0.00016	hsa-mir- 125b-5p	(1:0)	(A:)	()	0.1	0.322	-0.5771
245	TGGCAGTTTGTCAAATACC	0.00011	0.00011	hsa-mir-223- 3p	(0:-3)	(:)	(2:T>G)	0.1	1.467	-0.5979

表 5．憩室症に関連する結腸上皮組織の特定された sRNA バイオマーカー。

配列 番号	マーカー	重要性	imp_SE	sRNA_name	参照	ext	交換	選ば れた	thislbi	otherlbi
246	ACTGGACTTGGAGTCAGAAGGCA TAT	1.3057	0.12197	hsa-mir- 422a	(0:0)	(:ATAT)	(9:A>G,11:G>A)	1	1.458	-0.67
247	TCGACCGGACCTCGACCGGCTAG A	0.23143	0.11311	hsa-mir- 1307-5p	(0:2)	(:A)	(21:C>A)	0.4	1.008	-0.59
248	TCAGCACCAGGATATTGTTGGA	0.11606	0.05936	hsa-mir- 3065-3p	(0:-1)	(:)	()	0.4	1.535	-0.58
249	TGTAACCGCAACTCCATGTGGA	0.09378	0.05427	hsa-mir- 194-5p	(0:0)	(:)	(6:A>C)	0.3	1.788	-0.39
250	ACTGGACTTGGAGTCAGAAGGCA TTA	0.08715	0.04571	hsa-mir- 422a	(0:0)	(:ATTA)	(9:A>G,11:G>A)	0.3	1.098	-0.67
251	AACACTGTCTGGTAAAGATGGC	0.08212	0.0662	hsa-mir- 141-3p	(1:1)	(:)	()	0.2	1.265	-0.63
252	TGTAACATCCTACACTCTCAGCT TA	0.08206	0.03761	hsa-mir- 30c-5p	(0:1)	(:TA)	()	0.5	0.138	-0.69

10

20

30

40

50

【表 5 - 2】

配列番号	マーカ	重要性	imp_SE	sRNA_name	参照	ext	交換	選ばれた	thislbi	otherlbi
253	ACTGGACTTTGAGTCAGAAAGGCA	0.06028	0.04522	hsa-mir-422a	(0:0)	(:A)	(9:A>T,11:G>A)	0.3	0.671	-0.65
254	ACTGGACTTGGAGCCAGAAAGGCA A	0.05242	0.04482	hsa-mir-378f	(0:2)	(:AA)	(20:T>G)	0.2	0.921	-0.65
255	GTAACAGCAACTCCATGTGAAA	0.04186	0.02857	hsa-mir-194-5p	(1:1)	(:A)	()	0.2	0.92	-0.67
256	ACTGGACTTGGAGTCAGAAAGGCA ATA	0.03645	0.01948	hsa-mir-422a	(0:0)	(:AATA)	(9:A>G,11:G>A)	0.5	-0.038	-0.69
257	CTGGACTTGGAGTCAGAAAGGCA A	0.0346	0.02916	hsa-mir-378f	(1:2)	(:AGA)	(12:C>T,19:T>G)	0.2	0.159	-0.68
258	TGATATGTTTGATATATTAGGTTA	0.03153	0.02537	hsa-mir-190a-5p	(0:1)	(:A)	()	0.2	1.842	-0.53
259	TGAAATGTTTAGGACCACCTAGAA T	0.02779	0.02185	hsa-mir-203a-3p	(1:1)	(:AT)	()	0.2	0.309	-0.68
300	TGGACTTGGAGTCAGAAAGGCA	0.02407	0.01645	hsa-mir-378a-3p	(2:0)	(:AT)	()	0.2	0.622	-0.66
301	TGTAACAGCAACTCCATGTGGAC TA	0.01862	0.01862	hsa-mir-194-5p	(0:2)	(:A)	()	0.1	0.327	-0.58
302	TCGACCGGACCTCGACCCGGCTA	0.01749	0.01519	hsa-mir-1307-5p	(0:0)	(:A)	()	0.2	1.518	-0.6
303	TGAGATGAAGCACTGTAGCTCAT A	0.01455	0.01455	hsa-mir-143-3p	(0:1)	(:TA)	()	0.1	0.975	-0.61
304	TTTCAGTCGGATGTTTGCAGCAA	0.01444	0.01444	hsa-mir-30e-3p	(1:0)	(:AA)	(16:A>G)	0.1	0.141	-0.69
305	GACCTATGAATTGACAGCCAT	0.01188	0.00963	hsa-mir-215-5p	(2:1)	(:T)	(17:A>C)	0.2	1.014	-0.58
306	CCACTGCCCCAGGTGCTGCTGGA	0.01092	0.01092	hsa-mir-324-3p	(-2:0)	(:A)	()	0.1	0.692	-0.6
307	CTGACCTATGAATTGACAGCCAT GA	0.0102	0.0102	hsa-mir-192-5p	(0:1)	(:TGA)	()	0.1	0.583	-0.63
308	ACCACAGGGTAGAACCACCGGACG A	0.00927	0.00927	hsa-mir-140-3p	(1:2)	(:GA)	()	0.1	0.682	-0.58

