

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7545891号
(P7545891)

(45)発行日 令和6年9月5日(2024.9.5)

(24)登録日 令和6年8月28日(2024.8.28)

(51)国際特許分類		F I	
C 1 2 Q	1/06 (2006.01)	C 1 2 Q	1/06
C 1 2 Q	1/6881(2018.01)	C 1 2 Q	1/6881 Z
G 0 1 N	33/48 (2006.01)	G 0 1 N	33/48 M
G 0 1 N	33/50 (2006.01)	G 0 1 N	33/50 P
G 0 1 N	33/68 (2006.01)	G 0 1 N	33/68
請求項の数 24 (全93頁) 最終頁に続く			
(21)出願番号	特願2020-502329(P2020-502329)	(73)特許権者	503115205
(86)(22)出願日	平成30年7月19日(2018.7.19)		ザ ボード オブ トラスティーズ オブ
(65)公表番号	特表2020-527946(P2020-527946 A)		ザ レランド スタンフォード ジュニア
(43)公表日	令和2年9月17日(2020.9.17)		ユニバーシティ
(86)国際出願番号	PCT/US2018/042949		アメリカ合衆国 9 4 3 0 5 - 2 0 3 8
(87)国際公開番号	WO2019/018684		カリフォルニア州 スタンフォード メイ
(87)国際公開日	平成31年1月24日(2019.1.24)		ン クウッド ビルディング 1 7 0 サー
審査請求日	令和3年7月16日(2021.7.16)		ド フロア ビー. オー. ボックス 2 0
(31)優先権主張番号	62/535,645		3 8 6 オフィス オブ ザ ジェネラル
(32)優先日	平成29年7月21日(2017.7.21)	(74)代理人	100149294
(33)優先権主張国・地域又は機関	米国(US)		弁理士 内田 直人
		(72)発明者	ニューマン, アーロン エム.
			アメリカ合衆国, カリフォルニア州 9
			4 4 0 2, サン マテオ, パロット ドラ
			最終頁に続く

(54)【発明の名称】 混合細胞集団を分析するためのシステムおよび方法

(57)【特許請求の範囲】

【請求項1】

対象からの生体試料のデジタルサイトメトリーのための方法であって、

(a) 複数の異なる細胞型から生体高分子を含む生体試料を得ることと、

(b) 前記生体試料を処理して、前記生体高分子の特徴プロファイルを生成することであって、前記特徴プロファイルが、前記複数の異なる細胞型に関連する特徴を含む、生成することと、

(c) デコンボリューションモジュールを使用して、前記特徴プロファイルを処理して、前記生体試料中の前記複数の異なる細胞型のうちの少なくとも1つの存在量を定量化することであって、前記存在量を定量化することが、一括修正手順を適用して前記定量化における技術的変動を除去することを含む、定量化することと、を含み、

前記一括修正手順が、単一細胞参照モード（Sモード）またはバルク参照モード（Bモード）で適用され、

(i) 前記Sモードでの一括修正手順の適用が、

(1) 参照マトリックスB及び前記参照マトリックスBが誘導された単一細胞参照プロファイルのセットRが与えられたとき、混合試料の入力セットM内の細胞頻度の推定値F*を得ること、及び、

(2) 前記参照マトリックスBに一括修正手順を適用して、調整された参照マトリックスを得て、前記混合試料の入力セットMに前記調整された参照マトリックスを適用することにより、

単一細胞参照プロファイルから誘導された前記参照マトリックスBと、混合試料の入力セットMとの間の技術的変動を除去することを含み、

(i i) 前記Bモードでの一括修正手順の適用が、

(1) 混合試料の入力セットMの入力された細胞型プロポーションと、参照マトリックスBの対応するプロファイルとの線形結合を含む推定混合試料のセットM*を生成すること、及び、

(2) 前記混合試料の入力セットMに一括修正を実施して、前記混合試料の入力セットMと、前記推定混合試料のセットM*との間のバッチ効果を推定すること、により、

バルク参照プロファイルから誘導された参照マトリックスBと、混合試料の入力セットMとの間の技術的変動を除去することを含む、方法。

【請求項2】

前記生体試料を得ることが、前記生体試料からの細胞の物理的単離を含まない、請求項1に記載の方法。

【請求項3】

前記特徴プロファイルが、複数の遺伝子発現プロファイル(GEP)を含み、前記複数のGEPのそれぞれが、前記複数の異なる細胞型のうちの異なる細胞型に対応する、請求項1に記載の方法。

【請求項4】

前記複数のGEPが、前記複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生成される、請求項3に記載の方法。

【請求項5】

前記単一細胞遺伝子発現測定値が、単一細胞RNA配列決定(scRNA-Seq)によって生成される、請求項4に記載の方法。

【請求項6】

前記存在量が、前記生体試料中の前記複数の異なる細胞型のうちの前記少なくとも1つの部分的存在量である、請求項1から5のいずれか一項に記載の方法。

【請求項7】

前記存在量を定量化することが、前記特徴プロファイルと、第2の複数の異なる細胞型についての特徴シグネチャーの前記参照マトリックスBとの間の回帰を最適化することを含み、前記特徴プロファイルが、前記参照マトリックスBの線形結合としてモデル化される、請求項1から5のいずれか一項に記載の方法。

【請求項8】

前記回帰を最適化することが、前記回帰の回帰係数のセットを解くことを含み、その解が、線形損失関数およびL₂-ノルムペナルティ関数を最小化する、請求項7に記載の方法。

【請求項9】

前記線形損失関数が線形ε不感度損失関数である、請求項8に記載の方法。

【請求項10】

前記回帰を最適化することが、サポートベクトル回帰(SVR)または非負行列因子分解(NMF)を使用することを含む、請求項7に記載の方法。

【請求項11】

前記SVRが、ε-SVRまたはν(nu)-SVRである、請求項10に記載の方法。

【請求項12】

前記生体試料がバルク組織を含む、請求項1から5のいずれか一項に記載の方法。

【請求項13】

前記生体試料が、ホルマリン固定、パラフィン包埋(FFPE)組織または凍結組織を含む、請求項1から5のいずれか一項に記載の方法。

【請求項14】

前記生体試料が、血液試料を含むか、または固形組織試料に由来する、請求項13に記載の方法。

10

20

30

40

50

【請求項 15】

前記固形組織試料が、腫瘍試料を含む、請求項 14 に記載の方法。

【請求項 16】

前記生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む、請求項 1 から 5 のいずれか一項に記載の方法。

【請求項 17】

前記生体高分子が核酸を含む、請求項 16 に記載の方法。

【請求項 18】

前記核酸が RNA 分子である、請求項 17 に記載の方法。

【請求項 19】

前記核酸が DNA 分子である、請求項 17 に記載の方法。

【請求項 20】

前記一括修正手順が、前記単一細胞参照モード（S モード）で適用される、請求項 1 から 5 のいずれか一項に記載の方法。

【請求項 21】

前記一括修正手順が、前記バルク参照モード（B モード）で適用される、請求項 1 から 5 のいずれか一項に記載の方法。

【請求項 22】

同定された前記複数の異なる細胞型のうちの少なくとも 1 つに対して細胞型状態を生成することと、を更に含む、請求項 1 から 21 のいずれか一項に記載の方法。

【請求項 23】

前記細胞型状態が、高解像度の細胞型状態である、請求項 22 に記載の方法。

【請求項 24】

前記細胞型状態が、群レベルの細胞型状態である、請求項 22 に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願の相互参照

本出願は、2017 年 7 月 21 日に出願された米国仮特許出願第 62 / 535, 645 号の優先権の利益を主張し、そのすべてが参照により本明細書に組み込まれる。

【背景技術】

【0002】

組織は、様々な細胞型で構成された複雑な生態系であり得、その異なる発生起源および機能状態によって区別し得る。組織組成を研究するための戦略は、基礎生物学および医学に重要な洞察をもたらすが、細胞の不均一性の包括的な評価は容易ではない。

【発明の概要】

【0003】

本開示は、インシリコ組織解剖を実施して、無傷組織の RNA プロファイルから細胞型存在量および細胞型（例えば、細胞型特異的）遺伝子発現を正確に推定するための計算フレームワークを提供する。単一細胞実験または分類された細胞サブセットからの細胞型発現シグネチャーを活用することにより、本開示のシステムおよび方法は、物理的解離、抗体、または生体物質なしで組織組成の包括的なポートレートを提供できる。このようなアプローチには、例えば、データ正規化およびインシリコ細胞精製のクロスプラットフォーム技術を用いて、組織遺伝子発現プロファイル（GEP）から細胞組成を列挙する方法が含まれ得る。後者により、個々の目的の細胞型のトランスクリプトームを、物理的に分離せずにバルク RNA 混合物からデジタルに「精製」し得る。その結果、細胞分離や予備知識なしで、細胞型遺伝子発現の変化を推測することができる。

【0004】

CIBERSORTx と呼ばれるこのフレームワークの技術的性能および臨床的有用性を説明するために、本開示のシステムおよび方法を使用して、（1）単一細胞発現データ

10

20

30

40

50

でバルク組織組成をプロファイリングし、(2) 解離せずに、悪性および腫瘍関連の細胞型の GEP を精製し、(3) 抗癌療法に関連する細胞型 (例えば、細胞型特異的) シグネチャーを探索し得る。これらの結果は、CIBERSORTx が、研究および臨床環境での高解像度の細胞表現型検査の意味合いをもつ、複雑な組織を解読するための有用なツールであることを示す。

【0005】

CIBERSORTx プラットフォームの機能は、細胞型の列挙および遺伝子発現精製のための統合フレームワーク、クロスプラットフォームの変動を抑制するための専用の正規化スキーム、および RNA 混合物を細胞型 (例えば、細胞型特異的) 発現プロファイルに分離するためのアプローチの改善を含むことができる。これらの結果は、CIBERSORTx が、異なる供給源に由来する発現プロファイルを使用して、組織の不均一性に関する正確なポートレートを提供できることを示す。

10

【0006】

一態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、(a) 複数の異なる細胞型から生体高分子を含む生体試料を得ることと、(b) 該生体試料を処理して、該生体高分子の特徴プロファイルを生成することと、(c) デコンボリューションモジュールを使用して、該特徴プロファイルを処理して、該生体試料中の該複数の異なる細胞型のうちの少なくとも 1 つの存在量を定量化することと、該存在量を定量化することが、一括修正手順を適用して該定量化における技術的変動を除去することを含む、定量化することと、を含む。

20

【0007】

いくつかの実施形態では、該生体試料を得ることは、該生体試料からの細胞 (例えば、単一細胞) の物理的単離を含まない。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル (GEP) を含み、該複数の GEP のそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応する。いくつかの実施形態では、該複数の GEP が、該複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生成される。いくつかの実施形態では、該単一細胞遺伝子発現測定値が、単一細胞 RNA 配列決定 (scRNA-Seq) によって生成される。いくつかの実施形態では、該存在量が、該生体試料中の該複数の異なる細胞型のうちの該少なくとも 1 つの部分的存在量である。いくつかの実施形態では、該存在量を定量化することが、該特徴プロファイルと、第 2 の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスとの間の回帰を最適化することを含み、該特徴プロファイルが、該参照マトリックスの線形結合としてモデル化される。いくつかの実施形態では、該回帰を最適化することは、該回帰の回帰係数のセットを解くことを含み、該解は、線形損失関数および L_2 -ノルムペナルティ関数を最小化する。いくつかの実施形態では、該線形損失関数は、線形 ϵ 不感度損失関数である。いくつかの実施形態では、該最適化することが、サポートベクトル回帰 (SVR) または非負行列因子分解 (NMF) を使用することを含む。いくつかの実施形態では、SVR は、 ϵ -SVR または ν (nu)-SVR である。いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋 (FFPE) 組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記該酸が RNA 分子である。いくつかの実施形態では、該酸が DNA 分子である。いくつかの実施形態では、該一括修正手順が、単一細胞参照モード (S モード) で適用される。いくつかの実施形態では、該一括修正手順が、バルク参照モード (B モード) で適用される。

30

40

【0008】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法

50

を提供し、その方法は、（a）複数の異なる細胞型から生体高分子を含む生体試料を得ることと、（b）該生体試料を処理して、該生体高分子の特徴プロファイルを生成することと、（c）デコンボリューションモジュールを使用して、該特徴プロファイルを処理して、（i）該生体試料中の該複数の異なる細胞型のうちの少なくとも1つを同定し、（ii）該同定された該複数の異なる細胞型のうちの少なくとも1つに対して細胞型状態を生成することと、を含む。

【0009】

いくつかの実施形態では、該デコンボリューションモジュールを使用して、該特徴プロファイルを処理して、該生体試料中の該同定された該複数の異なる細胞型のうちの少なくとも1つの存在量を定量化することをさらに含む。いくつかの実施形態では、前記生体試料を得ることは、前記生体試料からの細胞（例えば、単一細胞）の物理的単離を含まない。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル（GEP）を含み、該複数のGEPのそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応し、該細胞型状態が、細胞型GEPを含む。いくつかの実施形態では、該複数のGEPが、該複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生成される。いくつかの実施形態では、該単一細胞遺伝子発現測定値が、単一細胞RNA配列決定（scRNA-Seq）によって生成される。いくつかの実施形態では、該存在量が、該生体試料中の該複数の異なる細胞型のうちの該少なくとも1つの部分的存在量である。いくつかの実施形態では、該存在量を定量化することが、該特徴プロファイルと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスとの間の回帰を最適化することを含み、該特徴プロファイルが、該参照マトリックスの線形結合としてモデル化される。いくつかの実施形態では、該回帰を最適化することは、該回帰の回帰係数のセットを解くことを含み、該解は、線形損失関数および L_2 -ノルムペナルティ関数を最小化する。いくつかの実施形態では、該線形損失関数は、線形 ϵ 不感度損失関数である。いくつかの実施形態では、該最適化することが、サポートベクトル回帰（SVR）または非負行列因子分解（NMF）を使用することを含む。いくつかの実施形態では、SVRは、 ϵ -SVRまたは ν （nu）-SVRである。いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋（FFPE）組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記該酸がRNA分子である。いくつかの実施形態では、該該酸がDNA分子である。いくつかの実施形態では、該存在量を定量化することが、一括修正手順を適用して該定量化における技術的変動を除去することを含む。いくつかの実施形態では、該一括修正手順が、単一細胞参照モード（Sモード）で適用される。いくつかの実施形態では、該一括修正手順が、バルク参照モード（Bモード）で適用される。いくつかの実施形態では、該細胞型状態が、高解像度の細胞型状態である。いくつかの実施形態では、該細胞型状態が、群レベルの細胞型状態である。

【0010】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、（a）第1の複数の異なる細胞型からの生体高分子を含む複数の生体試料を得ることと、（b）該複数の生体試料のそれぞれについて、該生体試料を処理して、該生体高分子の特徴プロファイルを生成することであって、該特徴プロファイルのそれぞれが、該第1の複数の異なる細胞型に関連する特徴を含み、それによって第1の複数の特徴プロファイルを生成することと、（c）デコンボリューションモジュールを使用して、該第1の複数の特徴プロファイルを処理して、該複数の生体試料のそれぞれについて、該第1の複数の異なる細胞型の中から選択される第2の複数の異なる細胞型の第2の複数の特徴プロファイルを決定し、それによって、該複数の生体試料を、該第2の複数の

異なる細胞型の該第2の複数の特徴プロファイルにデコンボリューションすることと、を含む。

【0011】

いくつかの実施形態では、方法は、該デコンボリューションモジュールを使用して、該特徴プロファイルを処理して、該生体試料中の該第2の複数の異なる細胞型のそれぞれのうちの少なくとも1つの存在量を定量化することをさらに含む。いくつかの実施形態では、該存在量が、該生体試料中の該複数の異なる細胞型のうちの該少なくとも1つの部分的存在量である。いくつかの実施形態では、該生体試料を得ることは、該生体試料からの細胞（例えば、単一細胞）の物理的単離を含まない。いくつかの実施形態では、該第1の複数の特徴プロファイルおよび該第2の複数の特徴プロファイルが、遺伝子発現プロファイル（GEP）を含み、該GEPのそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応する。いくつかの実施形態では、該複数のGEPが、該複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生成される。いくつかの実施形態では、該単一細胞遺伝子発現測定値が、単一細胞RNA配列決定（scRNA-Seq）によって生成される。いくつかの実施形態では、該存在量を定量化することが、該特徴プロファイルと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスとの間の回帰を最適化することを含み、該特徴プロファイルが、該参照マトリックスの線形結合としてモデル化される。いくつかの実施形態では、該回帰を最適化することは、該回帰の回帰係数のセットを解くことを含み、該解は、線形損失関数およびL₂-ノルムペナルティ関数を最小化する。いくつかの実施形態では、該線形損失関数は、線形ε不感度損失関数である。いくつかの実施形態では、該最適化することが、サポートベクトル回帰（SVR）または非負行列因子分解（NMF）を使用することを含む。いくつかの実施形態では、SVRは、ε-SVRまたはv(nu)-SVRである。いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋（FFPE）組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記該酸がRNA分子である。いくつかの実施形態では、該酸がDNA分子である。いくつかの実施形態では、方法は、該決定された第2の複数の特徴プロファイルを使用して、異なる細胞型特異的特徴を同定することをさらに含み、該同定することは生物学的または機能的グループ化の予備知識を含まない。いくつかの実施形態では、該存在量を定量化することが、一括修正手順を適用して該定量化における技術的変動を除去することを含む。いくつかの実施形態では、該一括修正手順が、単一細胞参照モード（Sモード）で適用される。いくつかの実施形態では、該一括修正手順が、バルク参照モード（Bモード）で適用される。

【0012】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、（a）第1の複数の異なる細胞型から生体高分子を含む生体試料を得ることであって、該生体試料を得ることが、該生体試料からの細胞（例えば、単一細胞）の物理的単離を含まない、得ることと、（b）該生体試料を処理して、該生体高分子の特徴プロファイルmを生成することであって、該特徴プロファイルが、該第1の複数の異なる細胞型に関連する特徴を含む、生成することと、（c）該特徴プロファイルmと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスBとの間の回帰を最適化することであって、該特徴プロファイルmが、Bの線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含むfを解くことを含む、最適化することと、（d）少なくとも該回帰係数のセットに基づいて、該第2の複数の異なる細胞型の中から選択される異なる細胞型の存在量を該生体試料中で定量化することと、を含む。

【0013】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、(a) 第1の複数の異なる細胞型から生体高分子を含む生体試料を得ることと、(b) 該生体試料がアーカイブ組織を含む、得ることと、(c) 該生体試料を処理して、該生体高分子の特徴プロファイルmを生成することと、該特徴プロファイルが、該第1の複数の異なる細胞型に関連する特徴を含む、生成することと、(d) 該特徴プロファイルmと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスBとの間の回帰を最適化することと、該特徴プロファイルmが、Bの線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含むfを解くことを含む、最適化することと、(e) 少なくとも該回帰係数のセットに基づいて、該第2の複数の異なる細胞型の中から選択される異なる細胞型の存在量を該生体試料中で定量化することと、を含む。

10

【0014】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、(a) 第1の複数の異なる細胞型から生体高分子を含む生体試料を得ることと、(b) 該生体試料を処理して、該生体高分子の特徴プロファイルmを生成することと、該特徴プロファイルが、該第1の複数の異なる細胞型に関連する特徴を含む、生成することと、(c) 該特徴プロファイルmと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスBとの間の回帰を最適化することと、該特徴プロファイルmが、Bの線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含むfを解くことを含む、最適化することと、(d) 少なくとも該回帰係数のセットに基づいて、該第2の複数の異なる細胞型の中から選択される異なる細胞型の存在量を前記生体試料中で定量化することと、該特徴プロファイルが、複数の遺伝子発現プロファイル(GEP)を含み、該複数のGEPのそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応する、定量化することと、を含む。

20

【0015】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、(a) 第1の複数の異なる細胞型から生体高分子を含む生体試料を得ることと、(b) 該生体試料を処理して、該生体高分子の特徴プロファイルmを生成することと、該特徴プロファイルが、該第1の複数の異なる細胞型に関連する特徴を含む、生成することと、(c) 該特徴プロファイルmと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスBとの間の回帰を最適化することと、該特徴プロファイルmが、Bの線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含むfを解くことを含む、最適化することと、(d) 一括修正手順を使用して、少なくとも該回帰係数のセットに基づいて、該第2の複数の異なる細胞型の中から選択される異なる細胞型の存在量を該生体試料中で定量化することと、該一括修正手順が、該参照マトリックスBと該生体試料との間の技術的変動を除去する、定量化することと、を含む。

30

【0016】

いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋(FFPE)組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記核酸がRNA分子である。いくつかの実施形態では、該核酸がDNA分子である。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル(GEP)を含み、該複数のGEPのそれぞれが、該第1の複数の異なる細胞型のうちの異なる細胞型に対応する。いくつかの実施形態では、該複数のGEPが、該第1の複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生

40

50

成される。いくつかの実施形態では、該単一細胞遺伝子発現測定値が、単一細胞RNA配列決定 (scRNA-Seq) によって生成される。いくつかの実施形態では、該一括修正手順が、単一細胞参照モード (Sモード) で適用される。いくつかの実施形態では、該一括修正手順が、バルク参照モード (Bモード) で適用される。

【0017】

別の態様では、本開示は、対象からの生体試料のデジタル分析のための方法を提供し、その方法は、(a) 第1の複数の異なる細胞型からの生体高分子を含む複数の生体試料Mのそれぞれについて、該生体高分子の特徴プロファイルmを得ることであって、該特徴プロファイルmのそれぞれが、該第1の複数の異なる細胞型に関連する特徴を含む、得ることと、(b) 係数Fの複数のセットを含む行列を得ることと、(c) ノイズ閾値手順を使用して、該複数の生体試料Mから、少なくとも該複数のセットの係数Fに基づいて、該第1の複数の異なる細胞型の中から選択された第2の複数の異なる細胞型の複数の代表的特徴Gを決定することであって、該ノイズ閾値手順が、該複数の代表的特徴Gの該特徴決定におけるノイズを低減する、決定することと、を含む。

【0018】

いくつかの実施形態では、(b) は、該特徴プロファイルmのそれぞれと、該第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスBとの間の回帰を最適化することを含み、該特徴プロファイルmが、Bの線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含むfを解くことを含み、それによって該複数のセットの係数Fを提供する。いくつかの実施形態では、該特徴決定が、信頼できない細胞型特異的遺伝子発現決定を含む。いくつかの実施形態では、該生体試料のそれぞれはバルク組織を含む。いくつかの実施形態では、(a) が、該生体試料からの細胞 (例えば、単一細胞) の物理的単離を含まない。いくつかの実施形態では、該生体試料のそれぞれはアーカイブ組織を含む。いくつかの実施形態では、前記生体試料のそれぞれは、ホルマリン固定、パラフィン包埋 (FFPE) 組織または凍結組織を含む。いくつかの実施形態では、該生体試料のそれぞれが、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記核酸がRNA分子である。いくつかの実施形態では、該核酸がDNA分子である。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル (GEP) を含み、該複数のGEPのそれぞれが、該第1の複数の異なる細胞型のうちの異なる細胞型に対応する。いくつかの実施形態では、該複数の遺伝子発現プロファイルが、該第1の複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生成される。いくつかの実施形態では、該単一細胞遺伝子発現測定値が、単一細胞RNA配列決定 (scRNA-Seq) によって生成される。

【0019】

別の態様では、本開示は、対象からの生体試料のデジタル分析のための方法を提供し、その方法は、(a) 第1の複数の異なる細胞型からの生体高分子を含む複数の生体試料Mのそれぞれについて、該生体高分子の特徴プロファイルmを得ることであって、該特徴プロファイルmのそれぞれが、該第1の複数の異なる細胞型に関連する特徴を含む、得ることと、(b) 複数のセットの係数Fを含むマトリックスを得ることと、(c) 該複数の生体試料Mのそれぞれについて、非負行列因子分解 (NMF) を実行して、少なくとも該複数のセットの係数Fに基づいて、該第1の複数の異なる細胞型の中から選択される第2の複数の異なる細胞型の複数の特徴プロファイルGを決定することであって、それによって、該Fおよび該Mを考慮して、該複数の生体試料Mを該Gにデコンボリューションする、決定することと、を含む。

【0020】

いくつかの実施形態では、(b) は、該特徴プロファイルmのそれぞれと、該第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスBとの間の回帰を最適

化することを含み、該特徴プロファイル m のそれぞれが、 B の線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含む f を解くことを含み、それによって該複数のセットの係数 F を提供する。

【0021】

別の態様では、本開示は、対象からの生体試料のデジタル分析のための方法を提供し、その方法は、(a) 第1の複数の異なる細胞型からの生体高分子を含む複数の生体試料 M のそれぞれについて、該生体高分子の特徴プロファイル m を得ることであって、該特徴プロファイル m のそれぞれが、該第1の複数の異なる細胞型に関連する特徴を含む、得ることと、(b) 複数のセットの係数 F を含むマトリックスを得ることと、(c) 該複数の生体試料 M のそれぞれについて、少なくとも該複数のセットの係数 F に基づいて、該第1の複数の異なる細胞型の中から選択される第2の複数の異なる細胞型の複数の細胞存在量 G を決定することであって、該決定することが、非負行列因子分解(NMF)を実行することを含み、該 F および該 M を考慮して、該複数の生体試料 M を該 G に分離する、決定することと、を含む。

【0022】

いくつかの実施形態では、(b)は、該特徴プロファイル m のそれぞれと、該第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックス B との間の回帰を最適化することを含み、該特徴プロファイル m のそれぞれが、 B の線形結合としてモデル化され、該最適化することが、該回帰の回帰係数のセットを含む f を解くことを含み、それによって該複数のセットの係数 F を提供する。いくつかの実施形態では、該NMFモデルは、部分的観察を含む。いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、(a)が、該生体試料からの細胞(例えば、単一細胞)の物理的単離を含まない。いくつかの実施形態では、該生体試料はアーカイブ組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋(FFPE)組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記該酸がRNA分子である。いくつかの実施形態では、該該酸がDNA分子である。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル(GEP)を含み、該複数のGEPのそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応する。いくつかの実施形態では、該NMFモデルは、分割統治アプローチを含む。いくつかの実施形態では、マトリックス Z は、該 G と該 F の積を含み、該 G および該 Z の非対角要素は不明である。いくつかの実施形態では、方法は、該決定された複数の特徴プロファイル G を使用して、異なる細胞型特異的特徴を同定することをさらに含み、該同定することは生物学的または機能的グループ化の予備知識を含まない。いくつかの実施形態では、該複数の特徴プロファイル G を決定することが、一括修正手順を利用して該決定における技術的変動を除去することを含む。いくつかの実施形態では、該一括修正手順が、単一細胞参照モード(Sモード)で適用される。いくつかの実施形態では、該一括修正手順が、バルク参照モード(Bモード)で適用される。

【0023】

別の態様では、本開示は、対象からの生体試料のデジタル精製のための方法を提供し、その方法は、(a) 複数の異なる細胞型から生体高分子を含む生体試料を得ることと、(b) 該生体試料を処理して、該生体高分子の特徴プロファイルを生成することであって、該特徴プロファイルが、該複数の異なる細胞型に関連する特徴の組み合わせを含む、生成することと、(c) 該生体試料中で、該複数の異なる細胞型の中から選択される異なる細胞型に対応する該特徴プロファイルの断片部分を同定することと、を含む。

【0024】

いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、(a)が、該生体試料からの細胞(例えば、単一細胞)の物理的単離を含まない。いくつ

かの実施形態では、該生体試料はアーカイブ組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋（FFPE）組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記該酸がRNA分子である。いくつかの実施形態では、該該酸がDNA分子である。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル（GEP）を含み、該複数のGEPのそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応する。

【0025】

別の態様では、本開示は、コンピュータシステムの1つ以上のプロセッサによって実行されると、該1つ以上のプロセッサに本明細書に開示の方法を実行させる命令を含む、該1つ以上のプログラムを格納する非一時的なコンピュータ可読記憶媒体を提供する。

【0026】

別の態様では、本開示は、1つ以上のプロセッサを含むシステムと、コンピュータシステムの1つ以上のプロセッサによって実行されると、該1つ以上のプロセッサに本明細書に開示の方法を実行させる命令を含む、該1つ以上のプログラムを格納するメモリと、を提供する。

【0027】

別の態様では、本開示は、個体の疾患を評価する方法を提供し、（a）疾患を有する、または疾患を有する疑いのある個体から生体試料を得ることと、（b）本明細書に開示の方法を実行することにより、該生体試料中の異なる細胞型の存在量を定量化することと、（c）該生体試料中の該異なる細胞型の該推定された存在量を、参照生体試料中の該異なる細胞型の参照存在量と比較することとであって、該参照生体試料は、該疾患を有する個体のコホートに由来し、該異なる細胞型の該推定された存在量は、該疾患の診断または予後の指標である、比較することと、を含む。

【0028】

いくつかの実施形態において、該参照量は、本明細書に開示の方法を実施することにより定量化される。

【0029】

別の態様では、本開示は、個体の疾患を評価する方法を提供し、（a）疾患を有する、または疾患を有する疑いのある個体から生体試料を得ることと、（b）本明細書に開示の方法を実行することにより、該生体試料中の異なる細胞型の特徴を決定することと、（c）該生体試料中の該異なる細胞型の該決定された特徴を、参照生体試料中の該異なる細胞型の参照特徴と比較することとであって、該参照生体試料は、該疾患を有する個体のコホートに由来し、該異なる細胞型の該特徴は、該疾患の診断または予後の指標である、比較することと、を含む。いくつかの実施形態では、該参照量は、本明細書に開示の方法を実施することにより決定される。

【0030】

別の態様では、本開示は、対象の疾患の進行を監視する方法であって、本明細書に開示の方法を第1の時点および第2の時点で実行することを含み、該第2の時点が、該第1の時点より後に発生する、方法を提供する。いくつかの実施形態では、該第1の時点は、該疾患について該対象を治療する前であり、該第2の時点は、該疾患について該対象を治療した後である。

【0031】

別の態様では、本開示は、疾患治療の臨床転帰を予測する方法を提供し、（a）本明細書に開示の方法を実施することにより、対象から得られた試料に存在する複数の異なる細胞型のうちの異なる細胞型の存在量を定量化し、該対象が疾患の治療を受けている、定量化することと、（b）該試料中の該異なる細胞型の該定量化された存在量と、該異なる細胞型と該治療の臨床転帰との所定の関連性とを比較することと、（c）少なくとも該比較

10

20

30

40

50

に基づいて、該治療の該臨床転帰を予測することと、を含む。

【0032】

別の態様では、本開示は、対象からの生体試料のデジタルサイトメトリーのための方法を提供し、その方法は、(a)複数の異なる細胞型から生体高分子を含む生体試料を得ることと、(b)該生体試料を処理して、該生体高分子の特徴プロファイルを生成することと、(c)該特徴プロファイルを処理して、該生体試料中の該複数の異なる細胞型のうちの少なくとも1つを同定することと、(d)デコンボリューションモジュールを使用して、該特徴プロファイルを処理して、該同定された該複数の異なる細胞型のうちの少なくとも1つに対して細胞型状態を生成することと、を含む。

【0033】

いくつかの実施形態では、該生体試料中の該同定された該複数の異なる細胞型のうちの少なくとも1つの存在量を定量化することをさらに含む。いくつかの実施形態では、該生体試料を得ることは、該生体試料からの細胞（例えば、単一細胞）の物理的単離を含まない。いくつかの実施形態では、該特徴プロファイルが、複数の遺伝子発現プロファイル（GEP）を含み、該複数のGEPのそれぞれが、該複数の異なる細胞型のうちの異なる細胞型に対応し、該細胞型状態が、細胞型GEPを含む。いくつかの実施形態では、該複数のGEPが、該複数の異なる細胞型のそれぞれの複数の細胞の単一細胞遺伝子発現測定値から生成される。いくつかの実施形態では、該単一細胞遺伝子発現測定値が、単一細胞RNA配列決定（scRNA-Seq）によって生成される。いくつかの実施形態では、該存在量が、該生体試料中の該複数の異なる細胞型のうちの該少なくとも1つの部分的存在量である。いくつかの実施形態では、該存在量を定量化することが、該特徴プロファイルと、第2の複数の異なる細胞型についての特徴シグネチャーの参照マトリックスとの間の回帰を最適化することを含み、該特徴プロファイルが、該参照マトリックスの線形結合としてモデル化される。いくつかの実施形態では、該回帰を最適化することは、該回帰の回帰係数のセットを解くことを含み、該解は、線形損失関数およびL₂-ノルムペナルティ関数を最小化する。いくつかの実施形態では、該線形損失関数は、線形ε不感度損失関数である。いくつかの実施形態では、該最適化することが、サポートベクトル回帰（SVR）または非負行列因子分解（NMF）を使用することを含む。いくつかの実施形態では、SVRは、ε-SVRまたはν(nu)-SVRである。いくつかの実施形態では、該生体試料はバルク組織を含む。いくつかの実施形態では、該生体試料は、ホルマリン固定、パラフィン包埋（FFPE）組織または凍結組織を含む。いくつかの実施形態では、該生体試料が、血液試料を含むか、または固形組織試料に由来する。いくつかの実施形態では、該固形組織試料が、腫瘍試料を含む。いくつかの実施形態では、該生体高分子が、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含む。いくつかの実施形態では、該生体高分子が、核酸を含む。いくつかの実施形態では、前記該酸がRNA分子である。いくつかの実施形態では、該酸がDNA分子である。いくつかの実施形態では、該存在量を定量化することが、一括修正手順を適用して該定量化における技術的変動を除去することを含む。いくつかの実施形態では、該一括修正手順が、単一細胞参照モード（Sモード）で適用される。いくつかの実施形態では、該一括修正手順が、バルク参照モード（Bモード）で適用される。いくつかの実施形態では、該細胞型状態が、高解像度の細胞型状態である。いくつかの実施形態では、該細胞型状態が、群レベルの細胞型状態である。

【0034】

本開示の追加の態様および利点は、本開示の例示的な実施形態のみが示され説明される以下の詳細な説明から当業者には容易に明らかになるであろう。理解されるように、本開示は、他の異なる実施形態が可能であり、そのいくつかの詳細は、本開示から逸脱することなく、様々な明らかな点について変更可能である。したがって、図面および明細書は、本質的に例示とみなされるべきであり、限定とみなされるべきではない。

【0035】

（参照による組み込み）

本明細書中で言及されるすべての刊行物、特許、および特許出願は、各個々の刊行物、特許、または特許出願が参照により組み込まれることが具体的かつ個別に示されているのと同じように、参照により本明細書に組み込まれる。

【0036】

本発明の新規な特徴は、添付の特許請求の範囲に詳細に記載されている。本発明の特徴および利点のより良い理解は、本発明の原理が利用される例示的な実施形態を記載する以下の詳細な説明、および添付の図面を参照することにより得られるであろう。

【図面の簡単な説明】

【0037】

【図1】インシリコサイトメトリーのフレームワークを示す。CIBERSORTxのワークフローには次のものが含まれる。(1) 単一細胞または分類された細胞垂集団からRNAをプロファイリングして、所与の組織型で目的の各細胞サブセットを識別することができるバーコード遺伝子からなる「シグネチャマトリックス」を定義すること、(2) 細胞型の割合を推測するために、シグネチャマトリックスおよび(3) 代表的な細胞型の発現シグネチャーをバルク組織RNAプロファイルに適用すること、(4) 関連組織試料(例えば、チェックポイント遮断に対するレスポnderおよび非レスポnderからのメラノーマ腫瘍生検)のコホートから各細胞型の複数のトランスクリプトームを精製すること、(5) 細胞型存在量および発現変動のデジタルプロファイルを分析して生物学的または臨床的疑問に対処すること、が含まれる。

【図2】単一細胞参照プロファイルを用いたバルク組織デコンボリューションを示す。(2a) 左: 6つの主要な白血球集団が示された、非小細胞肺癌(NSCLC)患者の末梢血からのscRNA-seqデータを埋め込んでいるt分布型確率的近傍埋め込み法(t-SNE)。右: シグネチャマトリックス遺伝子のヒートマップ。(2b) 一括修正の前後を示した、パネル(2a)のPBMCシグネチャマトリックスを使用した全血(n=12人の健常ドナー)の白血球頻度の列挙。フローサイトメトリーおよび全血球数(CBC)によって決定されるグラウンドトゥルス比率と比較して、パフォーマンスをRMSEで示し、各血液サンプルについて5つの細胞型を評価する。以下の細胞型を評価する。B細胞(B)、NK細胞(NK)、CD8 T細胞(CD8)、CD4 T細胞(CD4)、および単球(Mono)。

統計的有意性は、ウィルコクソン符号順位検定を用いて決定した。(2c) 23個の頭頸部扁平上皮癌腫瘍からのscRNA-seqデータのt-SNE図(左)および単一細胞のデコンボリューション性能のテスト手法(右)。Macはマクロファージを示し、DCは樹状細胞を示し、CAFは癌関連線維芽細胞を示し、Endoは内皮細胞を示し、Myoは筋細胞を示す。(2d) パネル(2c)からの20個のヘルドアウト(held-out)の頭頸部扁平上皮癌(HNSCC)腫瘍について、scRNA-seqおよびCIBERSORTxデコンボリューションによって測定された細胞型比率の一致。すべての腫瘍GEPは単一細胞データから再構築された。(2e) 多様なシグネチャマトリックス、組織、およびプラットフォームにわたる細胞サブセット列挙の分析。デコンボリューションは、Bモード(LM22、HNSCC、および黒色腫シグネチャマトリックス)またはSモード(5' PBMCおよび3' PBMCシグネチャマトリックス)一括修正で実行された(図7a~h)。

グラウンドトゥルス細胞比率は、scRNA-seq(HNSCC、黒色腫)またはフローサイトメトリーおよびコールターカウンター(血液)によって決定された。グレーバンド内の細胞サブセットは、ピアソン相関($P < 0.05$)によってグラウンドトゥルスと有意に一致している。シグネチャマトリックスは、以前に報告されたLM22を除き、図25に示す。HNSCC、頭頸部扁平上皮癌。細胞サブセットの略語: DC、樹状細胞。Mast、マスト細胞。Endo、内皮細胞。CAF、癌関連線維芽細胞。(2f) 左: 10人のドナーからの脾臓サブセットのt-SNE図。そのうち5人はscRNA-seq、バルクRNA-seq、およびIHCによってプロファイリングされた(図8f~iも参照のこと)。右: IHCに対するscRNA-seq(上)およびバルクRNA-seqのCIBERSORTxデコンボリューション(下)によって定量化された4つの主要な脾臓サブセットの

10

20

30

40

50

頻度の一致を示す散布図。

【図3】代表的な細胞型特異的トランスクリプトームプロファイルの精製を示す。(a) インシリコ精製およびグルーブレベルの細胞型特異的GEPの検証のためのアプローチ。(b) デジタルに精製された(x軸)および蛍光活性化細胞選別(FACS)精製された(y軸)B細胞のゲノム全体の発現プロファイルと濾胞性リンパ腫(FL)リンパ節からのCD8⁺T細胞とを比較する散布図。(c) 変動の幾何係数(c.v.)(x軸)によって捕捉された、パネル(b)(y軸)中の各遺伝子の予測された発現レベルをその不確実性の関数として示す散布図。(d)(b)と同じであるが、各細胞型のc.v.値のトランスクリプトーム全体の分布に基づいてノイズフィルタを適用した後である。(e) フィルタありおよびなしで試料サイズ(FL)の関数として評価された、3つのFL細胞サブセットの精製されたおよびグラウンドトゥールストランスクリプトームスピアマンの間の相関。各試料サイズについて、腫瘍を、より大きなコホート(n=319)からの置換なしで10回サブサンプリングした。データは、箱ひげ図として表される(中心線は中央値を示し、ボックス限界は、上部および下部四分位点を示し、ウィスカーは、1.5倍の四分位範囲を示し、点は、異常値を示す)。(f) 3つのFL免疫サブセットの推定したおよびグラウンドトゥールストランスクリプトーム発現プロファイルを、LM22遺伝子のセットを行として、免疫細胞型を列として比較するヒートマップ。発現されると予測されなかったか、または適応ノイズフィルタによって除去された遺伝子は、濃紺色で色付けられる。(g)(d)と同じであるが、免疫(CD45⁺)、上皮(上皮細胞接着分子陽性、EpCAM⁺)、および31NSCLC腫瘍のRNA-Seqプロファイルからデジタルに選別された間質(CD10⁺またはCD31⁺)細胞型である。グラウンドトゥールストランスクリプトームGEP(y軸)は、FACS精製集団から得られた。

【図4】細胞型特異的発現の高解像度精製を示す。(4a) 4つのランダムに混合された免疫サブセットのシリーズについて観察されていない(左)発現データおよび観察された(右)発現データの逸話的例を示す概略図であり、各々が、グラウンドトゥールストランスクリプトーム差動発現遺伝子(DEG)の1つ以上のセットを有する(色付きのブロックのようなパターンによって示される)。(4b) パネルaのバルク組織発現マトリックスに適用されたインシリコ精製の結果。(4c) 各細胞型について別々に行われ、bにおける各ブロックパターンに対応するクラスによって色分けされた(赤=最も左のブロック、青=最も右のブロック)、(b)における精製された発現マトリックスのt-SNE表現。(4d) 左: 4つのランダムに混合された免疫サブセットの合成GEPマトリックスであり、そのうちの1つがブルズアイ(単球)の形状のDEGを含有する。右: インシリコ精製の結果。(4e) 試料クラスが既知であるときの合成混合物中のDEG回収の感受性および特異性の分析。左: 所与の倍率変化を有するDEGをCD8⁺T細胞の参照プロファイルに追加し、これを所定の割合で3つの他の免疫サブセットのランダム混合物にスパイクした。結腸癌細胞株GEPを各混合物に添加して、約50%未知の含有量をシミュレートした。インシリコ精製後、既知のCD8⁺T細胞DEGを、既知のDEGクラスに対する倍率変化によって遺伝子をランク付けすることによって、各精製された細胞型において評価した。DEG回収の曲線下面積(AUC)を、倍率変化および細胞型スパイクイン画分の各組み合わせについて計算し、ヒートマップとして示した。パネル(4a)、(4b)、および(4d)におけるデータを、ヒートマップをレンダリングする前に、各遺伝子についてlog₂を調整し、中央値を中心にした。

