(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2024-528441
(P2024-528441A)
(43)公表日　令和6年7月30日(2024.7.30)

(51)国際特許分類　FI
G16B 30/00 (2019.01)　G16B 30/00
G16B 40/00 (2019.01)　G16B 40/00

審査請求　未請求　予備審査請求　未請求　（全65頁）

(21)出願番号　特願2023-578712(P2023-578712)
(86)(22)出願日　令和4年6月17日(2022.6.17)
(85)翻訳文提出日　令和6年2月19日(2024.2.19)
(86)国際出願番号　PCT/US2022/034068
(87)国際公開番号　WO2022/271566
(87)国際公開日　令和4年12月29日(2022.12.29)
(31)優先権主張番号　63/202,716
(32)優先日　令和3年6月22日(2021.6.22)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　508152917
ザ ボード オブ リージェンツ オブ ザ ユニバーシティー オブ テキサス システム
THE BOARD OF REGENTS OF THE UNIVERSITY OF TEXAS SYSTEM
アメリカ合衆国 ７８７０１ テキサス，オースティン， ウェスト ７番 ストリート ２１０
(74)代理人　230104019
弁護士　大野 聖二
(74)代理人　100149076
弁理士　梅田 慎介
(74)代理人　100173185
最終頁に続く

(54)【発明の名称】　多重疾患診断のためのＴＣＲレパートリーフレームワーク

(57)【要約】
幾何学的等長性に基づく抗原特異的ＴＣＲアラインメント（ＧＩＡＮＡ）の新規な方法が本明細書に記載される。ＧＩＡＮＡは、数千万の配列を効率的に取り扱うことができる抗原特異的ＴＣＲクラスター化法である。ＧＩＡＮＡは現存する全ての方法より高い感度および精度を達成し、既知の抗原に特異的なＴＣＲを高い正確度で検索することができる。新規の試料の超大スケールのＴＣＲクラスター化および高速のクエリーによって、参照に基づく新規なレパートリー分類フレームワークも可能になった。ＧＩＡＮＡはＴＣＲ領域が解明された単一細胞のＲＮＡ－ｓｅｑデータを解析することもでき、パブリックドメインにあるＴＣＲレパートリー試料の大きなデータベースに対する未知のデータにＴＣＲを探索要求し、共有された抗原特異性に対する新たな洞察を提供することが可能である。ＧＩＡＮＡは、大型Ｂ細胞受容体のシーケンシングデータをクラスター化または探索要求するために適用することもできる。
【選択図】図１

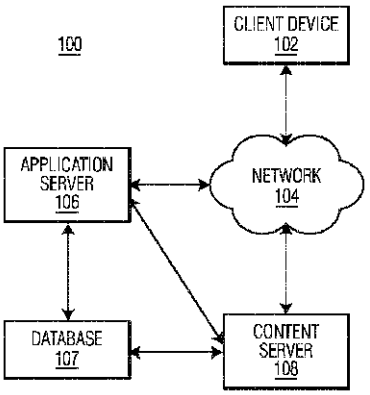


FIG. 1

【特許請求の範囲】
【請求項１】
Ｔ細胞受容体（ＴＣＲ）の比較のための計算効率を改善する方法であって、
計算デバイスによって、ただ１つのエピトープに特異的なＴＣＲからなる参照ＴＣＲ配列（ＴＣＲ－ｓｅｑ）データセットから相補性決定領域３（ＣＤＲ３）配列を同定するステップ、
前記計算デバイスによって、参照ＴＣＲ－ｓｅｑデータセットからのＣＤＲ３配列のそれぞれをＣＤＲ３配列のそれぞれにおけるアミノ酸の配列に対応する数値ベクトルにエンコードするステップ、
前記計算デバイスによって、数値ベクトルを高次元ユークリッド空間における座標に変換するステップ、
前記計算デバイスによって、
ニューラルネットワークによって、座標の相対的距離に基づいて数値ベクトルのツリーデータ構造を生成するように学習し、
ニューラルネットワークによって、相対的距離に基づいて座標をプレクラスターにグループ化することによって、
ニューラルネットワークを用いて予測モデルを生成するステップ、
前記計算デバイスによって、プレクラスター中のＣＤＲ３配列をフィルタリングするステップ、ならびに
前記計算デバイスによって、フィルタリングされたプレクラスターから抗原特異的ＣＤＲ３クラスターを同定するステップ
を含む、方法。
【請求項２】
前記計算デバイスによって、同定、エンコーディング、変換、生成、およびフィルタリングのステップを、既知の抗原特異的ＴＣＲ情報を有しないクエリーＴＣＲ－ｓｅｑデータセットについて実施するステップ、
前記計算デバイスによって、クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターを抗原特異的ＣＤＲ３クラスターと比較するステップ、および
前記計算デバイスによって、クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターが抗原特異的ＣＤＲ３クラスターと一致することを判定し、疾患の状態を診断および／または判定するステップ
をさらに含む、請求項１に記載の方法。
【請求項３】
前記計算デバイスによって、同一の座標をともに有するＣＤＲ３配列をグループ化するステップ
をさらに含む、請求項１に記載の方法。
【請求項４】
フィルタリングするステップが、
前記計算デバイスによって、プレクラスター中のＣＤＲ３配列のそれぞれの対のＴＣＲ可変（ＴＲＢＶ）アレルを比較してアラインメントスコアを決定するステップ、および
前記スコアが所定のレベルを超えれば、前記計算デバイスによって、プレクラスターを１つまたは複数の新たなプレクラスターに分割するステップ
を含む、請求項１に記載の方法。
【請求項５】
フィルタリングするステップが、
前記計算デバイスによって、プレクラスターのそれぞれについてスミス－ウォーターマンアラインメントを実施してアラインメントスコアを決定するステップ、および
前記スコアが所定のレベル未満であれば、前記計算デバイスによってプレクラスターを除去するステップ
をさらに含む、請求項４に記載の方法。