10

20

30

40

50

【表 5 - 3】

配列 番号	マーカ－	重要性	imp_SE	sRNA_name	参照	ext	交換	選ば れた	thislbi	otherlbi
309	TCGACGGGACCTCGACCCGGCTGA	0.00896	0.00896	hsa-mir-1307-5p	(0:0)	(:GA)	()	0.1	-0.463	-0.68
310	TGGCTCAGTTCAGCAGGAACAGG A	0.00641	0.00641	hsa-mir-24-3p	(0:2)	(:)	()	0.1	0.543	-0.6
311	AGCTTATCAGACTGATGTTGAAA	0.00487	0.00487	hsa-mir-21-5p	(1:0)	(:AA)	()	0.1	0.052	-0.66
312	ATCACATTGCCAGGGATAAAA	0.00469	0.00469	hsa-mir-23c	(0:3)	(:AA)	(13:T>G,17:T>A)	0.1	0.333	-0.66
313	TCAACAAAATCACTGATGCTGGA	0.0018	0.0018	hsa-mir-3065-5p	(0:0)	(:)	()	0.1	0.71	-0.53
314	ACATTGCCAGGGATTCCA	0.00084	0.00084	hsa-mir-23a-3p	(3:1)	(:)	()	0.1	1.31	-0.57
315	AACACTGTCTGGTAAAGATG	0.00065	0.00065	hsa-mir-141-3p	(1:1)	(:)	()	0.1	-0.094	-0.69

10

20

30

40

【 0 1 1 6 】

マルチクラス疾患分類

疾患分類器を、sRNAパネルの陽性または陰性マーカ－、同様に、対照、クローン病、潰瘍性大腸炎、及び憩室症について上で特定されたパネルにおけるsRNAの存在または不在に基づいてトレーニングした。クラスメトリックがすべて組み合わされた際の計算モデルの精度を評価するために、試験を、各クラスの参照試料に対するモデルの特定予測力を評価するために実行した。モデルは98%の正解率を有することが見いだされた。図4は、真の参照同一性に対する疾患クラスの正確な予測の割合を示すヒートマップを示す。これらの結果は、以下のマトリックスにも示される。

50

【表 6】

	参照			
予測	クローン病	対照	憩室症	潰瘍性大腸炎
クローン病	1 1 6	0	0	0
対照	0	1 7 9	0	0
憩室症	0	0	5 9	4
潰瘍性大腸炎	4	1	1	2 2 6

10

【0117】

実施例 2：スパイクインデータの使用

20

本実施例は、miRNeasy Serum/Plasma Advanced Kit (Qiagen) を使用して 137 個の 0.5 mL 脳脊髄液試料から抽出した sRNA を使用して、全シーケンシングランから取得したスパイクインデータの使用を説明する。

【0118】

プールした 5 つのキャリブレーションを含む RNA スパイクイン混合物を使用し、試料中の各スパイクの最終濃度が以下になるように、プールを各試料にスパイクした後にライブラリ調製した。

キャリブレーション 1 = $0.0001 \text{ amol} / \mu\text{L}$

キャリブレーション 2 = $0.001 \text{ amol} / \mu\text{L}$

キャリブレーション 3 = $0.01 \text{ amol} / \mu\text{L}$

キャリブレーション 4 = $0.1 \text{ amol} / \mu\text{L}$

キャリブレーション 5 = $1.0 \text{ amol} / \mu\text{L}$

30

【0119】

試料（スパイクイン混合物を含む）を、3' 及び 5' アダプターライゲーションを含むライブラリ調製、続いて、逆転写、その後、PCR 増幅に供して、Sciclone iQ NGS Workstation (PerkinElmer) 上で NextFlex Small RNA Library Preparation Kit v3.0 (BIOO) を使用して、固有のバーコードを各試料に添加した。