【図5】バルク腫瘍生検の高解像度発現プロファイルを示す。(5a)~(5c) 濾胞性リンパ腫(FL)腫瘍からのB細胞におけるCREBBP変異関連発現変化の分析。(5a) スキーマ。(5b) CIBERSORTxによって推定されるCREBBP変異体FL B細胞GEPにおけるMHCクラスII発現の喪失を確認するヒートマップ(左)および散布図(右)。プロットする前に、発現値は中央値が中心であった。(5c) CREBBP変異体FL腫瘍におけるより低い(左)またはより高い(右)B細胞発現に関連する公開された遺伝子セットの分析。スキャッタープロットは、中央値を中心にし、CREBBP変異状態によって平均化した後の、デジタルに選別されたB細胞GEPにおける各

遺伝子の対応する $\log 2$ 発現を示す。B および c におけるグループ比較は、ウィルコクソン符号順位検定によって評価した。WT、野生型 ($n=9$) ; MT、変異体 ($n=15$)。 (5 d) ~ (5 f) 非小細胞肺癌 (NSCLC) 腫瘍からの腫瘍細胞サブ集団の高解像度発現プロファイリング。 (5 d) TCGA からの 1, 020 個の NSCLC 腫瘍 GEP における上皮細胞 (Epith., EpCAM⁺)、免疫細胞 (Imm., CD45⁺)、線維芽細胞 (Fib., CD10⁺)、および内皮細胞 (Endo., CD31⁺) の発現シグネチャーをプロファイリングおよび検証するためのスキーマ、 (5 e) 左: 1, 020 個の NSCLC 腫瘍から推定された集団特異的転写多様性を示す tSNE プロットであり、腫瘍組織学および隣接する正常組織を示すために色付けされている。プロットを 10 に設定されたパープレキシティと共に作成した。右: 腫瘍組織学によって色付けされた、対応する FACS 精製集団からの GEP の PCA プロット。組織学的差はまた、陰影のある卵巣によっても強調されている。 (5 f) 20 名の NSCLC 患者からの対応する FACS 精製集団の TCGA (左) および RNA-Seq プロファイルに特定された上皮集団、免疫集団および線維芽細胞集団における DEG を示すヒートマップ (右)。中央値の中央化を各細胞集団に別々に適用した。LUAD、肺腺癌、LUSC、肺扁平上皮癌。

【図 6】 黒色腫におけるドライバ変異状態および免疫療法応答の細胞シグネチャーを示す。 (6 a) ~ (6 b) 黒色腫腫瘍細胞型における変異関連表現型状態の高解像度発現プロファイリング。 (6 a) 上部: scRNA-Seq によってプロファイリングされた 8 個の主細胞黒色腫腫瘍細胞型を示す tSNE プロット。下部: TCGA からの 301 バルク黒色腫腫瘍におけるコンテキスト依存的発現変異を特徴付けおよび検証するためのスキーマ。 (6 b) 黒色腫におけるキードライバ変異に関連する CIBERSORTx によって特定された細胞型特異的 DEG を示すヒートマップ。明確性のために、TCGA 腫瘍試料 (CIBERSORTx、左) および単一細胞 GEP (scRNA-Seq、右) にわたって発現値を平均した。基礎データを、図 19 に提供する。CIBERSORTx によって検出された DEG を有する細胞型および利用可能な単一細胞検証データのみが示される。CAF、癌関連線維芽細胞; MT、変異体; WT、野生型。 (6 c) 左: 正常な CD8 T 細胞参照プロファイルと以下の 3 つの黒色腫 CD8 TIL GEP との間のランクとして表される遺伝子発現の絶対差を示すヒートマップ: 「推定された (FF)」、TCGA によって生成された新鮮/凍結 (FF) 腫瘍の 473 個のバルク RNA-Seq プロファイルから推定されたグループレベル CD8 TIL GEP; 「推定された (FFPE)」、42 個のホルマリン固定パラフィン、包埋 (FFPE) 腫瘍から予測されたグループレベル CD8 TIL プロファイル; 「scRNA-Seq」、19 個の黒色腫患者の集約された単一細胞データに由来するグループレベル CD8 TIL プロファイル。ヒートマップは、「推定された (FF)」ベクターおよび正常な CD8 T 細胞 GEP のランクの差によって順序付けされる。すべての推定プロファイルは、適応ノイズフィルタによって処理された。FF コホートで予測されるが、FFPE コホートでは予測されない遺伝子 (ノイズフィルタによる) は、後者では灰色で色付けられる。選択された T 細胞疲弊遺伝子が示される。右: ヒートマップにおける遺伝子の順序に対して評価した、以前に公開された黒色腫 CD8 TIL 関連遺伝子セットの遺伝子セット濃縮分析 (GSEA)。 (6 d) 単一の参照プロファイルを使用して、バルク黒色腫腫瘍における PD CD1⁺ CTLA4⁺ CD8 T 細胞の臨床関連性を特徴付けるためのフレームワーク。PD CD1⁺ CTLA4⁺ CD8 T 細胞および残りの CD8 T 細胞は、橙色および濃紺色によって示される。 (6 e) 応答者および非応答者によって分層された、免疫療法を受けた黒色腫患者からのバルク腫瘍発現プロファイルにおける PD CD1⁺ CTLA4⁺ CD8 T 細胞の推定レベル。データは、データ範囲全体を包含するウィスカーを有する箱ひげ図として表される。統計的比較は、マン・ホイットニー検定を用いて行った。 (6 f) 治療前腫瘍における PD CD1⁺ CTLA4⁺ CD8 T 細胞の推定レベルが高いおよび低い黒色腫患者間の全生存率の差を示す Kaplan-Meier プロット。患者を、PD CD1⁺ CTLA4⁺ CD8 T 細胞の推定画分存在量中央値によって、高グループおよび低グループに分割し、その後、利用可能な全生存率データを有する 2 つの研究にわたってプールした。統計的

10

20

30

40

50

有意性は、ログ順位検定によって計算した。HRは、ハザード比を示す。95%のHR信頼区間を括弧に示す。

【図7】CIBERSORTxを用いたクロスプラットフォームデコンボリューションを示す。(7a) 所与の入力シグネチャーマトリックスに対して最も適切なCIBERSORTx一括修正戦略を選択するための決定木：(i) 単一細胞参照モード(「Sモード」、パネル(7b)、(7e)、または(ii) バルク参照モード(「Bモード」、パネル(7c)、(7f)～(7h))。(7b)～(7c) Sモード(7b) およびBモード(7c) 一括修正の概略図。(7d) 3'キットおよび5'キットを用いた10倍のゲノムによってプロファイリングされたPBMCの2つの一般に入手可能な単一細胞トランスクリプトームデータセットのt-SNE投影図。細胞標識を、Seurat (例えば、v1.4.0.16) および正規マーカー遺伝子評価を使用して手動で決定し、シグネチャーマトリックスを生成するために使用した。(7e) Sモードの評価であって、10倍のクロム(パネル(7d))によって生成されたPBMCサブセットの単一細胞参照プロファイルを使用した、健康なドナー(n=12個のRNA-Seqプロファイル)からの全血の白血球デコンボリューションを示す。Sモード一括修正の前後のピアソン相関(左)とRMSE(右)。グラウンドトゥルス細胞比率は、フローサイトメトリーおよび全血球数によって決定された。グループの差は、ウィルコクソン符号順位検定によって評価した。(7f) Bモード一括調整の前後の両方において、RNA-Seqによってプロファイリングされた固定黒色腫腫瘍試料中のLM22免疫サブセットを列挙するためにCIBERSORTxを適用した結果。グラウンドトゥルス比なしの性能を評価するために、選択された白血球サブ集団のデコンボリューション結果を、未調整のRNA-Seqデータにおける正規マーカー遺伝子の発現と比較した。CIBERSORTx結果を、以前に記載されたアプローチを用いて全免疫含有量によってスケーリングし、各腫瘍試料の「正規化された免疫指標」を計算した。(7g) 5プラットフォーム、3つの組織型、および新鮮/凍結対固定組織保存状態にわたるLM22(アフィメトリクスマイクロアレイ)による細胞サブセット列挙に対するBモード一括修正の影響：凍結保存されたPBMC対フローサイトメトリー(GSE65133)のIllumina Beadchipマイクロアレイ、新鮮な全血対フローサイトメトリー/CBCのバルクRNA-Seq(本明細書の他の箇所で開示されるように)、FFPE黒色腫(Mel.)腫瘍対代用細胞マーカーのバルクRNA-Seq(パネルbと同じ)、ならびにscRNA-Seqプロファイルからの再構築されたバルク黒色腫(Mel)およびHNSCC腫瘍(それぞれSmart-Seq2; 黒色腫瘍およびHNSCC腫瘍からの再構築された単一細胞発現データにおけるTILサブセット頻度)。点は、各細胞型の予測される画分と推測される画分との間のピアソン相関係数を表し、バープロットは、エラーバーとして表される95%信頼区間を有する中央値を示す。(7h) 左：一括修正後のランクの絶対差によってランク付けした、(c)の同じデータセットにおけるCIBERSORTx一括修正のトランスクリプトーム全体の影響を示すヒートマップ。右：左パネルのランク付けされた遺伝子リストおよび黒色腫における非造血細胞に高度に特異的である遺伝子のセット(n=139)に対する事前ランク付けされた遺伝子セット濃縮分析(GSEA)の適用。遺伝子セットは、非免疫細胞と免疫細胞との間のlog2空間において2を超えるシグナル対ノイズ比(SNR)を有する遺伝子の黒色腫scRNA-Seqデータを分析することによって導出された。この分析の前に、単一細胞データを集計して、細胞型ごとに5回のトランスクリプトーム複製を得た。NESは、正規化された濃縮スコアを示す。

【図8】単一細胞参照プロファイルを用いた固形腫瘍および脾島の解剖を示す。(8a) 図2dにおける結果の積み重ねられたバープロットの描写であり、単一細胞トランスクリプトームから再構築された20個のHNSCC腫瘍試料における推定(左)および予測(右)細胞サブセット頻度を示す。(8b)～(8e) 黒色腫腫瘍の単一細胞誘導型デコンボリューション。(b) 19個の黒色腫瘍からのscRNA-Seqデータのt-SNE可視化(左)および単一細胞デコンボリューション性能を試験するためのアプローチ(右)。(8c)～(8d) パネル(8b)に関連する、単一細胞データから再構築された1

10

20

30

40

50

7個のヘルドアウト (held-out) 黒色腫瘍について、CIBERSORTxデコンボリューションによって測定された細胞型比率とscRNA-Seqとの一致。観察された細胞サブセット比対予測された細胞サブセット比率の積み重ねられたバープロット (左) およびピアソン相関 (右) が示される。(8e) scRNA-Seqによって測定されるような黒色腫TIL比率の分析、パネル(8b)のシグネチャマトリックスを使用した皮膚皮膚黒色腫瘍のバルクRNA-Seqプロファイルのデコンボリューション、およびIHC定量化。(8f)~(8i) 脾島の単一細胞誘導デコンボリューション。(8f) 10個の頭頸部扁平上皮癌腫瘍からのscRNA-Seqデータのt-SNE可視化 (左) および単一細胞デコンボリューション性能を試験するためのアプローチ (右)。(8g) 8個のヘルドアウト (held-out) の島 (左) における9個の脾臓細胞型に関するscRNA-SeqおよびCIBERSORTx比率を、個々の細胞型の対応するピアソン相関係数 (右) と比較する散布図。散布図のデータポイントは、バープロット (右) に一致するために細胞型によって色付けされる。(8h) (8g) と同じであるが、バルクRNA-SeqおよびscRNA-Seqによってプロファイリングされた7個の島試料についてである。(g) における再構築された腫瘍を分析するために使用される同じ単一細胞由来のシグネチャマトリックスを、バルクRNA-Seq (x軸) によってプロファイルされる同じ島試験片に適用した。顕著な細胞型を楕円で強調する。(i) (1) scRNA-Seq (注釈された細胞標識を用いて)、(2) 単一細胞由来のシグネチャマトリックス (パネル(8f)) を用いたバルク島のCIBERSORTxデコンボリューション、および(3) 非分解島におけるIHC定量化によって測定されるように、脾島におけるアルファ、ベータ、ガンマ、およびデルタ細胞比率の比較。(8e) および(8i) のデータは、平均+/-標準誤差として表される。

【図9】バルク組織発現プロファイルにおける特定の細胞サブセットのデコンボリューションのためのデジタルゲーティングの評価を示す。(9a) PDCD1+/CTLA4+CD8 T細胞を、他の8個の主要な黒色腫瘍細胞型と共に列挙するための単一細胞由来のシグネチャマトリックスを作成および検証するためのスキーム (パネル(b)、左上に示す)。(9b) パネル(9a) におけるフレームワークを適用して、PDCD1+CD8 T細胞の性能を試験するもの、CTLA4+CD8 T細胞を評価するもの、および組織レジデントメモリ (TRM) CD8 T細胞を評価するものの3つの追加のシグネチャマトリックスを作製した。シグネチャマトリックス構築から保持された17個の再構築された黒色腫瘍に適用される各シグネチャマトリックスのデコンボリューション性能を示す棒グラフが示される。各細胞サブセットに対して、推定された細胞比率と既知の細胞比率との間のピアソン相関 (および対応する統計的有意性) が提供される。

【図10】単一細胞誘導型デコンボリューションに対するキーパラメータの影響を示す。(10a) パラメータスイープのスキーム。細胞比率を推定するためのデコンボリューション性能を、表現型あたりの単一細胞の数、シグネチャマトリックスあたりのドナーの数、および表現型あたりのマーカー遺伝子の範囲に関して評価した。(10b) 単一細胞RNA-Seqによってプロファイルされた2つの腫瘍型のパネル(10a) におけるパラメータスイープの結果: 頭頸部扁平上皮癌 (HNSCC) 腫瘍および黒色腫瘍。推定された細胞比率と既知の細胞比率との間の一致を、ピアソン相関によって各細胞型について評価し、各パラメータの組み合わせの中央値相関係数を表形式で示す。金色および灰色は、それぞれ、より高い相関係数およびより低い相関係数を示す。(10c) パラメータ範囲の反対端から定義されるシグネチャマトリックスの一貫性 (TCGAによってプロファイルされたバルク黒色腫瘍からのグループレベルGEPを解消するためのパネル(10b) 内の緑色および赤色の長方形によって示される)。ピアソン相関は、log₂調整された発現値上で行った。適応フィルタリング後に検出可能な発現を有する遺伝子のみを分析した。

【図11】グループレベル発現精製に対するキーパラメータの影響を示す。(11a) 図3dに関連して、302個のFL腫瘍試料から得られたB細胞、CD8 T細胞、およびCD4 T細胞の推定されたグループレベルGEPのブランドアルトマンプロット。整合

10

20

30

40

50

性は、線形回帰線および整合性相関係数（CCC）によって定量的に捕捉される。（11b）図3eに関連して、図3eと同じであるが、グループレベルGEP推定に使用される腫瘍の数の関数としてCCC値の分布を示す。（11c）図3eに関連するが、発現推定に使用されるFL腫瘍の数に対する検出可能な遺伝子（例えば、非ゼロ発現を有する遺伝子）の数を示す。データは箱ひげ図として表される（中央線、中央値；ボックス限界、上位および下位四分位数；ウィスカー、四分位範囲の1.5倍；点、異常値）。（11d）302個のFL GEPから濾胞性ヘルパーT細胞のトランスクリプトームを再構築するためのグループレベル発現精製の分析。プロファイルを、図3dにおけるB細胞およびCD8 T細胞について説明されるように生成した。パネルa、b、およびdにおける推定したGEPを、分析前に適応ノイズフィルタによって処理した。（11e）図3eと同じであるが、分析された混合物試料の数と、適応ノイズフィルタの前後に示される、大規模なPBMC試料コホート上の白血球トランスクリプトーム推定の性能との間の関係を示す。試料を、PBMCコホート全体（ $n=83$ ）から5回置換せずにサブサンプリングした。LM22シグネチャーマトリックスを、細胞型の列挙後およびトランスクリプトーム精製前に、10個の主な白血球サブセットに崩壊させた。白血球参照プロファイルは、健康なドナーから得られた選別された末梢血液集団に由来し、正規化された。（11f）（11e）と同じであるが、非ゼロ発現の遺伝子の回収に対するPBMC試料の数の影響を示す。（11e）および（11f）の線は、標準誤差を示すエラーバーでの手段として表される。（11g）および（11h）（11e）および（11f）と同じであるが、同じPBMCデータセット上の適応ノイズフィルタ（それぞれ、「フィルタされていない」および「フィルタされた」）の前後の両方において、トランスクリプトーム精製（11g）および非ゼロ発現（11h）を有するタンパク質コード遺伝子の画分の両方に対する推測された白血球存在量の影響を示す。（d）における線形回帰線を、より低い細胞画分のパフォーマンスへの影響をモデル化するために $x=0$ に推定した。

【図12】高解像度発現精製のためのCIBERSORTxフレームワークおよびアプローチを示す。（12a）（1）細胞比率の列挙（上部）、（2）代表的な細胞型特異的トランスクリプトームの推定（中央）、および（3）高解像度トランスクリプトーム精製（下部）の、CIBERSORTxフレームワーク内の主要な客観的機能。対応する主な図は右に示され、不明な量は橙色で色付けられる。（12b）高解像度CIBERSORTxの客観的機能。ここで、Mは既知であり、Fは以前に推定されており（パネル（a）、上部）、推定対象であるGは、 n 遺伝子 $\times c$ 混合物試料 $\times k$ 細胞型の3次元マトリックスである。（12c）簡単な逸話的例によって例示される高解像度CIBERSORTxの概略図。ここでは、単一遺伝子は、単球およびナイーブB細胞において相互排他的な発現パターンを有し、CD8 T細胞およびNK細胞において差次的発現を有さない4つの免疫サブセットによって発現される。4つのプロファイルを、バルク混合物GEP M_i にランダムに混合する。アルゴリズムは、5つの主なステップで進む：（i）バルクプロファイルを昇順に選別する。（ii）選別したベクターを反復 t によって分割し、ブートストラップドNNLSをランク付けされたベクターのいずれかの側に適用して、代表的な細胞型特異的発現係数（ g^t_1 および g^t_2 ）を推定する。回帰制約およびウィンドウ長 w を満たす t のすべての可能な値についてこのプロセスを繰り返す。（iii） t の各値について、 g^t_1 、 g^t_2 、およびFを使用して M_i を推測し、 M_i の最適な近似をもたらす g^t_1 、 g^t_2 対を選択する。（iv）遺伝子 i の発現推定値が各細胞型に対して有意であるかどうかを評価し、その場合、ステップiiiから最適な g_1 および g_2 対の各細胞型間で発現が有意に異なるかどうかを評価する。（v）ブートストラップされたNNLSおよびスライドウィンドウを使用して G^*_i の初期推定値を推定し、ステップivに基づいて g_1 、 g_2 を絞り込み、 G_i をアンカー係数（調整された g_1 、 g_2 ）に適合し、 G_i の推定値を最終化するために元の順序を復元する。

【図13】バルク組織におけるDEGの発現と比較した単一細胞において特定されたDEGの分析を示す。（13a）いくつかの主要な膵島細胞型（ x 軸）において、健康なドナーとII型糖尿病（T2D）を有する患者との間の、単一細胞（ y 軸）から再構築された

10

20

30

40

50

バルク島に対する差次的発現遺伝子の倍率変化を示す散布図。単離された細胞型におけるすべてのDEGおよび対応する倍率変化は、例えば、参照によりその全体が本明細書に組み込まれる、Segerstolpe et al. (Single Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes, Cell Metab 24, 593–607 (2016)) によって補足表6に記載されているように得ることができる。バルク島の倍率変化は、(例えば、Segerstolpe et al. によって記載されるように) 各膵島試料片中のすべての単一細胞GEPを合計し、得られたトランスクリプトームをTPMに正規化し、各DEGの健常試験片とT2D試験片との間のlog2倍変化を計算することによって得られた。全体として、細胞型特異的DEGの98%は、バルク島で分析されるか、または個々の細胞型で分析されるかにかかわらず、同じ方向で倍率変化を示した。(13b)(13a)と同じであるが、単離された細胞型(x軸)におけるDEGの倍率変化を、各遺伝子を昇順の発現に選別した後のT2Dドナーの相対歪み(y軸)と比較する。濃縮を計算するために、T2Dを有する患者が1として表され、健常なドナーが0として表されるように、二元マトリックスを定義した。次いで、各遺伝子について、上位50パーセンタイルの発現におけるT2D患者の画分を、0.5から差し引いた。したがって、歪みのない遺伝子は0の濃縮を受け、T2Dに対して歪んだ遺伝子は負の値を受け、健常なドナーに対して歪んだ遺伝子は正の値を受けた。これにより、x軸(例えば、健常/T2D)における倍率変化方向との均等な比較が可能になった。この分析は、各遺伝子をバルク混合物におけるその発現によって選別することが、細胞型特異的転写異質性を通知することができることを示す。この概念は、CIBERSORTxによる高解像度トランスクリプトーム精製の最初のステップに活用される(図12c)。

【図14】重複パターンを有するDEGの高解像度精製を示す。左：4つの免疫サブセットの400個の合成GEP混合物を示すヒートマップであり、それぞれ固有の細胞型特異的DEGパターン：標的、一連の対角線に整列された正方形、および各々の逆を含む。右：高解像度CIBERSORTx結果を示すヒートマップ。

【図15】高解像度発現精製のパラメータスイープを示す。(15a)8個の主な細胞型を含む再構築された黒色腫瘍におけるDEG回収を評価するためのパラメータスイープを描写するスキーム。(i)DEGを有する細胞型の総数(1、2、3、...、8)、(ii)腫瘍試料の総数(50、100、200)、および(iii)DEGのlog2倍変化(1、2、および3)を系統的に変化させることによって性能を評価した。図4eのようにDEG回収を評価した。細胞画分を、元のデータセットにおけるそれらの頻度に従って均一にサンプリングした。(15b)人工DEGを有する細胞型の数および(下部)DEGのlog2倍変化の関数(上部)として、インシリコ腫瘍における異なる数の既知の細胞型特異的DEGの回収の曲線下面積(AUC)を示すヒートマップ。(15c)AUCを除き、(15b)と同じであり、非ゼロ変異を有する推定した遺伝子に関して計算される。したがって、発現の証拠が不十分な遺伝子は考察から除外される。色付きバーの垂直線は、0.1のAUC間隔を示す。パネル(15b)に示されるように、DEG回収は非常に特異的であり、1つの細胞型からの既知のDEGが別の細胞型で評価されるとき、誤った検出率は、ランダムな確率(約0.5のAUC)以下である。より多くの腫瘍試料がプロファイリングされ、より高い倍率変化が評価されるにつれて、DEG検出率が増加する。性能の適度な低減は、DEGを有する細胞型の数の関数としてみなされ得る。しかしながら、これらの結果は、細胞型特異的発現の重要でない証拠を有する遺伝子を含む、すべての遺伝子を反映する。パネル15(c)に示されるように、可変発現を有する遺伝子間の性能(例えば、非ゼロ変異で推定される)を考慮すると、DEG回収のためのAUCは、一貫して高い(平均=0.93、標準偏差=0.04)。

【図16】DLBCL分子サブタイプの高解像度精製を示す。(16a)びまん性大細胞型B細胞リンパ腫(DLBCL)腫瘍(例えば、Lenz et al. (Stromal Gene Signatures in Large-B-Cell Lymphomas

、New England Journal of Medicine 359, 2313–2323 (2008)、参照によりその全体が本明細書に組み込まれる)によって記載されるような、GCB/ABC DLBCL、CHOPコホート)を、高解像度精製によって解剖するスキーマ。(16b)左:DLBCL腫瘍からデジタルに精製したB細胞トランスクリプトーム内のABCとGCB DLBCLとの間で特定されたDEGを示すヒートマップ。試料は、以前に注釈されたサブタイプによってグループ化される。右:左パネルと同じ順序で同じ遺伝子を示すヒートマップであるが、健常なドナーに由来する生殖中心細胞および活性化B細胞について示されている。両方のヒートマップにおける発現データを、ヒートマップをレンダリングする前に、各遺伝子について $\log_2$ を調整し、中央値を中心にした。(16c)バルクDLBCL腫瘍とデジタルに精製されたB細胞とを比較する、以前に公開されたマーカー遺伝子に対するGCBとABC DLBCLとの間の差次的発現の有意性($-\log_{10}p$ 値は、同等でない変異を有する両側非対応t検定から)。グループ間の統計的有意性は、両側対t検定を用いて決定した。デジタルに精製されたB細胞における非ゼロ発現を有するマーカー遺伝子のみを評価した(144個の遺伝子のうち $n=141$ 個; SPINK2、TOX2、およびFAM46Cを他の細胞型に割り当てた)。

10

【図17】 相関ベースのアプローチに対する高解像度精製の相対有用性を示す。(17a) マイクロアレイによってプロファイルされた活性化および生殖中心B細胞を区別するDEGの識別および評価のためのスキーマ。(17b) パネル(17a)に記載されているように識別されたDEGを示すヒートマップ。(17c) 150個のバルクDLBCL腫瘍(例えば、Lenz et al.によって記載されるような、GCB/ABC DLBCL、CHOPコホート)における(17b)(行)からの各遺伝子と最も相関する細胞型(10個の評価された免疫サブセット; 列)を示すヒートマップ。示された系統に崩壊したLM22画分を使用して、各遺伝子の非対数発現ベクターと、すべてのバルク腫瘍にわたる各細胞型の推定した画分存在量との間のピアソン相関として相関を計算した。遺伝子は、パネル(17b)および(17c)において同一に順序付けされる。(17d) および(17e) 以前に注釈された分子サブタイプ(ABC DLBCL、青色; GCB DLBCL、橙色)によって色付けされた、DLBCL腫瘍における5個の最も存在する細胞型間の転写変異($\log_2$ 調整)を示すPCAプロット。パネル(17d)における各PCAプロットは、パネル(17c)における各遺伝子をその最も相関する細胞型に比類なく割り当てることによって生成された(例えば、B細胞には、(17c)の最も左側の列で黄色で強調された遺伝子のみが使用された)。パネル(17e)における各プロットは、PCAを高解像度発現精製の出力に適用することによって生成し、これを相関ベースの戦略と同じ150個のバルクDLBCL腫瘍GEP上で実行し、フィルタリングして、分析を同じ1,100個のDEG(パネル(17a)に記載されるように特定した)に限定した。

20

30

【図18】 以前の方法に対する高解像度遺伝子発現精製のベンチマーキングを示す。(18a) 生殖中心B細胞(GCB)と活性化B細胞(ABC)DLBCLとの間のB細胞において識別されたDEGの分析(Lenz et al.によって記載されるように、 $n=150$ 個の腫瘍; GCB/ABC DLBCL、CHOPコホート)。DEGの偽発見率(FDR)(y軸)および数(x軸)に関連する高解像度精製およびcsSAMの性能を示す。LM22を用いて細胞比率推定値を得て、同じ推定値を両方の方法への入力として用いた。GCBおよびABC標識をcsSAMへの入力として提供したが、CIBERSORTxはクラス標識の事前知識なしに実行した。GCBとABC DLBCLとの間のDEGを特定するために、マイクロアレイの有意性分析(SAM)をCIBERSORTx結果に適用した。両方の方法について、Samを100個の順列で実行した。(18b) 健常なドナーに由来する活性化または生殖中心B細胞GEPにおけるそれらの濃縮についてCIBERSORTxで指定されたDEG(パネル(18a)から)の遺伝子セット濃縮分析。NES、正規化濃縮スコア。(18c) NSCLC腫瘍の18個のバルクRNA-Seqプロファイルからの上皮細胞の試料レベルGEP推定のためのISOpure

40

50

に対するCIBERSORTxのベンチマーキングのためのスキーマ。(18d)パネル(18c)からの各推定した上皮トランスクリプトームを、ピアソン相関によって同じ患者からのその対応するFACS精製GEPに対して比較し、結果を箱ひげ図(18個の試料)として示す。方法間の差異を、ウィルコクソン符号順位検定によって評価した。(18e)18個のバルクNSCLC腫瘍のバルクRNA-Seqプロファイルから精製された4個の主要細胞サブセットのトランスクリプトームを推測するための高解像度CIBERSORTxの精度。性能を評価するために、各試料レベルトランスクリプトームを、(i)対応する未選別バルク腫瘍および(ii)同じ患者由来の対応するFACS精製細胞サブセットのRNA-Seqプロファイルと比較した。細胞選別スキームは、図5dに提供される。グループ比較を、ウィルコクソン符号順位検定によって評価した。

10

【図19】黒色腫瘍細胞サブセットにおけるドライバ変異関連DEGの予測および検証を示す。図6bに関連して、バルク黒色腫瘍(TCGA)の高解像度精製によって識別されるように、(19a)BRAFまたは(19b)NRAS変異状態に関連する細胞型特異的DEGを示すヒートマップ。結果は、TCGAからのインシリコ精製されたトランスクリプトーム(左)および検証コホートにおけるscRNA-Seqプロファイル(右)について示される。遺伝子は、デジタルに精製されたトランスクリプトームにおける変異型グループと野生型グループとの間のそれらのlog₂倍変化によって、上から下に組織化される。上位300個の過剰発現または低発現遺伝子のみが示される。発現の希薄性のため、scRNA-Seqプロファイルもまた、離れた右の変異状態によって平均化される。プロットする前に、遺伝子レベル発現データを、log₂空間でZスコア正規化した。

20

【図20】新鮮／凍結されたおよび固定された黒色腫瘍から推定されたTIL GEPの比較を示す。白血球トランスクリプトーム細胞を、新鮮／凍結(TCGA; n=473個の腫瘍)の独立コホートからのCIBERSORTxで推定し、RNA-Seqによってプロファイリングされたバルク黒色腫瘍を固定した(n=42個の腫瘍)(例えば、参照によりその全体が本明細書に組み込まれる、Van Allen et al., Genomic correlates of response to CTLA-4 blockade in metastatic melanoma, Science 350, 207-211 (2015)によって記載されるように)。非ゼロ発現を有する遺伝子のみが示される。LM22シグネチャーマトリックスを、細胞型の列挙後およびトランスクリプトーム精製前に、11個の主な白血球サブセットに崩壊させた。両方のデータセットに適応ノイズフィルタを適用した。SKCMは皮膚皮膚黒色腫を示す。

30

【図21】黒色腫患者におけるPDCD1+/CTLA4+CD8⁺T細胞の生存率分析を示す。(21a)9個の腫瘍細胞サブセットを区別する単一細胞由来の黒色腫シグネチャーマトリックス(図9b、左上)と、nCounter NanoStringプラットフォームとの間の重複を示すベン図(例えば、その全体が参照により本明細書に組み込まれる、Chen et al., Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade, Cancer Discovery (2016)によって記載されるように)。(21b)図9b(左上パネル)と同じ分析であるが、パネル(21a)に関連するNanoString nCounterアッセイと重複する245個のシグネチャーマトリックス遺伝子を使用する。すべての腫瘍はシグネチャーマトリックス構築からヘルドアウト(held-out)単一細胞GEPから再構築された。(21c)~(21f)免疫チェックポイント遮断に対する応答を予測するための遺伝子発現シグネチャー対デコンボリューション。(21c)PDCD1+/CTLA4+CD8⁺T細胞の推定レベルと、データセットにおける全生存率との間の関係を示す Kaplan-Meier 生存率プロット(例えば、参照によりその全体が本明細書に組み込まれる、Van Allen et al. および Nathanson et al. (Somatic Muta

40

50

tions and Neoepitope Homology in Melanomas Treated with CTLA-4 Blockade, Cancer Immunology Research (2016) によって記載されるような)。各データセットにおけるそれらの中央値によって、PDCD1+／CTLA4+CD8 T細胞レベルを分層した。(21d) 抗CTLA4療法で治療した黒色腫患者における全生存率を予測するための、PDCD1+CTLA4+CD8 T細胞のデコンボリューションを代用遺伝子発現シグネチャーと比較する Kaplan-Meier プロット。遺伝子発現および推定した画分の両方を、中央値分割によって分層し、有意性を、ログランク検定によって評価した。遺伝子発現シグネチャーを、PDCD1、CTLA4、およびCD8A (中央) またはPDCD1およびCTLA4 (右) の幾何平均として計算した。(21c) および (21d) における統計的有意性は、ログ順位検定を用いて決定した。HRは、ハザード比を示す。95%のHR信頼区間を括弧に示す。(21e) 免疫療法に対する応答とPDCD1+CTLA4+CD8 T細胞レベルまたは発現シグネチャーとの二元関連性 (パネル (21d) と同じ)。有意性は、ウィルコクソンの順位和検定で評価し、 $-\log_{10} p$ 値として表す。(21f) 抗CTLA4療法に対する応答の持続時間と、データセット内の細胞型存在量 (左) または $\log_2$ FPKM (100万個あたりのトランスクリプトームのキロベースあたりの断片) 空間 (右) における選択された遺伝子の平均発現の CIBERSORTx 推定値との間のスピアマン相関 (例えば、Nathanson et al. によって記載されるような)。

10

【図22】本開示の方法を実装するようにプログラムまたはそうでなければ構成されたコンピュータシステムを示す。

20

【図23】CIBERSORTxを用いたクロスプラットフォームデコンボリューションを示す。(23a) 所与の入力シグネチャーマトリックスに対して最も適切なCIBERSORTx一括修正戦略を選択するための決定木：(1) 単一細胞参照モード (「Sモード」、パネル (23g) ~ (23i)、または (2) バルク参照モード (「Bモード」、パネル (23d) ~ (23f)。(23b~c) Sモード (23b) および Bモード (23c) 一括修正の概略図。(23d) 図2bからのシグネチャーマトリックスおよび混合物データセットを用いたデコンボリューション性能の分析であるが、技術的変動を克服するためのBモードおよびSモード一括修正の相対有用性を比較する。(23e) 3' キットおよび5' キットを用いた10倍のゲノムによってプロファイリングされたPBMCの2つの一般に入手可能な単一細胞トランスクリプトームデータセットのt-SNE投影図。細胞標識を、Seurat (例えば、v1.4.0.16) および正規マーカー遺伝子評価を使用して手動で決定し、2個のシグネチャーマトリックス、3' PBMC、および5' PBMCを生成するために使用した。(23f) パネル (23e) からのデータセット間のプラットフォーム特異的一括効果を示すt-SNEプロット。(23g) ~ (23j) 単一細胞トランスクリプトームデータに由来する細胞型参照プロファイル进行调整するためのSモードの評価。(23g) パネル (23e) 内のデータセットを使用して、異なる10倍のクロムキット (パネル (23f) に由来する参照プロファイル間の技術的変動を最小限にするためのSモードの適用を例示する、実験的アプローチの概略図。ここで、3' キットからのシグネチャーマトリックスを、5' キットに由来する再構成された血液試料に適用した。Sモード一括修正後、調整された3' 細胞型参照プロファイルは、ヒートマップ (23h) およびピアソン相関 (23i) によって示されるように、5' キットからのそれらの対応する参照プロファイルとより密接に整列される。(23j) (23g) と同じであるが、5' PBMC単一細胞データセットをマイクロアレイによってプロファイルされた選別された白血球サブセットで置き換える。細胞サブセットは、パネル (23i) のように色付けされる。パネル (23g) ~ (23j) では、3' PBMCシグネチャーマトリックス内の遺伝子のみを、各細胞型の参照プロファイルとして考慮した。(k) 10倍のクロムによって生成されたPBMCサブセットの単一細胞参照プロファイルを用いた、健常なドナー由来の全血の白血球デコンボリューション (n=12個のRNA-Seqプロファイル) を示す、Sモードの評価。Sモード一括修正の前後のピアソン相関 (左) とRMSE

30

40

50

(右)。グラウンドトゥールス細胞比率は、フローサイトメトリーおよび全血球数によって決定された。グループの差は、ウィルコクソン符号順位検定によって評価した。(23 l) 5プラットフォーム、3つの組織型、および新鮮／凍結対固定組織保存状態にわたるLM22 (アフィメトリクスマイクロアレイ) による細胞サブセット列挙に対するBモード一括修正の影響：凍結保存されたPBMC対フローサイトメトリー (GSE65133) のIllumina Beadchipマイクロアレイ、新鮮な全血対フローサイトメトリー／CBCのバルクRNA-Seq (本明細書の他の箇所で開示されるように)、FFPE黒色腫 (Mel.) 腫瘍対代用細胞マーカーのバルクRNA-Seq、ならびにscRNA-Seqプロファイルからの再構築されたバルク黒色腫 (Mel) およびHNSCC腫瘍 (それぞれSmart-Seq2；黒色腫瘍およびHNSCC腫瘍からの再構築された単一細胞発現データにおけるTILサブセット頻度)。点は、各細胞型の予測されるレベルと推測されるレベルとの間のピアソン相関係数を表し、バープロットは、エラーバーとして表される95%信頼区間を有する中央値を示す。評価可能な細胞型およびグラウンドトゥールス推定の詳細は、本明細書の他の箇所で提供される。FFPE黒色腫のグラウンドトゥールス比率なしの性能を評価するために、選択された白血球サブ集団のデコンボリューション結果を、未調整RNA-Seqデータ (PTPRC対推測免疫含有量、CD19対B細胞、CD3D対T細胞、CD8A対CD8 T細胞、CD68対単球／マクロファージ) における正規マーカー遺伝子の発現と比較し、CIBERSORTx結果を、総免疫によってスケールリングして、各腫瘍試料について「正規化免疫指標」を計算した。パネル(23 d)、(23 i)、および(23 j)におけるデータは、平均＋／－標準偏差として表される。

10

20

【図24】Sモード一括修正の概略図を示す。

【図25】発現推定に使用されるLM22表現型グループ化を示す。

【発明を実施するための形態】

【0038】

本発明の様々な実施形態が本明細書に示され説明される一方で、そのような実施形態が例としてのみ提供されることは当業者には明らかである。本発明から逸脱することなく、多数の変形、変化、および置換は当業者にとって思い当たり得る。本明細書に記載された本発明の実施形態の様々な代替が使用され得ることを理解するべきである。

【0039】

30

定義

本明細書に別段の定義がない限り、本明細書で使用する技術用語および科学用語はすべて、本発明の属する技術分野の当業者によって通常理解されるのと同じ意味を有する。本明細書に説明されるものと類似のまたは等価の任意の方法および材料を本発明の実施または検査に使用することができ、好ましい方法および材料を説明する。

【0040】

本明細書で言及されるすべての特許および刊行物 (そのような特許および刊行物内に開示されているすべての配列を含む) は、参照により明示的に組み込まれる。

【0041】

本明細書で提供される見出しは、本発明の様々な態様または実施形態の限定ではない。したがって、直下に定義される用語は、全体として本明細書を参照することでより完全に定義される。

40

【0042】

別段の定義がない限り、本明細書で使用する技術用語および科学用語はすべて、本発明の属する技術分野の当業者によって通常理解されるのと同じ意味を有する。それでも、明確さと参照の容易さのために、特定の用語を以下に定義する。

【0043】

「複数」には、少なくとも2つのメンバーが含まれる。特定の場合では、複数は、少なくとも10、少なくとも100、少なくとも1000、少なくとも10,000、少なくとも100,000、少なくとも10⁶、少なくとも10⁷、少なくとも10⁸、または少

50

なくとも10⁹以上のメンバーを有し得る。

【0044】

本明細書で使用される「物理システム」という用語は、要素（分子、細胞、組織、生物、電気回路、デバイス、電化製品、コンピュータ、車両、建物など）の集合を指し得、要素が直接的または間接的に、相互に機能的に関連している（例えば、システムの1つの要素の存在、位置、または活動が、システム内の別の要素の存在、位置、または活動に直接的または間接的に影響する）。いくつかの場合では、物理的システムは物理的試料である。

【0045】

本明細書で使用される「物理的試料」という用語は、採取物が得られた環境からコヒーレント単位として物理的に単離され得る物質のあらゆる採取物（例えば、液体、気体形態の固体）を指し得る。本明細書で使用される「生体試料」という用語は、生物から得るか、生物を含むか、またはその一部を含むか（例えば、生物の組織試料、生検、細胞試料）のいずれかの物理的試料（例えば、固体または液状）を指す。生体試料は、細胞培養物または生物から得られ得る。生体試料は、生体試料の1つ以上の成分を除去することにより精製され得る。

10

【0046】

本明細書で使用される「異なる成分」という用語は、あらゆる形態の物質（例えば、分子、化合物、タンパク質、核酸、細胞など）、またはその採取物を指し、経験的に決定された1つ以上の性質および／または機能的関連性に基づいて群に分類される。各異なる成分は、経験的に決定された1つ以上の性質および／または機能的関連性を共有する1つ以上のメンバーを有し得る。

20

【0047】

本明細書で使用される「細胞サブセット」という用語は、その存在がRNAレベルでの遺伝子発現、タンパク質発現、ゲノム変異、バイオマーカーなど、1つ以上の特徴によって特徴付けられる生体試料中のあらゆる細胞群を指し得る。細胞サブセットは、例えば、細胞型または細胞サブタイプであり得る。

【0048】

本明細書で使用される「細胞サブセットの推定相対比率」または「細胞サブセットの相対比率のベクトル」という用語は、細胞サブセットのそれぞれの相対比率を指し得る（例えば、主題の方法の実施形態によって推定されるように）。このように、各細胞サブセットの相対比率は、ベクトルとして理解され得る（各細胞サブセットは該ベクトルの異なる次元である）。本明細書で使用される場合、「細胞サブセットの相対比率」とは、別の細胞サブセット、他の細胞サブセット、または相対比率が推定されているすべての細胞サブセットに対する細胞サブセットの比率（例えば、存在量）を指す。