【請求項６】
エンコードするステップが、
前記計算デバイスによって、ＣＤＲ３配列のそれぞれについて一連のユニタリー変換を実施するステップを含む、請求項１に記載の方法。
【請求項７】
プロセッサーによって実行された場合に、
ただ１つのエピトープに特異的なＴＣＲからなる参照ＴＣＲ配列（ＴＣＲ－ｓｅｑ）データセットから相補性決定領域３（ＣＤＲ３）配列を同定するステップ、
前記参照ＴＣＲ－ｓｅｑデータセットからのＣＤＲ３配列のそれぞれをＣＤＲ３配列のそれぞれにおけるアミノ酸の配列に対応する数値ベクトルにエンコードするステップ、
前記数値ベクトルを高次元ユークリッド空間における座標に変換するステップ、
ニューラルネットワークによって、座標の相対的距離に基づいて数値ベクトルのツリーデータ構造を生成するように学習し、
前記ニューラルネットワークによって、相対的距離に基づいて座標をプレクラスターにグループ化することによって、
ニューラルネットワークを用いて予測モデルを生成するステップ、
プレクラスター中のＣＤＲ３配列をフィルタリングするステップ、ならびに
フィルタリングされたプレクラスターから抗原特異的ＣＤＲ３クラスターを同定するステップ
をプロセッサーに実行させる非一時的なコンピューター可読な命令を記憶するメモリーに作動可能に連結されたプロセッサー
を含む、計算デバイス。
【請求項８】
前記コンピューター可読な命令が、プロセッサーによって実行された場合に、
同定、エンコーディング、変換、生成、およびフィルタリングのステップを、既知の抗原特異的ＴＣＲ情報を有しないクエリーＴＣＲ－ｓｅｑデータセットについて実施するステップ、
クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターを抗原特異的ＣＤＲ３クラスターと比較するステップ、および
クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターが抗原特異的ＣＤＲ３クラスターと一致することを判定し、疾患の状態を診断および／または判定するステップ
をプロセッサーにさらに実行させる、請求項７に記載の計算デバイス。
【請求項９】
前記コンピューター可読な命令が、前記プロセッサーによって実行された場合に、
同一の座標をともに有するＣＤＲ３配列をグループ化するステップ
を前記プロセッサーにさらに実行させる、請求項７に記載の計算デバイス。
【請求項１０】
フィルタリングするステップが、
前記計算デバイスによって、プレクラスター中のＣＤＲ３配列のそれぞれの対のＴＣＲ可変（ＴＲＢＶ）アレルを比較してアラインメントスコアを決定するステップ、および
前記スコアが所定のレベルを超えれば、前記計算デバイスによって、プレクラスターを１つまたは複数の新たなプレクラスターに分割するステップ
を含む、請求項７に記載の計算デバイス。
【請求項１１】
フィルタリングするステップが、
前記計算デバイスによって、プレクラスターのそれぞれについてスミス－ウォーターマンアラインメントを実施してアラインメントスコアを決定するステップ、および
前記スコアが所定のレベル未満であれば、計算デバイスによってプレクラスターを除去するステップ

をさらに含む、請求項１０に記載の計算デバイス。
【請求項１２】
エンコードするステップが、
前記計算デバイスによって、ＣＤＲ３配列のそれぞれについて一連のユニタリー変換を実施するステップを含む、請求項７に記載の計算デバイス。
【請求項１３】
計算デバイスに付随するプロセッサーによって実行された場合に、
ただ１つのエピトープに特異的なＴＣＲからなる参照ＴＣＲ配列（ＴＣＲ－ｓｅｑ）データセットから相補性決定領域３（ＣＤＲ３）配列を同定するステップ、
前記参照ＴＣＲ－ｓｅｑデータセットからのＣＤＲ３配列のそれぞれをＣＤＲ３配列のそれぞれにおけるアミノ酸の配列に対応する数値ベクトルにエンコードするステップ、
前記数値ベクトルを高次元ユークリッド空間における座標に変換するステップ、
ニューラルネットワークによって、座標の相対的距離に基づいて数値ベクトルのツリーデータ構造を生成するように学習し、
前記ニューラルネットワークによって、相対的距離に基づいて座標をプレクラスターにグループ化することによって、
ニューラルネットワークを用いて予測モデルを生成するステップ、
前記プレクラスター中のＣＤＲ３配列をフィルタリングするステップ、ならびに
フィルタリングされたプレクラスターから抗原特異的ＣＤＲ３クラスターを同定するステップ
をプロセッサーに実行させる、コンピューター実行可能な命令が有形的にエンコードされた非一時的なコンピューター可読の記憶媒体。
【請求項１４】
前記コンピューター実行可能な命令が、プロセッサーによって実行された場合に、
同定、エンコーディング、変換、生成、およびフィルタリングのステップを、既知の抗原特異的ＴＣＲ情報を有しないクエリーＴＣＲ－ｓｅｑデータセットについて実施するステップ、
前記クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターを抗原特異的ＣＤＲ３クラスターと比較するステップ、および
クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターが抗原特異的ＣＤＲ３クラスターと一致することを判定し、疾患の状態を診断および／または判定するステップ
をプロセッサーに実行させる、請求項１３に記載の非一時的なコンピューター可読の記憶媒体。
【請求項１５】
コンピューター実行可能な命令が、プロセッサーによって実行された場合に、
同一の座標をともに有するＣＤＲ３配列をグループ化するステップ
をプロセッサーに実行させる、請求項１３に記載の非一時的なコンピューター可読の記憶媒体。
【請求項１６】
フィルタリングするステップが、
前記計算デバイスによって、プレクラスター中のＣＤＲ３配列のそれぞれの対のＴＣＲ可変（ＴＲＢＶ）アレルを比較してアラインメントスコアを決定するステップ、および
前記スコアが所定のレベルを超えれば、前記計算デバイスによって、プレクラスターを１つまたは複数の新たなプレクラスターに分割するステップ
を含む、請求項１３に記載の非一時的なコンピューター可読の記憶媒体。
【請求項１７】
フィルタリングするステップが、
前記計算デバイスによって、プレクラスターのそれぞれについてスミス－ウォーターマンアラインメントを実施してアラインメントスコアを決定するステップ、および

前記スコアが所定のレベル未満であれば、前記計算デバイスによってプレクラスターを除去するステップ
をさらに含む、請求項１６に記載の非一時的なコンピューター可読の記憶媒体。
【請求項１８】
エンコードするステップが、
前記計算デバイスによって、ＣＤＲ３配列のそれぞれについて一連のユニタリー変換を実施するステップを含む、請求項１３に記載の非一時的なコンピューター可読の記憶媒体。
【請求項１９】
共通の抗原特異性を用いてＴ細胞受容体（ＴＣＲ）データベースを組織化し探索要求する方法であって、
１つまたは複数のＴＣＲ非類似性メトリックスを用いる最近傍探索を実施して、共通の抗原特異性を有するＴＣＲの対を見出すステップを含む、方法。
【請求項２０】
前記１つまたは複数のＴＣＲ非類似性メトリックスが、スミス－ウォーターマン距離および高次元ユークリッド空間における埋め込みの１つもしくは複数、または他の任意の距離もしくは非類似性メトリックを含む、請求項１９に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願への相互参照
本出願は、２０２１年６月２２日出願の米国仮出願第６３／２０２，７１６号の３５Ｕ．Ｓ．Ｃ．セクション１１９（ｅ）の下での優先権の利益を主張する。先行出願の開示は、参照により全体として本出願の開示の一部であるとみなされ、本明細書に組み込まれる。
【０００２】
本開示は一般に、免疫レパートリーに基づく疾患の診断技術に関し、より詳細には、類似のＴ細胞受容体（ＴＣＲ）配列を効率的にグループ化し、疾患を有する患者を診断し、末梢血のＴＣＲレパートリーによって患者の疾患の状態を判定するための新規なシステムおよび方法に関する。
【背景技術】
【０００３】
適応免疫レパートリーは、多様なヒトの疾患の重要な調節因子であり、近年、１万を超えるＴＣＲレパートリーシーケンシング（ＴＣＲ－ｓｅｑ）試料が生成されている。しかし、ＴＣＲデータの解釈は、既知の抗原特異性が希少であることによって妨げられてきた。最近の研究によって、ＴＣＲの超可変相補性決定領域３（ＣＤＲ３）における類似性が抗原認識のための構造的類似性に関与していることが実証された。したがって、類似したＣＤＲ３のクラスター化が、抗原特異的受容体を同定するための重要な方法となってきた。
【発明の概要】
【課題を解決するための手段】
【０００４】
Ｔ細胞受容体（ＴＣＲ）の比較のための計算効率を改善するための方法、システム、および装置を本明細書に記載する。一例では、相補性決定領域３（ＣＤＲ３）の配列が、参照ＴＣＲ配列（ＴＣＲ－ｓｅｑ）データセットから同定され得る。参照ＴＣＲ－ｓｅｑデータセットは、ただ１つのエピトープに特異的なＴＣＲからなり得る。参照ＴＣＲ－ｓｅｑデータセットからのＣＤＲ３配列のそれぞれが数値ベクトルにエンコードされ得、数値ベクトルはＣＤＲ３配列のそれぞれにおけるアミノ酸の配列に対応する。数値ベクトルは高次元ユークリッド空間における座標に変換され得る。ニューラルネットワークを用いて予測モデルを生成することができる。ニューラルネットワークは、座標の相対的距離に基