【0120】

試料を 0.65 nM の最終濃度までプールし、1 方向あたり 101 bp でランする S2 フローセルを使用して NovaSeq 6000 Sequencing System (Illumina) で配列決定した。このスキーマを使用して、各試料を 12,000,000 リード以上の深度で配列決定した。トリミングアルゴリズムを使用してデータをトリミングした。

40

【0121】

スパイクイン参照ライブラリを使用してスパイクインをマッピングした。リードを TPM (100 万リードあたりのトリミングされたリード) に変換した。データをプロットし、 R^2 を計算した。図 5 は、137 個の試料の全ランのデータをプロットした結果を示す ($R^2 = 0.989$)。

【0122】

50

実施例 3：特発性肺線維症のサブタイプ化

特発性肺線維症（IPF）は、不可逆的な致命的疾患である。IPFの発症率は、米国、ヨーロッパ、及びアジアで10万人あたり2.5～16.0である。これらの発症率に基づいて、世界中で毎年100万人以上がこの疾患と闘っていると推定することができる。IPFは、呼吸困難、咳、及び経時的な肺機能の低下を症候的に呈する。IPFの診断は、多くの場合1年以上かかり、かつ臨床試験、気管支鏡検査、肺生検、及び組織学的検査を行う呼吸器科医、胸部放射線科医、及び病理医からなる多くの専門分野にわたるチームを必要とする複雑な手順である。

【0123】

IPF患者は予後不良であり、診断時から5年未満で死亡率が50%超に達する。IPF肺組織の病理学は、線維芽細胞の無制御増殖及び細胞外マトリックス分子の過剰蓄積に起因する肺構造の歪みを示す。しかしながら、全生存率は絶対的なものではなく、患者の経緯は様々であり、一部の患者では進行の遅い疾患であり、他の患者では急速に悪化する。したがって、異質性は、疾患ドライバーに影響を及ぼす遺伝的及び環境的要因、ならびに十分に理解されていない疾患維持に必要な他の遺伝子に関係している可能性がある。

【0124】

IPF患者における転帰を予測し、かつ疾患ドライバーをよりよく理解することができるバイオマーカーを特定するために、本明細書に記載の機械学習発見プラットフォームを使用して、血液ベースの低分子RNA（sRNA）バイオマーカーを発見することができると仮定した。この仮説を試験するために、観察的、多部位、前向き縦断プロファイル研究からのIPF試料を、本発明の実施形態に従って評価した。プロファイル研究は、123個の血清タンパク質の統計的相関を分析した。Maher TM, et al., PROFILEing idiopathic pulmonary fibrosis: rethinking biomarker discovery. European Respiratory Review 22, 148–152 (2013)、Maher, TM, et al., An epithelial biomarker signature for idiopathic pulmonary fibrosis: an analysis from the multicenter PROFILE cohort study. The Lancet Respiratory Medicine 5, 946–955 (2013)を参照されたい。本研究の目的は、IPFをさらに分類する（すなわち、サブタイプ化する）ことであつた。結果は、sRNAシグネチャ（86個のsRNAのパネルに基づく）が、100%の精度で対照試料からIPFを分類することができ、IPF試料をいくつかの異なるクラスにタイプ分けすることができることを示す。

【0125】

50個のIPF疾患試料及び170個の健常ドナー試料を使用した（PAXgene生物検体）。具体的には、一酸化炭素（DLco）、努力呼気量、第1の呼吸（Fev1）、努力肺活量（Fvc）、及びFev1:Fvc比メタデータに対応する肺拡散能力を有する170名の年齢及び性別が一致した対照を選択した。