30

【0049】

本明細書で使用される「特徴」という用語は、経験的に決定された、物理的試料または物理的システム（例えば、物理的試料、生体試料など）の特性を指し得る。いくつかの場合では、特徴の存在量（例えば、値）は、試料または物理的システム中の異なる成分の存在量を示し得る（例えば、生体試料中の1つ以上の細胞サブセットの存在量を示す）。いくつかの場合では、特徴は、mRNAでの遺伝子発現、タンパク質発現、特定の遺伝子型、バイオマーカー、またはそれらの組み合わせを含む。

40

【0050】

本明細書で使用される「特徴プロファイル」という用語は、物理的試料（例えば、生体試料）または物理的システム中の特徴の集合の測定値のセットを指し得る。物理的試料または物理システムが複数の異なる成分を含む場合、任意の所与の特徴の測定値は、存在する各異なる成分の量に応じて、特徴に対する各異なる成分の寄与の組み合わせ（例えば、線形和などの合計）であり得る。特徴プロファイルは、特徴のそれぞれの測定値に対応する要素をもつベクトルmで表し得る。

【0051】

本明細書で使用される生体試料の「特徴プロファイル」という用語は、生体試料の特徴

50

の集合の測定値を指し得る。試料の特徴プロファイルの例には、「遺伝子発現プロファイル」または「GEP」（例えば、マイクロアレイ分析によって得られる）、タンパク質発現プロファイル、遺伝子型プロファイル（例えば、不均一な腫瘍細胞を有する試料のもの）、バイオマーカープロファイル（例えば、試料中の遊離バイオマーカーまたは試料の細胞中／細胞上のバイオマーカー）、DNAメチル化プロファイル（例えば、メチル化のレベルまたは変化が異なるDNAを含む細胞を有する試料のもの）などが含まれる。特徴プロファイルは、1つ以上の細胞型、組織型、または細胞期の状態を表し得る。状態は、細胞型特異的、組織型特異的、または細胞期特異的であり得る。状態は、例えば、遺伝子発現プロファイル、タンパク質発現プロファイル、遺伝子型プロファイル、バイオマーカープロファイル、または特定の細胞型、組織型、または細胞ステージのDNAメチル化プロファイルに対応し得る。状態は、本明細書の他所に記載のとおり、群レベルの状態または高解像度の状態であり得る。例えば、DNAメチル化プロファイルは、細胞特異的、組織特異的、および／または細胞期特異的であり得、したがって、特徴プロファイルには、細胞型、組織型などにわたるDNAメチル化プロファイルが含まれ得る。

【0052】

本明細書で使用される「特徴シグネチャー」という用語は、実質的に純粋なまたは高度に濃縮された異なる成分のメンバーの集合の特性（または代表）である特徴プロファイルを指し得る。複数の異なる要素の特徴シグネチャーの群は、マトリックスBで表し得る。マトリックス内の1つの異なる要素は、特徴シグネチャーに、別の異なる要素と同じ特徴のセット有しても有さなくてもよい。

【0053】

本明細書で使用される「細胞サブセット参照プロファイル」または「細胞サブセット特徴プロファイル」という用語は、特定の細胞サブセットに関連する特徴プロファイル（例えば、特徴の値）を指し得る。参照プロファイルは、精製または濃縮された細胞サブセットの特徴を測定することで取得し得る。いくつかの場合では、本明細書で使用される「細胞サブセット特徴シグネチャーの参照マトリックス」という用語は、複数の細胞サブセットの予想される特徴値のマトリックスを指す。一部の参照プロファイルは、「多重共線性」すなわち異なる細胞サブセットの参照プロファイルが高度に相関し、信頼性の高いデコンボリューションを防ぐことができる現象を示す。

【0054】

本明細書で使用される「デコンボリューション」という用語は、細胞サブセットまたは細胞集団の混合物中の細胞サブセットまたは細胞集団の相対比率または存在量（例えば、絶対的または部分的存在量）を同定する（例えば、推定する）プロセスを指し得る。

【0055】

「部分的表現」、「相対比率」、および「寄与」という用語は、特徴シグネチャーの参照マトリックスに含まれる異なる成分のすべてに起因する特徴の合計値に対する、異なる成分に起因する特徴の測定値の部分を指すために交換可能に使用され得る。

【0056】

本明細書で使用される「サブセット」という用語は、初期マトリックスまたはベクトルの1つ以上の次元（例えば、特徴数）を減らすことにより得られるマトリックスまたはベクトル（例えば、特徴プロファイル）を指し得る。本明細書で使用される「スーパーセット」は、初期マトリックスまたはベクトルの1つ以上の次元を増加させることにより得られるマトリックスまたはベクトルを指し得る。「親マトリックス」または「親ベクトル」は、マトリックスまたはベクトルのスーパーセットを指し得る（例えば、「子」マトリックスまたはベクトル）。いくつかの場合では、親特徴プロファイルは、より多くの特徴を有することにより、親がスーパーセットである特徴プロファイルと異なる。

【0057】

本明細書で使用される「再構成された特徴プロファイル」または「デコンボリューション結果」という用語は、異なる成分（例えば、細胞サブセット）および既知の参照マトリックスの推定相対比率（または表示割合）に基づいて計算された特徴プロファイルを指し

10

20

30

40

50

得る。具体的には、再構成された特徴プロファイルは、異なる成分の推定相対比率、例えば、細胞サブセット（または「細胞サブセットの相対比率のベクトル」）と参照マトリックスの積から計算され得る。

【0058】

本明細書で使用する「サポートベクトル回帰」または「SVR」という用語は、サポートベクターマシン（SVM）のインスタンス、バイナリ分類問題の最適化方法のクラスを指し得、両方のクラスを最大限に分離する超平面が発見される。サポートベクトルは、超平面境界を決定する入力データのサブセットである。標準SVMとは異なり、SVRは入力データポイントに超平面を適合させて回帰を実行し、許容誤差 e および一意の線形誤差ペナルティ（例えば、 ϵ 不感度損失関数）の範囲内で回帰を実行して、異常値および過剰適合に比較的。SVRの2つの主要タイプは、「 nu サポートベクトル回帰」（または「 ν -SVR」）および「イプシロン-サポートベクトル回帰」（または「 ϵ -SVR」）である。 ν -SVRでは、 ν パラメータはトレーニング誤差 e の上限とサポートベクトルの希薄性の両方を便利に制御する。

【0059】

本明細書で使用する「モンテカルロサンプリング」という用語は、未知の確率論的エンティティにわたる分布を得るための反復ランダムサンプリングを指し得る。

【0060】

本明細書で使用する「有意値」という用語は、帰無仮説が真であると仮定して結果を得ることの確率を指し得る。特定の実施形態では、帰無仮説は、シグネチャーマトリックス内の細胞サブセットが生体試料の所与の特徴プロファイルに表されないことである。特定の態様では、有意値は「 p 値」であり得、それは本明細書で使用する場合、帰無仮説が真であると仮定して、実際に観察された結果に対して少なくとも極端なものとして、または近いものとして、試験統計結果を得ることの確率である。

【0061】

本明細書で使用する「差測定値」という用語は、2つの値またはベクトル間の関係（例えば、差、相関、偏差など）の測定値を指し得る。

【0062】

本明細書で使用する「誤差」という用語は、計算値または予測値からの値の偏差を指し得る。「二乗平均誤差」または「RMSE」という用語は、推定装置によって予測される値が推定される量と異なる量を指す。推定パラメータに関する推定装置のRMSEは、平均二乗誤差の平方根として定義される。

【0063】

本明細書で使用する「相関係数」という用語は、線形適合の尺度を指し得る。「ピアソンの積率相関係数」または「ピアソンの R 」は、2つの変数間の線形関係の強さと方向の尺度であり、変数の共分散を標準偏差の積で割ったものとして定義される。

【0064】

本明細書で使用する「RNAトランスクリプトーム」という用語は、生体試料中の細胞の凝集RNA発現レベルを指し得る。

【0065】

本明細書で使用する「白血球」または「白血球細胞」という用語は、単球、好中球、好酸球、好塩基球、およびリンパ球を含む任意の免疫細胞を指し得る。本明細書で 사용되는「リンパ球」という用語は、リンパに一般的に見られる細胞を指し、ナチュラルキラー細胞（NK細胞）、T細胞、およびB細胞を含む。上で列挙した免疫細胞型がさらなるサブセットに分割できることは、当業者には理解されよう。

【0066】

本明細書で使用する「腫瘍浸潤白血球」（TEL）という用語は、固形腫瘍に存在する白血球を指し得る。

【0067】

本明細書で使用する「対象」という用語は、ヒトまたは動物（例えば、非ヒト哺乳動

10

20

30

40

50

物、類人猿、サル、チンパンジー、爬虫類、両生類、または鳥)を指し得る。対象は、疾患または障害を有する人、または疾患または障害を有すると疑われる人、または疾患または障害を有していないかもしくは有すると疑われていない人であり得る。疾患または障害は、感染症、免疫障害または免疫疾患、癌、遺伝性疾患、変性疾患、生活習慣病、傷害、希少疾患、または加齢性疾患であり得る。感染症は、細菌、ウイルス、真菌、および／または寄生虫によって引き起こされ得る。

【0068】

本明細書で使用する「生体高分子」という用語は、一緒に共有結合したより小さな生体分子を含む大きな生体分子を指し得る。生体高分子は、核酸、タンパク質、代謝産物、炭水化物、糖、脂質、またはそれらの組み合わせを含み得る。核酸は、リボ核酸(RNA)分子またはデオキシリボ核酸(DNA)分子を含み得る。

【0069】

物理的試料および物理的システム

目的の物理的システムには、物理的システム内に複数の成分が存在し、物理的システムの特徴プロファイルに寄与する(例えば、線形モデルで近似できる方法で寄与する)任意の物理的システムが含まれる。物理的システムには、目的の成分が含まれる場合と含まれない場合がある。物理的システムには、任意の数の成分を含んでもよい。いくつかの場合では、物理的システムには、5以上(例えば、10以上、20以上)が含まれ、100以上(例えば、 10^3 以上、 10^4 以上(15, 000、20, 000、または25, 000以上など))が含まれ、例えば、経験的に決定された特性に関して相互に異なる、異なる(例えば、異なる)成分

【0070】

いくつかの場合では、物理的システムは、以下でさらに説明するように、複数の異なる成分を含む物理的試料である。いくつかの実施形態において、物理的システムは、生体内または生体外の細胞の集合(例えば、組織試料中の細胞の集合)、生物の組織の集合、生態系または社会内の生物の集合などである。いくつかの実施形態では、物理的システムは、デバイス内の電気回路の集合、部屋内のデバイスの集合、ネットワーク上のコンピュータの集合、建物内の電化製品の集合、都市またはその一部内の建物の集合、道路または高速道路システム上の車両の集合などである。

【0071】

物理的試料は、複数の異なる成分が物理的試料の特徴プロファイルに寄与する(例えば、線形モデルで近似できる方法で寄与する)異なる成分の混合を含む適切な試料であり得る。いくつかの実施形態では、以下でさらに説明するように、物理的試料は生体試料である。

【0072】

いくつかの場合では、物理的試料は、大気試料、水試料、または土壌試料などの環境試料である。環境試料は、限定されないが、河川、海、湖、雨、雪、貯水池、下水、下水処理流出、農業流出、工業流出、水道水、飲料水、排気システム(例えば、産業排気、車両排気など)、堆肥、埋め立て、都市開発サイト、または農地などの任意の適切な供給源から得られ得る。いくつかの場合では、物理的試料は、食品試料、例えば、動物、例えば、ヒト、犬、猫、鳥、魚などによる消費に適した、または消費のために準備されている材料である。いくつかの場合では、物理的試料は、合成化学混合、例えば、肥料、コーティング(例、塗料、ラッカーなど)、薬剤、洗剤などの市販の化学混合である。

【0073】

生体試料

本明細書で使用する「生体試料」(「試料」)という用語は、対象から得られた任意の試料、または対象から得られた試料に由来する任意の試料を指し得る。

【0074】

生体試料は、バルク組織またはアーカイブ組織(例えば、ホルマリン固定、パラフィン包埋(FFPE)組織、または凍結組織)を含み得る。生体試料は、血液試料を含むか、

10

20

30

40

50

固形組織試料（例えば、腫瘍試料）に由来し得る。

【0075】

本明細書で使用される「血液試料」という用語は、血漿、血液から単離された血液細胞などの血液から調製された任意の試料を指し得る。生体試料は、血液試料を含んでも、血液試料に由来してもよい。例えば、試料は、対象（例えば、癌患者）の末梢血から採取されてもよい。例えば、全血試料を対象（例えば、健常ドナー）から収集し、CLIA血液検査室でFACS（BD BTK6色IVD）およびコールターカウンターで白血球組成を数えるために処理し、その後のRNA抽出およびライブラリー調製のために、試料のアリコートにPAXgene血液RNAチューブ（Qiagen）に保存することができる。例えば、末梢血単核細胞（PBMC）は、Ficoll-Paque（GE Healthcare）密度勾配遠心分離を使用するなどして、対象の末梢血から採取することができる。生体試料は、解離により単細胞懸濁液に処理され得る。単細胞懸濁液は、例えば、FITC抗ヒトCD3クローンUCHT1（BD Biosciences）、パシフィックブルー抗ヒトCD4クローンRPA-T4（BD Biosciences）、APC抗ヒトCD14クローンM5E2（BD Biosciences）、AlexaFluor700抗ヒトCD8クローンRPA-T8（BD Biosciences）、APC-H7抗ヒトCD19クローンSJ25C1（BD Biosciences）、PE抗ヒトCD45ROクローンUCHL1（BD Biosciences）、PE-Cy7抗ヒトCD56クローンHCD56（BioLegend）、またはそれらの組み合わせで染色することによって処理され得る。単細胞懸濁液は、4%パラホルムアルデヒド（PFA）などで固定して処理できる。細胞試料は、FlowJo（FlowJo、LLC）を使用して行う分析によってFACS Aria IIuフローサイトメーター機器（BD Biosciences）などを用いた免疫表現型検査によって処理できる。免疫表現型検査を実行する場合、前方および側方散乱プロファイルに基づいてダブレットが除外され得る。

【0076】

本明細書で使用される「固形組織試料」または「バルク組織試料」という用語は、リンパ節、採取臓器、生検（例えば、腫瘍生検）などの固形組織から得られた試料を指し得る。試料自体は再構成され、懸濁され得る。生体試料はバルク組織を含み得る。バルク組織は、対象（例えば、癌患者）の新鮮に切除された外科用腫瘍試料から採取されてもよい。生体試料は、解離により単細胞懸濁液に処理され得る。単細胞懸濁液は、A700抗ヒトCD45クローンHI30（汎免疫細胞マーカー）、PE抗ヒトCD31クローンXWM59（内皮細胞マーカー）、APC抗ヒトEpCAMクローンX9C4（上皮細胞マーカー）、PE-Cy7抗ヒトCD10クローンXHI10a（線維芽細胞マーカー）、またはそれらの組み合わせなどの1つ以上の細胞マーカーでFACSソーティングすることによって処理され得る。

【0077】

本明細書で使用される「アーカイブ組織試料」という用語は、長期保存を受けた組織試料を指す。生体試料はアーカイブ組織を含み得る。アーカイブ組織は、（FFPE組織を得るための）ホルマリン固定およびパラフィン包埋、ならびに凍結組織を得るための凍結（例えば、凍結保存）などの1つ以上の保存方法を使用してバルク組織を処理することによって得ることができる。例えば、凍結保存腫瘍細胞懸濁液を解凍、処理、およびソートして、単細胞懸濁液（例えば、FL腫瘍細胞懸濁液からのCD5~CD19+B細胞）を単離してもよい。

【0078】

本明細書で使用される「精製試料」という用語は、1つ以上の細胞サブセットが濃縮される任意の試料を指し得る。試料は、サイズ、タンパク質発現等の特徴に基づいて細胞の除去または単離することによって精製され得る。

【0079】

本明細書で使用される「遺伝子発現プロファイリング」という用語は、一般に、遺伝子

10

20

30

40

50

発現を測定するための生体試料または単一細胞懸濁液上で測定方法を実行することを指す。遺伝子発現プロファイリングは、単一細胞RNA-Se q (s c RNA-Se q)、バルクRNA-Se q、マイクロアレイ、またはそれらの組み合わせを含んでよい。

【0080】

いくつかの実施形態では、生体試料は、細胞培養物または生物からインビトロで得られる。特定の態様では、生物は、霊長類（例えば、ヒト）、げっ歯類（例えば、マウス、ラット、ハムスター、モルモット）、ウサギ、または他の適切な動物などの動物であり得る。生物から採取された生体試料は、血液、脳、リンパ節、胸腺、骨髄、脾臓、骨格筋、心臓、結腸、胃、小腸、腎臓、肝臓、肺などの固形組織などの組織試料であり得る。組織試料は、臓器を収穫することによって、または当該技術分野で既知の生検を行うことによって得ることができる。特定の態様では、生体試料は、全血、血漿、または血液から得られた細胞などの血液試料である。

10

【0081】

特定の態様では、生体試料は腫瘍生検であり得る。生検とは、対象から得られた（例えば、切除、針吸引などによって）癌細胞を含有する任意の組織試料を指す。生検は、細胞懸濁液、薄い切片（例えば、スライド上に取り付けられた組織切片）の形態、または任意の他の適切な形態であり得る。

【0082】

特定の態様では、生体試料は、細胞分散溶液または懸濁液であり得る。この溶液は、5 ~ 25 mMなどの低濃度で許容される緩衝液と併せて、ウシ胎仔血清、ヒト血小板溶解物または他の因子を便利に補充した平衡塩溶液、例えば、生理食塩水、PBS、ハンクの平衡塩溶液等であり得る。便利な緩衝液としては、HEPES、リン酸緩衝液、乳酸緩衝液等が挙げられる。分離された細胞は、細胞の生存力を維持する任意の適切な培地に収集され得る。様々な培地は市販されており、多くの場合、ウシ胎仔血清またはヒト血小板溶解物で補充されているdMEM、HBSS、dPBS、RPMI、Isco ve培地等を含む、細胞の性質に従って使用されてもよい。他の態様では、生検は組織切片であり得る。例えば、生検は顕微鏡スライド上に取り付けられた薄い組織切片であり得る。上記実施形態のいずれかの生体試料は、当業者に既知のように固定および／または透過性であってもよい。

20

【0083】

試料は、例えば、粗形態の全試料であり得る。あるいは、試料は、分析の前に、例えば、密度勾配遠心分離、パニング、磁気ビーズソーティング、蛍光活性化セルソーティング (FACS) 等によって分画して、1つ以上の目的の細胞型について濃縮し得る。

30

【0084】

いくつかの場合では、生体試料は無細胞試料、例えば、細胞または組織ホモジネートである。いくつかの場合では、生体試料は単一の細胞を含む。

【0085】

特定の態様では、主題の方法は、生体試料中の細胞サブセットの相対比率を推定する前に、試料、例えば、生体試料（例えば、上述の）を得ることを含む。

【0086】

異なる細胞サブセット

40

本開示による生体試料の異なる細胞サブセット（例えば、細胞型）は、生体試料の特徴プロファイルに寄与する任意の異なる細胞型であり得る。

【0087】

いくつかの実施形態では、細胞の異なるサブセットは、異なる細胞周期段階の細胞のサブセットを含む。細胞のサブセットは、間期、有糸分裂期、またはサイトカイン形成を含むが、これらに限定されない、任意の適切な細胞周期段階にある細胞を含んでもよい。いくつかの実施形態では、細胞サブセット中の細胞は、分裂前期、中期、後期、または終期にある。いくつかの場合では、細胞のサブセットの細胞は、静止状態 (G₀期)、G₁チェックポイント (G₁期)、複製されたDNAであるが有糸分裂前 (G₂期)、またはD

50

NA複製（S期）中である。

【0088】

いくつかの実施形態では、異なる細胞サブセットは、1つ以上の細胞内の異なる機能的経路を含む。目的の機能的経路としては、細胞シグナル伝達経路、遺伝子調節経路、または代謝経路が挙げられるが、これらに限定されない。したがって、いくつかの実施形態では、本開示の方法は、シグナル伝達または代謝経路の複数の特徴を測定すること（例えば、シグナル伝達経路におけるタンパク質の活性化状態を測定すること、遺伝子調節ネットワークにおける遺伝子の発現レベルを測定すること、代謝経路における代謝産物のレベルを測定すること等）によって、細胞、細胞の集合、組織等の異なるシグナル伝達経路または代謝経路の相対活性を推定する方法であり得る。目的の細胞シグナル伝達経路としては、限定されないが、サイトカインシグナル伝達、死因子シグナル伝達、成長因子シグナル伝達、生存因子シグナル伝達、ホルモンシグナル伝達、Wntシグナル伝達、ヘッジホッグシグナル伝達、ノッチシグナル伝達、細胞外マトリックスシグナル伝達、インスリンシグナル伝達、カルシウムシグナル伝達、Gタンパク質結合受容体シグナル伝達、神経伝達物質シグナル伝達、およびこれらの組み合わせなどの任意の適切なシグナル伝達経路が挙げられる。代謝経路としては、限定されないが、糖分解、糖新生、クエン酸周期、発酵、尿素周期、脂肪酸代謝、ピリミジン生合成、グルタミン酸アミノ酸基合成、ポルフィリン代謝、アスパラギン酸アミノ酸基合成、芳香族アミノ酸合成、ヒスチジン代謝、分枝アミノ酸合成、リン酸ペントース経路、プリン生合成、グルコロン酸代謝、イノシトール代謝、セルロース代謝、スクロース代謝、デンプンおよび糖原代謝、ならびにこれらの組み合わせなどの任意の好適な代謝経路を挙げることができる。

10

20

【0089】

いくつかの実施形態では、細胞サブセットは、その存在が1つ以上の特徴（RNAレベルでの遺伝子発現、タンパク質発現、ゲノム突然変異、バイオマーカーなど）によって特徴付けられる生体試料中のあらゆる細胞群であり得る。細胞サブセットは、例えば、細胞型または細胞サブタイプであり得る。

【0090】

特定の態様では、1つ以上の細胞サブセットは白血球（例えば、白血球細胞またはWBC）であり得る。潜在的な白血球細胞サブセットとしては、単球、樹状細胞、好中球、好酸球、好塩基球、およびリンパ球が挙げられる。これらの白血球サブセットは、さらに細分することができ、例えば、リンパ球サブセットは、ナチュラルキラー細胞（NK細胞）、T細胞（例えば、CD8⁺ T細胞、CD4⁺ ナイーブT細胞、CD4⁺ メモリRO非活性化T細胞、CD4⁺ メモリRO活性化T細胞、濾胞ヘルパーT細胞、調節性T細胞など）、およびB細胞（ナイーブB細胞、メモリーB細胞、形質細胞）を含む。免疫細胞サブセットは、活性化（または刺激）状態に基づいてさらに分離され得る。

30

【0091】

特定の態様では、白血球は、血液癌、自己免疫疾患、骨髄異形成症候群などの白血球障害を有する個体からのものであり得る。血液疾患の例としては、急性リンパ球性白血病（ALL）、急性骨髄性白血病（AML）、慢性リンパ球性白血病（CLL）、慢性骨髄性白血病（CML）、急性単球性白血病（AMoL）、ホジキンリンパ腫、非ホジキンリンパ腫、および骨髄腫が挙げられる。自己免疫疾患の例としては、円形脱毛症、自己免疫溶血性貧血、自己免疫性肝炎、皮膚筋炎、糖尿病（1型）、糸状腎炎、グレイブス病、ギラン・バレー症候群、特発性血小板減少性紫斑、重症筋無力症、いくつかの形態の筋炎、多発性硬化症、ペンフィガス／ペンフィガイド、悪性貧血、結節性多発性動脈炎、多発性筋炎、原発性胆汁硬化症、乾癬、関節リウマチ、硬皮症／全身性硬化症、シェーグレン症候群、および全身性エリテマトーデスが挙げられる。

40

【0092】

特定の態様では、1つ以上の細胞サブセットには、腫瘍浸潤白血球（TIL）が含まれる。腫瘍浸潤性白血球は、生体試料中の癌細胞と混合し得るか、または上記もしくは当該技術分野で既知の任意の方法によって濃縮され得る。

50

【0093】

特定の態様では、1つ以上の細胞サブセットには、血液癌、乳癌、結腸癌、肺癌、前立腺癌、肝細胞癌、胃癌、膵臓癌、子宮頸癌、卵巣癌、肝臓癌、膀胱癌、尿路の癌、甲状腺癌、腎臓癌、癌腫、黒色腫、および脳腫瘍などの癌細胞が含まれ得る。

【0094】

目的の細胞サブセットとしては、神経細胞、星細胞、オリゴデンドロサイト、およびミクログリアを含む脳細胞、ならびにそれらの前駆細胞を挙げることができる。他の目的の細胞サブセットとしては、血液、脳、リンパ節、胸腺、骨髄、脾臓、骨格筋、心臓、結腸、胃、小腸、腎臓、肝臓、肺などの固形組織を含む、任意の生体組織の幹細胞、多能性幹細胞、および前駆細胞が挙げられる。

10

【0095】

特徴

目的の特徴としては、物理的試料（例えば、生体試料）、または1つ以上の異なる成分（例えば、細胞サブセット）の存在を示し得る物理的システムの任意の特徴が挙げられる。特定の態様では、特徴の存在量（例えば、値）は、1つ以上の異なる成分、例えば細胞サブセットの存在量を示し得る。特徴は、試料、例えば、mRNA、タンパク質、特定の遺伝子型、バイオマーカーなどの生体試料の特徴の集合であり得る。

【0096】

遺伝子発現および／または細胞遺伝子型などの特徴は、興味深い場合がある。例えば、細胞型および／または状態は、遺伝子発現によって分化され得る。別の例では、癌細胞は、突然変異に起因する遺伝的異質性に基づいて分化し得る。そのような特徴は、ポリメラーゼ連鎖反応（PCR）方法（例えば、RNAから合成された相補的DNA（cDNA）の定量的PCR）、RNA-seq、DNA-seq、DNAマイクロアレイ、タイリングアレイ、NanoString（登録商標）nCounter（登録商標）、ノーザンブロット、遺伝子発現の連続分析（SAGE）などを含む、当該技術分野で既知の任意の手段によって測定され得る。タンパク質発現などの特徴は、ウェスタンブロット、タンパク質マイクロアレイ、ELISA、他の免疫アッセイ、質量分析などの当該技術分野で既知の任意の手段によって測定され得る。

20

【0097】

いくつかの実施形態では、特徴プロファイルは、任意の適切な方法によって得られた、異なる化学化合物の適切な測定された特性を含む。場合によっては、特徴は、核磁気共鳴（NMR）（例えば、 ^1H 、 ^{13}C 、 ^2H 、 ^6Li 、 ^{10}B 、 ^{11}B 、 ^{14}N 、 ^{15}N 、 ^{17}O 、 ^{19}F 、 ^{23}Na 、 ^{29}Si 、 ^{31}P 、 ^{35}Cl 、 ^{113}Cd 、 ^{129}Xe 、または ^{195}Pt NMR）スペクトル、電磁放射線（例えば、紫外線、可視光線、赤外線）吸収および／または放出スペクトル、円形二色スペクトル、ラマンスペクトル、質量スペクトル、ならびにクロマトグラム（例えば、親和性クロマトグラフィー、液体クロマトグラフィー、サイズ排除クロマトグラフィー等から）を含む。

30

【0098】

特徴（無細胞バイオマーカーなど）は、ウェスタンブロット、酵素結合免疫吸着アッセイ（ELISA）、質量分析、クロマトグラフィー（例えば、薄層クロマトグラフィー、ガスクロマトグラフィー、液体クロマトグラフィーなど）を含む、当該技術分野で既知の任意の手段によって測定され得る。例えば、特徴は、クロマトグラフまたは質量スペクトル上で観察されるピークの強度であってもよい。

40

【0099】

特定の態様では、この方法は、物理的試料（例えば、生体試料）、または物理的システムから特徴測定値を取ることを含み得る。他の態様では、方法は、公的に入手可能なDNAマイクロアレイデータベース、RNA-seqデータベース、および／またはDNA-seqデータベースなどのデータベース、あるいは他の適切な特徴測定値のデータベースからの特徴測定値を得ることを含み得る。

【0100】

50

特徴シグネチャーの参照プロファイルは、精製または濃縮された異なる成分（例えば、異なる細胞サブセット）の特徴を測定することによって取得することができる。例えば、細胞サブセットは、上述のように、密度勾配遠心分離、パニング、磁気ビーズソーティング、蛍光活性化セルソーティング（FACS）等によって精製されてもよい。あるいは、細胞サブセットは、例えば、前駆体細胞の刺激および／または分化を介してインビトロで培養されてもよい。細胞サブセットの単離後、細胞サブセットの特徴を（例えば、上記のように、または当業者に既知のように）測定してもよい。特定の態様では、細胞サブセットの遺伝子発現を（例えば、DNAマイクロアレイ分析で）測定して、細胞サブセットの遺伝子発現プロファイル（GEP）を得ることができる。

【0101】

いくつかの場合では、細胞環境などで薬理的または誘導的にシグナル伝達経路を選択的に活性化することにより、シグナル伝達経路の成分を含む機能的な無細胞システムの特徴を測定することにより、シグナル伝達経路である異なる成分の特徴シグネチャーを得られる。いくつかの場合では、化学化合物である異なる成分の特徴シグネチャーは、化学化合物の実質的に純粋なまたは濃縮された試料の特徴を測定することによって得られる。

【0102】

特定の態様では、シグネチャーマトリックスは、上記の細胞サブセットのいずれかの特定のmRNA、タンパク質、遺伝子型、および／またはバイオマーカーのレベルを含む。シグネチャーマトリックスは、以前の研究において「塩基または塩基マトリックス」と呼ばれることが多く、例えば精製または濃縮された細胞集団の差次的発現分析によって得ることができる。遺伝子シグネチャーマトリックスは、変動またはノイズを入力する線形システムの安定性を測定する条件番号と呼ばれる固有のマトリックス特性を最小化することによって、より堅牢にすることができる。特定の態様では、シグネチャーマトリックスの安定性は、例えば、Rのカッパ関数で計算された2ノルム条件数を介して測定され得る。

【0103】

同じシグネチャーマトリックス内のいくつかの参照プロファイルは、複数の異なる成分、例えば細胞型の参照プロファイルが高度に相関する現象である「多重共線性」を呈し得る。多重共線性は、主題の方法における有意性値によって報告され得るように、物理的試料（例えば、生体試料）中の相対量の異なる成分（例えば、細胞サブセット）のデコンボリューションを防止し得るか、またはデコンボリューションの信用を低減し得る。主題の発明のシグネチャーマトリックス（VIF）によって測定されるような、本発明のシグネチャーマトリックスにおける2つの参照プロファイル間の多重共線性の厳格度は、1以上、例えば、2以上、5以上、10以上、15以上（20以上を含む）であってもよく、場合によっては、50以下、例えば、40以下、30以下、20以下、15以下（10以下を含む）であってもよい。

【0104】

場合によっては、参照マトリックスは、1以上、例えば、2以上、5以上、8以上、10以上、15以上、20以上、50以上、100以上、500以上（1,000以上を含む）の2ノルム条件数を有し、いくつかの実施形態では、104以下、例えば、103以下、500以下、250以下、200以下、150以下、100以下、50以下、30以下、20以下、15以下、8以下（5以下を含む）の2ノルム条件数を有する。いくつかの実施形態では、参照マトリックスは、1～5（例えば、5～8、8～10、8～15、10～15、15～20、20～30、20～50、50～100、100～150、100～200、100～250、100～500、500～1,000、1,000～10,000を含む）の範囲の2ノルム条件数を有する。

【0105】

参照マトリックスの条件数は、任意の適切な方法を使用して調整され得る。場合によっては、初期参照マトリックスの条件数は、マトリックスから1つ以上の特徴を追加または削除することによって低減され、それによって、より低い条件数を有する初期参照マトリックスのスーパーセットまたはサブセットを生成する。このプロセスは、最終参照マトリ

10

20

30

40

50

ックスの十分に低い条件数が得られるまで反復され得る。

【0106】

特定の態様では、参照マトリックスは、2以上、例えば5以上、10以上（15以上を含む）の異なる成分（例えば、細胞サブセット）に関連付けられる（によって発現される）少なくとも1つの特徴（例えば、遺伝子）、例えば少なくとも2、少なくとも3、少なくとも5、少なくとも10、少なくとも20、少なくとも50（少なくとも100を含む）の特徴を含み得る。場合によっては、参照マトリックスは、2～20、例えば2～15、2～12（5～10を含む）の異なる成分（例えば、細胞サブセット）に関連付けられる（によって発現される）少なくとも1つの特徴（例えば、遺伝子）、例えば少なくとも2、少なくとも3、少なくとも5、少なくとも10、少なくとも20、少なくとも50（少なくとも100を含む）の特徴を含み得る。場合によっては、参照マトリックスは、複数の異なる成分（例えば、細胞サブセット）に関連付けられる（例えば、によって発現される）10、000以下、例えば5、000以下、1、000以下、500以下、200以下（100以下を含む）の特徴（例えば、遺伝子）を含み得る。

10

【0107】

特定の態様では、シグネチャーマトリックスに含める候補特徴がフィルタリングされ得る。特定の実施形態では、低い値および／または分散を有する特徴は、シグネチャーマトリックスからフィルタリングされ得る。例えば、他の候補特徴と比較して90%、80%、75%、50%、または25%より低い値および／または分散を伴う特徴は、フィルタリングされ得る。別の例では、他の候補特徴と比較して90%、80%、75%、50%、または25%を超える値および／または分散を伴う特徴がシグネチャーマトリックスに含まれ得る。いくつかの実施形態では、シグネチャーマトリックスで表されない異なる成分が豊富な特徴は、シグネチャーマトリックスに含まれない。いくつかの実施形態では、シグネチャーマトリックスで表されない異なる成分の閾値より高値を有する特徴は、シグネチャーマトリックスに含まれない。

20

【0108】

目的の異なる成分、例えば細胞サブセットに対してより予測的である特徴は、シグネチャーマトリックスに含まれ得る。例えば、この方法は、所与の異なる成分（例えば、細胞サブセット、または物理的試料／物理的システム）における所与の特徴の濃縮スコア（ES）を、その特徴の他の異なる成分（例えば、細胞サブセット、または物理的試料／物理的システム）とのすべての対比較から得られる線形モデル係数の合計に基づいて計算することを含み得る。特定の態様では、特徴は、他の異なる成分、例えば細胞サブセットと比較した、異なる成分、例えば細胞サブセットの特徴の値の倍数変化に基づいてシグネチャーマトリックスに含めるために選択され得る。例えば、1つの異なる成分（例えば、細胞サブセット）において、任意の他の異なる成分（例えば、細胞サブセット）よりも2倍以上、5倍以上、10倍以上、または20倍以上の特徴が、シグネチャーマトリックスに含まれるように選択され得る。逆に、特徴は、多重共線性を低減するために、シグネチャーマトリックスから除外され得る。

30

【0109】

特定の態様では、主題の方法は、1つ以上の異なる成分（例えば、細胞サブセット）を単離すること、および1つ以上の異なる成分（例えば、細胞サブセット）の特徴を測定してシグネチャーマトリックスを得ることを含む。例えば、第1の生体試料中の細胞は、FACSによって別個の細胞サブセットに分離され得る。別個の細胞サブセットをDNAマイクロアレイによって分析して、別個の細胞サブセットの各々について遺伝子発現プロファイル（GEP）を得ることができる。次いで、各細胞サブセットのGEPをコンパイルして、各細胞サブセットのいくつかの遺伝子の発現の値を有するシグネチャーマトリックスを生成してもよい（例えば、図17に見られる）。

40

【0110】

試料の特徴プロファイルの例としては、「遺伝子発現プロファイル」または「GEP」（例えば、マイクロアレイ分析によって得られる）、タンパク質発現プロファイル、遺伝

50

子型プロファイル（例えば、不均一な腫瘍細胞を有する試料のもの）、バイオマーカープロファイル（例えば、試料中の遊離バイオマーカーもしくは試料の細胞中／細胞上のバイオマーカー）、DNAメチル化プロファイル（例えば、メチル化のレベルもしくは変化が異なるDNAを含む細胞を有する試料のもの）、またはそれらの組み合わせが挙げられる。例えば、DNAメチル化プロファイルは、細胞特異的、組織特異的、および／または細胞期特異的であり得、したがって、特徴プロファイルには、細胞型、組織型等にわたるDNAメチル化プロファイルが含まれ得る。DNAメチル化プロファイルは、例えば、試料の重亜硫酸塩配列決定によって測定され得る。

【0111】

試料の特徴プロファイルは、上述のように得ることができる（例えば、特徴は、生体試料から直接測定され得るか、または特徴プロファイルは、公的に入手可能なDNAマイクロアレイデータベースなどのデータベースから得ることができる）。上述のように、生体試料は、任意の細胞型を含んでもよい。特定の実施形態では、試料の特徴プロファイルは、ベンチマークデータセットであり得る。

10

【0112】

本方法は、物理的試料または物理的システムにおける異なる成分の画分表現を推定する高感度法を提供することができ、ここで異なる成分は、低画分で存在する。いくつかの実施形態では、物理的試料または物理的システムは、試料中に存在する第2の複数の異なる成分の総量の10%未満（例えば、8.0%未満、6.0%未満、4.0%未満、2.0%未満、1.0%未満を含む）の濃度で、場合によっては、0.01%以上（例えば、0.05%以上、0.1%以上、0.5%以上、1.0%以上を含む）の濃度で、特徴シグネチャーで表される少なくとも1つの異なる成分を含む。いくつかの実施形態では、物理的試料または物理的システムは、試料中に存在する第2の複数の異なる成分の総量の0.01%～10%（例えば、0.05%～8.0%、0.1%～6.0%、0.1%～4.0%、0.1%～2.0%を含む）の範囲の濃度で、特徴シグネチャーで表される少なくとも1つの異なる成分を含む。

20

【0113】

本方法は、シグネチャーマトリックスに表されていない異なる成分の存在下で、物理的試料または物理的システムにおける異なる成分の画分表現を堅牢に推定することを提供し得る。いくつかの実施形態では、特徴シグネチャーに表される異なる成分は、試料中の異なる成分の総量の50%以下、例えば45%以下、40%以下、35%以下、30%以下、25%以下、20%以下、10%以下、（5%以下を含む）、および場合によっては、1%以上、例えば5%以上、10%以上、20%以上、30%以上、35%以上（40%以上を含む）で試料中に存在する。いくつかの実施形態では、特徴シグネチャーに表される異なる成分は、1～50%、例えば5～50%、10～50%（20～45%を含む）の範囲で試料中に存在する。

30

【0114】

特定の態様では、生体試料は、シグネチャーマトリックスによって表されない細胞を含み得る。例えば、生体試料中の細胞の5%以上、10%以上、25%以上、50%以上、75%以上、5%～50%、5%以下、10%以下、25%以下、または50%以下は、シグネチャーマトリックス中の細胞サブセットによって表されない場合がある。

40

【0115】

あるいは、または加えて、生体試料は、10%以下、5%以下、2%以下、1%以下、0.5%以下、0.25%以下、0.1%以下、0.1%～10%、0.25%～2%などの低量で存在するシグネチャーマトリックスによって表される細胞サブセットを含み得る。

【0116】

序論

組織組成は、表現型変形の主要な決定因子であり、疾患転帰に影響を及ぼす重要な要因であり得る。scRNA-Seqは、細胞異質性を特徴付けるための強力な技術であり得

50

るが、それは、大きな試料コホートにとっては実用的ではなく、通常の臨床治療の一部として収集された固定標本に適用されない場合がある。これらの課題を克服するために、本開示は、バルク組織トランスクリプトームからの細胞型特異的遺伝子発現プロファイル（G E P）および細胞型存在量の同時推定を可能にするインシリコサイトメトリーのためのプラットフォームを提供する。インシリコ精製のために本明細書に開示される方法を使用して、バルク組織組成は、s c R N A - S e q由来の参照シグネチャーを使用して正確に推定され得る。開示された方法およびシステムは、無偏向細胞型発見を、大規模な組織解剖と関連付けることができる。C I B E R S O R T xと呼ばれるこの統合されたフレームワークの有用性は、黒色腫を含む複数の腫瘍型で実証することができ、単一細胞参照プロファイルを使用して、新鮮な、凍結された、および固定された臨床標本を解剖し、腫瘍浸潤白血球（T I L）と免疫療法応答との間の関連性を明らかにする。デジタルサイトメトリーは、単一細胞プロファイリングの努力を増強し、抗体、分解、または生存細胞なしで費用対効果の高い高スループット組織特徴付けを可能にすることができる。

【0117】

フローサイトメトリーおよび免疫組織化学（I H C）などの免疫表現型決定アプローチは、事前選択されたマーカー遺伝子の小さな組み合わせに依存し得、同時に照合され得る細胞型の数を制限し得る。対照的に、単一細胞m R N A配列決定（s c R N A - S e q）を、単一細胞懸濁液（s c R N A - S e q）からの数百～数千個の個々の細胞の無偏向転写プロファイリングに使用することができる。この技術の力にもかかわらず、大型試料コホートの分析は実用的でなくてもよく、多くの固定臨床標本（例えば、ホルマリン固定、パラフィン包埋（F F P E）試料）は、単一細胞懸濁液に解離されてもよい。さらに、細胞型表現に対する組織分解の影響はよく理解されていない場合がある。

【0118】

混合物試料のゲノムプロファイルから細胞含有量を直接解剖するための計算技術は、F A C S精製またはインビトロ分化／刺激された細胞サブセットに由来する細胞型特異的「バーコード」遺伝子（例えば、「シグネチャーマトリックス」）の専門的なナレッジベースに依存し得る。目的の細胞型が十分に定義されているときに有用であるが、そのような遺伝子シグネチャーは、新規の細胞型および細胞型G E Pの発見、ならびに複雑な組織における主要な細胞表現型の完全なスペクトルを捕捉するために最適ではない場合がある。