づいて数値ベクトルのツリーデータ構造を生成するように学習し、相対的距離に基づいて座標をプレクラスターにグループ化することができる。プレクラスター中のＣＤＲ３配列は、ノイズを低減させるための１つまたは複数の基準を用いてフィルタリングすることができる。フィルタリングされたプレクラスターから抗原特異的ＣＤＲ３クラスターを同定することができる。

【０００５】

別の実施形態では、未知のＴＣＲ－ｓｅｑ試料を存在する参照データに対して探索要求（クエリー）し、疾患を有する患者を診断して、末梢血のＴＣＲレパートリーによってその疾患状態を判定することができる。上記の同定、エンコーディング、変換、生成、およびフィルタリングのステップも、クエリーＴＣＲ－ｓｅｑデータセットについて実施することができる。クエリーＴＣＲ－ｓｅｑデータセットは、既知の抗原特異的ＴＣＲ情報を有しないことがある。クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターを抗原特異的ＣＤＲ３クラスターと比較することができる。クエリーＴＣＲ－ｓｅｑデータセットからのフィルタリングされたプレクラスターが抗原特異的ＣＤＲ３クラスターと一致することを判定することができる。

【０００６】

別の例では、大きなＴＣＲデータベースを探索要求し、共通の抗原特異性のＴＣＲクラスターにグループ化することができる。１つまたは複数のＴＣＲ非類似性メトリックスを用いる最近傍探索を実施して、共通の抗原特異性を有するＴＣＲの対を見出すことができる。１つまたは複数のＴＣＲ非類似性メトリックスには、スミス－ウォーターマン距離および高次元ユークリッド空間における埋め込みの１つもしくは複数、または他の任意の距離もしくは非類似性メトリックが含まれ得る。

【０００７】

以下に記載する図面は説明のためのみである。図面は本開示の範囲を限定することを意図していない。

【図面の簡単な説明】

【０００８】

【図１】本開示の一部の実施形態によるシステムのダイアグラムである。

【図２】本開示の一部の実施形態による、本明細書に記載した方法を実施するための構成要素を説明するブロックダイアグラムである。

【図３】本開示の一部の実施形態による、参照ＴＣＲ－ｓｅｑデータのＧＩＡＮＡ解析を説明するフローチャートである。

【図４】本開示の一部の実施形態による、多次元スケーリング（ＭＤＳ）に基づく等長埋め込みの性能を説明するチャートである。

【図５Ａ】本開示の一部の実施形態による、ＣＤＲ３ストリングについてのＧ６によってエンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す図である。

【図５Ｂ】本開示の一部の実施形態による、ＣＤＲ３ストリングについてのＧ６によってエンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す図である。

【図５Ｃ】本開示の一部の実施形態による、ＣＤＲ３ストリングについてのＧ６によってエンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す図である。

【図５Ｄ】本開示の一部の実施形態による、ＣＤＲ３ストリングについてのＧ６によってエンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す図である。

【図５Ｅ】本開示の一部の実施形態による、ＣＤＲ３ストリングについてのＧ６によってエンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す図である。

【図５Ｆ】本開示の一部の実施形態による、ＣＤＲ３ストリングについてのＧ６によって

エンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す図である。
【図６】本開示の一部の実施形態による、幾何学的等長性に基づく抗原特異的ＴＣＲアラインメント（ＧＩＡＮＡ）ワークフローの概観を説明する図である。
【図７】本開示の一部の実施形態による、積み上げＭＤＳベクトルを用いるＧＩＡＮＡワークフロー（ＧＩＡＮＡｓｖ）の概観を説明する図である。
【図８】本開示の一部の実施形態による、様々なＴＣＲクラスター化アルゴリズムについての時間計算量の比較を示すチャートである。
【図９】本開示の一部の実施形態による、時間計算量を評価する際の様々なＴＣＲクラスター化アルゴリズムのメモリー使用量を示すチャートである。
【図１０】本開示の一部の実施形態によるクラスター化の精度を説明するチャートである。
【図１１】本開示の一部の実施形態によるクラスター化の感度を説明するチャートである。
【図１２】ＴＣＲクラスター化の４つの方法の間の正規化された相互情報量（ＮＭＩ）の比較を説明するチャートである。
【図１３】パラメーター設定の範囲におけるＧＩＡＮＡの性能を測定する適合率－再現率の曲線を説明するチャートである。
【図１４】様々な置換行列を用いてＧＩＡＮＡの性能を測定する適合率－再現率の曲線を説明するチャートである。
【図１５Ａ】本開示の一部の実施形態による、大きくノイズの多いＴＣＲ配列（ＴＣＲ－ｓｅｑ）試料に適用した場合のＧＩＡＮＡの感度および特異性を説明するチャートである。
【図１５Ｂ】本開示の一部の実施形態による、大きくノイズの多いＴＣＲ配列（ＴＣＲ－ｓｅｑ）試料に適用した場合のＧＩＡＮＡの感度および特異性を説明するチャートである。
【図１５Ｃ】本開示の一部の実施形態による、大きくノイズの多いＴＣＲ配列（ＴＣＲ－ｓｅｑ）試料に適用した場合のＧＩＡＮＡの感度および特異性を説明するチャートである。
【図１５Ｄ】本開示の一部の実施形態による、大きくノイズの多いＴＣＲ配列（ＴＣＲ－ｓｅｑ）試料に適用した場合のＧＩＡＮＡの感度および特異性を説明するチャートである。
【図１５Ｅ】本開示の一部の実施形態による、大きくノイズの多いＴＣＲ配列（ＴＣＲ－ｓｅｑ）試料に適用した場合のＧＩＡＮＡの感度および特異性を説明するチャートである。
【図１５Ｆ】本開示の一部の実施形態による、大きくノイズの多いＴＣＲ配列（ＴＣＲ－ｓｅｑ）試料に適用した場合のＧＩＡＮＡの感度および特異性を説明するチャートである。
【図１６Ａ】本開示の一部の実施形態による、ＧＬＩＰＨ２についての感度および特異性の推定を示す図である。
【図１６Ｂ】本開示の一部の実施形態による、ＧＬＩＰＨ２についての感度および特異性の推定を示す図である。
【図１６Ｃ】本開示の一部の実施形態による、ＧＬＩＰＨ２およびＧＩＡＮＡについての陽性予測値（ＰＰＶ）の推定を示す図である。
【図１７】本開示の一部の実施形態による、等長変換に基づく高速ＧＩＡＮＡクエリーを説明するダイアグラムである。
【図１８】本開示の一部の実施形態による、様々な数のＴＣＲによる参照／クエリーデータを用いるＧＩＡＮＡクエリーモジュールの時間計算量評価を説明するチャートである。
【図１９】本開示の一部の実施形態による、参照データセットに対してクラスター化することによってクエリーＣＯＶＩＤ－１９患者が健常対照から分離される程度を説明するチ