【0126】

QIACube Connect（QIAGEN）自動液体ハンドラ上でPAXgene Blood RNA Extraction Kit（QIAGEN）を使用して血液RNAを抽出した。Qubit 4 Fluorometer（Thermo）上でRNA HS Assay Kit（Thermo）を使用してRNA量を評価した。LabChip GX Touch（PerkinElmer）上でLabChip RNA HS Assay Kit（PerkinElmer）を使用してRNA完全性スコア（RIN）を評価した。各試料から250ugの総RNAを96ウェルプレートにアリコートした。スパイクインキャリブレーションのカクテルを各試料に添加して、分析中に品質管理を監視し、下流正規化を容易にした。各試料に固有のi7/i5デュアルインデックスを組み込んで多重配列決定を支援するSciclone iQ NGS Workstat

10

20

30

40

50

ion (PerkinElmer) 上で NextFlex Small RNA Library Prep Kit v3 (BIOO) を使用して次世代配列決定 (NGS) ライブラリを調製した。Qubit 4 Fluorometer (Thermo) 上で 1x dsDNA HS Assay Kit (Thermo) を使用してライブラリを定量化した。LabChip DNA 3K NGS Assay Kit (PerkinElmer) を使用してライブラリ断片化分析を評価した。ライブラリを 1.0 nM の濃度でプールした。NovaSeq 6000 Sequencing System (Illumina) 上で S2 Flow Cell Kit (Illumina) を使用して、プールしたライブラリを 1 試料あたり 4,000 万の対合末端リードの標的深度で配列決定した。

10

【0127】

低分子 RNA 配列決定データ品質を、FASTQC を使用して評価した。フィルタを通過したリード (Qスコア 00% 超) を処理し、低分子 RNA に注釈を付けるように設計された一連のトリミング及びショートリードアライメントアルゴリズムを使用して注釈を付けた。このショートリードアライメントアプローチは、低分子 RNA の 5' 末端及び 3' 末端におけるテンプレート化及び非テンプレート化ヌクレオチド付加の注釈付けを可能にし、エクソソームへの遺伝子標的及び細胞局在に関する情報を提供する。このショートリードアライメントアプローチは、マイクロ RNA の注釈付きライブラリと比較して 10,000 倍超多くの固有の低分子 RNA 遺伝子のマッピングも可能にする。分析は、分析に使用した 17~43 塩基対長のマッピングされたリードにわたって一貫したプロファイルを示した。

20

【0128】

IPF 試料及び CTL 試料を各々、モンテカルロ交差検証ランで使用するために、90:10 の比率 (トレーニング:試験) でトレーニングセット及び試験セットに無作為に分けた。モンテカルロラン後、教師あり及び教師なし機械学習 (ML) を利用した一連の人工知能アルゴリズムを使用してデータを分析して、予測 sRNA シグネチャを特定した。ML アルゴリズムは、トレーニング試料セットを使用してモデルを作成し、その後、試験試料セットを使用して精度を測定した。

【0129】

具体的には、トレーニング試料中の最小クラス頻度が 5% 超であった sRNA を選択した。エラスティックネットアルゴリズムを使用して、sRNA 遺伝子ファミリー及び 3' 非テンプレート化ヌクレオチド付加などの超特徴を使用してパネルを減少させた。サポートベクトルマシン (SVM) を使用して試験試料を分析し、その後、受信者動作特性 (ROC) を使用して、曲線下面積、精度、感度、特異度、正の予測値、負の予測値、及び F1 スコアを測定した。

30

【0130】

96 回のモンテカルロ交差検証ランでは、86 個の低分子 RNA 遺伝子の sRNA シグネチャが特定され、IPF 試料を CTL 試料と区別する際に 99.3% の精度 (95% 信頼区間 98.5~100%、 $p < 0.00001$) を提供した。疾患予測モデルは、試料毎に 0.0~1.0 の SVM スコアももたらした。0.5 超のスコアを疾患として分類した。CTL 試料の疾患確率スコアのおよそ 94% が 0.0~0.1 であった。試料の厳密なグループ化は、CTL 試料が同種群であることを示した。対照的に、IPF 試料は、異質性を示唆するいくつかの異なるピークを有する広い平坦な領域に広がる分布を示した。

40

【0131】

IPF 試料を CTL 試料と区別した 86 個の低分子 RNA 遺伝子が sRNA シグネチャに存在した。このシグネチャでは、CTL 試料と比較して、37 個 (43%) の sRNA が上方制御され、49 個 (57%) の sRNA が下方制御された。このシグネチャは、71 個の miRNA アイソフォーム、タンパク質コード遺伝子のイントロン及びエクソンにマッピングする 9 つの遺伝子間由来の sRNA、3 つの rRNA 由来の sRNA、2 つの piRNA アイソフォーム、及び 1 つの yRNA 由来の sRNA から成った。CTL と比