【0119】

本開示は、無傷組織のR N Aプロファイルから細胞型存在量および細胞型（例えば、細胞型特異的）遺伝子発現を正確に推定するための計算フレームワークを提供する（図1）。単一細胞実験または分類された細胞サブセットからの細胞型発現シグネチャーを活用することにより、本開示のシステムおよび方法は、物理的解離、抗体、または生体物質なしで組織組成の包括的なポートレートを提供できる。そのようなアプローチには、例えば、クロスプラットフォームデータ正規化およびインシリコ細胞精製の技術を用いて、組織G E Pから細胞組成を列挙するための方法が含まれ得る。後者により、個々の目的の細胞型のトランスクリプトームを、物理的に分離せずにバルクR N A混合物からデジタルに「精製」し得る。その結果、細胞分離や予備知識なしで、細胞型特異的遺伝子発現の変化を推測することができる。C I B E R S O R T xと呼ばれるこのフレームワークの技術的性能および臨床的有用性を説明するために、本開示のシステムおよび方法を使用して、（1）単一細胞発現データでバルク組織組成をプロファイリングし、（2）解離せずに、悪性および腫瘍関連の細胞型のG E Pを精製し、（3）抗癌療法に関連する細胞型特異的シグネチャーを探索することができる。これらの結果は、C I B E R S O R T xが、研究および臨床環境での高解像度の細胞表現型検査の意味合いをもつ、複雑な組織を解読するための有用なツールであることを示す。

【0120】

C I B E R S O R T xプラットフォームは、インシリコ組織解剖に使用することができる。プラットフォームの機能は、細胞型の列挙および遺伝子発現精製のための統合フレームワーク、クロスプラットフォームの変動を抑制するための専用の正規化スキーム、およ

びRNA混合物を細胞型特異的発現プロファイルに分離するためのアプローチの改善を含むことができる。CIBERSORTxは、末梢血、脾臓標本、および悪性腫瘍（例えば、そのうち350はプロファイリングされる）の分析に使用され得る。これらの結果は、CIBERSORTxが、異なる供給源に由来する発現プロファイルを使用して、組織の不均一性に関する正確なポートレートを提供できることを示す。

【0121】

本明細書に記載されるシステムおよび方法を使用して、包括的なヒト細胞アトラスを定義することができる。scRNA-Seqデータセットを組み合わせるための新興の技術と併せたデータ生成の急速なペースを考慮すると、単一細胞参照マップを広範に適用する方法は、特に組織が制限され、固定され、または分解することが困難な設定において重要であり得る。腫瘍組織および健康な組織の研究では、単一細胞参照プロファイルを使用して、組織組成の詳細な照合を行うことができ、対象間の不均一性は、デコンボリューションの結果に影響を及ぼす主要な要因ではない。CIBERSORTxは、プラットフォームまたは保存状態とは関係なく、複雑な組織からの細胞サブセットGEPの堅牢な分子プロファイリングを可能にするための現在のアプローチを上回ることができる。発現分析のための複数の技術を柔軟に組み込むことによって、CIBERSORTxは、表現型カテゴリが既知のときの細胞型GEPの迅速な評価（図1、ステップ3、図3a~g）、および追加の詳細が所望されるときに発現変動の高解像度プロファイリングを容易にすることができる（図1、ステップ4、図4a~e）。これらの特徴を有すると、CIBERSORTx分析は、特定の細胞型を標的化することに依存する診断アプローチおよび治療アプローチを通知するための意味をもつ、腫瘍微小環境内の異種型相互作用の改善された理解を提供し得る。

【0122】

加えて、CIBERSORTxを使用して、複雑な組織における細胞不均一性を復号化することができる。この戦略を使用して、単一細胞トランスクリプトームからの目的の細胞サブセットを「デジタルゲート」し、バルク組織GEP（例えば、臨床試験からの固定標本）のコホートにおけるこれらの細胞の同一性および発現パターンをプロファイルし、ゲノム特徴および臨床転帰を含む様々なメタデータとのそれらの関連性を系統的に決定することができる。したがって、CIBERSORTxは、生物学的発見の統計的なパワーを増加させ、それによって再現可能な研究を促進するための有用性を有することができる。CIBERSORTxアプローチの一般性および他の技術とシームレスに統合する能力を考慮して、インシリコサイトメトリーでは、ヒト、マウス、および他のメタゾンにおける多細胞系の分析を強化することができる。

【0123】

本明細書で使用される「scRNA-Seq」という用語は、一般に、個々の細胞の発現プロファイルを得るための単一細胞RNA配列決定方法を指す。例えば、単一細胞ライブラリーは、Chromiumをv2化学（10x Genomics）と共に使用して、PBMC単一細胞懸濁液（例えば、癌患者から）から調製することができる。そのような単一細胞ライブラリー（例えば、NextSeq 500（Illumina））を配列決定することができる。配列決定読み取り値は、例えば、Cell Ranger 1.2（10x Genomics）を使用して、アラインメント、フィルタリング、重複解除、および／またはデジタルカウントマトリックスへの変換によって処理され得る。外れ値細胞は、（1）異常に高い／低いミトコンドリア遺伝子発現（例えば、10%超もしくは1%未満のミトコンドリア含有量を有する細胞が除去され得る）および／または（2）細胞あたりによって検出される発現遺伝子の数と、細胞あたり検出される固有の分子識別子（UMI）の数とを比較することによって識別されるように、潜在的な二重項／多重項に基づいて（例えば、3,500超および500未満の発現遺伝子を有する細胞が除去され得る）、識別およびフィルタリングされ得る。クラスターは、（1）UMIの数およびミトコンドリア含有量の割合に対する遺伝子発現の依存性を回帰させること、ならびに（2）データの主成分の適切な数（例えば、1、2、3、4、5、6、7、8、9、10、

または10を超える)上で「FindClusters」を実行することによって(例えば、Seurat v. 1. 4. 0. 16を使用して)特定されてもよい。細胞標識は、正規白血球マーカー遺伝子(例えば、MS4A1高=B細胞;CD8A高およびGNLY低=CD8 T細胞;CD3E高、CD8A低、およびGNLY低=CD4 T細胞;GNLY高およびCD3E低=NK細胞;GNLY高およびCD3E高=NKT細胞;CD14高=単球)の発現に従って割り当てられてもよい。Chromium v2(5'および3'キット)によってプロファイルされた健康なドナーからの公的に入手可能なPBMCデータセットは、以下の軽微な修正を伴ってダウンロードされ(表1)、上記のように前処理されてもよい。品質管理中、5'PBMCについて5000超の発現遺伝子、3'PBMCについて4000超の発現遺伝子、および200未満の発現遺伝子を有する細胞が除外され得る。Seurat「FindClusters」は、解像度パラメータが0. 6に設定された最初の20の主コンポーネントに適用され得る。細胞標識は、上述のように割り当てられ得る。加えて、骨髄細胞は、高CD68発現によって定義され得、巨核球は、高PPBP発現によって定義され得、樹状細胞は、高FCER1A発現によって定義され得る。

【0124】

本明細書で使用される「バルクRNA-Seq」という用語は、一般に、バルク細胞集団または組織の発現プロファイルを得るためのバルクRNA配列決定方法を指す。例えば、全RNAは、製造業者の推奨に従って、例えばPAXgene Blood RNAキット(Qiagen)を使用して、例えばPAXgeneチューブに保存された血液試料から単離されてもよい。RNAは、例えば、2100 Bioanalyzer(Agilent)を使用して定量化し、品質を評価してもよい。ライブラリー調製は、例えば、製造業者の推奨に従ってRNAエキソームキット(Illumina)を使用して行われ得る。RNA-Seqライブラリーは、一緒に多重化され、例えば、2x150bp読み取り値を使用する単一のHiSeq 4000レーン(Illumina)を使用して配列決定されてもよい。例えば、全RNAは、製造業者の推奨に従ってTRIzol(Invitrogen)を使用して、PBMC試料から(例えば、NSCLC癌患者から)単離されてもよい。例えば、RNA 6000 Picoチップ(Agilent)を備えた2100 Bioanalyzer(Agilent)を使用して、RNA分子を定量化し、品質を評価することができる。RNA分子のライブラリー調製は、例えば、メーカーの推奨に従ってSMARTer Stranded Total RNA-Seq-Picoキット(Takara Biosciences)を使用して実施することができる。ライブラリーは、例えば、Qubit 3. 0蛍光計(Thermo Fisher Scientific)を使用するdsDNA HSアッセイキット(Thermo Fisher Scientific)を用いて定量化され得る。ライブラリーの品質は、例えば、D1000スクリーンテープを備えた4200 TapeStation Instrument(Agilent)を使用して評価されてもよい。RNA-Seqライブラリーは、好適な配列決定機器(例えば、2x150塩基対(bp)読み取り値を使用するNextSeq 500(Illumina))上で配列決定され得る。別の例として、全RNAは、例えば、AllPrep DNA/RNAマイクロキット(Qiagen)を使用して、バルク腫瘍(例えば、NSCLC)およびソートされた細胞集団(例えば、約100、約200、約300、約400、約500、約1, 000、約5, 000、約10, 000、約15, 000、約20, 000、約25, 000、または25, 000を超える細胞の範囲)から抽出されてもよい。全RNAの量(例えば、約10ナノグラム(ng)、約20ng、約30ng、約40ng、約50ng、または50ng超)は、例えば、Ovation RNA-SeqシステムV2(NuGEN)を使用して増幅されてもよい。得られた相補的DNA(cDNA)を(例えば、超音波処理(Covaris S2システムによって)150~200bpの平均サイズに剪断してもよく、(例えば、NEBNext DNAライブラリー調製マスターミックス(New England Biolabs)を使用して)DNAライブラリーを構築するために使用してもよい。ライブ

10

20

30

40

50

ラリーを、好適な配列決定機器（例えば、HiSeq 2000（Illumina）上で配列決定して、試料あたり平均1億（M）個の読み取り値を有する100bp対の末端読み取り値を生成することができる。

【0125】

デコンボリューション分析の文脈における線形性を最大化するために、生のFASTQ読み取り値は、GENCODE v23参照転写物、`-biasCorrect`フラグ、および他のデフォルトパラメータを使用して（例えば、`Salmon v0.8.265`で）処理され得る。RNA-Seq定量化結果は、Rパッケージ`tximport66`を使用して単一の遺伝子レベルTPMマトリックスにマージされ得る。

【0126】

マイクロアレイを使用し、マイクロアレイを使用してグラウンドトゥールズ参照プロファイルを生成し得る。全RNAは、バルクFL標本およびソートされたB細胞から抽出され得、収率および品質について評価され得る。相補的RNA（cRNA）は、線形増幅後の100ngの全RNA（3'IVT発現、`Affymetrix`）から調製し、次いで製造業者のプロトコルに従ってHGU133 Plus 2.0マイクロアレイ（`Affymetrix`）にハイブリダイズしてもよい。得られたCELデータファイルは、FLリンパ節からFACSソートされたCD4およびCD8腫瘍浸潤リンパ球（TIL）を含む公的に入手可能な`Affymetrix`データセットプールすることができる（`GSE2792840`）。結果として得られたデータセットを、`Bioconductor`の「`affy`」パッケージを使用してRMA正規化し、カスタムチップ定義ファイル（例えば、`Brainarrayバージョン21.0`、`http://brainarray.mbni.med.umich.edu/Brainarray/`）を使用してNCBI Entrez遺伝子識別子にマッピングし、HUGO遺伝子シンボルに変換してもよい。ソートされた細胞サブセットの複製を組み合わせ、発現値の幾何平均を使用して、グラウンドトゥールズ参照プロファイル（例えば、図3bに示される）を作成した。

【0127】

外部データセットは、正規化設定を使用してダウンロードおよび分析される次世代配列決定（NGS）データセットを含んでもよい。そのような外部データセットは、100万あたりの転写物（TPM）、100万あたりの転写物1キロベースあたりの読み取り値（RPKM）、または100万あたりの転写物1キロベースあたりの断片（FPKM）空間のうちの1つ以上を含んでもよい。`log2`空間における分析の場合、`log2`調整の前に、1の値を発現値に追加してもよい。`Affymetrix`マイクロアレイデータセットは、マイクロアレイで説明されるように、バルク組織およびグラウンドトゥールズ細胞サブセットが同じ`Affymetrix`プラットフォーム上でプロファイルされた場合にRMAを使用し、そうでなければMAS5正規化を使用して要約および正規化され得る。`NanoString nCounter`データは、任意の追加の前処理なしに、非ログ線形空間でバッチ補正を用いてダウンロードおよび分析され得る。

【0128】

本明細書で使用される「単一細胞シグネチャーマトリックス」という用語は、一般に、生体試料中に表され得る細胞型の表現型を含むマトリックスを指す。入力データセットからの発現データは、上述のように要約され得る。細胞型表現の変動性およびscRNA-Seqデータにおける固有のノイズを考慮して、単一細胞差次的発現解析における偶発的ドロップアウトに対処し、細胞レベルスケリング因子を推定し、技術的および生物学的ノイズ成分をモデル化するための技術を開発してもよい。単一細胞表現型を定義するのに有用であるが、細胞標識が割り当てられた後にそのような技術を適用するとき、デコンボリューション性能の有意な増加は観察されない場合がある。scRNA-Seqデータにおける細胞下位集団の発見は焦点ではない場合があるため、以下のアプローチに記載されるように、前処理ステップが限定され得る。

【0129】

単一細胞発現値は、まず100万あたりの転写物（TPM）に正規化され、1細胞あた

10

20

30

40

50

りの転写物の数をよりよく近似させるために10で除してもよい。各細胞表現型について、 $\log_2$ 空間における低平均発現を有する遺伝子は、品質管理フィルタとして0に設定され得る。より低密度の遺伝子カバレッジのため、フィルタは10x Chromiumによって生成されたデータに適用されない場合がある。少なくとも3個の単一細胞によって表される各細胞型について、すべての利用可能な単一細胞GEPの50%は、置換なしにランダムサンプリングを使用して選択されてもよい（3個しか利用可能でなかった場合、2個の細胞がサンプリングされたように、画分試料サイズを切り上げてよい）。プロファイルは、非ログ線形空間における合計によって集計され得、各集団レベルのGEPは、TPMに正規化され得る。このプロセスは、細胞型ごとに集約されたトランスクリプトーム複製物（例えば、2、3、4、5、または5を超える）を生成するために繰り返されてもよい。例えば、scRNA-SeqおよびバルクRNA-Seqシグネチャマトリックスは、以下の典型的なパラメータ：細胞型あたりの遺伝子の最小数=300、細胞型あたりの遺伝子の最大数=500、q値0.01、および分位正規化なしで、前述のように生成されてもよい。

【0130】

例えば、HNSCC scRNA-Seqデータに由来するシグネチャマトリックスは、腫瘍ID MEEI18（原発腫瘍）およびMEEI25（原発腫瘍および対リンパ節転移）に関してトレーニングされてもよい（図2c）。別の例として、scRNA-Seqデータに由来するシグネチャマトリックスは、腫瘍ID80および88（図6a、8b~e、9、15、および19）上でトレーニングされてもよい。以前に注釈された細胞表現型標識は、CD8 T細胞を除いて両方の場合に用いてもよく、CD8 T細胞は、CD8AまたはCD8Bを発現するTリンパ球（非対数TPM>0）として定義されてもよい。膵島シグネチャマトリックス（図2fおよび8f~i）は、以前に割り当てられた細胞表現型を使用して、試料ID HP1508501T2DおよびHP1504901に関してトレーニングされてもよい。肥満細胞および主要組織適合性複合体（MHC）II細胞は、トレーニングコホート内の単一細胞（例えば、3細胞未満）が不十分である場合、島シグネチャマトリックスから省略され得る。 α 細胞における高発現マーカー遺伝子であるグルカゴンは、膵島トレーニングデータにおけるデコンボリューション性能に悪影響を及ぼし得るため、省略され得る。

【0131】

本明細書で使用される「デジタルゲーティング」という用語は、一般に、細胞集団を（例えば、マーカー遺伝子の高/低発現または高/中/低発現に基づいて）2つ以上の別個のセットに（例えば選択、包含、または除外によって）ソーティングすることを指す。図6dおよび9に示されるように、単一細胞は、複数（例えば、7）の以前に注釈された表現型に従って事前に標識されてもよい。B細胞、T細胞、NK細胞、マクロファージ、内皮細胞、癌関連線維芽細胞、および悪性黒色細胞。Tリンパ球は、CD8AまたはCD8B（非対数TPM>0）の発現によって、CD4およびCD8 T細胞にさらに分割されてもよい。例えば、CD8 T細胞は、以下のように別個の下位集団にさらに細分されてもよい。PD1および/またはCTLA4発現に陽性の候補CD8 T細胞を単離するために、CD8 T細胞は、PD1（CD8 T細胞間のPD1発現値の非対数TPM>5パーセントイル）および/またはCTLA4（CD8 T細胞間のCTLA4発現値の非対数TPM>5パーセントイル）の発現によって分割され得る。候補組織レジデントメモリ（TRM）CD8 T細胞を単離するために、CCR7（非対数TPM=0）ではなく、ITGAE（非対数TPM>0）およびCD69（非対数TPM>0）を発現するCD8 T細胞を選択してもよい。

【0132】

t分布型確率的近傍埋め込み法（t-SNE）によるscRNA-Seqデータの視覚化は、デフォルトパラメータを有するRにおけるRtSneパッケージを使用して行われ得る。図2fおよび8fに示されるように、Combatは、t-SNE予測の前にドナー固有の変動を排除することができる。

10

20

30

40

50

【0133】

CIBERSORTx分析フレームワークの概要

様々な計算方法を開発して、細胞型存在量、細胞型特異的GEP、または両方をバルク組織発現プロファイルから推測してもよい。これらの方法は、生体混合物試料を線形方程式のシステムとしてモデル化することができ、ここで、 n 遺伝子を有する単一の混合物トランスクリプトーム m が、 H 、 k 個の異なる細胞型にわたる同じ n 個の遺伝子の発現プロファイルからなる $n \times k$ 細胞型発現マトリックス、および細胞型混合比率からなるサイズ k のベクター f の産物として表される。

【0134】

CIBERSORTx内でこのモデルを使用して細胞型存在量を推測する方法の一例として、 M を n 個の遺伝子および c 個の混合物GEPを有する $n \times c$ マトリックスとし、マトリックス B を k 個の細胞サブセット（例えば、シグネチャーまたはベースマトリックス）のそれぞれについて差別的マーカー遺伝子を含有する H のサブセットとする。 M および B を与えられた場合、次いで、以下の方程式を使用して、カラム $[f_1, f_2, \dots, f_c]$ を有する $k \times c$ 画分存在量マトリックスである F を推定することができる。

$$B \times F \cdot, j = M \cdot, j, 1 \leq j \leq c \quad (1)$$

式中、すべての i, j について $F_{ij} \geq 0$ である場合、システムは過剰に決定され（例えば、 $n > k$ ）、 m および B の発現データは、非対数線形空間で表される。 $(M_{ij}, \cdot)$ および $M \cdot j$ は、マトリックス M の行 i および列 j をそれぞれ示すことができることに留意する）いくつかのアプローチは、各混合物試料について、推測される混合係数が合計して1になるように、 F を正規化するか、または F に追加の制約を課すことができ、 F を（ B における細胞サブセットに関して）細胞型比率として直接解釈できるようになる。例えば、CIBERSORT法を使用して、 v -サポートベクトル回帰の実装を使用して F を推定してもよく、これは（例えば、参照によりその全体が本明細書に組み込まれる米国特許公開第US2016/0217253号に記載されるように）細胞型参照プロファイル間のノイズ、不明な混合物含有量、および共線性に堅牢であり得る機械学習技術である。CIBERSORT法を使用して、例えば、クロスプラットフォーム分析のための適切なバッチ補正技術で F を推定することができる。

【0135】

群レベルの細胞型特異的遺伝子発現精製

F が M について決定されると、発現デコンボリューションまたは以前の知識（例えば、コールターカウンター、フローサイトメトリー）のいずれかを通じて、 F における各細胞型に対する代表的な推定GEPは、以下の一次方程式のシステムを解くことによって推定することができる。

$$H_{ij} \cdot \times F = M_{ij} \cdot, 1 \leq i \leq n \quad (2)$$

式中、 H は、 n 個の遺伝子および k 個の細胞型の $n \times k$ 発現マトリックスであり、すべての i, j について $H_{ij} \geq 0$ であり、 F は、上記のように、各混合物試料について相対細胞画分が合計して1になるという制約で定義される。上記の方程式1と同様に、システムは過剰に決定されてもよく（ $c > k$ ）、 c と k との間のより大きなデルタは、一般にGEP推定の改善をもたらす（図3eおよび11）。遺伝子発現の生物学的に現実的な推定を確実にするために、非負最小二乗回帰（NNLS）は、最適化フレームワークを用いて、非負制約で最小二乗問題を解いてもよい。NNLSは、単純な混合物およびトイの例では堅牢であり得るが、実際の組織試料内に固有のより複雑な混合物に対するその性能は、線形システムにおけるノイズ、不精度、および欠損したデータによって影響され得る。様々なデータ正規化およびフィルタリング技術が、そのような問題を軽減するのに役立つように適用され得る（図3b～d）。

【0136】

高解像度発現精製

バルク組織（例えば、図3a～gにおけるインシリコ精製における）からの細胞型特異的遺伝子発現を推定するための方程式2の有用性にもかかわらず、各細胞型に対する単一

10

20

30

40

50

の代表的な G E P の推定に限定され得る。したがって、複数の目的の状態間の細胞型特異的差次的発現遺伝子 (D E G) を探索するために、細胞型トランスクリプトームは、各条件 (例えば、所与の療法に対する応答者対非応答者、早期対進行期疾患等) に対して再生成する必要がある場合がある。この問題により広範に対処するために、方程式 2 におけるモデルは、「高解像度」インシリコ細胞精製のための方法 (例えば、C I B E R S O R T x) 内で拡張され得る。

【0137】

そのようなアプローチは、バルク組織 G E P (例えば、M) のマトリックスを、等しいサイズの k 遺伝子発現マトリックス (F における各細胞サブセットに対して 1 つ) に分解することを含んでよい。この方法は、表現型クラス構造に対して非依存であり得、部分的な観察を伴う非負行列分解 (N M F) 問題として式化することができる。例えば、M および F を上記のように定義し (それぞれ $n \times c$ および $c \times k$ マトリックス)、後者が C I B E R S O R T によって推定されると仮定する。次いで、各遺伝子 i について、k 細胞型による c 混合物試料の $c \times k$ 発現マトリックスとして本明細書で定義される発現マトリックス G_i を決定することができる。M および F を固定して G について求解すると、以下の制約されたマトリックス分解問題が生じる場合がある。

$$F \times G_i, \cdot, \cdot = P_i, \cdot, \cdot, 1 \leq i \leq n \quad (3)$$

式中、

$$d i a g (P_i, \cdot, \cdot) = M_i, \cdot, \cdot, 1 \leq i \leq n \quad (4)$$

およびすべての i、j、k については、 $G_i, j, k \geq 0$ である。方程式 1 および方程式 2 の線形システムとは異なり、3 次元 (3 D) $n \times c \times k$ マトリックスであってもよい G に対する閉形式解は存在し得ず、特に追加の制約 (例えば、規則化) なしで、非負行列因子分解 (N M F) のための既存の技術は、生物学的に合理的な推定を生じる可能性が低い場合がある。したがって、図 1 4 に概略的に示されるヒューリスティックアルゴリズムを使用して G を推定することができる。

【0138】

G を推測するためのアプローチは、生物学的に合理的な解を生成しながら、問題の取り扱い性を向上させ得る 2 つの異なる仮定を行ってもよい。第 1 に、各遺伝子を独立して分析できると仮定することができる。遺伝子間共分散関係を無視することは、いくつかの遺伝子の解像力に影響を及ぼし得るが、この仮定は、実際に有効であると見出され得る (図 4 a ~ e、5、6 a ~ b、および 1 3)。第 2 に、所与の遺伝子に関して、細胞型特異的差次的発現の少なくともいくつかの証拠が、統計的に有意でない場合であっても、バルク組織試料において検出可能であると仮定することができる。

【0139】

この仮説は、以前に公開された細胞型特異的 D E G のデータセットを使用して評価され得る。具体的には、D E G は、膵臓内分泌島細胞型 (例えば、5 細胞型) 内の発現を考慮し (各細胞型は、全内分泌膵島内の約 5 % ~ 約 4 4 % の中央画分存在量を含む)、非糖尿病正常対象 (例えば、6 対象) と 2 型糖尿病 (T 2 D) 患者 (例えば、4 患者) との間の差次的発現レベルを比較するときに、1. 2 倍を超える変化を呈する遺伝子として以前に定義されてもよい。重要なことに、これらの D E G は、s c R N A - s e q プロファイリングによって同定され、インシリコで再構築されたバルク島における差次的発現の証拠の試験を可能にした。既知の表現型クラスによってグループ化されると、T 2 D に関連する D E G の 9 8 % は、再構築された島における正しい配向の少なくともいくつかの倍率差を示す (図 1 3 a)。この結果は、細胞型特異的 D E G が、表現型クラス標識の予備知識なしにバルク組織で認識され得ることを示唆し得る。実際に、各遺伝子が発現レベルによって再構築された島において独立して順序付けされ、各ベクターが高発現群および低発現群に均等に分割される (例えば、5 0 パーセントイル分割) とき、糖尿病患者からのデータは、以前に定義された細胞型特異的 D E G の 8 0 % について低発現または高発現に歪み得る。これらの遺伝子のうち、濃縮方向は、9 9 % の症例において既知の倍率変化の配向に一致した (図 1 3 b)。したがって、バルク遺伝子発現データの変動を活用して、基礎と

なる細胞下位集団における潜在的表現型クラス構造を推測することができる。

【0140】

これらの基本的な仮定を混合マトリックスMおよび細胞型画分存在量マトリックスFと共に考慮すると、高解像度精製のためのヒューリスティックアルゴリズムを開発することができる。このヒューリスティックの概要を以下に概説し、対応するグラフィック要約を図12cに示す。

【0141】

各遺伝子*i*について、その対応する混合ベクトル M_i 、 $\cdot$ は、監督されていないアプローチを使用してバルク発現の違いに基づいて、2つのクラスに分層されてもよい（図12c、ステップ*i*、*i i*）。細胞型特異的遺伝子発現係数 g_1 および g_2 は、各クラスについて推定されてもよく（図12c、ステップ*i i*）、データに最適な適合性を有する g_1 、 g_2 対が見出されてもよい（図12c、ステップ*i i i*）。

10

【0142】

次に、統計試験を行って、所与の遺伝子が各細胞型によって発現されるかどうかを判定し、そうである場合、 g_1 と g_2 との間で有意に異なるかどうかを評価してもよい（図12c、ステップ*i v*）。

【0143】

次に、前述のステップを満たす遺伝子／細胞型の組み合わせについて、連続発現値は、回帰分割統治アルゴリズム戦略によって推定され得る。遺伝子が有意に発現されるが、混合物試料にわたって有意に変動しない場合、定常発現値が推定されてもよく、そうでなければ、それは発現しないものに設定されてもよい。マトリックスGには、細胞型特異的発現データが取り込まれ得る（図12c、ステップ*v-i*）。

20

【0144】

次に、 g_1 および g_2 係数は、有意性試験の結果に従って改変され得る（例えば、所与の細胞型について、発現が g_1 と g_2 との間で有意に異なる場合、 g_1 および g_2 における係数は、同じ値に設定され得る。図12c、ステップ*v-i i*）。

【0145】

次に、Gにおける連続発現ベクターを調整して、以前に決定した調整された g_1 および g_2 係数とのそれらの合致性を最大化してもよい（図12c、ステップ*v-i i i*）。

【0146】

最後に、最終推定細胞型特異的発現マトリックス（例えば、G）は、データベースに保存されてもよい（例えば、ディスクに書き込まれてもよい）。

30

【0147】

このフレームワーク内で、細胞型特異的遺伝子発現ベクターは、NNLSを使用して方程式2（上記）で推定され得る。 M_i 、 $\cdot$ 全体にわたる発現の変動を捕捉するために、方程式2は、同様の発現値によってグループ化された混合試料のサブセット上で反復求解され得る。各サブセットのサイズは、長さwの摺動ウィンドウによって制御され得る。NNLS制約を満たすために、かつ表現型クラスの重複を避けるために、wは、区間 $k < w \leq (c/2)$ によって境界されてよく、式中、kは細胞型の数を示し、cは試料の数を示す。このアプローチの良好な性能は、幅広い範囲のw値にわたって観察され得る。しかしながら、群レベルの発現精製の飽和度分析（図3e、11）内で、より大きなw値で観察される限界増加を考慮して、wはkよりも4～5倍大きく設定されてもよく、実際の考慮事項と性能を均衡させてもよい。線形システムにおける潜在的な不安定性に対処し、堅牢な標準誤差および信頼区間を有する発現係数を推定するために、NNLSは、ブートストラップを使用して実行されてもよい。

40

【0148】

CIBERSORTx一括修正

一括修正技術は、発現プロファイリングにおける技術的変動を最小限に抑えるために開発されてもよく、遺伝子発現デコンボリューションに適用されてもよい。以前の方法を使用して、そのような一括修正は、特定の状況では不可能ではないが、非常に困難であり得

50

る。例えば、TCGA腫瘍試料からの対マイクロアレイとRNA-Seqプロファイルとの間の遺伝子レベルの技術的変動をモデル化するために、4自由度を有する立方体スプラインを使用する以前のアプローチは、対試料が利用可能である場合に潜在的に有用であるが、新しい発現データセットおよびアーカイブ発現データセットがそのような対を欠いている場合には、非常に困難であり得る。別の例として、細胞標準プロファイルとバルク組織GEPとの間の技術的変動を排除するために適用され得る経験的ベイズ法であるComBatは、バッチが同じ組織型からのGEPを表すときに潜在的に有用であるが、バッチが異なる組織型からのGEPを表すとき、細胞存在量推定の著しい歪みを導入するために、不可能ではないが、非常に困難であり得る。

【0149】

発現プロファイリングにおける技術的変動の問題を克服するために、本開示は、（例えば、シグネチャマトリックスと混合物試料の入力セットとの間の）技術的差を最小限に抑えるための一括修正手順を提供する。例えば、そのようなアプローチは、Mと、細胞参照プロファイルから推定される混合物試料M*の一致するセットとの間で一括修正を行ってもよい。クロスプラットフォームの技術的変動またはクロスサンプルの技術的変動などの技術的変動は、低減、最小化、または除去され得る。一括修正を行うことは、試料型がバッチ間で比較可能であることを確実にするのに役立ち、したがって、潜在的な推定アーチファクトと両方のプラットフォーム上でプロファイルされた対試料の必要性を回避することができる。

【0150】

デコンボリューション方法（例えば、異なる細胞型の混合物から細胞型状態を識別または定量化するための）は、（例えば、シグネチャマトリックスとバルク混合物プロファイルとの間の）技術的変動を低減するために一括修正手順を実行することを含み得る。例えば、バルク参照モード（例えば、Bモード）一括修正は、以下のように行われてもよい。一般に、単一細胞トランスクリプトームプロファイリングに由来する参照表現型を含む、デコンボリューション方法（例えば、CIBERSORT）がRNA-Seqに適用され得るが、そのような方法は、シグネチャマトリックスとバルク混合物プロファイルとの間の技術的変動を明示的に処理しない場合がある。技術的変動には、クロスプラットフォーム技術的変動またはクロスサンプル技術的変動が含まれ得る。例えば、技術的変動は、異なるプラットフォーム（例えば、RNA-Seq、scRNA-Seq、マイクロアレイ、10x Chromium、SMART-Seq2、液滴ベースの技術、UMIベースの技術、非UMIベースの技術、3' / 5' 偏向技術）および／または異なる試料型（例えば、新鮮な／凍結された試料、FFPE試料、単一細胞試料、バルクソートされた細胞集団もしくは細胞型、および細胞集団もしくは細胞型の混合物を含む試料）にわたるシグネチャマトリックスの特徴プロファイルおよびバルク混合物の特徴プロファイルの取得から生じる場合がある。例えば、同じタイプの発現データ（例えば、GEP）を有する特徴プロファイルが異なるプラットフォームを使用して取得される場合、クロスプラットフォーム技術的変動が生じる場合がある。技術的変動は、デコンボリューション結果を可変的に混乱させる可能性があるため、少なくとも2つの異なる戦略を含み得る正規化ワークフロー（図7a～h）を適用して、プラットフォーム（例えば、RNA-Seq、マイクロアレイ）および組織貯蔵タイプ（例えば、新鮮な／凍結された対FFPE）にわたって遺伝子発現デコンボリューションを確実に適用することができる。例えば、最も適切な戦略を選択する際にユーザを導く決定木を図7aに提供する。決定木は、実行されるバルクモード一括修正（例えば、Bモード）手順および／または単一細胞一括修正（例えば、Sモード）手順の選択を支援するために使用され得る。

【0151】

バルク参照モード（例えば、Bモード）一括修正は、バルクソートされた参照プロファイル（例えば、バルクRNA-Seqまたはマイクロアレイ）に由来するシグネチャマトリックスと混合物試料の入力セットとの間の技術的差異を除去することができる（図7c）。例えば、一括修正は、異なる細胞型の混合物に対する1つの細胞型の特徴プロファ

10

20

30

40

50

イル間の技術的差異を除去し得る。この技術はまた、転写物がバルク混合物 G E P（例えば、S M A R T - S e q 2 によってプロファイルされた U M I なしの完全長転写物）と類似して測定されることを条件として、s c R N A - S e q プラットフォームに由来するシグネチャーマトリックスに適用され得る。混合物試料 M のセットを与えられた場合、バルク参照モード一括修正アプローチは、一連の推定混合物試料 M * を作成することを含んでよく、ここで後者は、（非対数線形空間において）M における推定細胞型比率と対応するシグネチャーマトリックスプロファイルとの線形組み合わせを含む。戦略は一般的であってもよく、異なるデコンボリューションおよび一括修正方法（図 7 f ~ h）に柔軟に対応することができるが、C o m B a t（経験的ベイズ法）を使用して、l o g 2 調整後の M と M * との間のバッチ効果を排除してもよい。バッチ効果が除去されると、細胞比率は、非対数線形空間における調整混合物試料を使用して再推定され得る（図 7 f ~ g）。細胞型列挙に加えて、このアプローチはまた、細胞型特異的遺伝子発現精製に適用され得、それによって、発現レベルが低いか、またはシグネチャーマトリックスにおける細胞型の集合から欠如している遺伝子のダウンウェイトを可能にする（図 7 h）。この技術は、一括修正手順を実行するために、少なくとも約 3 ~ 5 個の混合物試料を必要とし得る。

【0152】

上記の B モード一括修正技術にかかわらず、過剰なクロスプラットフォーム技術的変動が存在する場合（例えば、存在することが予想される細胞型がバルク G E P 試料中で脱落することが観察される場合）、デコンボリューションが不良に行われ得るか、または失敗し得る。例えば、この現象は、10x C h r o m i u m 由来のシグネチャーマトリックスを使用して観察され得る（図 2 b および 7 e）。U M I ベースの方法および 10x などの 3' / 5' 偏向方法と、S M A R T - S e q 2 などの U M I なしの完全な転写物を捕捉する方法との間のトランスクリプトーム表現の大きな差異を考慮すると、そのような不一致は驚くほどではない場合がある。B モード一括修正が細胞ドロップアウトに適切に対処できず、したがって、そのような変動を克服するのに不十分であり得るため、単一細胞参照モード一括修正が、10x C h r o m i u m に由来するシグネチャーマトリックスに関与するデコンボリューション方法および他の液滴ベースの技術または U M I ベースの技術に適用され得る（図 7 b）。単一細胞参照モード（例えば、S モード）一括修正は、以下のように（図 2 4 の概略図によって例示されるように）行われてもよい。

【0153】

B モード一括修正と同様に、S モードの目的は、技術的変動を最小限に抑えながら、c 試料を含有する混合物 G E P M のセットから細胞頻度 F を得ることであってもよい。高い全体性能の場合、このアプローチは、少なくとも 2 つの手順、すなわち初期細胞頻度推定相（相 1）、および混合マトリックスではなくシグネチャーマトリックスが調整される改良相（相 2）の連続的な適用を含んでもよい。相 1 において、目標は、（1）シグネチャーマトリックス B および（2）B が導出された単一細胞参照プロファイルのセット R を考慮して、混合物試料 M のセット内の細胞頻度 F * の初期推定値を得ることであってもよい（図 7 b）。これを達成するために、まず、R 内の単一細胞 G E P から再構成された c 人工混合物プロファイル M * のシリーズを作成してもよい。混合係数は、正規分布 N (μ , σ) に従って決定してもよく、ここで μ は、R { μ_1 , μ_2 , ..., μ_n } 中の B からの各細胞型の画分存在量に設定され、 σ は、{ $2\mu_1$, $2\mu_2$, ..., $2\mu_n$ } に設定される。次いで、各細胞サブセットからの単一細胞トランスクリプトームを、この分布に従ってランダムにサンプリングし、T P M 空間における c バルク発現プロファイルに凝集してもよい。C o m B a t を、その後 l o g 2 空間で M および M * に適用して、それぞれ調整された混合物 M ^{a d j} および M ^{a d j} * を得ても良く、推定された細胞頻度 (F *) は、元のシグネチャーマトリックス (B) を使用して M ^{a d j} (l o g 空間内) から推定されてもよい。

【0154】

第 2 の相では、細胞比率推定値は、（1）シグネチャーマトリックス B 上で一括修正を実行すること、および（2）調整されたシグネチャーマトリックスを元の混合物セット M

に適用することによってさらに精製されてもよい。相 1 と同様に、プロセスは、単一細胞をサンプリングして R から c 人工混合物を作成することによって開始されてもよい。しかしながら、この場合、細胞型は、相 1 から得られたそれらの推定頻度 F^* に従ってサンプリングされてもよく、推定混合物 M^{**} の新しいマトリックスをもたらす。次いで、ComBat を、 $\log_2$ 空間で M および M^{**} に適用して、それぞれ調整混合物 M^{adj} および M^{adj**} を得ることができる。 M^{adj**} を非対数線形空間に変換した後、B における各遺伝子の細胞型発現係数を再構築するために、 M^{adj**} および F^* の両方を NNLS への入力として使用して（方程式 2 を参照）、調整されたシグネチャーマトリックス B^{adj} を得ることができる。特に、グループモード発現精製（図 1）とは異なり、S モードにおける NNLS のこの使用は、適応ノイズフィルタリングを必要としない場合がある。これは、（1） M^{adj**} における各人工混合物の細胞混合比率が F^* から導出され、したがって 100% 確実に知られてよく、（2）ComBat は、各遺伝子の $\log_2$ 発現プロファイルに線形形質転換のみを適用してよく、それによってバルク発現値の相対順序を保持するからである。したがって、 $M^{adj**} \sim B$ における調整を伝播するために NNLS を適用するための条件は理想的であり得る。 B^{adj} が得られると、元の混合物 M からセル周波数を推定することができる。B モード一括修正と同様に、S モード内で ComBat を適用するには、少なくとも約 3 ～ 5 個の混合物試料が必要であり得る。NNLS を実行するためには、細胞型よりも多くの混合物試料が必要とされ得る（式 2 を参照）が、この要件を満たさないデータセットについては、 F^* からランダムに描かれた周波数ベクトルを使用して R から追加の人工混合物を作成し得る。各追加の偽混合物は、バッチ修正の前に、その対応する混合物試料と M で対合され得る。

【0155】

ソフトウェアの実装

CIBERSORTx は、R および PHP に基づいてそのバックエンドを有する Web フレームワーク内で開発され得る（例えば、<http://cibersortx.stanford.edu> によって記載されるように）。このウェブフレームワークは、特定のハードウェア、ソフトウェアパッケージおよびライブラリー、およびファイルシステム属性への固有の依存を最小限に抑え得る。ユーザは、図 1 の各ステップを含む、本明細書に記載される結果の再現を可能にし得るいくつかの段階的チュートリアルを用いた詳細なガイドを提示され得る。このインターフェースを通じて、CIBERSORTx は、ユーザが、（1）対象の各細胞型の発現プロファイルを定義する遺伝子を列挙する署名遺伝子ファイルと共に、異なる細胞型のバルク混合物を表す遺伝子発現データを処理することを可能にし得る。後者の場合、ユーザは参照細胞型に既存／キュレートされたシグネチャーマトリックスを使用し得るか、純粋な細胞集団の参照遺伝子発現プロファイルを提供することでカスタムシグネチャー遺伝子ファイルを作成し得る。具体的には、カスタムシグネチャー遺伝子マトリックスを作成するために、ユーザは、目的の単一細胞型または細胞集団の表現型同一性と共に、バルク選別試料からの単一細胞 RNA 配列決定データまたはデータを提供することができる。

【0156】

これらの入力ファイルを考慮すると、CIBERSORTx は、（2）その前駆体と同様に、混合物中に存在する各細胞型の画分表現の埋め込みを可能にし得る。CIBERSORTx はまた、上記の重要なバッチ修正方法を実装する際に、バルク RNA-Seq データからの復元をサポートする。CIBERSORTx はまた、上述の 2 つの異なるモード（例えば、グループレベルおよび高分解）でシリコ精製細胞型における個体の GEP の埋め込みを可能にする。次いで、得られた埋め込まれた画分および埋め込まれた GEP は、視覚化およびダウンロードのためのヒートマップ、テーブル、および積み重ねられたバースプロットとしてレンダリングされてもよい。

【0157】

インタラクティブな CIBERSORTx ユーザインターフェースは、jQuery JavaScript（登録商標）ライブラリーおよびオープンソースライブラリー（p

10

20

30

40

50

hpMailer、idiorm、blueimp jQuery-File-Upload、DataTables、phpExcelおよびmPDFを含む)を使用して実装され得、Twitter Bootstrap 2.3.2を使用して実装されるウェブサイトのグラフィカルユーザインターフェース(GUI)を有する。サイトは、仮想マシン上のApacheサーバ上で実行され得、ユーザおよびジョブデータをMySQLデータベースに記憶し得る。しかしながら、ユーザはデータを完全に制御し、必要に応じて削除することができる。各ユーザの環境は、ベンチマーキングに使用される例示的なデータセット、CIBERSORTxの使用および入力データの準備のためのチュートリアル、および他の例示的なファイルを含む。