ャートである。
【図２０Ａ】本開示の一部の実施形態による、単一予測因子としてＣＯＶＩＤ－１９分率を用いる受信者動作特性（ＲＯＣ）曲線を説明するチャートである。
【図２０Ｂ】本開示の一部の実施形態による、単一予測因子としてＣＯＶＩＤ－１９分率を用いる受信者動作特性（ＲＯＣ）曲線を説明するチャートである。
【図２０Ｃ】本開示の一部の実施形態による、単一予測因子としてＣＯＶＩＤ－１９分率を用いる受信者動作特性（ＲＯＣ）曲線を説明するチャートである。
【図２１Ａ】本開示の一部の実施形態による、様々な数の参照ＴＣＲによるＣＯＶＩＤ－１９分率の変動係数を説明するチャートである。
【図２１Ｂ】本開示の一部の実施形態による、様々な数の参照ＴＣＲによるＣＯＶＩＤ－１９分率の変動係数を説明するチャートである。
【図２１Ｃ】本開示の一部の実施形態による、様々な数の参照ＴＣＲによるＣＯＶＩＤ－１９分率の変動係数を説明するチャートである。
【図２１Ｄ】本開示の一部の実施形態による、様々な数の参照ＴＣＲによるＣＯＶＩＤ－１９分率の変動係数を説明するチャートである。
【図２２Ａ】本開示の一部の実施形態による、ＴＣＲ共クラスター化に基づくＴＣＲ－ｓｅｑ試料の類似性のグラフ表現である。
【図２２Ｂ】本開示の一部の実施形態による、ＴＣＲ共クラスター化に基づくＴＣＲ－ｓｅｑ試料の類似性のグラフ表現である。
【図２２Ｃ】本開示の一部の実施形態による、ＴＣＲ共クラスター化に基づくＴＣＲ－ｓｅｑ試料の類似性のグラフ表現である。
【図２２Ｄ】本開示の一部の実施形態による、ＴＣＲ共クラスター化に基づくＴＣＲ－ｓｅｑ試料の類似性のグラフ表現である。
【図２３Ａ】本開示の一部の実施形態による、様々なカテゴリーのＴＣＲクローン頻度の分布を説明するビースウォームプロットである。
【図２３Ｂ】本開示の一部の実施形態による、様々なカテゴリーのＴＣＲクローン頻度の分布を説明するビースウォームプロットである。
【図２４Ａ】本開示の一部の実施形態による、重症急性呼吸器症候群コロナウイルス－２（ＳＡＲＳ－ＣｏＶ－２）感染の経過の間のＴＣＲクローン頻度の動的変化を説明するグラフである。
【図２４Ｂ】本開示の一部の実施形態による、重症急性呼吸器症候群コロナウイルス－２（ＳＡＲＳ－ＣｏＶ－２）感染の経過の間のＴＣＲクローン頻度の動的変化を説明するグラフである。
【図２５Ａ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチャートである。
【図２５Ｂ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチャートである。
【図２５Ｃ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチャートである。
【図２５Ｄ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチャートである。
【図２５Ｅ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチャートである。
【図２５Ｆ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチ

ャートである。
【図２６Ａ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格な方法を用いるＲＯＣ曲線を説明するチャートである。
【図２６Ｂ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格な方法を用いるＲＯＣ曲線を説明するチャートである。
【図２６Ｃ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格な方法を用いるＲＯＣ曲線を説明するチャートである。
【図２６Ｄ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格な方法を用いるＲＯＣ曲線を説明するチャートである。
【図２６Ｅ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格な方法を用いるＲＯＣ曲線を説明するチャートである。
【図２６Ｆ】本開示の一部の実施形態による、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格な方法を用いるＲＯＣ曲線を説明するチャートである。
【図２７】本開示の一部の実施形態による、参照ＴＣＲ－ｓｅｑ試料のクロスコホート類似性を説明するチャートである。
【図２８Ａ】本開示の一部の実施形態による、がん、ＣＯＶＩＤ－１９、多発性硬化症（ＭＳ）患者、および健常対照（ＨＣ）のクラス分率の分布を説明するバイオリンプロットである。
【図２８Ｂ】本開示の一部の実施形態による、がん、ＣＯＶＩＤ－１９、多発性硬化症（ＭＳ）患者、および健常対照（ＨＣ）のクラス分率の分布を説明するバイオリンプロットである。
【図２８Ｃ】本開示の一部の実施形態による、がん、ＣＯＶＩＤ－１９、多発性硬化症（ＭＳ）患者、および健常対照（ＨＣ）のクラス分率の分布を説明するバイオリンプロットである。
【図２８Ｄ】本開示の一部の実施形態による、がん、ＣＯＶＩＤ－１９、多発性硬化症（ＭＳ）患者、および健常対照（ＨＣ）のクラス分率の分布を説明するバイオリンプロットである。
【図２９Ａ】本開示の一部の実施形態による、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートである。
【図２９Ｂ】本開示の一部の実施形態による、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートである。
【図２９Ｃ】本開示の一部の実施形態による、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートである。
【図２９Ｄ】本開示の一部の実施形態による、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートである。
【図２９Ｅ】本開示の一部の実施形態による、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートである。
【図２９Ｆ】本開示の一部の実施形態による、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートである。
【発明を実施するための形態】
【０００９】
従来の多くの研究が、疾患の進行または免疫療法処置の間の抗原特異的なＴ細胞応答を検討するためにＴＣＲクラスター化を応用してきた。多くの研究からの多数のＴＣＲ－ｓｅｑ試料を統合することによって、免疫と疾患との相互作用へのより多くの洞察がもたらされ、予後および診断のための新規な機会が創成されることが推測される。それにも関わらず、高いクラスター化特異性はＣＤＲ３配列とＴＣＲ可変遺伝子（ＴＲＢＶ）アレルの両方におけるペアワイズなスミス－ウォーターマンアラインメントを必要とし、これは、ＴＣＲレパートリー試料のスケール（１００Ｋ配列以上）にまでスケールアップすることが通常はできない二次方程式計算の複雑性を有している。モチーフに基づくクラスター化は高速を達成できるが、特異性が非常に低くなる。したがって、現在のＴＣＲクラスター

化法はいずれも、大きなコホートのＴＣＲ－ｓｅｑ試料を解析することには適していない。

【００１０】

監視されていないＴＣＲクラスター化は、免疫レパートリーデータの基礎的な解析である。理想的なシナリオでは、同じエピトープに対して特異的な全てのＴＣＲが同じクラスターに含まれるべきである。しかし、特異性が共有されたＴＣＲ配列における想定される多様性のため、これは配列類似性またはモチーフに基づくクラスター化のアプローチについては実現不能である。そのような多様性は、Ｔ細胞受容体の独特のドッキング戦略によって惹き起こされる。例えば、インフルエンザＧＩＬエピトープに特異的なＴＣＲは通常、ＣＤＲ３領域に古典的なＲＳＳ／ＲＳＡモチーフを含むが、関連研究はＬＧＧＷモチーフも異なる方向からのＧＩＬへの強い結合を誘発することを報告している。そのような構造的変動は、単純なスミス－ウォーターマンアラインメントまたはモチーフグループ化によっては捕捉できない。結果として、類似していないモチーフを有するＣＤＲ３はその特異性を共有しているにも関わらず、小さなクラスターに断片化されることになり、これは現在の方法に対する一般的な限界である。