50

較して10倍超の過剰発現を有する4つのmiRNAアイソフォームがIPF試料に存在し、CTLと比較して10倍未満の下方制御を有する7つのmiRNAアイソフォーム及び3遺伝子間sRNAがIPF試料に存在した。

【0132】

教師なし階層的クラスタリングの場合、予測sRNAシグネチャから86個のsRNA遺伝子を使用してユークリッド距離を計算した。完全結合凝集型クラスタリングを使用して試料をグループ化した。結果は、3つのIPFサブタイプを明らかにし、86個の予測低分子RNA遺伝子がすべてのIPF試料において均一に分布及び発現していないことを示した。

【0133】

主成分(PC)分析は、教師なし階層的クラスタリング分析から割り当てられたサブタイプ群を使用してIPF試料を分離することを示した。単位分散スケリングを適用し、補完を伴う特異値分解(SVD)を使用して主成分を計算した。PC1(29%)及びPC2(19%)を使用して試料をプロットした。予測楕円は、同じ群からの新たな観察結果が楕円内に入る確率0.95を示す。

【0134】

目標予測アルゴリズムを使用して、sRNAシグネチャにおける86個の低分子RNA遺伝子の標的を特定した。目標予測プロセスは、99.3%の精度でIPFとCTLを分類し、かつIPF試料をサブグループに層別化したsRNAシグネチャからの86個の低分子RNA遺伝子の各々を分析することから開始した。これらの86個の遺伝子内で、40個の固有の「シード」が発見された。これらの40個のシードを使用して、目標予測アルゴリズムは、 $p < 0.01$ 及び $FDR < 0.05$ を有する14,280個の予測遺伝子をもたらした。3つの交差検証参照検索を使用して、重み付け予測を行った。生物学的方向性を適用して、機能的に関連する標的を解析した。「細胞成分」のための遺伝子オントロジー用語濃縮を使用して、低分子RNA遺伝子及び標的を解析した。

【0135】

この研究の結果は、99.3%の精度でIPF試料をCTL試料と区別することができ、IPF試料を3つの主要なサブタイプに層別化することもできたsRNAシグネチャを特定した。sRNAシグネチャは、86個の低分子RNA遺伝子のパネルを含む。sRNAシグネチャの生物学的有意性を分析することにより、いくつかの生物学的経路の調節不全が予測された。

【0136】

実施例4：候補sRNAの減少

特発性肺線維症(IPF)と診断された511名の患者及び221名の正常健常対照(CTL)対象のPAXgene Blood RNAから得られた低分子RNA配列決定データを、機械学習を使用して分析して、IPFまたはCTLを分類することができるバイオマーカーを特定した。3つの異なる分類ランを試験し、分類器が、(1)すべての低分子RNA特徴、(2)ヒトゲノムに完全にマッピングし、かつ低分子RNAの遺伝子間マッピングを許可しない低分子RNAのみ、及び(3)スワップなしのマイクロRNAアイソフォーム、トランスファーRNA由来の断片、リボソームRNA由来の断片のみを選択することを許可した。

【0137】

いずれの場合にも、モデルを49個のIPF試料及び182個のCTL試料でトレーニングし、462個のCTL試料及び39個のCTL試料で試験した。いずれの場合にも、分類器が最小トレーニングセット頻度10%で1クラスあたり最大3,000個の低分子RNA特徴を選択することを許可した。いずれの場合にも、エラスティックネットは、最終バイオマーカーパネルを1モデルあたり最大96個の低分子RNAに減少させた。

【0138】

結果は、スワップなしのマイクロRNAアイソフォーム、トランスファーRNA由来の断片、リボソームRNA由来の断片のみを許可するように事前選択フィルタを制限するこ

10

20

30

40

50

とにより、71.2のAUC及び92.6%の精度で最良の性能がもたらされたことを示す。事前選択フィルタにすべての低分子RNAを許可することにより、66.7%のAUC及び18.3%の精度がもたらされた。ヒトゲノムに完全にマッピングする低分子RNAのみを許可し、かつ低分子RNAの遺伝子間マッピングを許可しないように事前選択フィルタを制限することにより、69.3のAUC及び45.8%の精度がもたらされた。
【0139】