【0158】

NSCLC細胞サブセットのグループレベル精製

図3gで分析したNSCLC試料上で酵素的リボソームRNA(rRNA)枯渴を使用したにもかかわらず、TPMスケールリードの大部分(30~90%)は、ミトコンドリアrRNA、特にMT-RNR1(ミトコンドリアコード12S RNA)およびMT-RNR2(ミトコンドリアコード16S RNA)に由来し得る。この混乱変数を除去するために、これらの転写物を除去してよく、データマトリックスをTPM空間に再規格化してよい。発現精製のためのNSCLCシグネチャーマトリックス(図3g)を生成するために、上皮(EpCAM⁺)、造血(CD45⁺)、内皮(CD31⁺)、および間質(CD10⁺)細胞のGEPを分析し、これらのそれぞれを、ドナーから得られた一次NSCLC腫瘍生検からFACSによって選別した。シグネチャーマトリックスパラメータは、本明細書の他の箇所でシグネチャーマトリックス構築のために記載されるものと同一であってもよいが、scRNA-Seqに用いられる前処理ステップなしであってもよい。何百もの遺伝子(一例では、933遺伝子)からなる得られたシグネチャーマトリックスを、別のドナーセット(一例では、n=26)から得られたバルクNSCLC腫瘍試料の全転写プロファイルに適用してもよく、結果を分析して、グループレベルの細胞型特異的GEPを求めることができる。グラウンド真実細胞型GEPは、第2のドナーセットから分類されたEpCAM⁺、CD45⁺、CD31⁺、およびCD10⁺細胞FACSの複製RNA-Seqプロファイルの幾何平均を取ることによって誘導され得る。

【0159】

合成GEP混合物の生成および分析

図4aにおける模擬GEP混合物は、LM22に基づくソースデータから得られたk=4細胞型(ナイーブB細胞、CD8⁺T細胞、NK細胞、および単球)の参照プロファイルを使用して作製してもよい。コンセプトの最初の証拠として、分析は、転写体からランダムに選択された1,000個の遺伝子、ならびにLM22シグネチャーマトリックスからのすべての547個の遺伝子に限定され得る。次に、4つの細胞型発現ベクターは、log₂調整され得、それらのそれぞれはm=400回コピーされ得る。可変性は、4つの細胞型にわたる各遺伝子jの標準偏差s_{dj}を計算することによって、および各遺伝子の発現値vを、分布N(v, s_{dj}/5)からランダムな値を描画することによって置き換えることによって導入され得る。生物学的分類をシミュレートするために、第1の200および第2の200試料が2つのクラスに分層されるように、DEGを各細胞型に添加してもよい。重複DEGパターンは、図4a(左パネル)にそれぞれオレンジ細胞、緑細胞、青色細胞、および紫色細胞によって描写されるように、単球、B細胞、CD8⁺T細胞、およびNK細胞GEPに挿入されてもよい。DEGは、5倍高い発現レベルを反映するように作製され得る。次に、m×k行列は、N(1/k, 1/(2k))を有する正規分布からランダムに描かれた混合物係数から生成されてよい。各混合物の係数は、合計1に正規化され得る。混合する前に、合成細胞型GEPは、非対数線形空間に変換され得る。次に、それらをそれらの対応する混合係数に従って組み合わせてもよく、400個の模擬組織GEPを得てもよい。高分解能CIBERSORTxは、例えば、単球、B細胞、CD8⁺T細胞、およびNK細胞のバーコード遺伝子を含むLM22シグネチャーマトリックスのサブセットである20(=5×k)のウィンドウサイズおよびLM4を使用し

10

20

30

40

50

て、得られた混合物データセット上で行われ得る。

【0160】

図4dにおける合成データセットについて、上記のアプローチは、 $m=11,845$ (LM22が由来するソースデータセット内の遺伝子の完全なリスト) および以下の変化で繰り返され得る。まず、上方制御されたDEGは、単球GEPに「ブルセイ」標的(3つの同心円)の形状で埋め込まれ得る。バルク混合物中のDEG画像をマスクするために、DEGは、22個の遺伝子ごとに単球GEPに添加されてもよく、逆(等しい倍率変化の下方制御DEG)は、交互順序(例えば、66個の遺伝子ごと)で残りの細胞型GEP($n=3$)に添加されてもよい。次に、埋め込まれたブルズアイを回収するために、上記のように高解像度のCIBERSORTx方法を適用し得る。

10

【0161】

図14における合成データセットについて、分析がLM22におけるすべての遺伝子と共に3,000個の遺伝子のランダム選択に限定されることを除き、図4dに使用されるアプローチを繰り返してもよい。次に、2つの上方制御されたDEGパターンを埋め込んでもよい: 標的を単球GEPに埋め込んでもよい(図4dに示されるが、他の遺伝子ごとに挿入されている)、および10個の対角線で整列された正方形のセットをCD8⁺T細胞GEPに埋め込んでもよい(すべての遺伝子に添加されている)。これらのDEGを混合物プロファイルに混乱させるために、下方制御されたDEGは、それぞれのパターンを補完する相互方法でb細胞およびNK細胞に挿入されてもよい。上記のように高分解能CIBERSORTxを適用することができる。

20

【0162】

高解像度CIBERSORTxによる細胞型特異的DEG検出の分析

5つの細胞型を有する合成腫瘍は、以下のように分析され得る。模擬混合物は、本明細書の他の箇所に記載されるように作製されてもよいが、この場合、 k は50個の混合物に設定されてもよく、 m は3,000個の遺伝子(LM22を含有する)に設定されてもよく、DEGは、CD8⁺T細胞プロファイルに排他的に挿入されてもよい(試料26~50、すべてのDEGは、ベースラインレベルと比較して過剰発現された)。現実的な設定の範囲にわたるパフォーマンスを評価するために、(1)DEG倍率変化(1.5~5倍)および(2)CD8⁺T細胞の平均分数存在量(1%~25%)の異なる組み合わせについて、別個の50個の混合物のセットを評価してよい(図4e)。高解像度CIBERSORTx法は、例えば、長さ16(=4 k)のウィンドウを使用して、LM4シグネチャマトリックス(上述)を使用して、すべての混合物試料上で行われ得る。不明な混合物含有量(例えば、腫瘍中の悪性細胞)の存在を評価するために、分析は、各混合物データセットに平均50%のHCT116(結腸癌細胞株)を添加することによって、およびバッチ修正を伴う高分解能CIBERSORTxを適用することによって繰り返され得る。CD8⁺T細胞およびHCT116の平均比率がそれらの標的存在量に設定され得ることを除き、本明細書の他の箇所に記載されるように、混合係数を模擬するために同じ戦略が適用され得る。前述と同じHCT116 GEPデータセットおよび発現正規化ステップを使用し得る。曲線下面積は、既知のDEGクラス(それぞれ25個の試料)間のlog₂空間における遺伝子レベル倍率変化を計算することによって、および既知のDE遺伝子(陽性)と残りの遺伝子(陰性)との間の結果として生じる量を比較することによって評価され得る。特異性を評価するために、このプロセスは、各埋め込まれた細胞型Gepについて繰り返され得る。

30

40

【0163】

8細胞型を有する模擬黒色腫瘍は、以下のように分析され得る。黒色腫患者由来の単細胞RNA-Seqプロファイルは、本明細書の他の箇所に記載されるように得られ得、以前に注釈された7つの細胞表現型に従って標識され得る。B細胞、T細胞、NK細胞、マクロファージ、内皮細胞、癌関連線維芽細胞、および悪性細胞。Tリンパ球は、合計 $k=8$ の評価可能な細胞型について、CD8AまたはCD8B(非対数TPM>0)の発現によってCD8およびCD4⁺T細胞に分割されてもよい。次に、黒色腫瘍をシミュレ

50

ートするために、解離された腫瘍にわたる細胞型頻度の分布を反映するアプローチを使用してもよい。具体的には、各合成腫瘍を構築するために、各細胞型は、1回に1回繰り返され得、その細胞型に対応するすべての単一細胞は、元のコホート内のランダムに選択された腫瘍から選択され得る。

【0164】

このプロセスは、複数の腫瘍（例えば、50、100、または200）のコホートを作製するために反復され得る。各合成腫瘍について、非対数線形空間における単細胞Gepは、合計によって凝集され、TPMによって正規化され得る。各コホートについて、以下が変化し得る：（1）人工DEGを有する細胞型の総数（ $d=1, 2, 3, \dots, 8$ ）、および（2）DEGの $\log_2$ 倍変化（ $fc=1, 2$ 、および3）（図15a）。 d および fc の各値について、例えば、0.5の平均 fc および sd で、正規分布から描かれた発現値を、ランダムに選択された複数の（例えば、約2,000）遺伝子のセットに加えることによって、人工DEGを作製してもよい。差動発現は、 d 、DEGを有する細胞型の数に応じて異なるシミュレーションを行うことができる。 $d=1$ の場合、細胞型GEPの半分は、人工DEGを有するクラスに割り当てられ得、半分は、それらのベースライン値を保持し得る。 $d>1$ の場合、目的は、塊状腫瘍転写産物におけるDEGの回復を妨げる方法でDEGをシミュレートすることであってもよい。これを行うために、クラスサイズは、 d で割った腫瘍の数よりも25%大きいように設定されてよく、得られるDEGクラスは、細胞型にわたって均等にタイルをかけてよい。細胞比率を列挙するために、同じ8種類の細胞を捕捉する2つの腫瘍に由来する黒色腫シグネチャーマトリックスを使用してもよい（図8b）。すべての合成腫瘍は、残りの17個の保留腫瘍に由来し得る。長さ32（=4k）のウィンドウは、高解像度転写トーム精製のために使用され得る。5種の細胞型を有する合成腫瘍について、上述のように精度を評価してもよい。DEG回復を最大化する $\log_2$ 倍変化を得ることができる（例えば、95%信頼区間+/-0.009を有する0.11の平均）。

【0165】

高分解能CIBERSORTxを有する原発性腫瘍検体の分析

リンパ腫図5a~cおよび16について、長さ40（=4k）のウィンドウを使用し、LM22を、細胞型の列挙および前の転写体精製後に $k=10$ 個の主な白血球サブセットに折り畳んだ。

【0166】

NSCLC。図5d~fおよび18c~eについて、長さ20のウィンドウを使用して、本明細書の他の箇所に記載されるシグネチャーマトリックスを使用して $k=4$ 細胞型を列挙した。TCGA（Cancer Genome Atlas）試料の高解像度精製のために、LUADおよびLUSC腫瘍（および隣接する正常組織）のRNA-Seqプロファイルを一様に処理し、正規化し、単一の発現マトリックスにマージし、異常値およびバッチ効果の証拠についてt-SNEによって視覚化してもよい。推定される異常値試料を特定し、さらなる分析から省略することができる。高解像度精製後、細胞型特異的DEGは、LUAD腫瘍試料とLUSC腫瘍試料との間で特定され得る。有意性についてDEGをフィルタリングするために、例えば、追加のストリンジェンシーについて0.5未満のQ値閾値と連結された模擬黒色腫腫瘍（図15a~c）におけるDEG回復の感受性および特異性を最大化する、経験的に由来する閾値（例えば、0.1超の $\log_2$ 倍変化）を適用し得る。

【0167】

黒色腫図6bについて、長さ32（=4k）のウィンドウを使用してもよく、 $k=8$ 種の主な細胞型を標的とする単一細胞由来の黒色腫シグネチャーマトリックスを使用してもよい（図8b）。DEGは、NSCLCについて上述したように濾過されてもよい。

【0168】

バルク黒色腫生検におけるPD $CD1^+$ /CTLA4 $^+$ CD8 $^+$ T細胞の分析

CIBERSORTx法は、PD $CD1^+$ /CTLA4 $^+$ CD8 $^+$ T細胞を8種の他の

主要細胞型と共に列挙するように設計された黒色腫シグネチャーマトリックスを使用して、一括修正を用いて適用して、RNA-Seqによってプロファイルされた2つおよびNanoStringによってプロファイルされた3つのデータセットからのバルク黒色腫腫瘍における細胞組成物を列挙してよい(図9)。NanoString nカウンターによって尋問される遺伝子の数が少ないことを考慮して、シグネチャーマトリックス内の重複遺伝子は、単一細胞から再構築された黒色腫腫瘍上でベンチマークされ得、良好な性能を示すことが見出され得る(図21a~b)。

【0169】

統計分析

既知の細胞型特徴と予測される細胞型特徴(例えば、割合またはGEP)との間の線形一致は、示されるように、ピアソン相関(r)またはスピアマン相関(r_{ho})によって決定されてもよい。コンコーディネーション相関係数(CCC)は、例えばRパッケージDescToolsからのCCC関数を使用して予測および予想される $\log_2$ 調整された発現プロファイルを比較することによって決定され得る。データが通常分布されるとき、グループ比較は、不等しい変異を有する両側t検定または適切に対t検定を使用して決定されてもよく、そうでなければ、両側ウィルコクソン検定を適用してもよい。複数の仮説試験は、例えば、ベンジャミニおよびホッチバーグ法を使用して行われ得る。 $P < 0.05$ の結果は有意とみなされ得る。統計分析は、RおよびPrism v7(GraphPad Software, Inc.)で実施され得る。サンプルサイズの推定は、事前に指定された効果サイズを検出するのに十分な電力を確保するために行われない場合がある。

【0170】

コンピュータシステム

本開示は、本開示の方法を実装するようにプログラムされたコンピュータシステムを提供する。図22は、例えば、生体試料から特徴プロファイルを生成し、特徴プロファイルと、異なる細胞型に関する特徴シグネチャーの参照マトリックスとの間の回帰を最適化し、生体試料中の異なる細胞型の存在量を定量化するようにプログラムされるか、またはそれ以外の方法で構成されるコンピュータシステム2201を示す。

【0171】

コンピュータシステム2201は、例えば、生体試料から特徴プロファイルを生成すること、特徴プロファイルと、異なる細胞型の特徴署名の参照マトリックスとの間の回帰を最適化すること、および生体試料中の異なる細胞型の豊富性を定量化することなど、本開示の分析、計算、および生成の様々な態様を調節することができる。コンピュータシステム2201は、ユーザの電子デバイスまたは電子デバイスに対して遠隔に位置するコンピュータシステムであってもよい。電子デバイスは、モバイル電子デバイスであり得る。

【0172】

コンピュータシステム2201は、単一のコアまたはマルチコアプロセッサ、または並列処理のための複数のプロセッサであり得る中央処理ユニット(CPU、本明細書において「プロセッサ」および「コンピュータプロセッサ」も含む)2205を含む。コンピュータシステム2201はまた、メモリまたはメモリ位置2210(例えば、ランダムアクセスメモリ、読み取り専用メモリ、フラッシュメモリ)、電子記憶ユニット2215(例えば、ハードディスク)、1つ以上の他のシステムと通信するための通信インターフェース2220(例えば、ネットワークアダプタ)、ならびにキャッシュ、他のメモリ、データ記憶装置、および/または電子ディスプレイアダプタなどの周辺デバイス2225を含む。メモリ2210、記憶ユニット2215、インターフェース2220、および周辺デバイス2225は、マザーボードなどの通信バス(ソリッドライン)を介してCPU2205と通信する。記憶ユニット2215は、データを記憶するためのデータ記憶ユニット(またはデータリポジトリ)であり得る。コンピュータシステム2201は、通信インターフェース2220を用いて、コンピュータネットワーク(「ネットワーク」)2250に動作可能に結合され得る。ネットワーク2230は、インターネット、インターネ

ットおよび／またはエクストラネット、またはインターネットと通信しているイントラネットおよび／またはエクストラネットであってもよい。

【0173】

いくつかの場合では、ネットワーク2230は、通信および／またはデータネットワークである。ネットワーク2230は、クラウドコンピューティングなどの分散型コンピューティングを可能にすることができる1つ以上のコンピュータサーバを含むことができる。例えば、1つ以上のコンピュータサーバは、例えば、生体試料から特徴プロファイルを生成すること、特徴プロファイルと、異なる細胞型に関する特徴シグネチャーの参照マトリックスとの間の回帰を最適化すること、および生体試料中の異なる細胞型の存在量を定量化することなど、ネットワーク2230（「クラウド」）を介したクラウドコンピューティングが、本開示の分析、計算、および生成の様々な態様を実行することを可能にし得る。このようなクラウドコンピューティングは、例えば、Amazon Web Services（AWS）、Microsoft Azure、Googleクラウドプラットフォーム、およびIBMクラウドなどのクラウドコンピューティングプラットフォームによって提供され得る。ネットワーク2230は、いくつかの場合、コンピュータシステム2201を支援して、ピアツーピアネットワークを実装することができ、これは、コンピュータシステム2201に結合されたデバイスがクライアントまたはサーバとして動作することを可能にし得る。

10

【0174】

CPU2205は、1つ以上のコンピュータプロセッサおよび／または1つ以上のグラフィックス処理ユニット（GPU）を含んでよい。CPU2205は、プログラムまたはソフトウェアに具体化され得る一連の機械可読命令を実行することができる。命令は、メモリ2210などのメモリ位置に保存され得る。命令は、CPU2205に向けられ得、CPU2205はその後、本開示の方法を実装するためにCPU2205をプログラムまたは別様に構成することができる。CPU2205によって実行される動作の例は、フェッチ、デコード、実行、および書き戻しを含むことができる。

20

【0175】

CPU2205は、集積回路などの回路の一部であり得る。システム2201の1つ以上の他の構成要素を回路に含めることができる。いくつかの場合では、回路は、特定用途向け集積回路（ASIC）である。

30

【0176】

記憶ユニット2215は、ドライバ、ライブラリー、および保存されたプログラムなどのファイルを記憶することができる。記憶ユニット2215は、ユーザデータ、例えば、ユーザの好みおよびユーザプログラムを記憶することができる。いくつかの場合では、コンピュータシステム2201は、イントラネットまたはインターネットを介してコンピュータシステム2201と通信しているリモートサーバ上に位置するなど、コンピュータシステム2201の外部にある1つ以上の追加のデータ記憶ユニットを含むことができる。

【0177】

コンピュータシステム2201は、ネットワーク2230を介して1つ以上のリモートコンピュータシステムと通信することができる。例えば、コンピュータシステム2201は、ユーザのリモートコンピュータシステムと通信することができる。リモートコンピュータシステムの例としては、パーソナルコンピュータ（例えば、ポータブルPC）、スレートまたはタブレットPC（例えば、Apple（登録商標）iPad（登録商標）、Samsung（登録商標）Galaxy Tab）、電話、スマートフォン（例えば、Apple（登録商標）iPhone（登録商標）、Android対応デバイス、BlackBerry（登録商標））、またはパーソナルデジタルアシスタントが挙げられる。ユーザは、ネットワーク2230を介してコンピュータシステム2201にアクセスすることができる。

40

【0178】

本明細書に記載される方法は、例えばメモリ2210または電子記憶ユニット2215

50

などのコンピュータシステム 2201 の電子記憶位置に記憶された機械（例えば、コンピュータプロセッサ）実行可能コードによって実装され得る。マシン実行可能コードまたはマシン可読コードは、ソフトウェアの形式で提供することができる。使用中、コードはプロセッサ 2205 によって実行され得る。いくつかの場合では、プロセッサ 2205 がすぐにアクセスできるように、コードをストレージユニット 2215 から取り出してメモリ 2210 に保存することができる。いくつかの状況では、電子記憶ユニット 2215 が除外され得、機械実行可能命令がメモリ 2210 上に記憶される。

【0179】

コードは、コードを実行するように適合されたプロセッサを有するマシンと共に使用するために事前にコンパイルされ、構成され得るか、またはランタイム中にコンパイルされ得る。コードは、コードが事前にコンパイルされた、またはコンパイルされた状態で実行できるように選択され得るプログラミング言語で供給することができる。

【0180】

コンピュータシステム 2201 などの本明細書に提供されるシステムおよび方法の態様は、プログラミングに具体化することができる。本技術の様々な態様は、典型的には、機械（またはプロセッサ）実行可能コードおよび／または機械可読媒体のタイプ上に担持されるか、またはそのタイプに具体化される関連データの形態である「製品」または「製品」と考えられ得る。機械実行可能コードは、メモリ（例えば、読み取り専用メモリ、ランダムアクセスメモリ、フラッシュメモリ）またはハードディスクなどの電子記憶ユニットに記憶され得る。「ストレージ」タイプの媒体は、ソフトウェアプログラミングのためにいつでも非一時的なストレージを提供し得る、コンピュータ、プロセッサなどの有形メモリのいずれかまたはすべて、または様々な半導体メモリ、テープドライブ、ディスクドライブなどのそれらの関連するモジュールを含み得る。ソフトウェアのすべてまたは一部は、インターネットまたは様々な他の通信ネットワークを介して通信され得る。このような通信は、例えば、管理サーバまたはホストコンピュータからアプリケーションサーバのコンピュータプラットフォームにソフトウェアを別のコンピュータまたはプロセッサにロードすることを可能にし得る。したがって、ソフトウェア要素をもつことができる別のタイプの媒体は、ローカルデバイス間の物理的インターフェースにわたって、有線および光学的固定線ネットワークを介して、および様々な空気リンクを介して使用されるなどの光学、電気および電磁波を含む。有線または無線リンク、光学リンクなどのそのような波浪を運ぶ物理的要素は、ソフトウェアを運ぶ媒体としても考えられ得る。本明細書で使用される場合、非一時的で有形の「記憶」媒体に限定されない限り、コンピュータまたは機械「可読媒体」などの用語は、実行のためにプロセッサに命令を提供することに関与する任意の媒体を指す。

【0181】

したがって、コンピュータ実行可能コードなどの機械可読媒体は、有形記憶媒体、キャリア波形媒体または物理伝送媒体を含むが、これらに限定されない多くの形態をとってもよい。不揮発性記憶媒体としては、例えば、図面に示されるデータベース等を実装するために使用され得る任意のコンピュータ（複数可）等の記憶装置のうちのいずれかなどの光ディスクまたは磁ディスクが挙げられる。揮発性記憶媒体は、そのようなコンピュータプラットフォームのメインメモリなどのダイナミックメモリを含む。有形伝送媒体としては、同軸ケーブル、コンピュータシステム内のバスを含むワイヤを含む銅線および繊維光学系が挙げられる。キャリア波伝送媒体は、電気信号または電磁信号、または無線周波数（RF）および赤外線（IR）データ通信中に生成されるものなどの音波または光波の形態をとってもよい。したがって、コンピュータ可読媒体の一般的な形態は、例えば、フロッピーディスク、可撓性ディスク、ハードディスク、磁気テープ、任意の他の磁気媒体、CD-ROM、DVDまたはDVD-ROM、任意の他の光学媒体、パンチカード紙テープ、孔のパターンを有する任意の他の物理的記憶媒体、RAM、ROM、PROMおよびEPROM、FLASH-EPROM、任意の他のメモリチップまたはカートリッジ、データまたは命令を輸送するキャリア波形、そのようなキャリア波形を輸送するケーブルまた

10

20

30

40

50

はリンク、またはコンピュータがプログラミングコードおよび／またはデータを読み取ることができる任意の他の媒体を含む。コンピュータ可読媒体のこれらの形態の多くは、実行のためにプロセッサに1つ以上の命令の1つ以上のシーケンスを担持することに関与し得る。

【0182】

コンピュータシステム2201は、例えば、生体試料からの特徴プロファイル、特徴プロファイルと、異なる細胞型の特徴シグネチャーの参照マトリックスとの間の回帰のソリューション、ならびに生体試料中の異なる細胞型の存在量を示す視覚ディスプレイを提供するためのユーザインターフェース（UI）2240を備える電子ディスプレイ2235を含み得るか、またはそれと通信し得る。UIの例としては、グラフィカルユーザインターフェース（GUI）およびウェブベースのユーザインターフェースが挙げられるが、これらに限定されない。

【0183】

本開示の方法およびシステムは、1つ以上のアルゴリズムによって実装され得る。アルゴリズムは、中央処理ユニット2205によって実行されるとソフトウェアによって実装され得る。アルゴリズムは、例えば、生体試料から特徴プロファイルを生成し、特徴プロファイルと、異なる細胞型の特徴署名の参照マトリックスとの間の回帰を最適化し、生体試料中の異なる細胞型の存在量を定量化することができる。

【0184】

デコンボリューションモジュールは、1つ以上のデコンボリューションアプリケーション、例えば、サポートベクトル回帰（SVR）アプリケーションまたは非負行列因数分解（NMF）アプリケーションを含み得る。これは、メモリに格納された1つ以上のプログラムを含み、本開示の1つ以上の実施形態による方法を実行する命令を含み得る。復元アプリケーション（例えば、SVRアプリケーションまたはNMFアプリケーション）は、以下の例示的なモジュールまたはそのサブセットもしくはスーパーセットのいずれかを含んでもよい。

【0185】

いくつかの事例では、復元モジュールは、本開示の1つ以上の実施形態に従って、異なる構成要素、例えば、細胞サブセット（例えば、細胞タイプ）、特徴署名の参照マトリックスを使用して、物理試料中の異なる構成要素、例えば、細胞サブセット（例えば、細胞タイプ）の相対比率または存在量を定量化するために、線形損失関数およびL2標準ペナルティ関数を最小限にするサポートベクトル回帰または任意の他の回帰アルゴリズムを物理試料（例えば、生体試料）の特徴プロファイルに適用するように構成され得る。

【0186】

復元モジュールは、本開示の1つ以上の実施形態に従って、署名マトリックスに含める特徴を選択（またはフィルタリング）し、かつ／または特徴プロファイル（複数可）を選択するように構成され得る選択モジュールを含み得る。

【0187】

復元モジュールは、本開示の1つ以上の実施形態に従って、 nu の異なる値にわたって最も低い誤差で結果を判定するように構成され得るRMSEモジュールを含み得る。

【0188】

復元モジュールは、有意性値モジュールを含んでもよく、これは、a) 親特徴プロファイルからランダムに選択される特徴を含有するランダム特徴プロファイル m^* を生成し、親特徴プロファイルを含み、 m^* と m^* が同じユークリッドノルムを有することと、b) m^* と参照マトリックス B との間の回帰を最適化し、 m^* が B の線形結合としてモデル化され、最適化することが、回帰の回帰係数のセットを含む f^* を解くことを含み、解が、線形損失関数と、L2ノルムペナルティ関数とを最小化することと、c) f^* と参照マトリックス B の乗積を計算して、再構成された特徴プロファイルを生成することと、d) ランダム特徴プロファイル m^* と再構成された特徴プロファイルとの間の回帰を決定することと、e) 回帰係数のセットから決定された測定値の分布に基づいて有意性値を決定するこ

10

20

30

40

50

とと、を含み、 i が、1 以上である。有意性値モジュールは、ステップ b) でデコンボリューションモジュール、例えば、SVRモジュールを用いてもよい。重要性値モジュールは、本開示の他の実施形態のいずれかを実行するようにさらに構成され得る。

【0189】

復元モジュールは、本開示の他の実施形態のいずれかを実行するための追加のモジュールをさらに含んでもよい。特定の態様では、デコンボリューションモジュールは、コンピュータとは別のポータブルコンピュータ可読記憶媒体に保存され得る。

【0190】

デコンボリューションモジュールは、ユーザによって構成可能であり得る。例えば、デコンボリューションモジュールは、ユーザが、デコンボリューションアルゴリズム（例えば、SVR）を n の値に適用するように特徴プロファイルおよび／または署名マトリックスなどの1つ以上の設定を決定することを可能にするように構成されたユーザインターフェースモジュール、特徴が選択モジュールによって選択される基準、有意性値モジュールによって実行される反復の数、または本開示の1つ以上の実施形態を可能にする任意の他の設定を含んでもよい。

【実施例】

【0191】

CIBERSORT x 方法は、フレームワークとして、バルク組織のゲノムプロファイルから細胞組成を堅牢に評価するための機械学習技術を使用することができる。CIBERSORT x 方法は、シリコ解剖の概念を蛍光活性化セルソーティング（FACS）に拡張し、新たな単一細胞配列決定の努力を活用し、それによって細胞型の列挙および細胞型特異的遺伝子発現精製のためのフレームワークを提供し得る（図1に示されるように）。以下の実施例では、このアプローチを説明し、検証する。まず、バルク組織における細胞型比率を推定するための単一細胞参照プロファイルを使用し、続いて、新鮮／凍結および固定組織標本のRNAプロファイルから細胞型特異的遺伝子発現シグネチャーを精製するための分析技術を使用する。

【0192】

実施例1：scRNA-Seqによる組織解剖

CIBERSORT x プラットフォームは、単一細胞参照プロファイルを含む、様々な供給源に由来する細胞シグネチャーを使用して、大規模な組織特徴付けを可能にすることができる（図1、ステップ1および2）。単一細胞またはバルクソートされた転写データからシグネチャーマトリックスを導出するための分析技術を開発し、混乱した技術的変動の原因としてバッチ効果を最小限に抑えた（図7a～hおよび図8a～i）。次いで、10個のGenomics Chromium（液滴ベース）およびSMART-Seq2（プレートベース、Clontech Laboratories, Inc.）を含む、バルク組織由来のRNA混合物における細胞比率を列挙するためのscRNA-Seqプラットフォームの有用性を調査した。

【0193】

Chromium v2（3'アッセイ）を使用して、非小細胞肺癌（NSCLC）患者から得られた2つの連続性末梢血単核細胞（PBMC）試料から単一細胞RNA-Seqライブラリーを生成した。監督されていないクラスタリングおよび正規マーカー遺伝子評価は、両方の採血における6つの主要な白血球サブセット（B細胞、CD4⁺T細胞、CD8⁺T細胞、NK細胞、NK細胞、および単球；図2aに示されるように）を明らかにした。デコンボリューション性能を評価するために、シグネチャーマトリックスを構築して、これら6つの細胞サブセットを区別し、シグネチャーマトリックスを、12名の健康な成人から得られた血液のバルクRNA-Seqプロファイルの検証コホートで試験した。

【0194】

直接サイトメトリーおよび蛍光免疫表現型によって決定されるグラウンドトゥルス細胞比率と比較して、未補正のデコンボリューション結果は、バルク混合物中のいくつかの

10

20

30

40

50

細胞型について明確な推定バイアスを示した（図2 bおよび6）。そのようなバイアスは、例えば、ライブラリー調製中の固有の分子識別子の可変使用、または転写物に沿った差次的3'バイアスによって導入され得るように、シグネチャマトリックスとバルクRNA-Seqデータとの間のプラットフォーム特異的変動によって駆動され得る。実際に、一括修正スキーム（図6 a）の適用後、デコンボリューション結果は、グラウンドトゥールズ細胞比率と比較して有利である（図2 bおよび6 e）。公的に入手可能なChromium v2 PBMCデータ（3'および5'キット）、ならびにマイクロアレイを使用してプロファイリングされた精製白血球サブセット（図6 d～g）を含む、他のデータセットおよびシグネチャマトリックスを分析する際の一括修正によって、性能の同様の増加を得た。これらのシステムの改善を考慮して、別段の記載がない限り、一括修正を、その後のすべてのクロスプラットフォーム分析に適用した。これらのデータは、デコンボリューション実験におけるクロスプラットフォーム変動を最小限に抑える値、およびバルク組織組成を解剖するためのscRNA-Seqプロファイルの約束を強調した。

【0195】

次に、分析を固形腫瘍生検に拡張し、単一細胞をSMART-Seq 2によってプロファイリングした。頭頸部扁平上皮癌（HNSCC）および黒色腫（それぞれn=18および19患者）に焦点を当て、デコンボリューション性能を最初に、表現型定義における解離関連人工物および異質性を制御する設定で試験した。HNSCC患者からの18個の原発腫瘍および5個のリンパ節転移のデータセットでは、2個の原発腫瘍標本および1個のリンパ節生検を含むトレーニングコホートからシグネチャマトリックスを作成した。このマトリックスは、これらの腫瘍において、悪性細胞、CD4およびCD8 T細胞、B細胞、マクロファージ、樹状細胞、肥満細胞、内皮細胞、筋細胞、ならびに癌関連線維芽細胞（図2 c）の全10種類の主要細胞を区別した。注釈された表現型を有する単一細胞から再構築された保留腫瘍を使用して評価したとき、デコンボリューション結果は、グラウンドトゥールズ比と高度に一致した（図2 dおよび図7 a）。さらに、これらの結果は、シグネチャマトリックスを構築するために使用される腫瘍およびリンパ節の特定の組み合わせに依存しなかった。また、稀であるか、または単離することが難しい細胞下位集団内の、免疫療法応答に潜在的に関連する浸潤性黒色腫の異なるCD8 T細胞サブセットなどを含む、腫瘍型（HNSCCまたは黒色腫）および細胞型にわたって強力な性能を維持した（図7 b～dおよび8）。

【0196】

これらの好ましい結果を考慮して、CIBERSORTxデコンボリューションに対するキーシグネチャマトリックス関連パラメータの影響（表現型あたりの細胞数および分析されたドナー試料の数を含む）を検討した。パラメータ値の範囲にわたって、細胞比率推定に対する適度な影響が観察された（図9 a～b）。さらに、その供給源にかかわらず、CIBERSORTxシグネチャマトリックスは、プラットフォーム、データセット、または組織にわたって適用されるかにかかわらず、強力な一般化性を示した（図2 e）。

【0197】

次に、単一細胞誘導型デコンボリューションの技術的態様を検証した後、その有用性を、様々な一次組織を使用して、実際の生物学的混合物の特徴付けについて調査した。上述の黒色腫生検からの単一細胞由来のシグネチャマトリックスを活用して、CIBERSORTxを適用して、Cancer Genome Atlas（TCGA）からの黒色腫RNA-Seqプロファイルを解剖した。予期せず、バルク腫瘍における予測細胞型比率と元のscRNA-Seq結果を比較すると、B/Tリンパ球およびマクロファージの画分表現において顕著な差が観察された（図7 e）。これらの細胞サブセットは、scRNA-Seqデータセットにおいて互いに選択されていないため、そのような組成的歪みは、単一細胞単離および配列決定による技術的人工物、またはデコンボリューションアプローチ自体のいずれかから生じる可能性がある。前者を裏付けるために、CIBERSORTx推定値は、独立した黒色腫コホートにおける免疫組織化学（IHC）による直接TILの現場列挙と高度に一致した（図7 e）。さらに、scRNA-Seq（SMART-

10

20

30

40

50

Seq 2)、バルクRNA-Seq、およびIHCによってプロファイリングされたヒト脾島のデータセットにおいて、同じ歪み現象が観察された(図2fおよび7f~i)。一致する島標本の直接比較では、バルク組織におけるIHCによって決定された細胞画分は、バルク島デコンボリューション結果と強く相関したが、scRNA-Seqとは相関しなかった(図2fおよび7i)。これらのデータは、CIBERSORTxをさらに検証し、解離関連歪みを軽減しながら、固形組織組成を特徴付けるためのその値を強調する。

【0198】

実施例2：物理的な細胞単離なしの細胞型特異的遺伝子発現

単一細胞またはバルクソートされた集団に由来する細胞型特異的トランスクリプトームプロファイルは、細胞同一性および機能に関する貴重な洞察を提供することができる。しかしながら、そのようなプロファイルは、例えば、大きなコホートおよび固定された臨床試料に関して取得することが困難であり得る。精製された細胞型が利用可能であっても、組織の解離および保存条件は、下流分析を不明瞭にする遺伝子発現の非生物学的変化を引き起こし得る。バルク組織RNAプロファイルを細胞型特異的トランスクリプトームに数学的に分離することは、これらの問題を克服する可能性があるが、そのような技術は、実際の組織試料を使用して厳格な検証を必要とし得る。したがって、高度に最適化されたマーカー遺伝子を含むシグネチャマトリックスを使用して、新鮮な／凍結されたおよび固定された腫瘍を含む非分解組織試料から生物学的に有意なトランスクリプトームプロファイルを学習することができるかどうかを評価した(図3a)。

【0199】

代表的な細胞型特異的トランスクリプトームが、新鮮な／凍結された原発腫瘍生検から正確に推測できるかどうかを試験するために、302個の腫瘍を未治療濾胞リンパ腫(FL)患者からプロファイリングし、共通のアプローチを試験し、各細胞型の比率を使用して混合試料群から各細胞型の単一代表的GEPを推測する。B細胞、CD8 T細胞、およびCD4 T細胞は、FL腫瘍細胞の大部分を含み得、FACSによって容易に精製され得るため、これら3個のサブセットを使用して、アプローチの精度を評価した。CIBERSORTxはまた、単一細胞由来のシグネチャマトリックスをサポートしているが、LM22(22ヒト造血細胞サブセットを区別するためのマイクロアレイ由来のシグネチャマトリックス)および非負最小二乗回帰(NNLS)を用いて、以前の方法(CIBERSORTと称される)を適用して、技術的因子について直接制御しながら免疫比率を列挙した。重要なことに、LM22は、対試料を使用して同じ3個のFL細胞サブセットを正確に分解する能力について、フローサイトメトリーに対して以前に検証された。加えて、この実験におけるすべての発現データは、Affymetrixマイクロアレイによってプロファイリングされ、プラットフォーム特異的因子とは関係なくシリコ発現精製の実現可能性の評価を可能にした。最初に、B細胞およびCD8 T細胞を、FLリンパ節における高存在量および低存在量の細胞型の例として検査した(それぞれ>50%対約5~10%)。

【0200】

推定された細胞型トランスクリプトームおよびFACS精製細胞型トランスクリプトームは、合理的に良好に相関したが、かなりのノイズが、広範な発現レベルにわたって多くの遺伝子の発現推定値を歪めた(図3b)。したがって、各細胞型に関する信頼性のない推定遺伝子を除去するための適応ノイズフィルタ(図3c)を定義した。この戦略は、シリコ精製トランスクリプトームとFACS精製細胞由来のトランスクリプトームとの間の相関を著しく改善した(図3dおよび11a)。

【0201】

次いで、トランスクリプトーム再構築の精度に影響を及ぼす主要な要因を調査した(図3eおよび11a~h)。302個のFL GEPのセットを使用して、プロファイリングされた腫瘍の数とB細胞、CD4 T細胞、およびCD8 T細胞にわたるトランスクリプトーム推定の精度との間の予測可能な関係が観察され、各々が、FL腫瘍間で10倍以上の頻度で変化し得る(図3eおよび11b)。最大の増加は、細胞型より少なくとも4

～5倍多い混合物試料を分析するときに達成されたが（図3e）、濾過スキームは、コホートサイズ（c）、細胞型の数（k）、および全体的な細胞型存在量（ $c > k$ を前提とする、図3e、11b～eおよび11g）に関係なく、以前のアプローチと比較して均一に性能を向上させた。以前に同定された細胞同一性マーカーを解像するために有利な性能が観察された（図3f）。FLリンパ節における濾胞ヘルパーT細胞（図11d）、PBM Cにおける白血球サブセット（図11e～h）、ならびにNSCLC腫瘍のRNA-Se qプロファイルからの上皮、免疫、および間質サブセット（図3g）を含む、他の細胞型について同様の結果を得た。さらに、同一の細胞型を照合することを条件として、異なる単一細胞由来のシグネチャマトリックスの使用は、結果に有意に影響しなかった（図10b～c）。

【0202】

実施例3：高解像度での細胞型特異的発現精製

細胞型特異的差次的発現をバルク標本から推測することができる。いくつかの場合では、上述の戦略は、混合試料群（例えば、「群レベル」のGEP、図1、工程3）が与えられた各細胞型に対する単一の代表的発現プロファイルの学習に限定され得る。そのようなプロファイルは、試料特異的ではなく、それらの間の差次的発現遺伝子（DEG）の同定などの下流分析を行うために、目的の各条件の生成を必要とし得る。試料レベルのデコンボリューションについて公開されたアプローチが説明されているが、これらは2つの細胞成分を有する混合物のみを考慮することができる。CIBERSORTx法は、バルク遺伝子発現および推定された細胞比率を入力として使用して、部分的な観察を有する非負行列因子分解問題として遺伝子発現デコンボリューションをモデル化することができる（図12a～c）。このアプローチは、混合物GEPの単一マトリックスを、基礎となる細胞型特異的発現マトリックスのセットに分離する（図4a～b）。これらの発現プロファイルが得られた後、それらを事後分析して、個々の目的の細胞型について、遺伝子発現の試料レベルの変動およびパターンに関する洞察を得ることができる。数値近似を含んでよいマトリックス要因分解問題を解決するために、アルゴリズム（例えば、分割および征服アルゴリズム）を使用して、生物学的に現実的な解決策を生成することができる（図12a～cおよび図13a～b）。

【0203】

各々が1つ以上の細胞型でDEGを含有する一連の合成混合物を作成して、方法の「高解像度」細胞精製能力を試験した（図1、ステップ4）。これらのシミュレートされたDEGには、ブロック様パターン（実際の組織で見られるものを連想させる）、および非線形の幾何学形状が含まれ、これらのすべては、以前の計算技術によって確認することが困難であり得る。顕著には、この方法は、隠された標的（「ブルズアイ」）を含む、すべての試験されたケースにおいて予想されるDEGパターンを回復した（図4b～dおよび14）。さらに、グループモード（図1、ステップ3）とは異なり、得られた高解像度プロファイルは、監督されていない分析およびクラス発見のための標準的な方法（例えば、図4c）に本質的に適合した。したがって、高解像度精製は、（例えば、生体的または機能的に定義されたグループ化の）事前の知識なしに、多様な試料レベルの発現パターンをマスク解除するための汎用的なアプローチであり得る。

【0204】

次に、モデル化腫瘍混合物中のDEGの検出方法の分析性能（例えば、感受性および特異性）を評価した。シミュレートされたDEGを、CD8 T細胞トランスクリプトームに「スパイク」して、2つの既知のCD8表現型クラスを作成した。次いで、これらのCD8 GEPを、モデル化された腫瘍において他の3つの免疫サブセットとインシリコでランダムに混合し、結腸癌細胞株を含めて、50%の未知の含有量をシミュレートした（図4e、左）。高解像度精製後、細胞型特異的トランスクリプトームを定義されたDEGクラスにグループ化して、性能を評価した。広範囲の細胞スパイクインレベルおよび発現倍率変化にわたって、以前に定義されたDEGを、高感受性および特異性でCD8 T細胞中で回収した（図4e、右）。さらに、12.5%および5%の画分存在量で、予想さ