【００１１】

この課題に対処するため、ＣＤＲ３配列を変形し、配列アラインメントおよびクラスター化の課題を高次元ユークリッド空間における最近傍探索に変換する新規なフレームワークを開発した。この変形は、ＴＣＲのペアワイズ比較の計算効率を顕著に改善し、10^6～10^7配列にスケールアップすることができる。従来のシステムおよび方法によってはプールすることができない数千個のＴＣＲレパートリー試料をプールすることによって、本明細書に記載した新規な方法は、新規な疾患関連ＴＣＲを同定することができる。本明細書でさらに記載するように、これは新たな多重疾患診断のプラットフォームに新たな進路を開く可能性がある。

【００１２】

ここで本開示を、添付した図面を参照してこの後、より詳しく記載する。添付した図面はその一部を形成し、非限定的な説明として、ある種の例を示す。しかし、主題は種々の異なる形態で記載される可能性があり、したがって包含されるまたは特許を請求する主題は本明細書で説明するいずれの例にも限定されないと解釈すべきことを意図している。とりわけ、主題は方法、デバイス、成分、またはシステムとして記載され得る。したがって、例はハードウェア、ソフトウェア、ファームウェア、またはそれらの任意の組合せ（ソフトウェアそれ自体以外）の形態を取り得る。したがって、以下の詳細な説明は、限定的な意味で受け取られることを意図していない。

【００１３】

一般に、用語は少なくとも部分的に、文脈における用法から理解され得る。例えば、本明細書で用いられる用語、例えば「および」、「または」、または「および／または」は、それらの用語が用いられる文脈に少なくとも部分的に応じた種々の意味を含み得る。典型的には、「または」は、Ａ、Ｂ、もしくはＣ等のリストに付随して用いられる場合には、包括的な意味で用いられるＡ、Ｂ、およびＣ、ならびに排他的な意味で用いられるＡ、Ｂ、またはＣを意味することを意図している。さらに、少なくとも部分的に文脈に応じて本明細書で用いられる用語「１つまたは複数」は、任意の特徴、構造、もしくは特性を単数の意味で記載するために用いられ、または特徴、構造、もしくは特性の組合せを複数の意味で記載するために用いられ得る。同様に、例えば「１つの」または「その」等の用語も、少なくとも部分的に文脈に応じて単数の用法を伝えまたは複数の用法を伝えると理解され得る。さらに、用語「基づく」は、排他的な要素の組を伝えると必ずしも意図しないことが理解され、その代わりに、これも少なくとも部分的に文脈に応じて、必ずしも明示的に記載されていないさらなる要素の存在が認められ得る。

【００１４】

本開示を、方法およびデバイスのブロックダイアグラムおよび操作説明を参照して以下に記載する。ブロックダイアグラムまたは操作説明のそれぞれのブロック、およびブロッ

クダイアグラムまたは操作説明におけるブロックの組合せは、アナログまたはデジタルのハードウェアおよびコンピュータープログラムの命令によって実行され得ることが理解される。これらのコンピュータープログラムの命令は、その機能を本明細書に詳細を記載したように変更するために多目的コンピューターのプロセッサーに、専用コンピューターに、ＡＳＩＣに、またはその他のプログラム可能なデータ処理装置に提供され、それにより、コンピューターまたはその他のプログラム可能なデータ処理装置のプロセッサーを介して実行される命令が、ブロックダイアグラムまたは操作ブロックで特定された機能／動作を実施することができる。一部の代替の実施では、ブロックに注記された機能／動作は、操作説明書に注記された順序とは異なって生じることがある。例えば、関与する機能／動作に応じて、継続して示された２つのブロックが実質的に同時に実行されることがあり、時にはブロックが逆の順で実行されることがある。

【００１５】

本開示の目的のため、非一時的コンピューター可読媒体（またはコンピューター可読の記憶媒体）がコンピューターデータを記憶し、このデータは、コンピューターによって実行可能なコンピュータープログラムコード（またはコンピューター実行可能な命令）を、機械可読な形態で含むことができる。例として、限定としてではなく、コンピューター可読媒体は、データの有形のもしくは固定された記憶のためのコンピューター可読の記憶媒体、またはコードを含むシグナルの一時的な解釈のための通信媒体を含み得る。本明細書で用いられるコンピューター可読の記憶媒体は物理的または有形の記憶（シグナルと対立する）を意味し、コンピューター可読な命令、データ構造、プログラムモジュール、またはその他のデータ等の情報の有形の記憶のための任意の方法または技術において実行される揮発性および非揮発性の、除去可能および非除去可能な媒体を限定なく含む。コンピューター可読の記憶媒体には、これらに限らないが、ＲＡＭ、ＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリーもしくはその他の固体メモリー技術、ＣＤ－ＲＯＭ、ＤＶＤ、もしくはその他の光学ストレージ、クラウドストレージ、磁気カセット、磁気テープ、磁気ディスクストレージ、もしくはその他の磁気記憶デバイス、または所望の情報もしくはデータもしくは命令を有形的に記憶するために用いることができ、コンピューターもしくはプロセッサーによってアクセスすることができる他の任意の物理的もしくは物質的な媒体が含まれる。

【００１６】

本開示の目的のため、用語「サーバー」は、プロセシング、データベース、および通信の設備を提供するサービス点を意味すると理解されたい。例として、限定としてではなく、用語「サーバー」は、付随する通信およびデータ記憶およびデータベースの設備を有する単一の物理的プロセッサーを意味することがあり、またはプロセッサーおよび付随するネットワークおよび記憶デバイスのネットワーク化もしくはクラスター化された複合体、ならびにサーバーによって提供されるサービスを援助する動作ソフトウェアおよび１つもしくは複数のデータベースシステムおよびアプリケーションソフトウェアを意味することもある。クラウドサーバーはその例である。

【００１７】

本開示の目的のため、「ネットワーク」は、例えばワイヤレスネットワークを介して連結されたワイヤレスデバイスの間を含む、サーバーとクライアントデバイスまたはその他の型のデバイスとの間等の、通信が交換され得るようにデバイスを連結するネットワークを意味すると理解されたい。ネットワークには、大容量記憶装置、例えばネットワーク接続ストレージ（ＮＡＳ）、ストレージエリアネットワーク（ＳＡＮ）、コンテンツ配信ネットワーク（ＣＤＮ）、またはその他のコンピューターもしくは機械可読の媒体の形態も含まれ得る。ネットワークには、インターネット、１つもしくは複数のローカルエリアネットワーク（ＬＡＮ）、１つもしくは複数のワイドエリアネットワーク（ＷＡＮ）、ワイヤライン型の接続、ワイヤレス型の接続、セルラー、またはそれらの任意の組合せが含まれる。同様に、異なるアーキテクチャを採用するか、または異なるプロトコルに適合もしくは互換できるサブネットワークを、大きなネットワークの中で相互運用してもよい。