加えて、事前選択は、miRNAシード配列に関する情報を用いることができる。低分子RNA配列決定データを、合計693個のがん性(LUAD)及び231個の正常隣接組織(CTL)肺生検試料を含む4つの研究(GSE110907、GSE62182、GSE83527、及びTCGA-LUAD)から集計した。これらの試料を、LUADまたはCTL組織を分類するように設計された交差検証を用いた機械学習を使用して分析した。

【0140】

例示的な調査では、このシステムは、GSE62182、GSE83527、及びTCGA-LUADからの645個のLUAD試料及びCTL試料でトレーニングし、GSE110907からの48個のLUAD試料及びCTL試料で試験した。2回目の調査では、このシステムは、GSE110907及びTCGA-LUADからの563個のLUAD試料及び101個のCTL試料でトレーニングし、GSE2182及びGSE83527からの130個のLUAD試料及びCTL試料で試験した。いずれの場合にも、事前選択アルゴリズムが2,000個のsRNA特徴または6,000個のsRNA特徴のいずれかの選択を許可された50回のブートストラップ試験を行った。その後、一致するシード配列(低分子RNA特徴の5'末端からヌクレオチド2~8)に基づいて選択されたsRNAが凝集したか、または凝集しないままであった。シード凝集特徴セットまたは非凝集特徴セットを、最大96個の低分子RNAを許可するエラスティックネットアルゴリズムを使用して減少させた。減少した特徴セットを使用して、GSE110907またはGSE62182及びGSE83527からの試料を試験したサポートベクトルマシンをトレーニングした。

【0141】

結果は、2,000個及び6,000個のsRNAの事前選択により、試験した試料に対して同等の精度がもたらされたことを示した。その一方で、シード凝集特徴セットからの値でトレーニングしたサポートベクトルマシンは、非シード凝集研究と比較して向上した分類性能をもたらした。図7を参照されたい。

【表7】

研究	肺がん	対照
GSE110907	48	49
GSE62182	94	94
GSE83527	36	36
TCGA-LUAD	515	52
合計	693	231

【0142】

引用文献及び代替実施形態

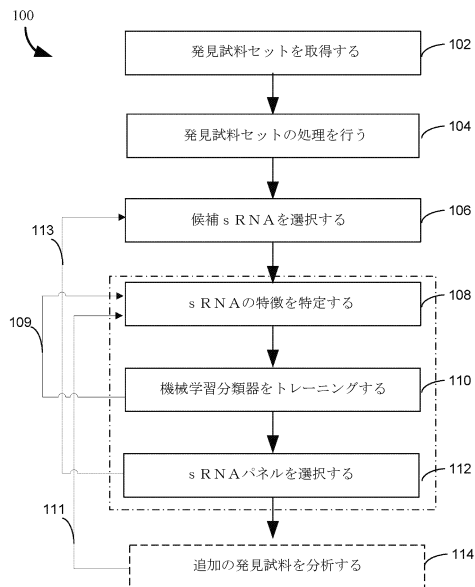
本明細書で引用されるすべての参考文献は、個々の刊行物または特許または特許出願が各々、すべての目的のために参照によりその全体が組み込まれると具体的かつ個別に示された場合と同じ程度に、すべての目的のために参照によりそれらの全体が本明細書に組み込まれる。

【 0 1 4 3 】

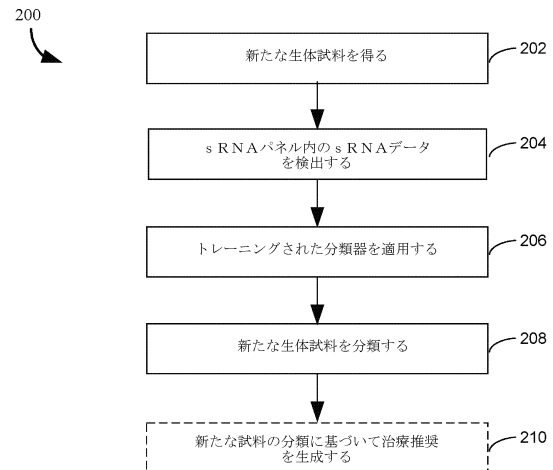
本発明は、非一時的なコンピュータ可読記憶媒体に組み込まれたコンピュータプログラム機構を備えるコンピュータプログラム製品として実装することができる。例えば、このコンピュータプログラム製品は、図 1 と図 2 の任意の組み合わせで示される及び／または説明されるプログラムモジュールを含み得る。これらのプログラムモジュールは、CD-ROM、DVD、磁気ディスク記憶装置製品、USB キー、または他の非一時的なコンピュータ可読データまたはプログラム記憶装置製品に記憶することができる。