10

20

30

40

50

れるDEGは、CD8⁺T細胞において、それぞれわずか2倍(AUC=0.84)または4倍(AUC=0.85)の発現変化で検出可能であった(図4e、右)。同様の分析感受性および特異性が別の実験で観察され、ここで黒色腫瘍をプールされたscRNA-Seqデータを使用してシミュレートし、コホートサイズ、DEGを有する細胞型の数、および差次的発現の大きさに関して性能を評価した(図15a~c)。

【0205】

実施例4：様々な腫瘍下位集団の高解像度プロファイリング

高解像度GEP精製の強力な技術的性能は、腫瘍を含む実際のバルク組織試料を解剖するための有用性を示唆している。したがって、悪性細胞における分子異質性のパターンが以前に説明されている2つの癌から始まる、いくつかのヒト腫瘍型における細胞状態を解決する能力を検査した。

【0206】

びまん性大細胞型B細胞リンパ腫(DLBCL)は、B細胞分化状態の違いに基づいて2つの主要な分子サブタイプに分類することができる：生殖中心様(GCB)および活性化B細胞様(ABC)DLBCL。重要なことに、これらのB細胞サブセットの下にあるGEPは、精製された原発DLBCL腫瘍B細胞および腫瘍細胞株において明らかであり得るが、別個のシグネチャーを有する他のTILにおいては明らかでない。以前に注釈された細胞起源サブタイプを有する150個のDLBCLリンパ節生検のGEPを使用して、B細胞に既知のGCB/ABC差を正しく付与することにおけるその性能について、10個の主要な白血球サブセットの高解像度プロファイリングを評価した。このアプローチは、クラス標識に対して盲目であったが、B細胞におけるサブタイプ特異的発現差は、正常なGC細胞および活性化B細胞のサブタイプと高度に一致し、かつバルクDLBCL腫瘍よりも平均で約9倍有意であった(図16a~c)。これらの結果を、2つの代替方法と比較した。(1)細胞存在量との相関に基づいてバルク組織発現パターンを個々の細胞型に割り当てるための一般的なアプローチ、および(2)表現型クラスおよび細胞型頻度が既知であるときに、細胞型特異的DEGを推定するための前述の技術。いずれの場合においても、CIBERSORTxは、細胞型特異性と所与の有意性閾値で検出可能なDEGの数の両方に関して優れた性能を示した(図17a-eおよび18a-b)。

【0207】

FLは、非常に一般的な非ホジキンリンパ腫であり得、FL腫瘍におけるCREBBP突然変異は、B細胞における抗原提示の喪失と関連付けられ得る。この関連付けは、FACS精製FLB細胞を使用して観察されたが、これらの腫瘍における正常B細胞または他の細胞型ではないため、高解像度精製を、バルク腫瘍GEPおよび対腫瘍遺伝子型からこの結果を再現する能力について評価した(図5a)。実際に、CREBBP突然変異状態によって腫瘍を分層した後、前述のシグネチャーは、CREBBP突然変異腫瘍におけるMHC-II発現の喪失を含む、デジタルソートされたB細胞発現プロファイルにおいて容易に検出可能であった(図5b~c)。DLBCLに関して観察されたように(図17a~e)、これらの遺伝子の大部分は、B細胞の存在量と相関せず、デコンボリューションなしのバルク組織におけるそれらの発見を妨げた。したがって、2つの異なる悪性腫瘍において、CIBERSORTxは、物理的な細胞単離または表現型のグループ化に関する予備知識なしに、既知の状況依存的な細胞状態を容易に確認し、新規生物学を明らかにする約束を強調した。

【0208】

分析複数の細胞型にも拡張しながら本出願を例示するために、RNA-Seqライブラリーを、同じ標本からFACS精製された上皮、造血、内皮、および線維芽細胞サブセットを含む、22回の外科的に切除されたNSCLCバルク腫瘍生検から生成した(図5d)。4患者のサブセットからシグネチャーマトリックスを導出した後、高解像度プロファイリングを適用して、18腫瘍GEPの検証セットからこれら4つの集団をデジタル解剖した。一致する患者試料上のFACS精製された細胞集団との直接比較において、シリコプロファイルは、精製の成功の強力な証拠を示した(図18c~e)。加えて、CIBE

10

20

30

40

50

R S O R T x は、バルク腫瘍から上皮細胞の G E P を精製するための以前の方法と比較して、有意な利点を示した（図 1 8 c ~ d）。

【 0 2 0 9 】

したがって、同じシグネチャーマトリックスを適用して、TCGAからの518個の肺腺癌（LUAD）腫瘍、502個の肺扁平上皮細胞癌（LUSC）腫瘍、および110個の隣接する正常組織のバルクRNA-Seqプロファイルから上皮、造血、内皮、および線維芽細胞GEPを解像した（図5d）。T-SNEを使用して、インシリコ精製された細胞集団における分布を視覚化し、探索した（図5e）。以下の推定外れ値試料を同定し、さらなる分析から省略した。TCGA.44.6775.01C.02R.A277.07、TCGA.44.3918.01A.01R.A278.07、TCGA.44.2656.01A.02R.A278.07、TCGA.44.4112.01A.01R.A278.07、TCGA.44.2665.01B.06R.A277.07、TCGA.44.2668.01A.01R.A278.07、TCGA.44.2666.01A.01R.A278.07、TCGA.44.6775.01A.11R.A278.07、TCGA.44.6147.01A.11R.A278.07、TCGA.44.2665.01A.01R.A278.07、TCGA.44.2662.01A.01R.A278.07、TCGA.44.6146.01A.11R.A278.07、TCGA.44.5645.01A.01R.A278.07、TCGA.44.3917.01A.01R.A278.07、TCGA.44.6147.01B.06R.A277.07、TCGA.44.2666.01B.02R.A277.07、TCGA.44.4112.01B.06R.A277.07、TCGA.44.3918.01B.02R.A277.07、TCGA.44.3917.01B.02R.A277.07、TCGA.44.2668.01B.02R.A277.07、TCGA.44.6146.01B.04R.A277.07、TCGA.44.2662.01B.02R.A277.07、TCGA.44.5645.01B.04R.A277.07、およびTCGA.44.2656.01B.06R.A277.07。これにより、癌関連線維芽細胞（CAF）における別個の表現型シフトを含む、各NSCLC組織学的サブタイプ（例えば、LUAD対LUSC）に対する差次的発現パターンの監督されていない発見が可能になった。対照的に、腫瘍関連内皮細胞および隣接する正常組織は、組織学的サブタイプに関係なく一緒にクラスター化した。これらの結果を検証するために、それらを20名の患者からのFACS精製されたNSCLC細胞型と比較した。全体トランスクリプトームレベルにおいて、および単離された細胞サブセットにおける差次的発現のパターンに関して強い整合性が観察され、インシリコ所見を検証した（図5e~f）。さらに、結果は、NSCLCにおける細胞不均一性の他の報告されたscRNA-Seqプロファイルと一致した。

【 0 2 1 0 】

NSCLC患者からのバルク腫瘍およびFACS精製された細胞サブセットの識別子を表1に示す。

【 0 2 1 1 】

10

20

30

40

50

【表 1】

腫瘍ID	組織学	バルクRNA-Se qプロファイルを有する試料 (1=はい、0=いいえ)					シグネチャーマトリックス
		バルク腫瘍	EpCAM+ (epith.)	CD45+ (immune)	CD10+ (fibro.)	CD31+ (endo.)	
T2	NOS	0	1	1	1	1	1
T3	LUSC	1	1	1	1	1	1
T8	LUAD	1	1	1	1	1	1
T9	ADSQ	1	1	1	1	1	1
T5	LUAD	1	1	1	1	1	0
T6	LUAD	1	1	1	1	1	0
T7	LUAD	1	1	1	1	1	0
T10	LUAD	1	1	1	1	1	0
T12	LUAD	1	1	1	1	1	0
T15	LUSC	1	1	1	1	1	0
T17	LUSC	1	1	1	1	1	0
T18	LUSC	1	1	1	1	1	0
T20	LUSC	1	1	1	1	1	0
T26	LUAD	1	1	1	1	1	0
T27	LUAD	1	1	1	1	1	0
T28	LUAD	1	1	1	1	1	0
T35	LUAD	1	1	1	1	1	0
T36	LUAD	1	1	1	1	1	0
T38	LUAD	1	1	1	1	1	0
T39	LUSC	1	1	1	1	1	0
T40	LUSC	1	1	1	1	1	0
T4	LUSC	1	1	1	1	1	0
T11	LUAD	1	1	1	1	1	0
T30	LUAD	1	1	1	1	1	0
T34	LUSC	1	1	1	1	1	0
T37	Other	1	1	1	1	1	0

【0212】

総合的に、これらのデータは、発現変動を学習するためのデジタルセルソーティングの有用性を示し、バルク標本中の分子経路を解剖するためのその値を強調する。

【0213】

実施例5：シリコサイトメトリーの黒色腫への適用

デジタルサイトメトリーの分析ワークフロー（図1）を確立した後、対応する技術を含む的なツールキット（<http://cibersortx.stanford.edu>）に統合した。次いで、黒色腫を有する患者から切除された腫瘍生検における細胞異質性を特徴付けるために、CIBERSORTxの3つの潜在的な適用を分析した（図6a～h）。方法の多様な能力を例示するために、以下の技術の各々を順次適用した：高解像度発現精製（図1、ステップ4、図6a～b）、群レベルの発現精製（図1、ステップ3、図6c）、および単一細胞参照プロファイルを使用する多様なプラットフォームにわたる細胞組成の列挙（図1、ステップ1および2、図6d～f）。

【0214】

癌性BRAF突然変異は、黒色腫の半分を超えて発生し得、承認された標的療法によって阻害され得るが、NRAS突然変異は、非BRAF突然変異型黒色腫腫瘍の約半分にお

いて発生し得るが、そのような療法を欠く場合がある。重要な突然変異が細胞状態にどのように影響するかを理解することは、新規の治療戦略につながる可能性がある。黒色腫瘍からの単一細胞参照プロファイルを使用して、シグネチャマトリックスを構築し、高解像度発現精製を使用して、TCGAによってプロファイリングされた301個のバルク黒色腫瘍のトランスクリプトームから8種の主要細胞型を解剖した(図6a)。インシリコ精製細胞サブセット内では、BRA FまたはNR A S突然変異状態に従って黒色腫を区別する悪性細胞およびC A F内で多くの有意なD E Gが発見された(図6b)。これらの所見を、突然変異データが入手可能であった原発性黒色腫からのs c R N A - S e qデータを使用して検証し、個々の悪性細胞および／またはC A F内のB R A FおよびN R A S遺伝子型に関連するG E Pの確認を可能にした(図6bおよび19a～b)。これらのデータは、単一細胞レベルでアプローチを検証し、ヒト腫瘍標本における体細胞変異と細胞表現型との間の理解不十分な関連性を解明する可能性をさらに例示する。

【0215】

腫瘍浸潤C D 8 T細胞は、慢性抗原刺激によって、または炎症シグナルへの過剰曝露によって、「消耗」状態に駆動され得る。現行および新興の癌免疫療法に対するこれらの細胞の重要性を考慮して、次にC I B E R S O R T xを使用して、消耗表現型を特徴付ける遺伝子発現の変化を検討した。健康な末梢血白血球に由来するL M 2 2を使用して、TCGA27によってプロファイリングされた新鮮な／凍結された黒色腫瘍における免疫組成を列挙した。群レベルの発現精製を行って、代表的なC D 8 T I L発現プロファイルを推定した。正常な末梢血C D 8 T細胞のベースライン参照プロファイルに対して推定C D 8 T I L G E Pを順位付けすることによって、P D C D 1およびC T L A 4を含む、重要な消耗マーカーが推定C D 8 T I L G E Pにおいて強烈に過剰発現したことが観察された(図6c)。加えて、C D 8 T I L特異的遺伝子は、単一細胞R N A - S e qおよびF A C Sによって黒色腫瘍から単離されたC D 8 T細胞について観察された遺伝子と一致した(図6c)。独立したコホートからのF F P E腫瘍に対する分析を繰り返すときにも、同様の結果が得られた(図6cおよび20)。したがって、C I B E R S O R T xは、単一細胞プロファイリングを必要とすることなく、固定腫瘍などの複雑な組織に浸潤する細胞下位集団の表現型状態の変化を正確に学習することができる。

【0216】

別に、転移性黒色腫のレジメンは、消耗したT細胞上のP D - 1および／またはC T L A 4発現を標的とする単剤または併用剤免疫療法によるチェックポイント遮断を用いることができる。患者のサブセットは、耐久性のある抗腫瘍T細胞応答を達成し得るが、臨床転帰は異質のままであり、効果的な予測バイオマーカーが欠如し得る。高レベルのP D C D 1(P D - 1をコードする)またはC T L A 4を発現するC D 8 T I Lは、これらの療法の重要な標的であり、両方のマーカーを発現するC D 8 T I Lは、P D 1遮断を受ける黒色腫患者におけるフローサイトメトリーによって最近示されたように、応答と関連し得ることを示唆する。

【0217】

この仮説を試験するために、黒色腫瘍の単一細胞参照プロファイルを使用して、8つの他の主要な腫瘍細胞型と併せてP D C D 1+／C T L A 4+C D 8 T細胞を含有するシグネチャマトリックスを構築した(図6d、9b、および21a～b)。シグネチャマトリックスを一括修正で適用して、公的に入手可能な3つの黒色腫発現データセットを照合した(図6d)。これらには、R N A - S e qまたはN a n o S t r i n gによってプロファイリングされ、P D - 1またはC T L A 4遮断の前、間、または後に得られたF F P Eおよび新鮮な／凍結された黒色腫瘍のバルク発現データが含まれていた。我々の仮説を裏付けるために、P D C D 1+／C T L A 4+C D 8 T細胞の推定レベルは、3つすべての試験において応答と有意に関連付けられた(図6e)。さらに、これらの細胞の検出は、このメタ分析における全生存期間を有意に分層し、個々のデータセットにおける生存期間曲線を分離し、かつ重要なマーカー遺伝子よりも生存期間および応答に有意に関連していた(図6fおよび21c～f)。これらのデータは、単一細胞参照プロファイ

ルに由来する候補バイオマーカーの迅速な評価のためのCIBERSORTxの有用性を示し、黒色腫患者における腫瘍浸潤PDCD1+／CTLA4+CD8 T細胞のさらなる分析が必要であることを示唆する。

【0218】

実施例6：代表的な細胞型特異的発現プロファイルの推定

以下に示す方程式(5)～(8)を使用して、CIBERSORTxアルゴリズムを実行する(例えば、代表的な細胞型特異的発現プロファイルを推定する)ことができる。

$$B \times F_{\cdot, j} = M_{\cdot, j}, \quad 1 \leq j \leq c \quad (5)$$

$$H_{i, \cdot} \times F = M_{i, \cdot}, \quad 1 \leq i \leq n \quad (6)$$

$$F \times G_{i, \cdot} = P_{i, \cdot}, \quad 1 \leq i \leq n \quad (7)$$

$$\text{diag}(P_{i, \cdot}) = M_{i, \cdot}, \quad 1 \leq i \leq n \quad (8)$$

【0219】

B(m個の遺伝子×k個の異なる細胞型からなるm×kシグネチャーマトリックス(例えば、LM22)であり得る)およびM(n個の遺伝子×c個の試料からなるGEPのn×cの混合物であり得る)を与えられた場合、方程式5の目標は、F(Mにおける各試料のk個の細胞サブセットの画分存在量からなるk×cマトリックスであり得る)を推定することである。方程式6の目標は、FおよびMを与えられた場合、H(代表的な細胞型特異的発現シグネチャーのn×kマトリックスであり得る)を推定することである。方程式7および8の目標は、FおよびMを与えられた場合、G(n個の遺伝子、c個の試料、およびk個の細胞型からなるn×c×kマトリックスであり得る)を推定することである。

【0220】

Rにおけるnnlsパッケージを使用する非負最小二乗回帰(NNLS)を用いて、FおよびMを与えられた場合、Hについて方程式6を解く(例えば、回帰係数のセットを解くことによって回帰を最適化する)。NNLSは、非負制約を施行することによって生物学的に合理的な溶液を生成することができるため、選択された。交換なしのサブサンプリングを組み込んで、結果の信頼性を向上させ、標準誤差および信頼区間を計算した。各遺伝子に対する回帰係数の有意性を推定するために、通常の線形最小二乗に従来使用される同一のアプローチをNNLSに適応した。ここでは、回帰係数の有意性は、まずそれを分析的に導出された標準誤差で除し、次いで得られた商を学生t分布を使用してp値に変換することによって決定される。各細胞型について、ベンジャミニ・ホックバーグ法を使用して複数の仮説試験のためにp値を補正した。所与の遺伝子iおよび細胞型jに関して調整されたp値(例えば、q値)が非常に有意であった(デフォルトではq値<10⁻⁵)場合、遺伝子iにおいてq値>0.25を有する他のすべての細胞型は、わずかに検出されるとみなされた(発現を0に設定)。交絡ノイズをさらに低減するために、遺伝子を、サブサンプリング回帰係数の自然対数を使用して計算したそれらの幾何変動係数(幾何c.v.)に基づいてフィルタリングした。幾何c.v.閾値を適応アプローチを使用して定義した。具体的には、各細胞型について、各幾何c.v.を昇順にランク付けし、次いで各c.v.を2次元座標系にマッピングして、すべてのc.v.値の1パーセンタイルまたは75パーセンタイルのいずれかの最大値として定義される、最大c.v.(「c v_{max}」)に達するまで最も低いc.v.(「c v₁」)が座標(1, c v₁)になり、第2の最も低いc.v.が(2, c v₂)になる。次いで、ユークリッド空間内の(0, c v_{max})から最も遠い点として定義される、c.v.曲線において、曲折点を特定した。この閾値に対応する幾何c.v.を、適応細胞型特異的ノイズ閾値として使用した(図3b～eおよび11a～h)。

【0221】

実施例7：高解像度細胞型特異的発現プロファイルの推定

高解像度細胞型特異的発現プロファイルの推定のためのアルゴリズムは、図12cに提示される対応するスキーマと平行な以下の主要なステップを使用して実行され得る。

【0222】

混合マトリックスMにおける各遺伝子iについて、このアプローチは、その発現ベクタ

一 M_i , $\cdot$ (以下、簡潔性のために M_i と示す) を昇順にソートすることによって開始し、ベクター M^*_i をもたらす (図 12 c、ステップ i)。バルク混合物中に検出可能なシグナルを有する任意の細胞型特異的 DEG は、この順序に影響を及ぼし得、最も顕著なシグナルは、分布の片側に歪み得る (例えば、図 13 b)。

【0223】

次に、F における混合試料 (F^* と示す) を、 M^*_i における試料の順序に一致するように再順序付けし、反復 t を定義し、これは w と等しく、方程式 6 を解くときに使用する混合試料の数を指定するパラメータであった。[w , $(1+c-w)$] における t の各整数値について、方程式 6 を 2 回: 1 回は、指数 $1 \sim t$ におよぶ M^*_i における混合物試料について、1 回は、指数 $(t+1) \sim c$ を有する M^*_i における混合物試料について求解した。したがって、 t の各値について、このプロセスは、代表的な細胞型特異的遺伝子発現係数の 2 つのベクター、それぞれ g_1 および g_2 の 2 つのベクターを生成する (簡潔性のために、ここでは指数 i が省略されることに留意されたい。図 12 c、ステップ i i)。

【0224】

G^t_i は、 g_1 の t 行、続いて g_2 の $c-t$ 行からなる $c \times k$ マトリックスである。 G^t_i は、2 自由度 (遺伝子あたり 2 つの可能な値) で細胞型特異的発現推定値を捕捉するため、 $G_i, \dots$ (以下、 G_i と示す) の粗略近似値である。 G^t_i および F^* 由来の再構成混合物と元のソートされた混合物 M^*_i との間の適合性を評価するために、残留二乗和を計算し、以下の方程式を最小化する t の値を決定した。

【数 1】

$$\min_{\forall t} \sum (\text{diag}(F^* \times G^t_i) - M^*_i)^2 \quad (9)$$

【0225】

方程式 9 を最小限に抑える t の値は、マトリックス $G^{t'}_i$ および基礎となる発現係数ベクトルの対、 g'_1 および g'_2 を定義し、これらは F を与えられた場合に M^*_i を最適に再構成する (図 12 c、ステップ i i i)。

【0226】

次に、各遺伝子/細胞型対を一連の統計的試験を受けて、遺伝子が (1) 有意に発現され、(2) g'_1 と g'_2 との間で有意に異なるかどうかを確認する (図 12 c、ステップ i v)。前者について、100 回の反復を用いたブートストラップを使用して、NNLS をすべての混合物試料に適用した。各遺伝子 i について、各細胞型の中央発現値は、ベクター g_{global} として保存される。各遺伝子/細胞型対の経験的 p 値は、係数が 0 に等しい回数を 100 で除したものとして計算される。このアプローチの十分に確立された利点にもかかわらず、不適切な試料サイズまたは不十分なブートストラップドロウに起因する不正確性の影響を受ける可能性がある。例えば、後者は、経験的 p 値に下限値を課し、それは他の例外的な統計的支持を有する遺伝子 (例えば、細胞型マーカー遺伝子) について著しく過小評価され得る。したがって、実施例 5 に記載される同一のアプローチを使用して、複数の仮説試験のために調整された分析的に導出された p 値を考慮した。後者が所与の遺伝子 i および細胞型 j に関して非常に有意である (デフォルトでは、 $q \text{ 値} < 10^{-5}$) 場合、遺伝子 i において $q \text{ 値} > 0.01$ を有するすべての他の細胞型は、重要でないとみなされる (q 値を 1 に設定) が、他のすべての q 値は変更されない。両方の試験の出力を組み合わせるために、フィッシャーの方法を使用して、それらの整合性に基づいて p 値を増加またはペナルティ化した。これらの試験は独立していないため、フィッシャーの方法の出力はメタスコアであり、 p_{global} として示す。 $p_{\text{global}} < \alpha$ (デフォルトでは $= 0.05$) を有する遺伝子/細胞型対は、著しく検出可能であると考えられる。

【0227】

有意に検出可能な発現を有する遺伝子/細胞型対について、次いで、 g'_1 と g'_2 と

の間の差次的発現がないヌル仮説を試験した。これは、以下の2つの異なる統計試験を使用して達成される。まず、ブートストラップを介して得られる g'_{1} および g'_{2} (それぞれ、 $s d_{1}$ および $s d_{2}$ として示す) の標準偏差に基づく分布がガウスであるという仮定の単純化を行う。次いで、 g'_{1} と g'_{2} との間の差を $s d_{1}$ および $s d_{2}$ の二乗和の平方根で除したものと、差次的発現のZスコアを計算することは些細なことである。結果として得られるZスコアは、各細胞型に対して1つの両側p値に変換され、ベクター p_{1} として記憶される。

【0228】

別に、順列分析を使用して差次的発現を試験した。ここで、 F^{*} (F^{rand} と示される) の行 (例えば、試料) をシャッフルし、方程式6は、 $(t' + 1) \sim c$ の指数を有する F^{rand} および混合物試料を使用して求解し、 g'_{2} (g^{r}_{2} と示される) のランダム化バージョンをもたらす。差分ベクトル $|g^{r}_{2} - g'_{1}|$ を保存し、プロセスを50回繰り返す。次いで、差分ベクトルのヌル分布を試験統計と比較し、 $|g'_{2} - g'_{1}|$ および経験的p値を、この試験統計を超えるランダム化されたデルタの分数として計算する。得られたp値は、ベクトル p_{2} として記憶される。各試験は、非常に異なる様式で差次的発現を評価するため、 p_{1} および p_{2} は、フィッシャーの方法を使用してメタスコア p^{deg} に組み合わされる。このメタ分析アプローチの有用性にもかかわらず、入力p値は形式的に独立していないため、フィッシャーの方法の出力はメタスコアと呼ばれ、p値ではない。

【0229】

次に、グローバル係数 g^{global} および差次的細胞型発現ベクター g'_{1} および g'_{2} は、上記の統計フィルタリングステップを反映するように更新される (図12c、ステップv-ii)。まず、 $F^{*}_{i} \leftarrow F^{*}$ を各遺伝子 i に対して定義し、 F^{*}_{i} における画分存在量ベクトルを、遺伝子 i における g'_{1} と g'_{2} との間の重要でない発現を有する細胞型に対してゼロに設定する。次いで、 F^{*}_{i} および g^{global} を計算するための上記の同じ手順を使用して、各遺伝子について g^{global} を再計算する。 g'_{1} および g'_{2} を更新するには、 F^{*} ではなく F^{*}_{i} を使用し、方程式9を使用して新しい最適な切断点 t' および対応するマトリックス $G^{t'}_{i}$ を識別する。

【0230】

次いで、メタスコアベクトル p^{deg} に基づいて、 g'_{1} 、 g'_{2} をさらに絞り込むために一連のステップを実行する。具体的には、所与の遺伝子 i および細胞型 j に関して、 g'_{1} 、 g'_{2} における対発現係数間の差は、各細胞型の差次的発現における推定統計的信頼性を反映する方法で低減される (図12c、ステップv-ii)。これを達成するために、まず、各細胞サブセット j の分数と発現ベクター M_{j} との間の相関 (および対応する $-\log_{10} p$ 値、 p^{cor} と示される) を、シグネチャマトリックス内の残りの細胞サブセットを回帰させた後に計算する。次いで、 l を、 p^{deg} (上記) の $-\log_{10} p$ 値に等しく設定する。あるいは、細胞サブセットが0.05のp値閾値で有意でない場合、 l を p^{cor} に等しく設定する。 l が $-\log_{10}$ 空間内にあるため、それをその最大値 l_{max} で除し (閾値 $-\log_{10} (0.05)$ で有意である場合)、最大値は、 g'_{1} での発現が g'_{2} よりも高い細胞サブセットについて別々に計算され、逆も同様である。これにより、所与の細胞サブセットに関する発現係数間の倍率変化が、差次的発現における推定統計的信頼性を捕捉する l^{*} に比例して低減するように、 g'_{1} および g'_{2} を調節するために使用される新しいベクトル l^{*} が生じる。調整後、(デフォルトでは) 1.1倍未満の倍率変化を有する g'_{1} および g'_{2} の対発現係数は、わずかに異なると考えられ、それらの対応するグローバル係数 (g^{global}) に設定される。ヒューリスティックであるが、このアプローチは、最も有意な遺伝子/細胞型対に対するデルタを同時に保持しながら、推定された発現における誤りの差を大幅に低減することができる。 g'_{1} および g'_{2} が調整されると、 $G^{t'}_{i}$ は方程式9を使用して再推定される。

【0231】

これまでの改善にもかかわらず、 $G^{t'}_{i}$ は、依然として遺伝子/細胞型対あたり最大2つの異なる値に限定され得る。複数の自由度を有する平滑発現係数を推定するために、

方程式 6 を、 M^*_i におけるすべての可能なウィンドウ位置について、長さ w の摺動ウィンドウ内で順次求解し、マトリックス Q_i を得る (図 12 c、ステップ $v-i$)。 Q_i は、 G_i より $w-1$ 行が少ないため、 Q_i の 2 つのバージョンが作成され、 1 つは Q_i の上部に追加された行 1 の $w-1$ コピー (Q'_i と示す) を有し、 1 つは Q_i の下部に追加された行 c の $w-1$ コピー (Q''_i と示す) を有する。 次いで、これら 2 つのマトリックスを平均化し、 $Q_i \leftarrow (Q'_i + Q''_i) / 2$ を設定する。 この手順は、自由度 c を有する平滑発現ベクトルをもたらすが、各ウィンドウ内の混合物試料が、最小限の細胞型特異的発現変動で十分に類似しているという仮定に依存し得る。 言い換えれば、所与の細胞型に関する試料固有の発現の違いは、ウィンドウ間で捕捉されているにもかかわらず、所与のウィンドウ内で平均化され得る。 この問題は、回帰フレームワーク内の数値の不安定性のリスクにもかかわらず、 w のサイズを減らすことによって部分的に対処することができる。

【0232】

加えて、 M^*_i におけるバルク発現がソートされているため、低発現を有する領域をカバーするウィンドウは、不適切な情報コンテンツによって不正確な推定をもたらし得る。 同じ問題は、サイズが不十分な窓から生じる可能性がある。 これらの潜在的な問題に対処するのに役立つために、 Q_i における発現係数は、 $G^{t'}_i$ 内の g'_{1} および g'_{2} の最終値に基づいて調整される (図 12 c、ステップ $v-ii$)。 具体的には、各細胞型 j について、 $Q_{i,j}$ の発現係数は、通常の最小二乗回帰を使用して、 $G^{t'}_{i,j}$ に個々に適合させる。 これにより、 g'_{1} および g'_{2} がアンカーポイントとして機能すること、および順序付け混合物にわたって細胞型特異的発現の差 (例えば、高〜低、または低〜高) の方向性が維持されるよう保証することが可能になる。 他のアプローチを使用してもよいが、最大 2 つの異なる値を有する剛性ベクトルを適合する際にある程度の柔軟性を可能にするため、 RLR が用いられる。 得られたマトリックスは、 Q'_i と示される。

【0233】

最後に、各遺伝子について、 Q'_i における混合物試料の元の順序を復元し、これを G_i として保存する (図 12 c、ステップ $v-iv$)。 全ての遺伝子にわたって G_i を 3D マトリックス G に凝集した後、各細胞型特異的発現マトリックスをディスクに書き込む。

【0234】

実施例 8：デコンボリューションの一括修正

10 × Chromium を使用した 12 個の全血 RNA-Seq 試料の細胞型列挙を行った。 図 23 d に示されるように、結果は、S モード補正を使用し、細胞レベルの性能の著しい改善を示す。 加えて、図 23 f に示されるように、バッチ効果から生じるプラットフォーム特異的な技術的変動は、3' キットおよび 5' キットを使用して 10 × Genomics によって PBMC の 2 つの公的に入手可能な単一細胞トランスクリプトームデータセットをプロファイリングすることによって生成されたデータセット間で観察された (図 23 e)。 図 23 g の概略図に示されるように、S モード一括修正をこれらのデータセットに適用して、異なる 10 × Chromium キットに由来する参照プロファイル間の技術的変動を最小限に抑えた。 3' キットからのシグネチャーマトリックスを、5' キットに由来する再構築血液試料に適用した。 S モード一括修正後、調整された 3' 細胞型参照プロファイルは、ヒートマップ (図 23 h) およびピアソン相関 (図 23 i) によって実証されるように、5' キットからのそれらの対応する参照プロファイルとより密接に整列される。 図 23 j に示されるように、5' PBMC 単一細胞データセットをマイクロアレイによってプロファイリングされた選別された白血球サブセットで置き換えたとき、図 23 g と類似した結果が観察された。

【0235】

実施例 9：腫瘍遺伝子型決定

バルク RNA-Seq を、以下のようにマイクロアレイを使用して行った。 バルク FL 標本およびソートされた B 細胞から全 RNA を抽出し、収率および品質について評価した。 直線増幅後の 100 ng の全 RNA から cRNA を調製し (3' IVT 発現、Affymetrix)、次いで製造業者のプロトコルに従って HG U133 Plus 2.0 マ

イクロアレイ (Affymetrix) にハイブリダイズした。すべてのCELファイルを、FLリンパ節からFACSソートされたCD4およびCD8 TILを含む公的に入手可能なAffymetrixデータセット (GSE27928) とプールした。次いで、結果として得られたデータセットを、Bioconductorの「affy」パッケージを使用してRMA正規化し、カスタムチップ定義ファイル (Brainarrayバージョン21.0、<http://brainarray.mbni.med.umich.edu/Brainarray/>) を使用してNCBI Entrez遺伝子識別子にマッピングし、HUGO遺伝子シンボルに変換した。ソートされた細胞サブセットの複製を組み合わせて、発現値の幾何平均を使用して、グラウンドトゥールズ参照プロファイル (例えば、図3b) を作成した。

10

【0236】

Deep配列決定によるCancerパーソナライズプロファイリング (CAPP-Seq) を、専用リンパ腫標的配列決定パネルおよび試料処理プロトコルを使用して、このコホートからの12個の冷凍保存されたFL腫瘍生検に適用した (例えば、参照によりその全体が本明細書に組み込まれる、Newman et al., An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage, Nature Medicine 20, 548 (2014) に記載されている)。異型呼び出しは、生殖細胞系対照として用いられた一致する末梢血液白血球で行った。Illumina HiSeq 2500機器上で2×100bpの読み取り値を使用して、高スループット配列決定を行った。CREBBPホットスポット突然変異は、リジンアセチルトランスフェラーゼ (KAT) ドメイン内で生じ得るため、KATドメイン内の非サイレント異型 (単一ヌクレオチド異型および挿入/欠失) のみを「変異体」とみなした。試料サイズを増加させるために、FLコホートからの12名の以前に遺伝子型決定された患者のCREBBP突然変異状態も分析した。

20

【0237】

実施例10：単一セル由来参照プロファイルのパラメータスイープ

デコンボリューションパフォーマンスに対するキーシグネチャーマトリックスパラメータの影響を系統的に評価するために、パラメータスイープを行い、以下を評価した。(1) 表現型あたりの細胞数 (3、5、10、20、および40)、(2) 細胞型あたりのシグネチャー遺伝子数 (50~150、150~300、および300~500の範囲)、ならびに(3) ドナー試料数 (1、2、または3) (図10a)。我々は、トレーニング/検証フレームワーク内で、黒色腫およびHNSCC腫瘍について別個の単一細胞由来のシグネチャーマトリックスを作成した (図10a)。結果の堅牢さを確実にするために、各パラメータの組み合わせをランダムサンプリングによって3回試験した (図10a)。加えて、分析は、悪性細胞 (HNSCC、黒色腫)、CD8およびCD4 T細胞 (HNSCC、黒色腫)、B細胞 (黒色腫)、および線維芽細胞 (HNSCC) のすべての評価されたパラメータ設定にわたるトレーニングデータセットおよび検証データセットの両方における十分な表現を確実にするためにそれぞれ選択された以下の細胞表現型に限定された。ドナーを、細胞表現型あたり少なくとも20個の細胞を有する試料のセットからサンプリングした (HNSCC腫瘍の場合、ドナーID HN16__T、HN17__T、HN25__LN、HN25__T、HN18__T、黒色腫瘍の場合、ドナーID79、80、88、89)。4つのHNSCC試料 (HN10__T、HN13__T、HN24__LN、およびHN7__T) は、各々が評価された表現型にわたって合計50個未満の細胞を含有していたため除外された。1名のドナーのみをサンプリングする場合、細胞が不十分 (50個未満) であるために黒色腫ドナー88も除外された。総合的に、データセットごとに126個のシグネチャーマトリックスを試験した。性能を、シグネチャーマトリックス構築から保持された再構築腫瘍試料からのグラウンドトゥールズ細胞比率に対する中央細胞レベル相関として定量化した (図10b)。

30

40

【0238】

50

加えて、3つのscRNA-Seq法(SmartSeq2、3'ーおよび5'ーバイアスされた10×Chromium発現プロファイリングキット)および3つの組織型(再構築された黒色腫腫瘍、再構築されたHNSCC腫瘍、および末梢血)にわたって、単一細胞誘導型デコンボリューションを評価した。すべての場合において、推定された比率を、フローサイトリー/コールターカウンター(血液試料)またはscRNA-Seq(黒色腫、HNSCC)によって定量化されたグラウンドトゥールズ頻度と比較した。健康なドナーに由来する22個の造血サブセットを区別するためのマイクロアレイ由来のシグネチャーマトリックスLM22も試験した。シグネチャーマトリックスおよび組織試料の両方に存在する細胞型のみを考慮し、比較前にすべての細胞画分を1個に正規化した(図2e)。HNSCCおよび黒色腫腫瘍におけるCD8/CD4 T細胞を除き、以前に注釈された細胞表現型のみを評価した。これらは、本明細書の他の箇所に記載されるように定義され、黒色腫シグネチャーマトリックスを使用してHNSCC腫瘍において評価され、逆も同様である。すべての場合、10×由来のシグネチャーマトリックスに対するSモード一括修正および残りのシグネチャーマトリックスに対するBモード一括修正を使用して、デコンボリューションを一括修正で適用した。

【0239】

実施例11：シグネチャーマトリックス

細胞画分の推定に続くが、発現推定の前に、LM22デコンボリューション結果を、B欄に示される対応する図を用いて、図25に示される主要な表現型群に合計することによって崩壊させた。注目すべきことに、LM22-merge10において、両方のサブセットにおける高CD8発現レベルおよびFACSソートされた細胞集団に由来する公的に入手可能なCD8 T細胞グラウンドトゥールズプロファイルにおけるgd T細胞枯渇の不在のために、gd T細胞をCD8 T細胞とマージした(GSE28491、GSE27928)。

【0240】

本発明の好ましい実施形態が本明細書に示され説明される一方で、そのような実施形態が例としてのみ提供されることは当業者には明らかである。本発明は、本明細書内に提供される特定の実施例によって制限されることは意図されない。本発明は、前述の明細書を参照して説明されているが、本明細書の実施形態の説明および例示は、限定的な意味で解釈されることを意図するものではない。本発明から逸脱することなく、多数の変形、変化、および置換は、ここで当業者にとって思い当たるであろう。さらに、本発明のすべての態様は、様々な条件および変数に依存する本明細書に記載される具体的な描写、構成、または相対的な比率に限定されないことを理解されたい。本明細書に記載された本発明の実施形態の様々な代替が、本発明の実施において使用され得ることを理解するべきである。したがって、本発明はまた、任意のそのような代替、修正、変形、または等価物を包含することも企図される。以下の特許請求の範囲は、本発明の範囲を定義し、これらの特許請求の範囲内の方法および構造、ならびにそれらの等価物をそれによって包含することが意図される。