【００１８】
本開示の目的のため、「ワイヤレスネットワーク」はクライアントデバイスをネットワークに連結することと理解されたい。ワイヤレスネットワークとしては、スタンドアローンアドホックネットワーク、メッシュネットワーク、ワイヤレスＬＡＮ（ＷＬＡＮ）ネットワーク、セルラーネットワーク、その他を採用してよい。ワイヤレスネットワークとしては、Ｗｉ－Ｆｉ、ロングタームエボリューション（ＬＴＥ）、ＷＬＡＮ、ワイヤレスルーター（ＷＲ）メッシュ、または第２、第３、第４、もしくは第５世代（２Ｇ、３Ｇ、４Ｇ、または５Ｇ）のセルラー技術、ブルートゥース（登録商標）、８０２．１１ｂ／ｇ／ｎ、その他を含む複数のネットワークアクセス技術をさらに採用してよい。ネットワークアクセス技術は、例えば様々な程度の可動性を有するクライアントデバイス等のデバイスのための広範囲のカバレッジを可能にし得る。

【００１９】
手短に言えば、ワイヤレスネットワークは、それによってシグナルがデバイスの間、例えばクライアントデバイスまたは計算デバイスの間、ネットワークの間もしくはその中、その他で通信される、実質的に任意の型のワイヤレス通信機構を含み得る。

【００２０】
計算デバイスは、例えば有線もしくは無線のネットワークを介してシグナルを送信または受信することができ、または例えば物理的メモリー状態としてメモリーの中にシグナルを処理または記憶することができ、したがってサーバーとして動作することができる。即ち、サーバーとして動作することができるデバイスには、例えば専用のラックマウント方式のサーバー、デスクトップコンピューター、ラップトップコンピューター、セットトップボックス、上記のデバイスの２つ以上の特徴等の種々の特徴を組み合わせた統合されたデバイス、その他が含まれ得る。

【００２１】
ここで図１を参照してシステム１００を示す。図１は、本明細書で論じるシステムおよび方法がその中で実施される一般的な環境の構成要素を説明している。本開示を実施するために全ての構成要素が必要なのではなく、本開示の精神および範囲から逸脱することなく、構成要素の配置および型における変形を行ってよい。

【００２２】
図１のシステム１００はネットワーク１０４を含み、これは上で論じたように、それだけに限らないが、ワイヤレスネットワーク、ローカルエリアネットワーク（ＬＡＮ）、ワイドエリアネットワーク（ＷＡＮ）、インターネット、またはそれらの組合せを含んでよい。

【００２３】
ネットワーク１０４は、別のネットワークまたはデバイスとともに、例えば１つもしくは複数のクライアントデバイス１０２、アプリケーションサーバー１０６、コンテンツサーバー１０８、およびデータベース１０７、ならびにそれらの構成要素と連結してよい。ネットワーク１０４は、１つまたは複数のクライアントデバイス１０２、アプリケーションサーバー１０６、コンテンツサーバー１０８、およびデータベース１０７のためのインフラストラクチャーを指向した連結を提供するために、スタンドアローンアドホックネットワークにさらに重層する種々のワイヤレスサブネットワーク等として構成してよい。ネットワーク１０４は、１つの電子デバイスから別の電子デバイスに情報を通信するために、任意の形態のコンピューター可読媒体またはネットワークを採用するように構成してよい。

【００２４】
１つまたは複数のクライアントデバイス１０２は、例えばデスクトップコンピューターまたはポータブルデバイス、例えば携帯電話、スマートフォン、ディスプレイページャー、ラジオ周波数（ＲＦ）デバイス、赤外（ＩＲ）デバイス、近距離通信（ＮＦＣ）デバイス、個人用デジタル端末（ＰＤＡ）、ハンドヘルドコンピューター、タブレットコンピューター、ファブレット、ラップトップコンピューター、セットトップボックス、ウェアラ

ブルコンピューター、スマートウォッチ、上記のデバイスの特徴等の種々の特徴を組み合わせた統合型もしくは分配型デバイス、その他を含んでよい。

【００２５】

１つまたは複数のクライアントデバイス１０２は、別の計算デバイスからコンテンツを受信するように構成された少なくとも１つのクライアントアプリケーションを含んでもよい。１つまたは複数のクライアントデバイス１０２は、ネットワーク１０４を通して他のデバイスまたはサーバーと通信してよく、そのような通信には、メッセージの送信および／または受信、ＴＣＲデータの生成および提供、ＴＣＲデータの探索、表現、および／もしくは共有、または他の形態の種々の通信のいずれかが含まれ得る。１つまたは複数のクライアントデバイス１０２は、例えば物理的メモリー状態としてメモリーの中にシグナルを処理または記憶することができ、したがってサーバーとして動作することができる。

【００２６】

アプリケーションサーバー１０６およびコンテンツサーバー１０８は、任意の型または形態のコンテンツを、ネットワークを介して別のデバイスに提供および／または生成するように構成された１つまたは複数のデバイスを含んでよい。アプリケーションサーバー１０６および／またはコンテンツサーバー１０８として作動し得るデバイスには、パーソナルコンピューター、デスクトップコンピューター、マルチプロセッサーシステム、マイクロプロセッサー系またはプログラム可能な家庭用電子製品、ネットワークＰＣ、サーバー類、その他が含まれ得る。アプリケーションサーバー１０６およびコンテンツサーバー１０８は、それぞれのデバイスによって提供されるコンテンツおよびサービスに関連する種々の型のデータを、データベース１０７に記憶することができる。

【００２７】

ユーザー（例えば患者、医師、技師、その他）は、アプリケーションサーバー１０６およびコンテンツサーバー１０８によって提供されるサービスにアクセスすることができる。これには、例えば１つまたは複数のクライアントデバイス１０２を用いるネットワーク１０４を介するアプリケーションサーバー、認証サーバー、検索サーバー、交換サーバーが含まれてもよい。即ち、例えばアプリケーションサーバー１０６は、種々の型のアプリケーションならびにアプリケーションデータおよびユーザープロファイル情報を含むアプリケーションに関連する情報を記憶することができる。

【００２８】

図１はアプリケーションサーバー１０６とコンテンツサーバー１０８とをそれぞれ単一の計算デバイスとして説明しているが、本開示はそれに限定されない。例えば、アプリケーションサーバー１０６とコンテンツサーバー１０８の１つまたは複数の機能を、区別できる１つまたは複数の計算デバイスに分配してよい。別の例では、アプリケーションサーバー１０６とコンテンツサーバー１０８は、本開示の範囲から逸脱することなく、単一の計算デバイスに統合してよい。

【００２９】

ここで図２を参照して、本明細書に記載した方法を実施するための構成要素を説明するブロックダイアグラムを示す。図２は、ＴＣＲエンジン２００、ネットワーク１０４、およびデータベース１０７を含む。ＴＣＲエンジン２００は、特殊目的の機械またはプロセッサーであってよく、アプリケーションサーバー１０６、コンテンツサーバー１０８、ウェブサーバー、サードパーティサーバー、ユーザーの計算デバイス、その他の１つまたは複数の中に含まれてもよい。