【 図 面 】

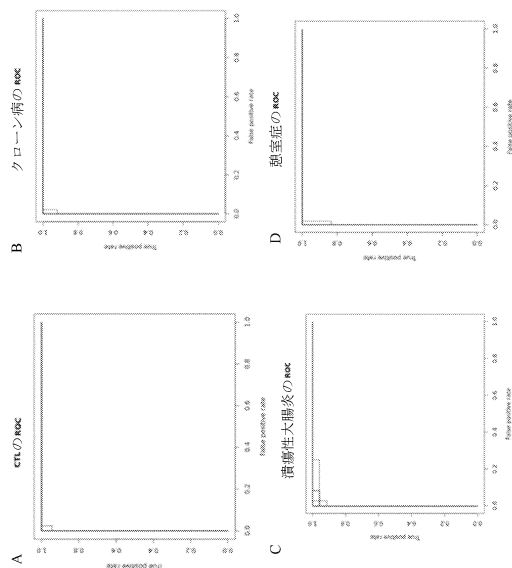
【 図 1 】



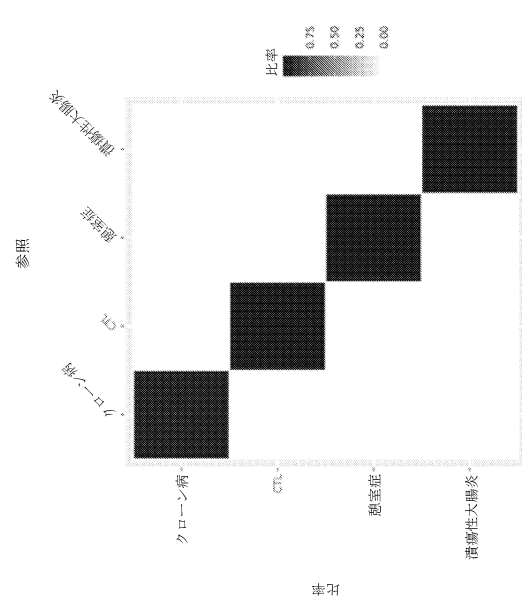
【 図 2 】



【 図 3 】



【 図 4 】



10

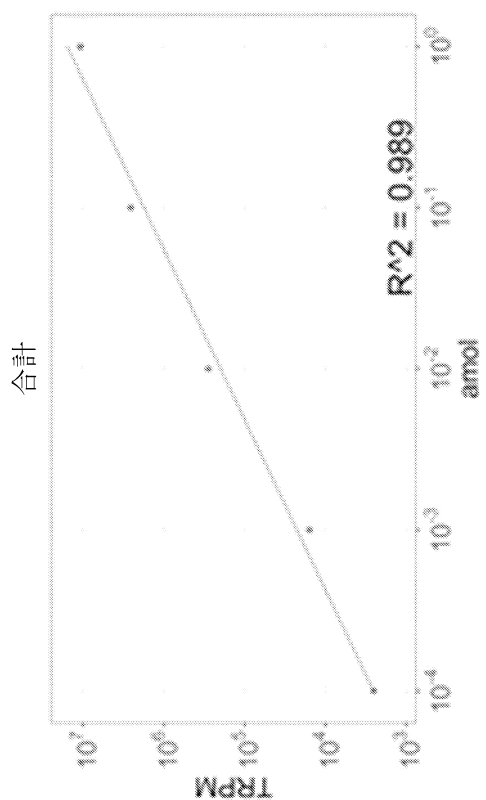
20

30

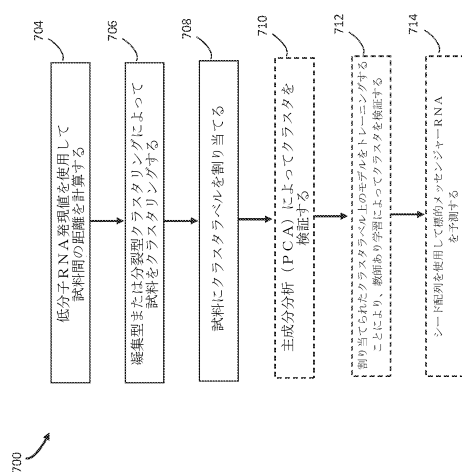
40

50

【図 5】



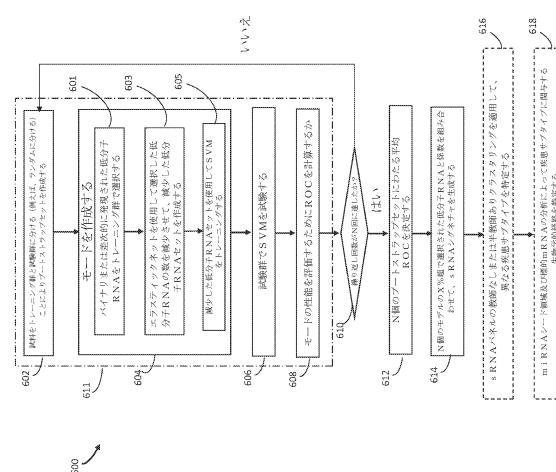
【図 6 B】



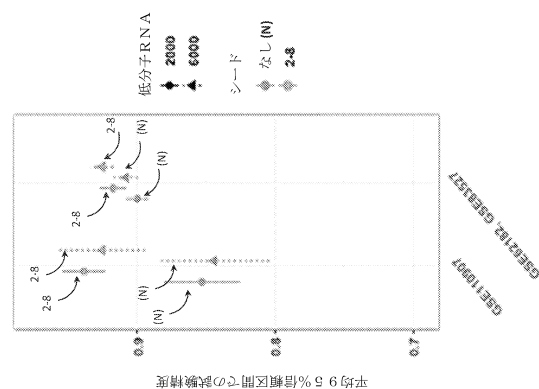
【配列表】

2023511368000001.xml

【図 6 A】



【圖 7】



30

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 21/14755

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C07H 21/04, C12N,15/00, C12Q 1/68, C12Q 1/6883 (2021.01)

CPC - C07H 21/04, C12Q 1/6809, C12Q 1/6883, C12Q 2600/178

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	US 2018/0305763 A1 (SRNALYTICS, INC.) 25 October 2018 (25.10.2018) para [0005]-[0010]; [0033]-[0037]; [0040]-[0044]; [0057]-[0058]; [0075]; [0094]; [0106]-[0107]	1-21
A	WO 2019/014375 A1 (SRNALYTICS, INC.) 17 January 2019 (17.01.2019) abstract; pg 3-5; claims 1, 21, 33, 35, 51, 57	1-21
A	US 2013/0317083 A1 (THOMAS JEFFERSON UNIVERSITY) 28 November 2013 (28.11.2013) abstract; claims 1-7, 11-13; para [0013]	1-21

☐ 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

29 April 2021

Date of mailing of the international search report

MAY 24 2021

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

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 21/14755

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.: 22-74
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:

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.:

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.

フロントページの続き

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,N
E,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,
CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,IT,JO,JP,K
E,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,N
G,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

セッツ ニーダム リンドバーグ アベニュー 202 シー／オー ゲートハウス バイオ インコー
ポレイテッド

(72)発明者 フォスター ニール シー.

アメリカ合衆国 02494 マサチューセッツ ニーダム リンドバーグ アベニュー 202 シー
／オー ゲートハウス バイオ インコーポレイテッド

(72)発明者 レイ ネイサン エス.

アメリカ合衆国 02494 マサチューセッツ ニーダム リンドバーグ アベニュー 202 シー
／オー ゲートハウス バイオ インコーポレイテッド

(72)発明者 メルコニアン テラン

アメリカ合衆国 02494 マサチューセッツ ニーダム リンドバーグ アベニュー 202 シー
／オー ゲートハウス バイオ インコーポレイテッド

F ターム (参考) 4B063 QA01 QA13 QA19 QQ03 QQ08 QQ52 QR08 QR42 QR55 QR62
QS25 QS34 QX01