10

20

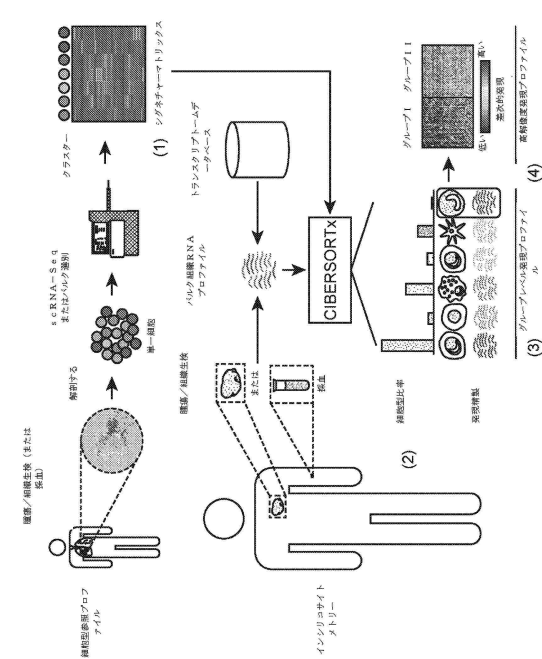
30

40

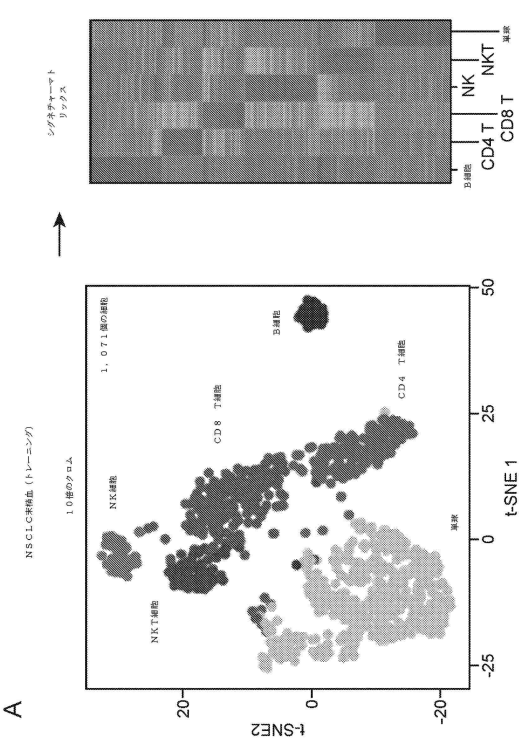
50

【図面】

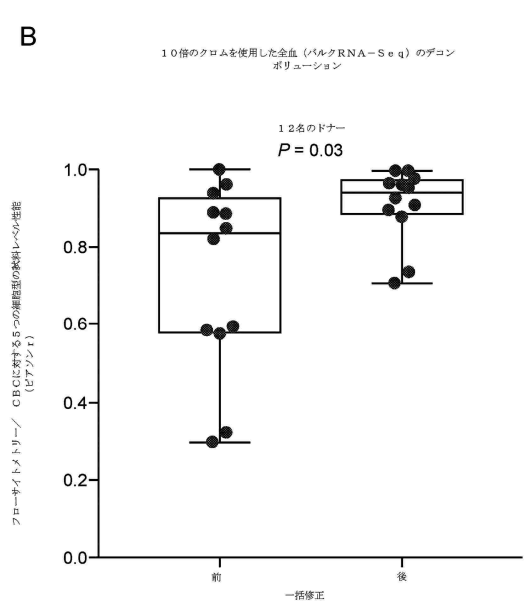
【図 1】



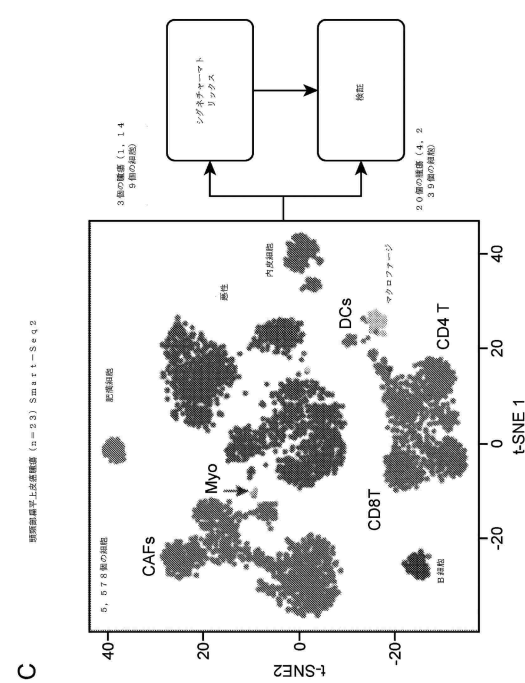
【図 2 - 1】



【図 2 - 2】



【図 2 - 3】



10

20

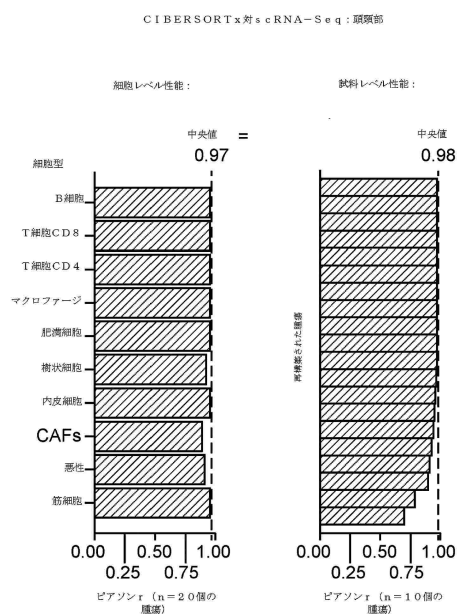
30

40

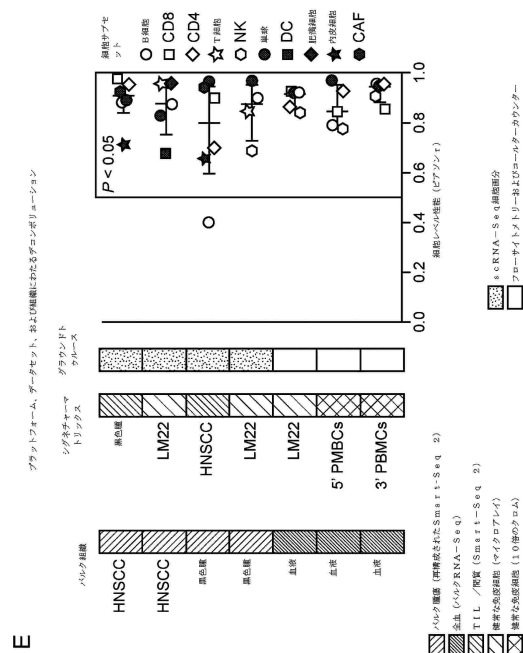
50

【圖 2-4】

D



【图 2-5】

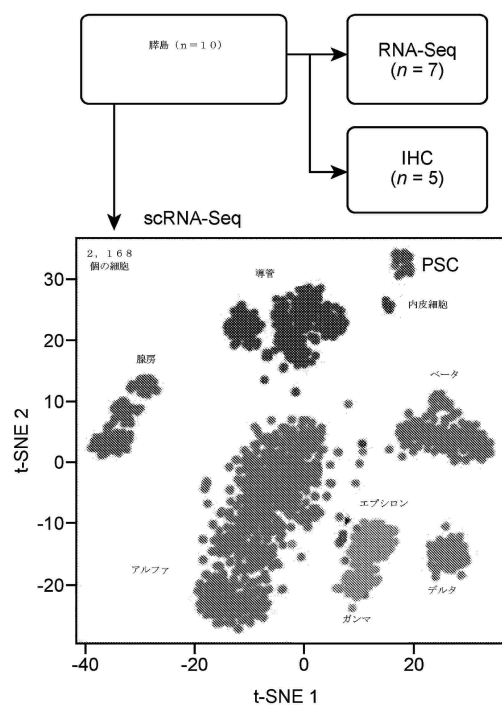


10

20

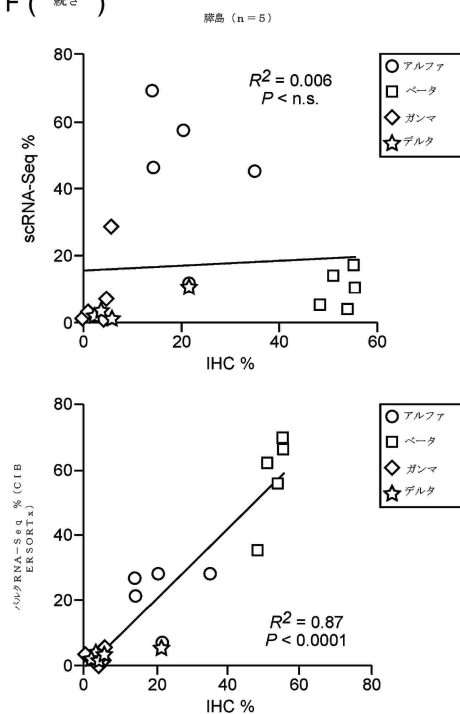
【圖 2-6】

F



【圖 2-7】

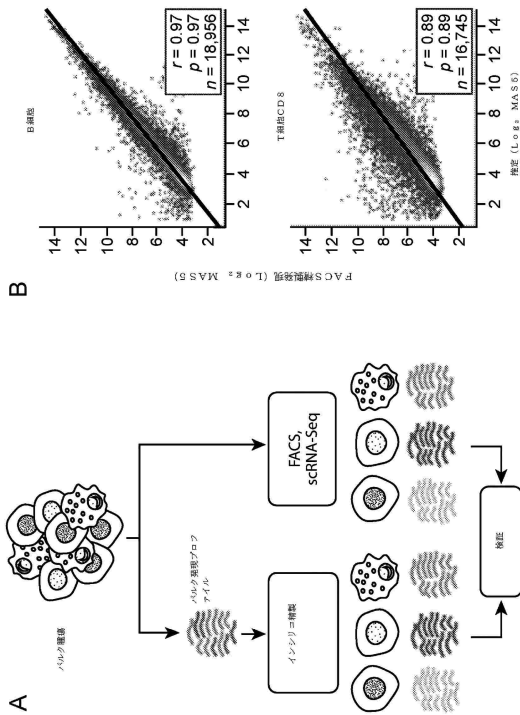
F (続き)



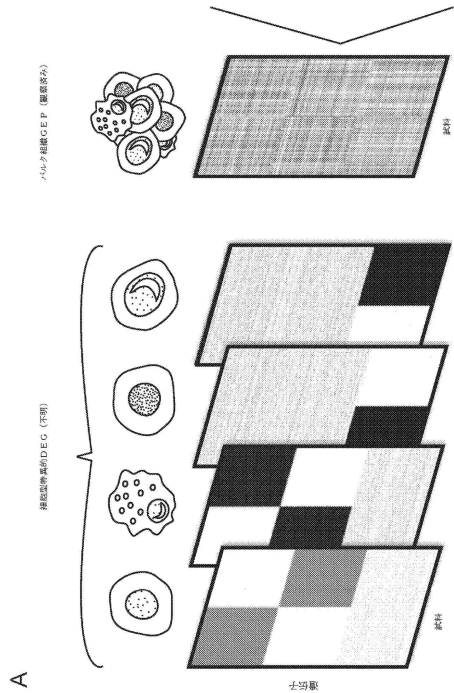
30

40

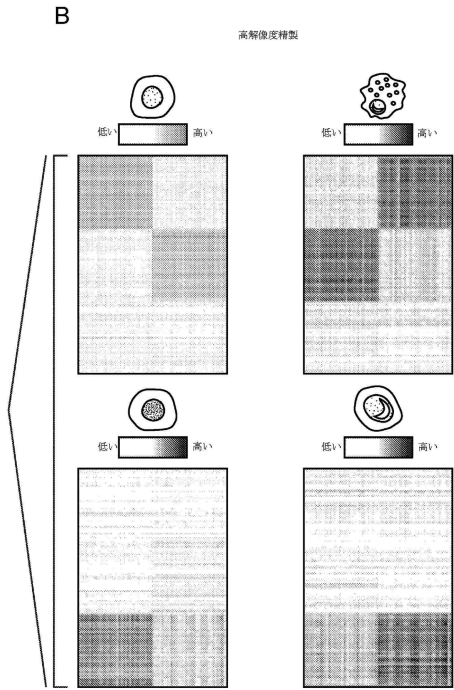
【図 3 - 1】



【図 4 - 1】



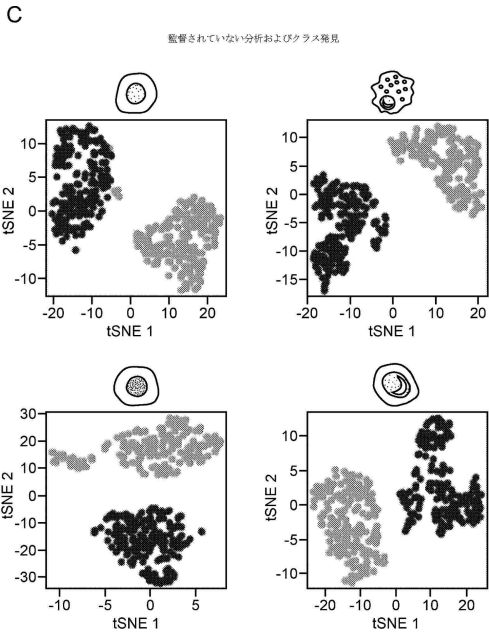
【図 4 - 2】



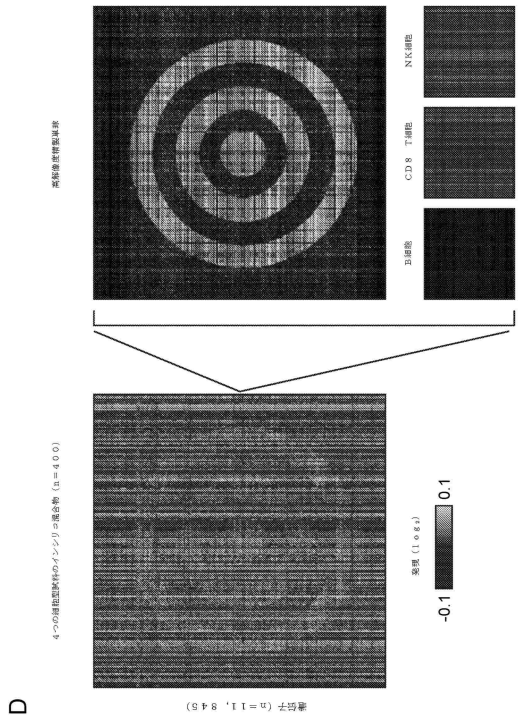
10

20

【図 4 - 3】



【図 4 - 4】

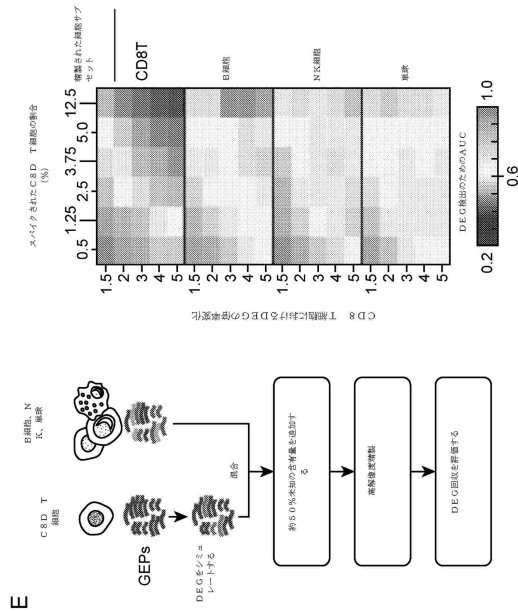


30

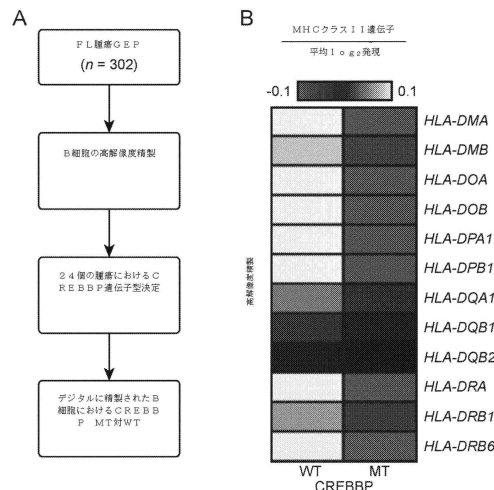
40

50

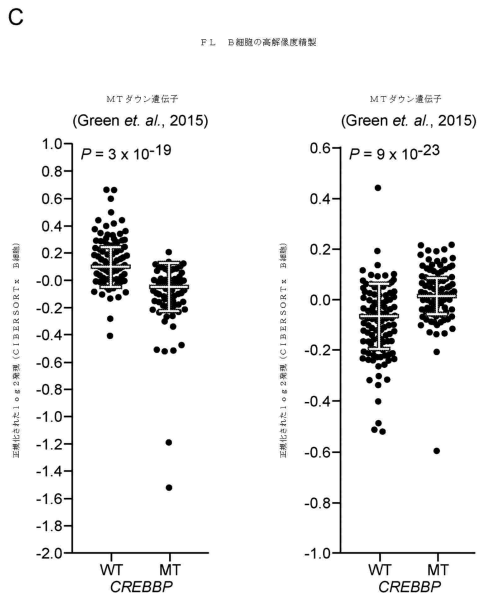
【図 4 - 5】



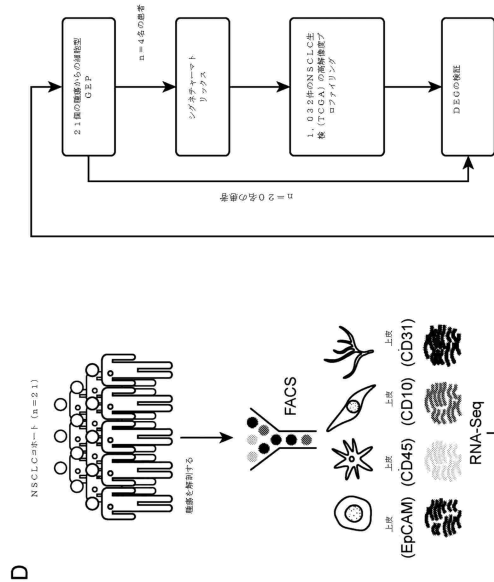
【図 5 - 1】



【図 5 - 2】



【図 5 - 3】



10

20

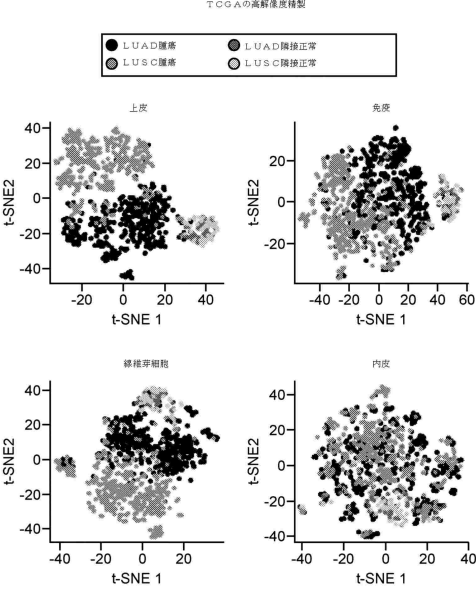
30

40

50

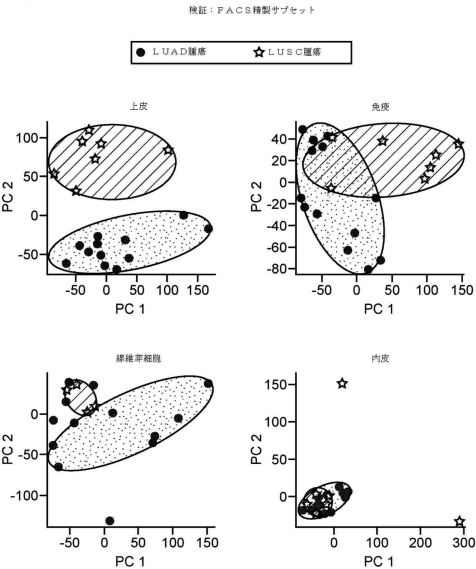
【図 5 - 4】

E



【図 5 - 5】

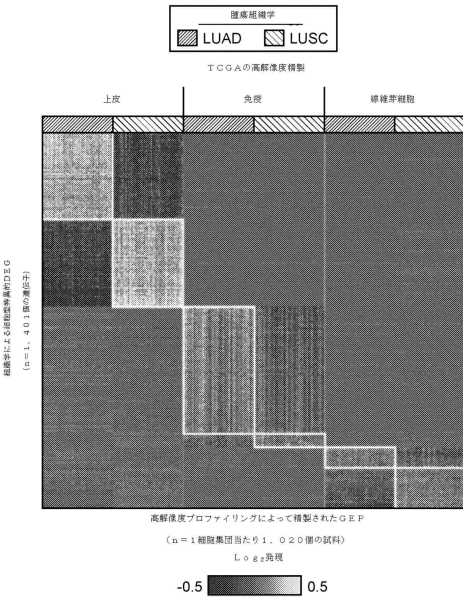
E (続き)



10

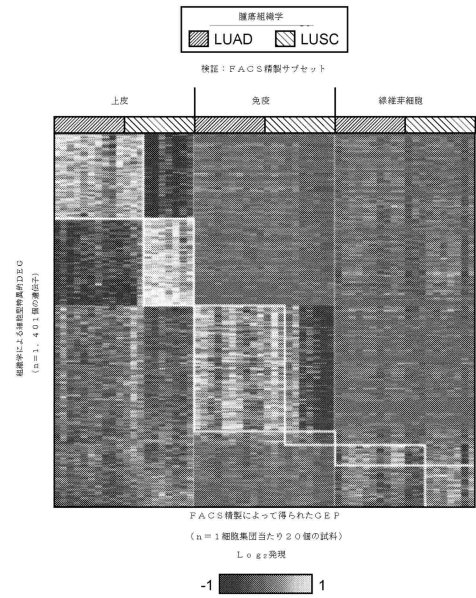
【図 5 - 6】

F



【図 5 - 7】

F (続き)



20

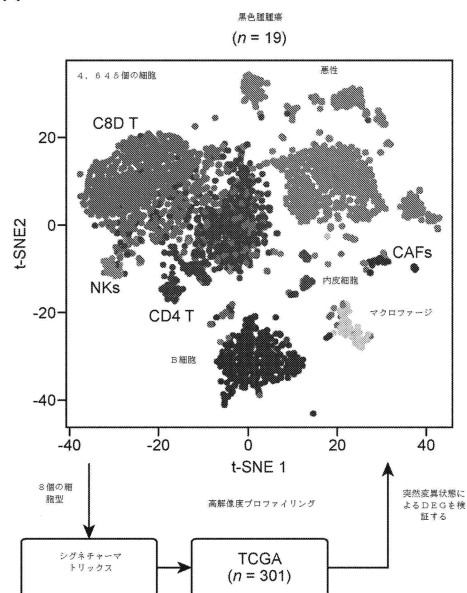
30

40

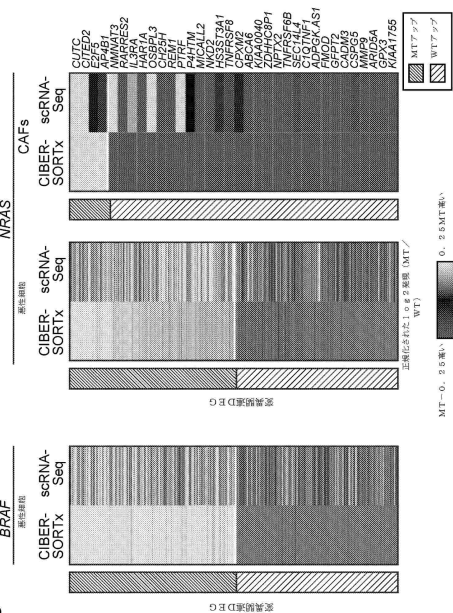
50

【図 6 - 1】

A

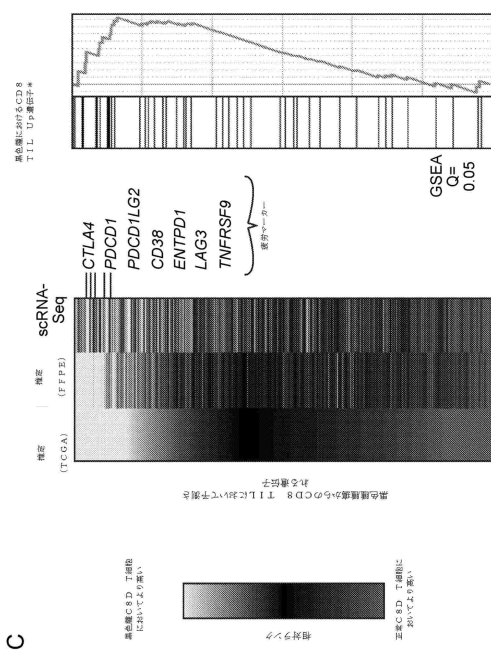


【圖 6-2】



10

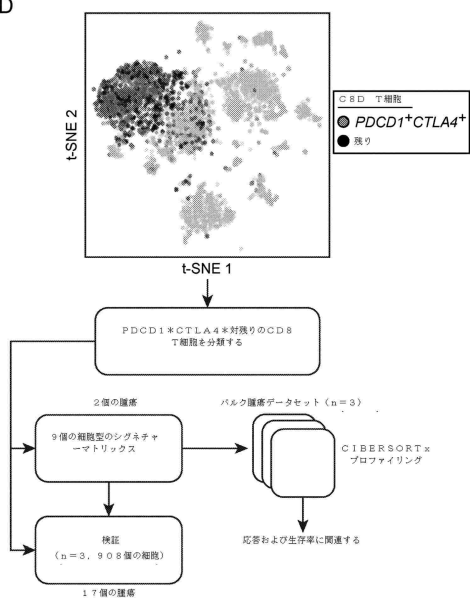
【図 6-3】



U

【図 6-4】

D



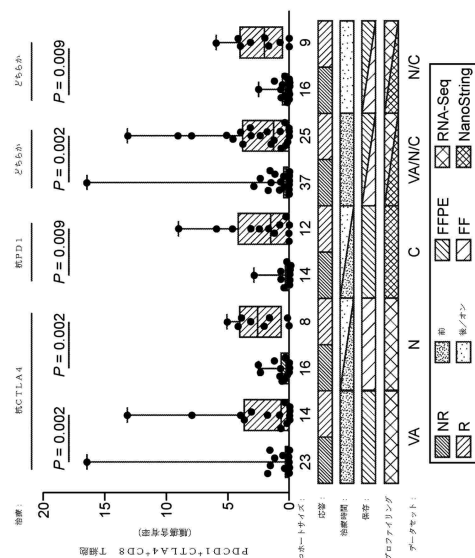
20

30

40

50

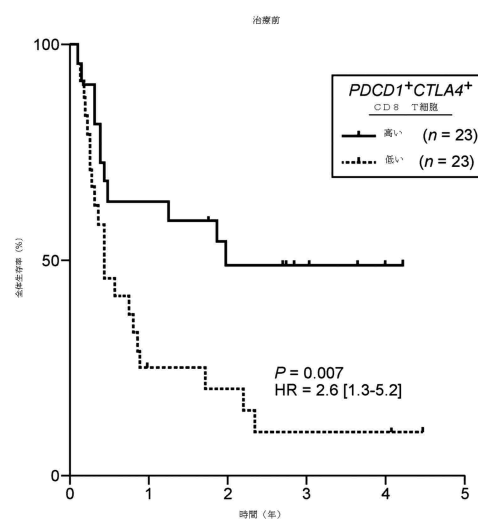
【圖 6-5】



W

【図 6 - 6】

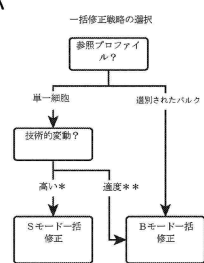
F



10

【図 7 - 1】

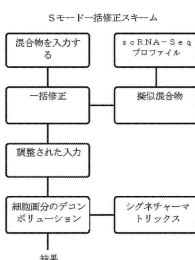
A



*例えば、3' / 5' 偏向転写物、UMI
**例えば、SMART-Seq

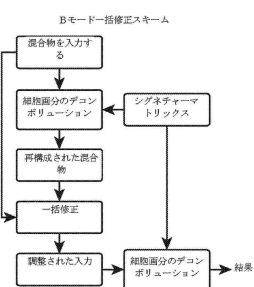
**例えば、SMART-Seq

B



結果

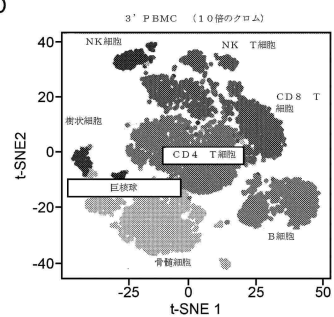
C



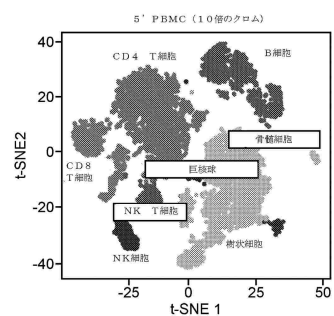
● 結果

【図 7 - 2】

D



20

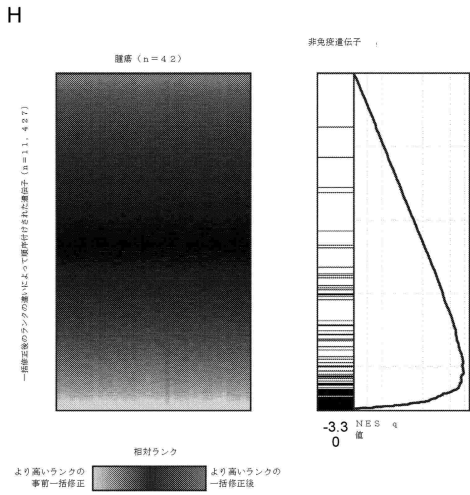


30

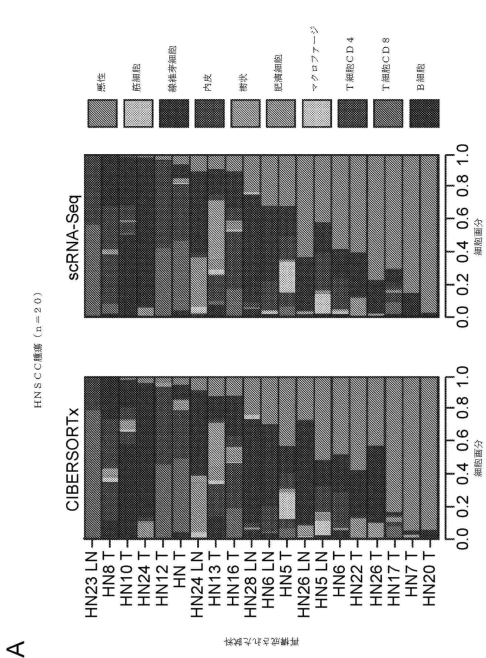
40

50

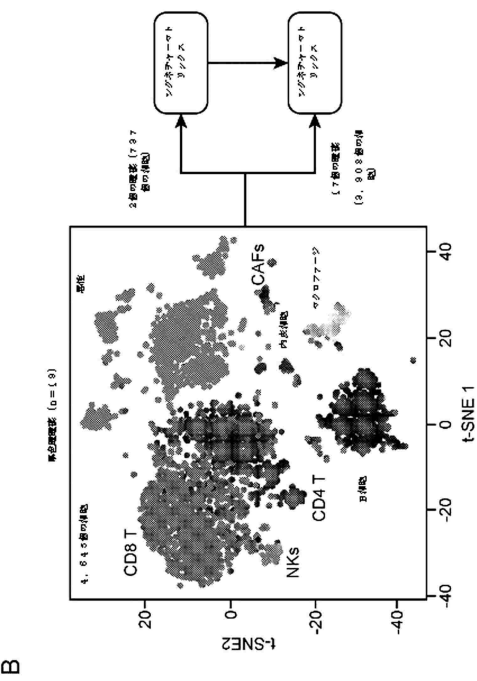
【図 7 - 7】



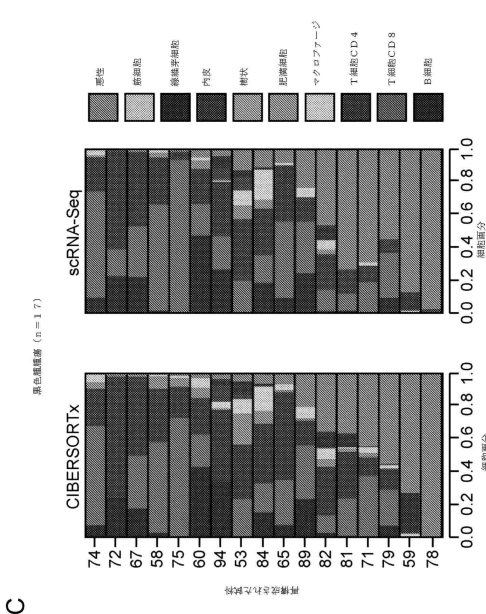
【図 8 - 1】



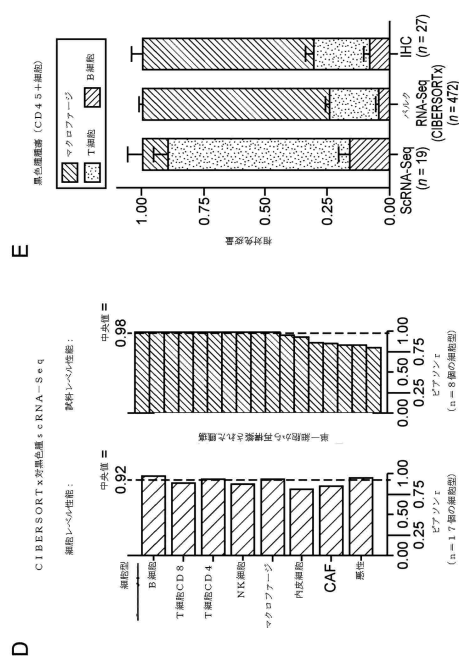
【図 8 - 2】



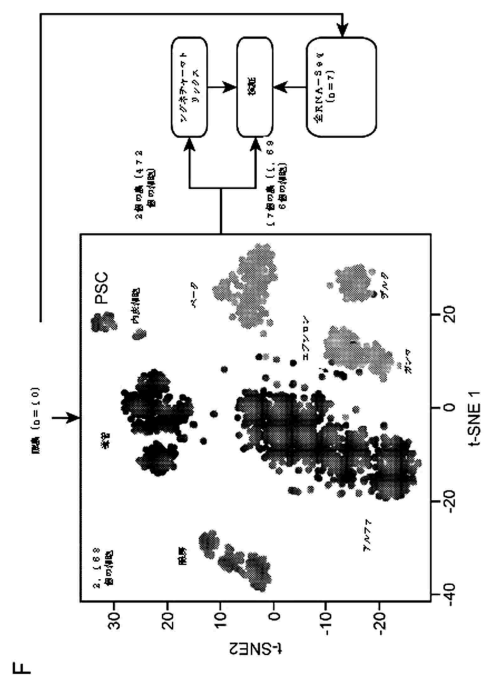
【図 8 - 3】



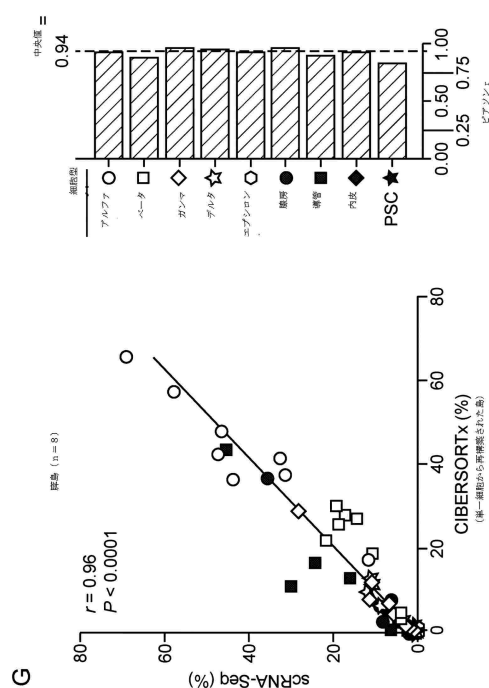
【圖 8-4】



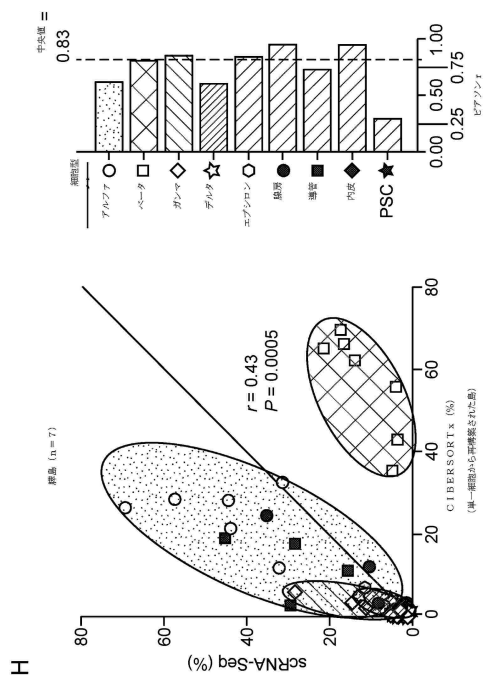
【図 8-5】



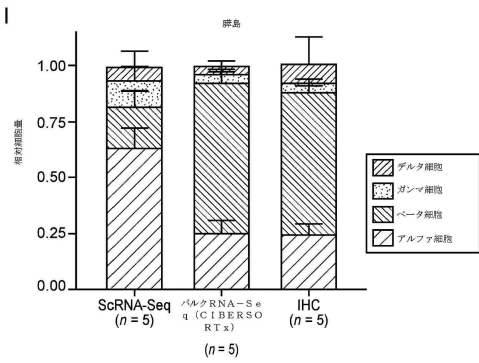
【図 8-6】



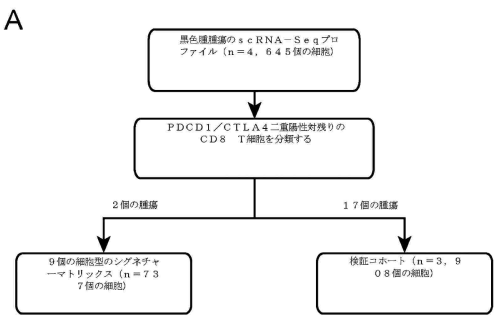
【図 8-7】



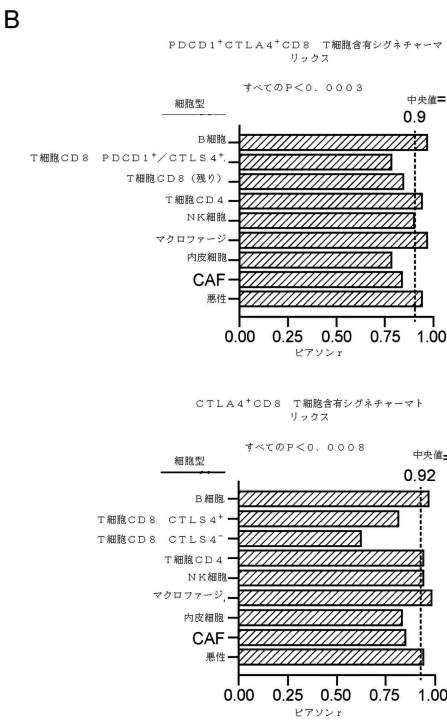
【図 8 - 8】



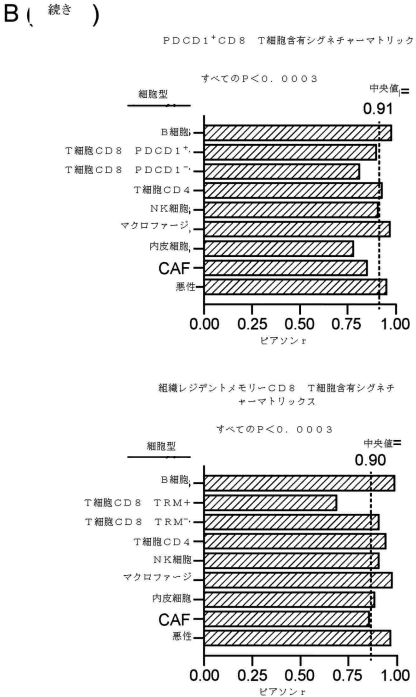
【図 9 - 1】



【図 9 - 2】



【図 9 - 3】



10

20

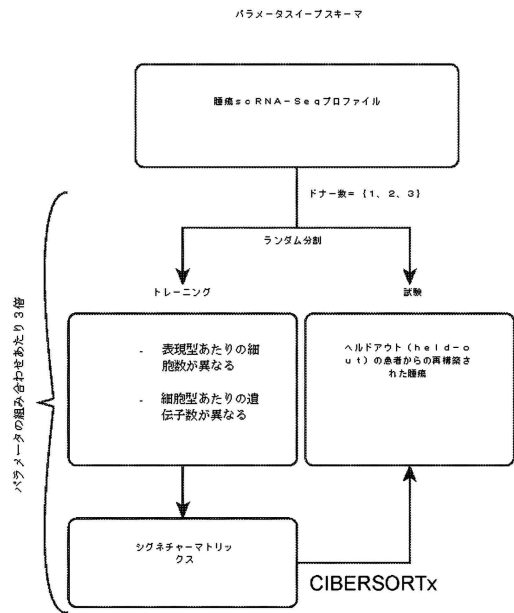
30

40

50

【図 10-1】

A



【図 10-2】

ヘルドアウト (held-out) の患者における性能 (中央値ピアソンr)

黒色腫

ドナー	細胞	遺伝子範囲			
		50-150	150-300	300-500	
1	3	0.818	0.829	0.823	
	5	0.815	0.853	0.846	
	10	0.804	0.809	0.811	
	20	0.869	0.887	0.890	
2	3	0.780	0.747	0.746	
	5	0.887	0.878	0.875	
	10	0.876	0.899	0.900	
	20	0.876	0.890	0.893	
3	3	0.814	0.779	0.797	
	5	0.853	0.867	0.859	
	10	0.898	0.903	0.897	
	20	0.867	0.903	0.894	
	40	0.875	0.878	0.889	

HNCC

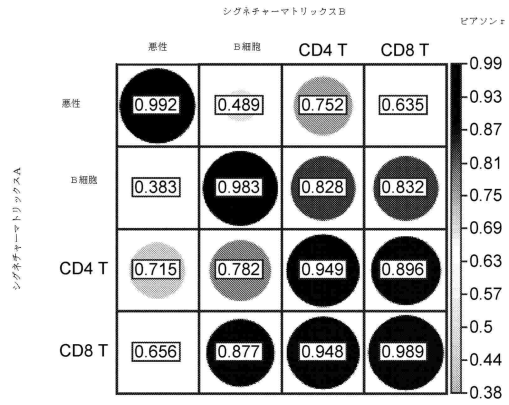
ドナー	細胞	遺伝子範囲			
		50-150	150-300	300-500	
1	3	0.885	0.912	0.916	
	5	0.857	0.909	0.925	
	10	0.917	0.942	0.942	
	20	0.906	0.910	0.925	
2	3	0.891	0.920	0.921	
	5	0.868	0.930	0.925	
	10	0.874	0.923	0.929	
	20	0.904	0.934	0.941	
3	3	0.897	0.923	0.932	
	5	0.894	0.922	0.934	
	10	0.919	0.925	0.931	
	20	0.917	0.918	0.921	
	40	0.925	0.952	0.953	

異なるドナー、表現型あたり300~500細胞の遺伝子
ヘルドアウト (held-out) の患者、細胞型あたり300~500細胞の遺伝子
異なるドナー、細胞型あたり400細胞の遺伝子

【図 10-3】

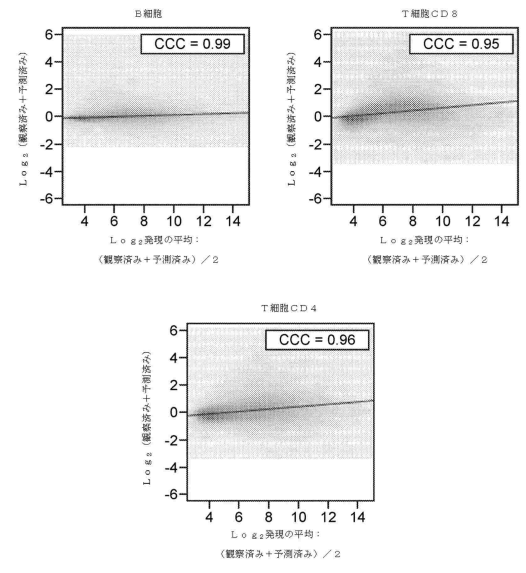
C

黒色腫腫瘍から推定されたグループレベルGEP間の一致 (TCGA RNA-Seq)