【００３０】

一例では、ＴＣＲエンジン２００は従来のパーソナルコンピューターであってよく、以下に記載する方法は、ＣＰＵ上の単一のスレッドを用いて実施してよい。別の例では、１千万配列の参照データをクラスター化する場合、ＴＣＲエンジン２００は高性能計算（ＨＰＣ）スーパークラスター（例えばメモリーアロケーションを１２８Ｇ、ＣＰＵノードを８とする）であってよい。

【００３１】

ＴＣＲエンジン２００は、デバイス（例えばユーザーデバイスまたはシステム／ウェブに連結されたサーバー／デバイス）上で実行されるスタンドアローンアプリケーションであってよい。別の例では、ＴＣＲエンジン２００は、デバイスにインストールされたアプリケーション、および／またはネットワークを通じてデバイスによってアクセスされるウェブ系アプリケーションとして機能してよい。ＴＣＲエンジン２００は、増補されるスクリプト、プログラム、またはアプリケーション（例えばプラグインまたは拡張）として、別のアプリケーション、例えば患者に関連するデータを集約し共有する健康管理アプリケーションにインストールしてよい。

【００３２】

データベース１０７は任意の型のデータベースまたはメモリーであってよく、ネットワーク上のサーバー（例えばアプリケーションサーバー１０６およびコンテンツサーバー１０８）またはユーザーのデバイス（例えば１つまたは複数のクライアントデバイス１０２）に付随していてよい。データベース１０７は、ユーザー、サービス、アプリケーション、コンテンツ、その他に関連するローカルおよび／またはネットワークの情報に付随するデータおよびメタデータのデータセットを含んでよい。そのような情報は、データベース１０７に独立に、および／またはリンクしたもしくは付随したデータセットとして、記憶し、索引付けしてよい。本明細書で論じるように、データベース１０７におけるデータ（およびメタデータ）は、本開示の範囲から逸脱しなければ、既知であってもこれから既知になるものでも、任意の型の情報および型であってよいことを理解されたい。

【００３３】

データベース１０７は、ユーザーのためのデータ（例えばユーザーデータ）を記憶することができる。記憶されるユーザーデータには、例えば参照ＴＣＲ－ｓｅｑデータに付随する情報、患者のがんの診断、患者の染色体の情報、患者のＤＮＡの情報、患者の血液の情報、患者の人口学的情報、患者の経歴情報、その他、またはそれらのいくつかの組合せが含まれ得る。

【００３４】

データベース１０７におけるデータ（およびメタデータ）は、本開示の範囲から逸脱しなければ、既知であってもこれから既知になるものでも、ＴＣＲ－ｓｅｑデータ、患者、医師、コンテンツ、デバイス、アプリケーション、サービス提供者、コンテンツ提供者に関連する任意の型の情報であってよい。

【００３５】

データベース１０７に記憶されるデータは、例えば２５６ビットの暗号化を用いて暗号化してよく、それにより、データは１９９６年のＨｅａｌｔｈ Ｉｎｓｕｒａｎｃｅ Ｐｏｒｔａｂｉｌｉｔｙ ａｎｄ Ａｃｃｏｕｎｔａｂｉｌｉｔｙ Ａｃｔ（ＨＩＰＰＡ）に従って非公開となり、管理される。

【００３６】

データベース１０７は、データおよびメタデータの連結された組として情報を記憶および索引付けすることができ、データとメタデータの関係はｎ次元のベクトルとして記憶することができる。そのような記憶は、ハッシュツリー、キュー、スタック、ＶＬｉｓｔ、またはその他の任意の型の既知もしくは既知となる動的メモリー割り当ての手法もしくは技術を含むがこれらに限らない、既知もしくは既知となるベクトルまたはアレイストレージを通して実現することができる。クラスター解析、データマイニング、ベイジアンネットワーク解析、隠れマルコフモデル、人工のニューラルネットワーク解析、論理モデル、および／またはツリー解析、その他であるがこれだけに限らない、任意の既知もしくは既知となる計算解析手法もしくはアルゴリズムが、患者および／または医療提供者についてのベクトル情報を決定、誘導、またはその他同定するために適用し得ることを理解されたい。

【００３７】

図１を参照して上で論じたように、ネットワーク１０４は、ワイヤレスネットワーク、ローカルエリアネットワーク（ＬＡＮ）、ワイドエリアネットワークネットワーク（ＷＡ

Ｎ）、インターネット、またはそれらの組合せ等であるがこれらに限らない任意の型のネットワークであってよい。ネットワーク１０４は、ＴＣＲエンジン２００と記憶されたリソースのデータベース１０７との接続を容易にすることができる。まさに、図２に示したように、ＴＣＲエンジン２００とデータベース１０７は、そのようなデバイスとリソースとの間を連結しおよび／または通信を可能にする既知もしくは既知となる任意の方法によって直接連結してよい。

【００３８】

本明細書における特殊目的機能に従ってプログラムされるハードウェアを含む主要なプロセッサー、サーバー、またはデバイスの組合せは、便宜上、ＴＣＲエンジン２００と称してよい。ＴＣＲエンジン２００は、試料モジュール２０２、ＡＩモジュール２０４、エンコーディングモジュール２０６、フィルタリングモジュール２０８、同定（ＩＤ）モジュール２１０、および変換モジュール２１２を含んでよい。論じたシステムおよび方法の例にさらなるまたはより少ないエンジンおよび／またはモジュール（またはサブモジュール）が適用可能であるので、本明細書で論じるエンジンおよびモジュールは限定的である。それぞれのモジュールの動作、構成、および機能、ならびに本開示の例の中でのそれらの役割について、以下に論じる。

【００３９】

本明細書に記載した原理は、多くの異なる形態で具現化され得る。抗原と反応するＴ細胞は種々の疾患に対する免疫の中心的なメディエーターかつ免疫療法の重要な目標であり、大部分のがん抗原が未知であるので、がん関連のＴ細胞の実験的検出はいまだに困難である。ディープ免疫レパートリーシーケンシング（ＴＣＲ－ｓｅｑ）技術の最近の発展は、そのようなＴ細胞の同定をさらに強調してきた。それは、これががん患者の非侵襲的臨床診断、予後および長期的な免疫モニタリングのための新たな機会を開く可能性があるからである。しかし、ヒトの免疫レパートリーは、多様な抗原に特異的なパブリックＴ細胞、ナイーブＴ細胞、およびメモリー／エフェクターＴ細胞を含み、この複雑性が、従来のシステムでは解決できない課題（例えば、ＴＣＲ－ｓｅｑデータにおいてがん関連のＴ細胞を同定すること）に加わる。

【００４０】

がん患者のＴＣＲレパートリーに関する以前の研究は、多様性およびクローン性等の単純な統計がある種の条件下で臨床的転帰に関連し、潜在的な予後因子としてのレパートリーデータの有用性を実証していることを報告している。しかし、免疫療法の急速な進歩およびＴＣＲ－ｓｅｑデータの迅速な集積により、基礎的な免疫ゲノムリサーチとがん患者に恩恵を与える臨床応用との間のギャップに橋掛けするコンピューターによるツールがもっと必要である。

【００４１】

開示するシステムおよび方法は、３チェインＴＣＲ－ｓｅｑデータを用いてがん関連免疫レパートリーのデノボ予測を提供するアンサンブル機械学習ソフトウェア（ＴＣＲブーストと称する）を実行する新規なフレームワークを通して、これらの必要なツールを提供する。