【図 11-1】

A



10

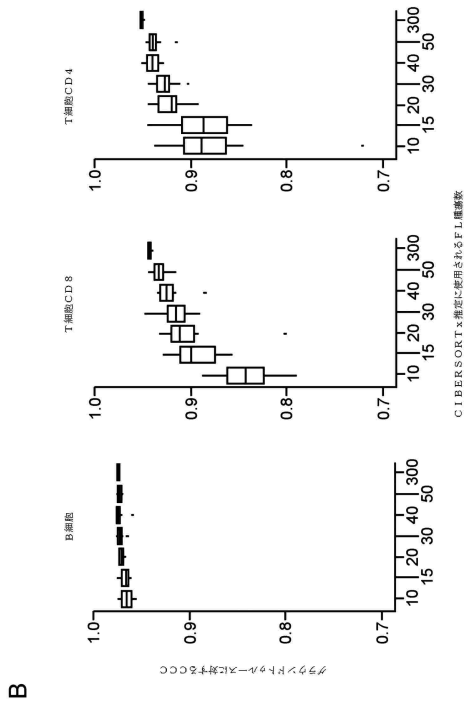
20

30

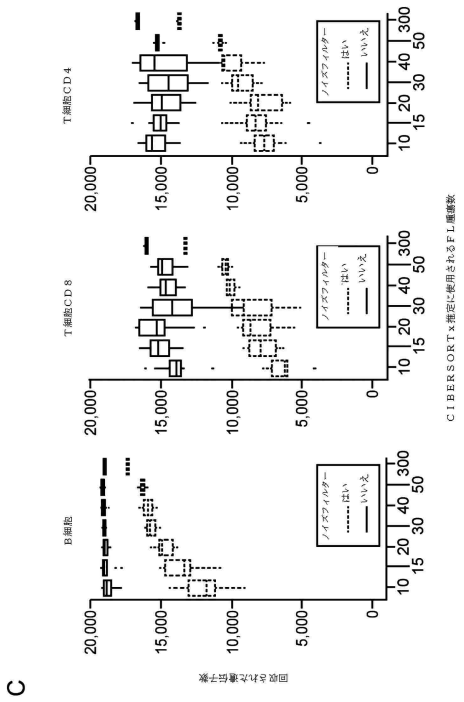
40

50

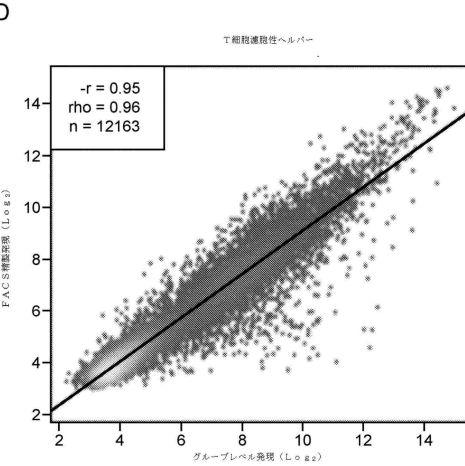
【図 1 1 - 2】



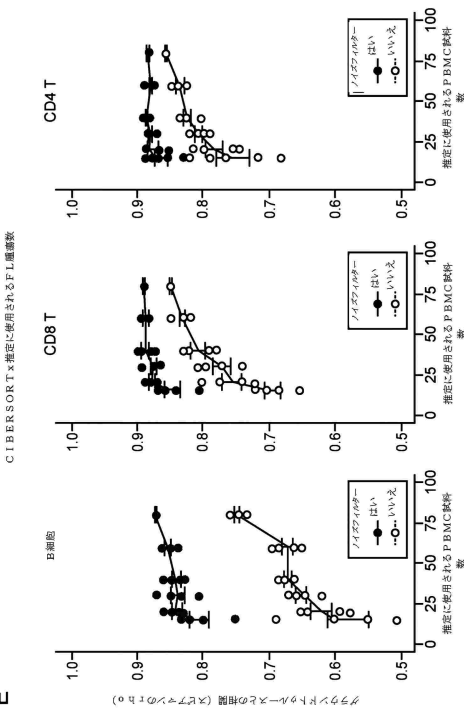
【図 1 1 - 3】



【図 1 1 - 4】



【図 1 1 - 5】



10

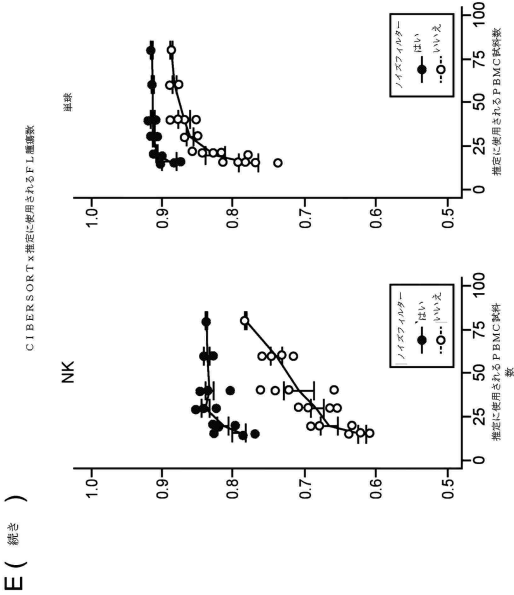
20

30

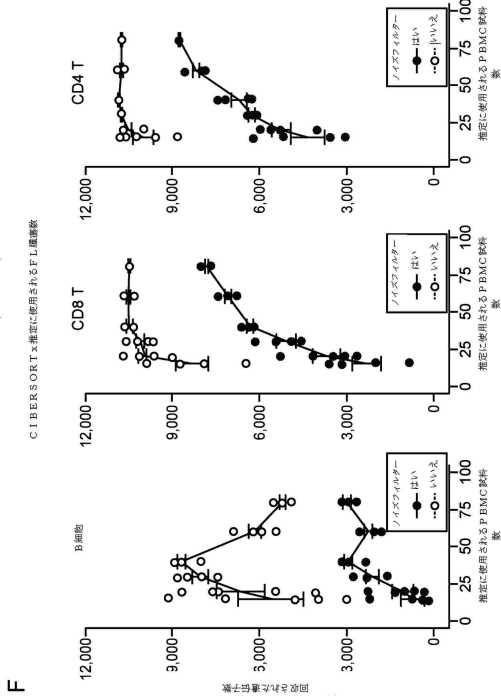
40

50

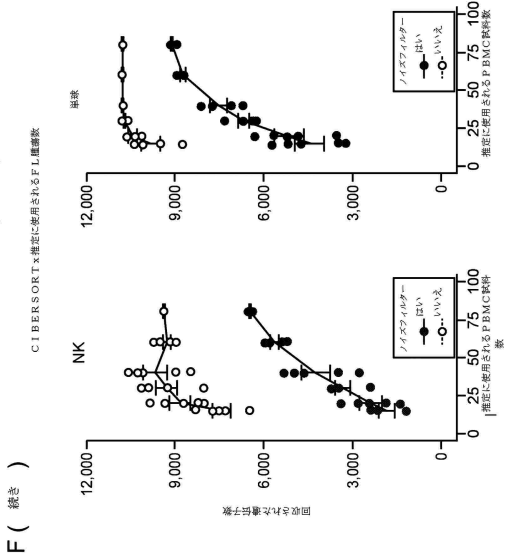
【図 1 1 - 6】



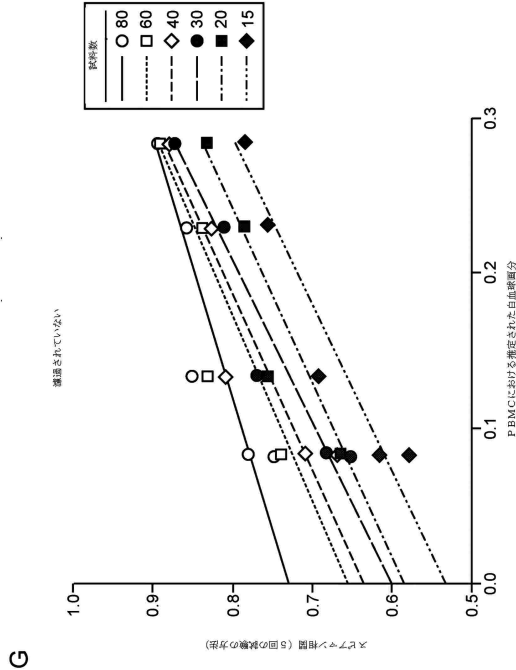
【図 1 1 - 7】



【図 1 1 - 8】

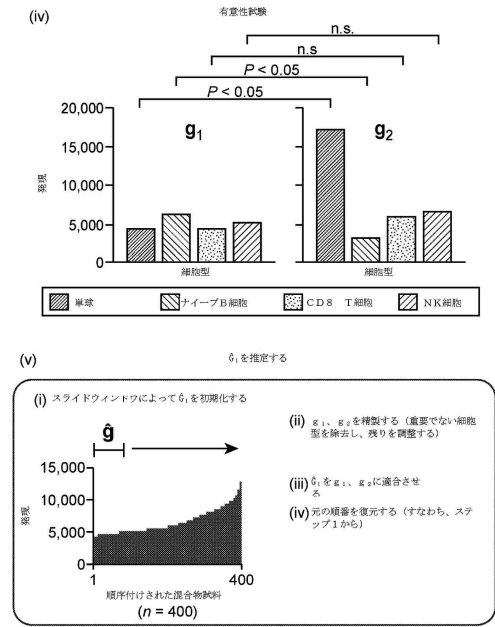


【図 1 1 - 9】

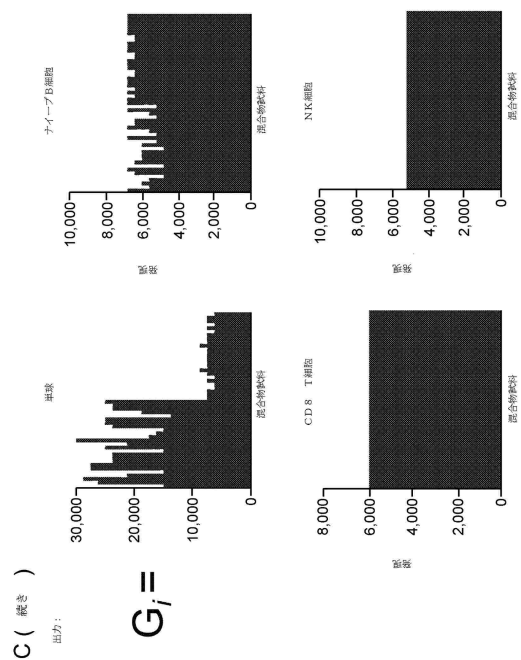


【図 1 2 - 6】

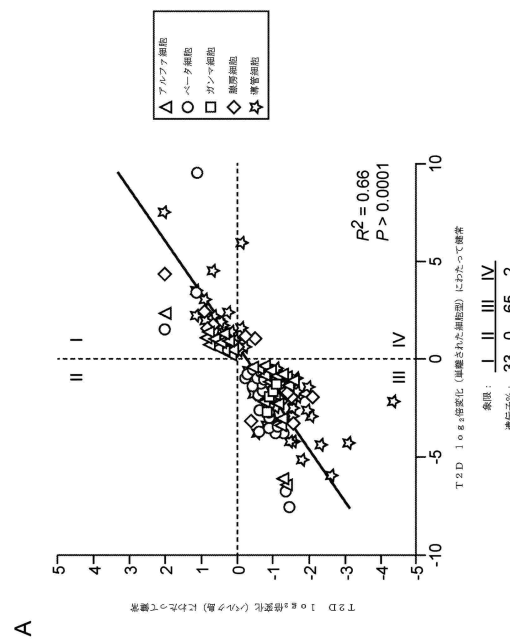
C (続き)



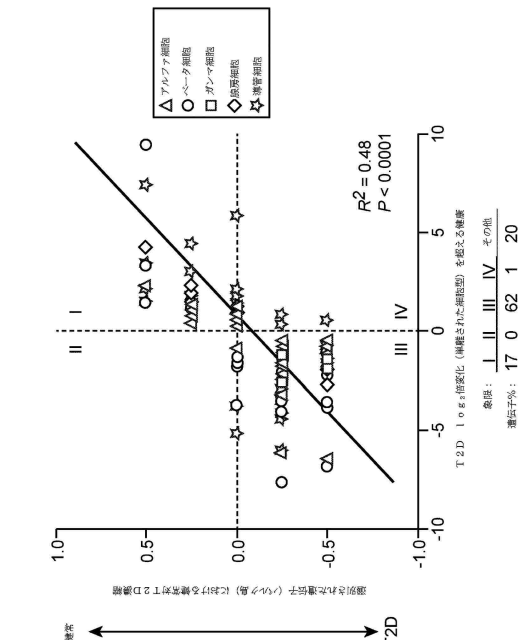
【図 1 2 - 7】



【図 1 3 - 1】



【図 1 3 - 2】



10

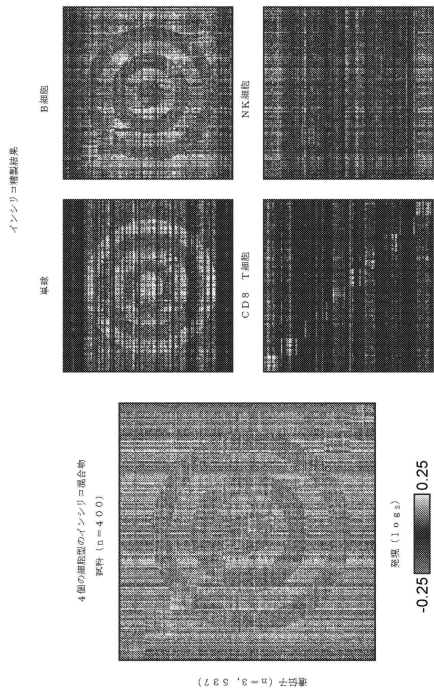
20

30

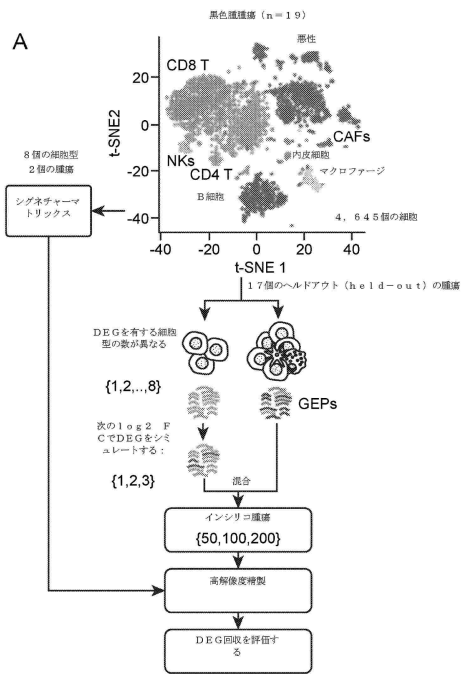
40

50

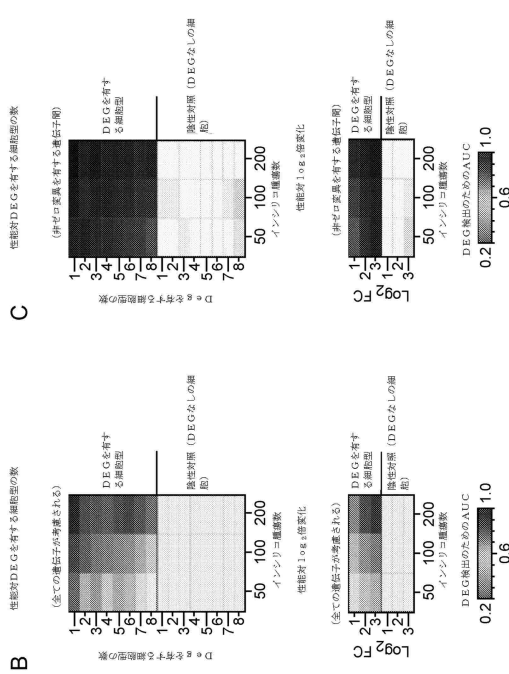
【図 1 4】



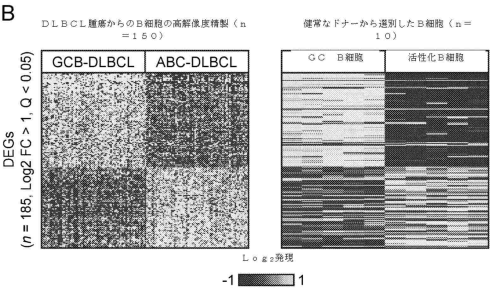
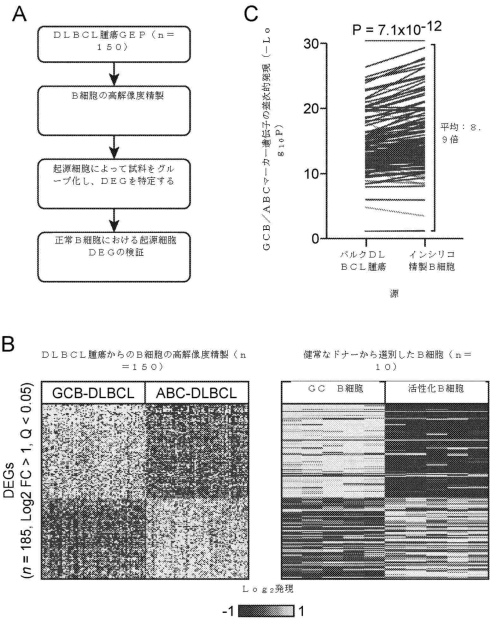
【図 1 5 - 1】



【図 1 5 - 2】



【図 1 6】



10

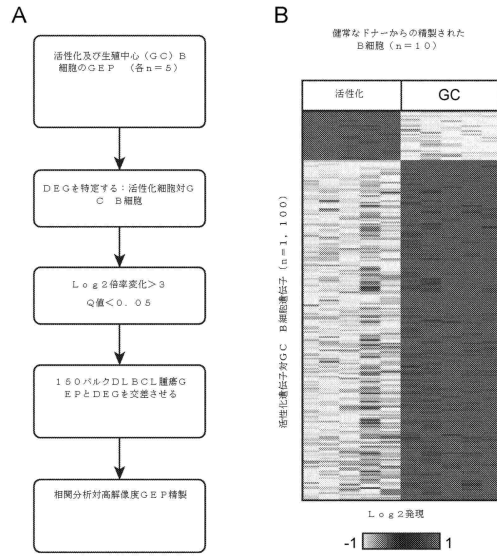
20

30

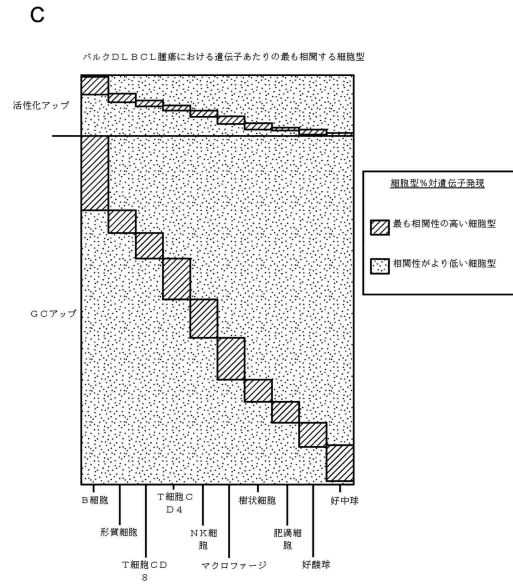
40

50

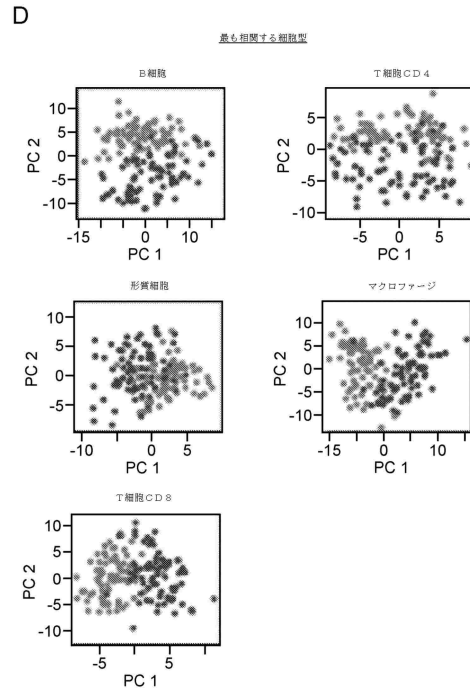
【図 17-1】



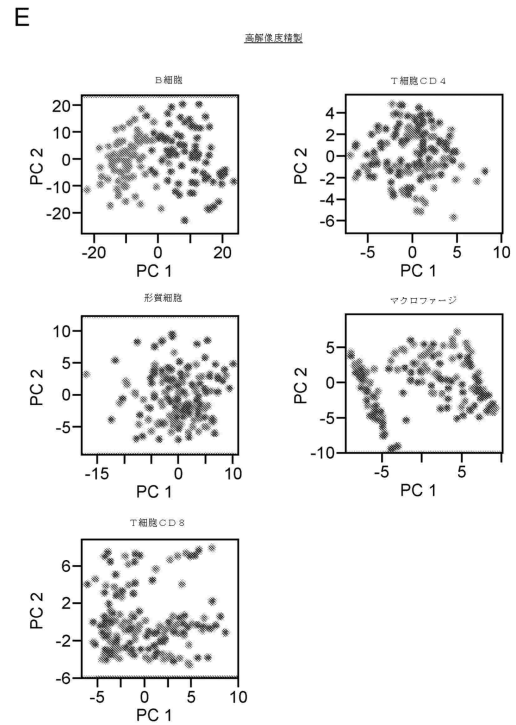
【図 17-2】



【図 17-3】



【図 17-4】



10

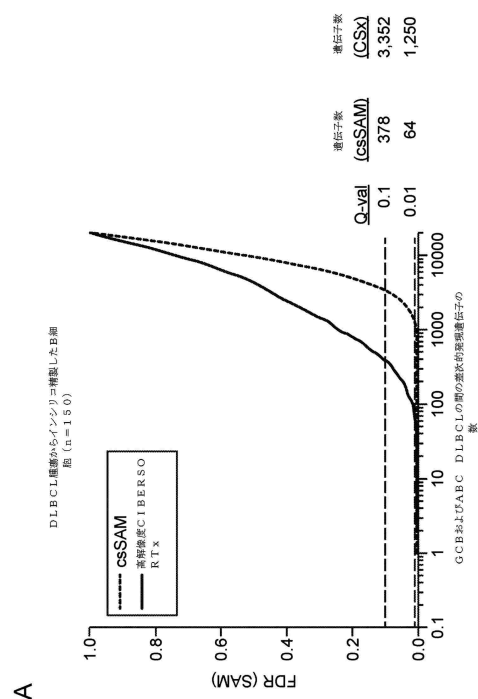
20

30

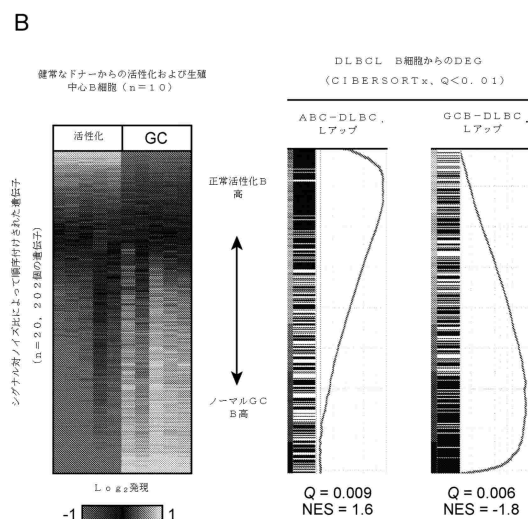
40

50

【图 18-1】



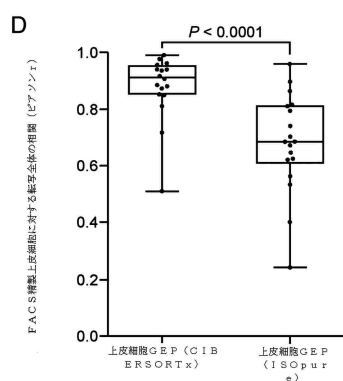
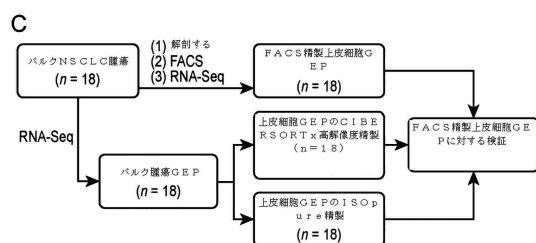
【图 18-2】



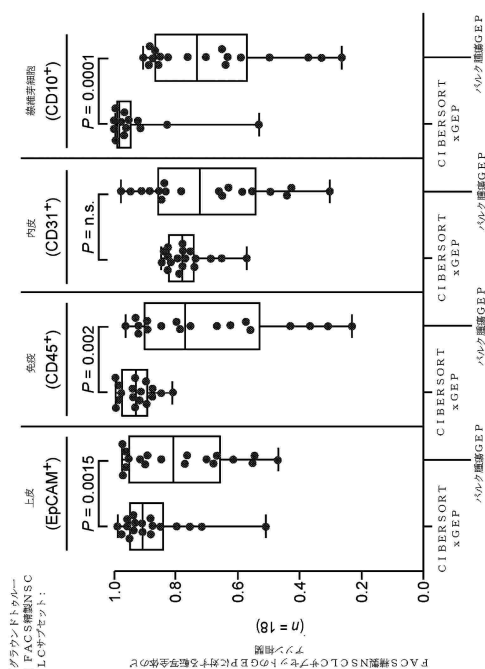
10

20

【圖 18-3】



【圖 18-4】



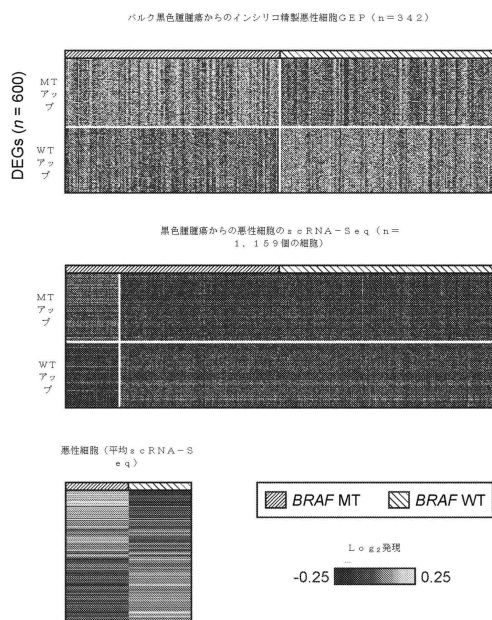
30

40

50

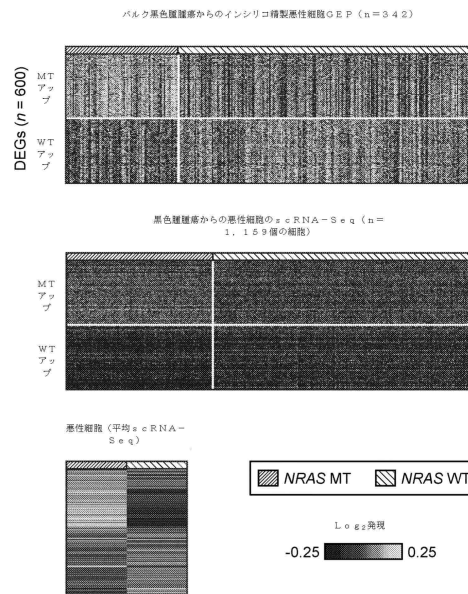
【圖 19-1】

A



【圖 19-2】

B

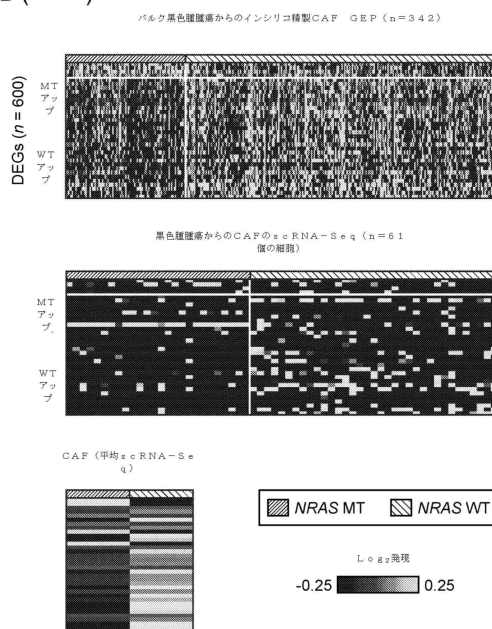


10

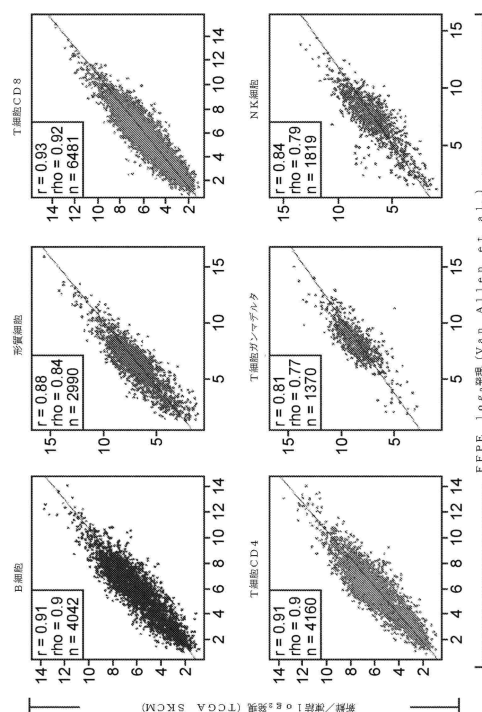
20

【圖 19-3】

B (続き)



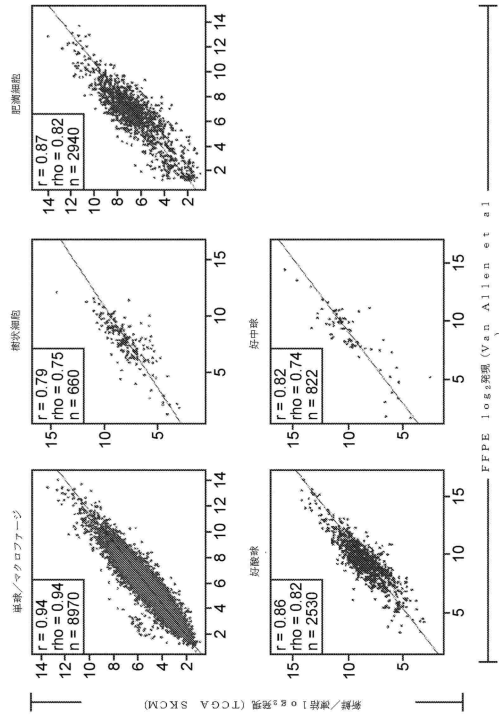
【図 20-1】



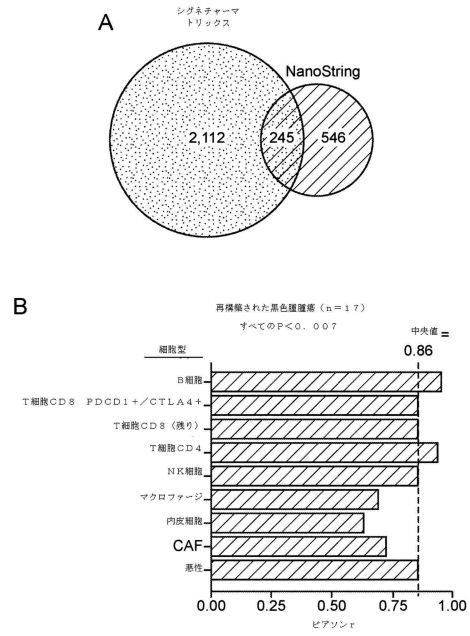
30

40

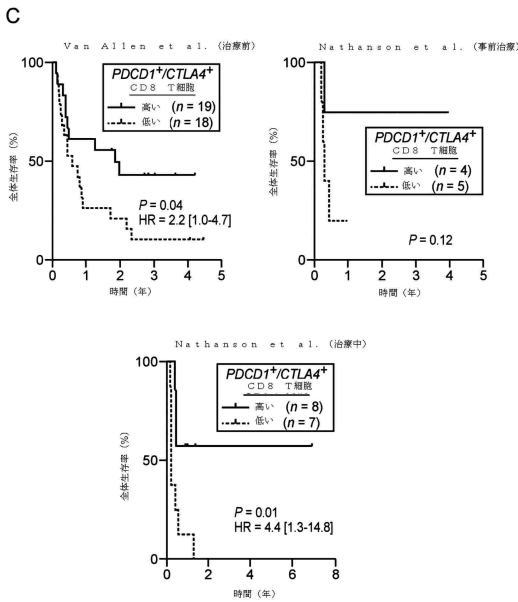
【図 20-2】



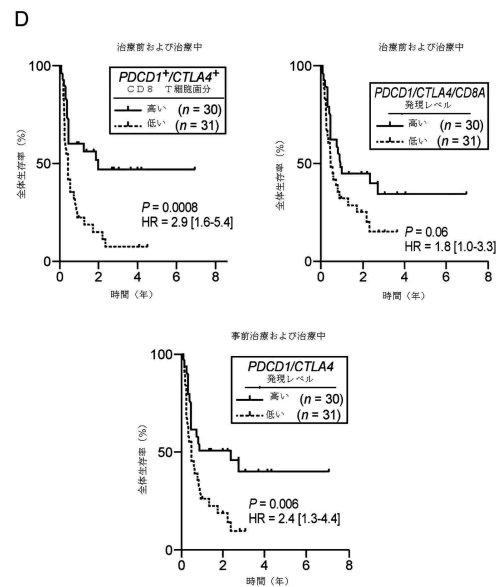
【図 21-1】



【図 21-2】



【図 21-3】



10

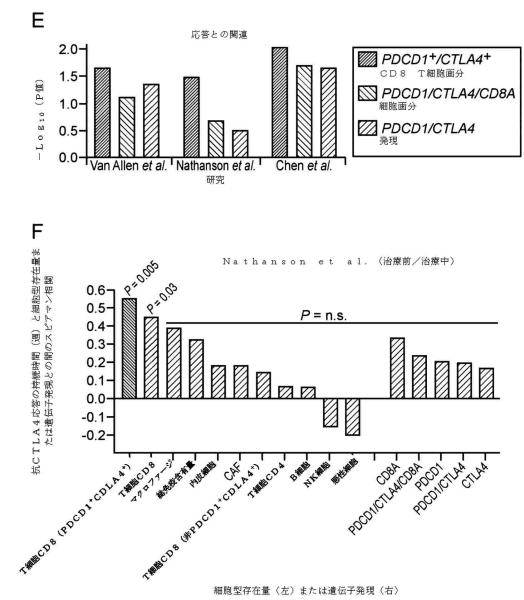
20

30

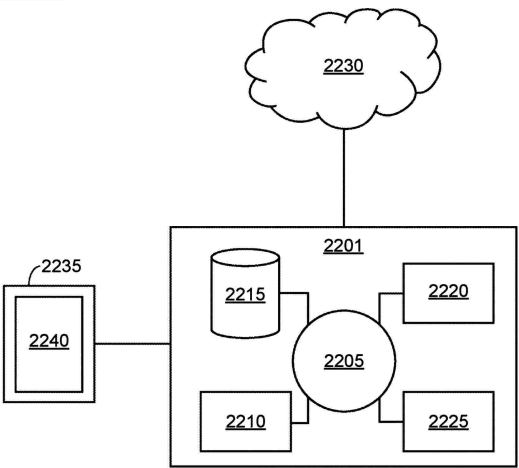
40

50

【図 2 1 - 4】

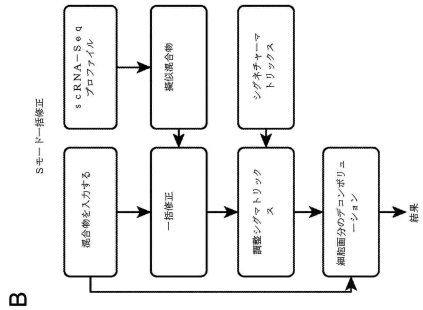


【図 2 2】

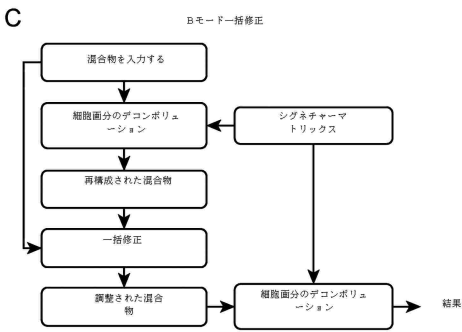


10

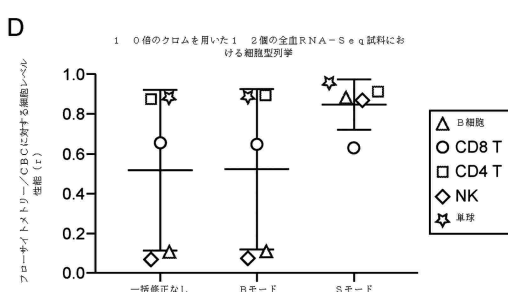
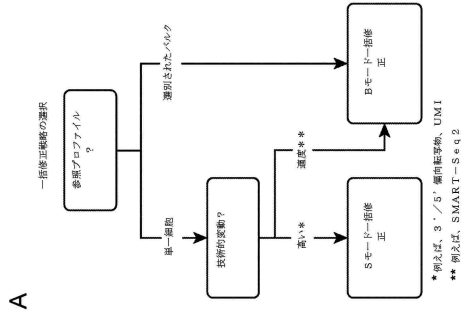
【図 2 3 - 1】



【図 2 3 - 2】



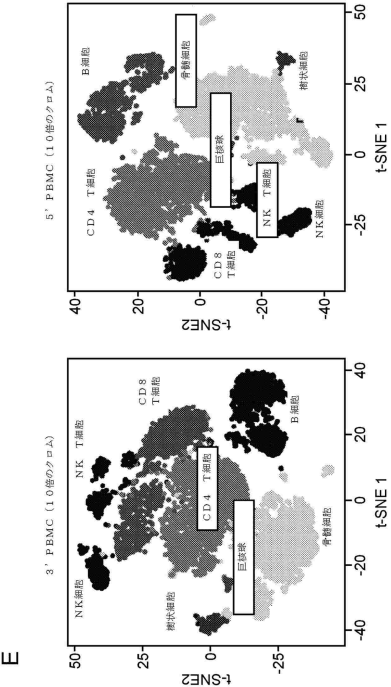
30



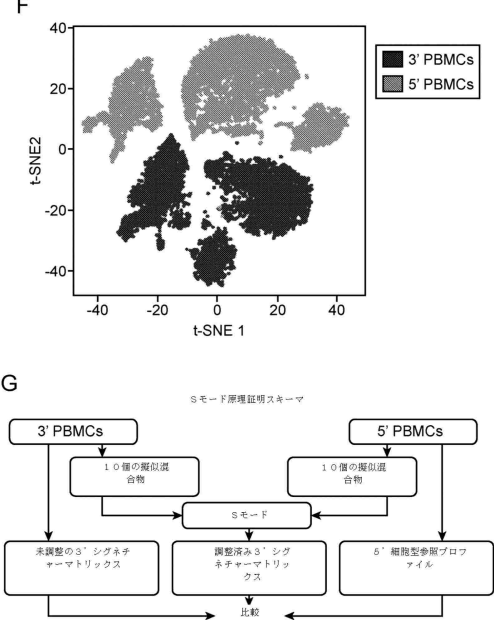
40

50

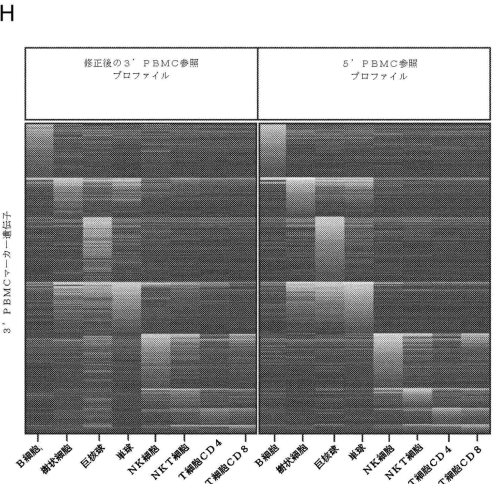
【図 2 3 - 3】



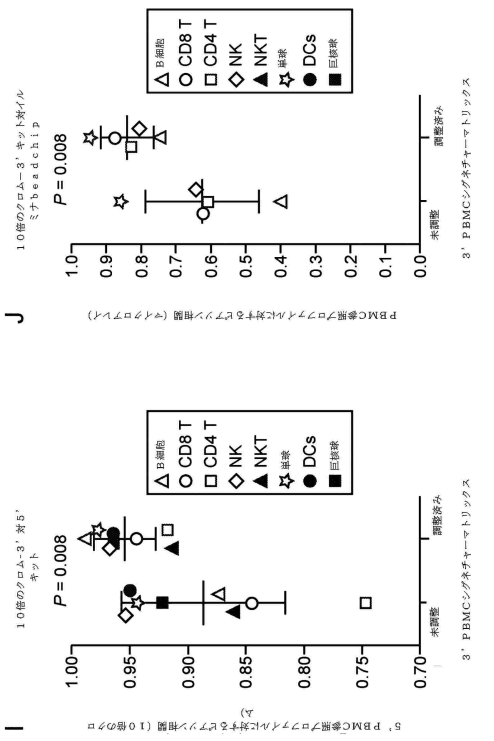
【図 2 3 - 4】



【図 2 3 - 5】



【図 2 3 - 6】



10

20

30

40

50

【図 25 - 2】

免疫細胞の種類											
免疫細胞の種類	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球
	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球	好中球

10

20

30

40

50

フロントページの続き

(51)国際特許分類		F I			
C 1 2 N	15/09 (2006.01)	C 1 2 N	15/09	Z	
C 1 2 M	1/34 (2006.01)	C 1 2 M	1/34	D	

イブ 1 6 8 4

(72)発明者 アリザデ, アラシュ アッシュ
 アメリカ合衆国, カリフォルニア州 9 4 4 0 3, サン マテオ, グリーンフィールド アベニュー
 3 2 8

審査官 吉門 沙央里

(56)参考文献 米国特許出願公開第 2 0 1 6 / 0 2 1 7 2 5 3 (U S, A 1)
 J. J-Y. Wang, et al., BMC Bioinformatics, 2013年, 14:107, p.1-11

(58)調査した分野 (Int.Cl., D B名)
 C 1 2 Q 1 / 0 0 - 3 / 0 0
 C 1 2 M 1 / 0 0 - 3 / 0 0
 C A p l u s / M E D L I N E / E M B A S E / B I O S I S (S T N)
 J S T P l u s / J M E D P l u s / J S T 7 5 8 0 (J D r e a m I I I)