【００４２】

類似したＴＣＲ配列のグループ化は、共有された抗原特異性に関係しており、新規な治療目標を発見するために用いることができる。従来の方法には、コンピューターの費用が高く、免疫レパートリーデータセットの規模にスケールアップすることができないという欠点がある。本明細書に記載した幾何学的等長性に基づく抗原特異的ＴＣＲアラインメント（ＧＩＡＮＡ）は、従来の方法（例えばＴＣＲｄｉｓｔ）の約６００倍の速度で、従来の方法より良好な精度および感度で、速度と予測正確度との間のギャップを閉じるために用いられ得る。ＧＩＡＮＡは、大きな参照コホートの超高速クエリーも可能にし、３分以内に１０００億回を超える配列比較を処理することができる。一例では、ＧＩＡＮＡは３分以内に10^{7}の参照配列に対して10^{4}のＴＣＲを比較することができる。大スケールのＴＣＲデータセットをクラスター化するためにＧＩＡＮＡを適用することによって、疾

患特異的な受容体の新規な洞察が明らかになり、レパートリー分類タスクの新たな解決が提供され得る。ＧＩＡＮＡを用いて既存の参照に対して未知のＴＣＲ－ｓｅｑ試料を探索要求することによって高い正確度が達成され、がん、感染性疾患、および自己免疫障害を識別するために用いることができる。ＧＩＡＮＡは、ＴＣＲに基づく非侵襲的多重疾患診断プラットフォームとして用いることができる。

【００４３】

図３を参照して、参照ＴＣＲ－ｓｅｑデータのＧＩＡＮＡ解析を説明するフローチャートを示す。図３に示すステップは、図２を参照して上に記載したＴＣＲエンジン２００によって実施され得ることに留意されたい。

【００４４】

ステップ３０２で、試料モジュール２０２はＴＣＲ－ｓｅｑデータセットからＣＤＲ３配列を同定することができる。試料モジュール２０２は、例えばデータベース１０７からＴＣＲ－ｓｅｑデータセットを受信することができる。一例では、ＴＣＲ－ｓｅｑデータセットは、ただ１つのエピトープに特異的なＴＣＲからなる参照ＴＣＲ－ｓｅｑデータセットを含んでよい。ステップ３０４で、エンコーディングモジュール２０６は、ＴＣＲ－ｓｅｑデータセットからのＣＤＲ３配列のそれぞれを数値ベクトルにエンコードすることができる。数値ベクトルは、ＣＤＲ３配列のそれぞれにおけるアミノ酸の配列に対応することができる。

【００４５】

ステップ３０６で、変換モジュール２１２は、数値ベクトルを高次元ユークリッド空間における座標に変換することができる。ステップ３０８で、ＡＩモジュール２０４は、ニューラルネットワークを用いて予測モデルを生成することができる。ニューラルネットワークは、座標の相対的距離に基づいて数値ベクトルのツリーデータ構造を生成するように学習し、次いで相対的距離に基づいて座標をプレクラスターにグループ化することができる。ステップ３１０で、フィルタリングモジュール２０８は、プレクラスター中のＣＤＲ３配列をフィルタリングすることができる。ステップ３１２で、ＩＤモジュール２１０は、フィルタリングされたプレクラスターから抗原特異的ＣＤＲ３クラスターを同定することができる。ＧＩＡＮＡプロセスについては、以下にさらに詳細に記載する。

【００４６】

ここで図４を参照して、多次元スケーリング（ＭＤＳ）に基づく等長埋め込みの性能を説明するチャートを示す。ＧＩＡＮＡは、ＭＤＳを用いるＢＬＯＳＵＭ６２行列の等長埋め込みへの近似解によって開始され、これはタンパク質を構成する２０個のアミノ酸のそれぞれに対するベクトルを生成することができる。それぞれのアミノ酸は数値ベクトルによって表すことができる。一例では、２０種全てのベクトルを、Ｐｙｔｈｏｎで入手可能な非計量多次元スケーリングアルゴリズムを用いて計算してよい。アミノ酸のそれぞれの対の間のユークリッド距離を計算し、全部で１９０対が得られる。１９０個全ての距離（二乗された）を、ＢＬＯＳＵＭ６２行列中の対応するスコアと目視で比較してよい。二乗された距離を、対応する変換されたＢＬＯＳＵＭ６２非類似性スコアと比較してよい（４－ＢＬＯＳＵＭ６２スコア、対角値を０とする）。例えば、アミノ酸ＷおよびＦは、ＢＬＯＳＵＭ６２スコア　１を有し得る。等長埋め込みベクトルの計算によるそれらの距離は、ほぼ２．３であり得る。したがって、点（２．３，１）が、図４に示す散布図に表示され得る。

【００４７】

スピアマン相関を計算して、２つの測定の間の類似性を評価することができる。続いて、ＣＤＲ３ストリングをＭＤＳベクトル上の連続的で非交換的な線形変換としてモデル化し、高次元空間の座標として表現してよい。ユニタリー変換行列は、ＣＤＲ３配列の典型的な長さに関連する十分に大きな任意の次数の巡回群、例えば６次の巡回群（G_6）の要素であってよく、これはストリングの対のユークリッド距離とそれらのアラインメントスコアとの間のほぼ完全な線形相関を生じ得る。

【００４８】

ここで図５Ａ～５Ｆを参照して、ＣＤＲ３ストリングについてのＧ６によってエンコードされる等長距離とスミス－ウォーターマンアラインメントスコアとの比較を説明するチャートを示す。デフォルトの等長距離カットオフ（－ｔ）を１０として、高いスミス－ウォーターマンアラインメントスコアを有する全てのＴＣＲ対は、下流のクラスターに含まれている。図５Ａ～５Ｆは、それぞれ１２～１７個のアミノ酸の長さを有するＣＤＲ３についての解析を示す。それぞれのチャートにおいて、ｘ軸はＣＤＲ３の対の間の等長距離（例えばユークリッド距離の二乗）を表し、ｙ軸は置換行列としてＢＬＯＳＵＭ６２を用いた対応するスミス－ウォーターマンアラインメントスコアを表すことができる。等長距離は、Ｇ６エンコーディングの後の数値ベクトルの対の間のユークリッド距離の二乗と定義してよい。図５Ａ～５Ｆは、異なる長さのカテゴリー（即ち１２～１７アミノ酸）に分割した１０，０００個のＣＤＲ３配列の解析を示す。それぞれの長さのカテゴリーにおける全ての配列についての数値ベクトル表現が得られ、ペアワイズ距離が計算される。ＣＤＲ３配列のそれぞれの対の配列類似性を、古典的なスミス－ウォーターマンアラインメントアルゴリズムを用いて評価してよく、これはアミノ酸の置換行列（例えばＢＬＯＳＵＭ６２）に依拠している。高いアラインメントスコアは、高い配列類似性を示す。２つの同一の配列について、最大スコアの４＊長さに到達する。高いアラインメントスコアは高い類似性と関係し、これは短い距離に対応するので、スピアマンの相関値はマイナスとして示される。これは、図４で用いた非類似性スコアとは異なる。

【００４９】

高い計算効率をもってＣＤＲ３プレクラスター（即ち、高い類似性および共通と想定される抗原特異性を有するＴＣＲ）を同定するために、索引に基づく高速の最近傍探索および再帰的セントロイドグループ化を座標について実施してよい。最近傍探索の方法には、１つまたは複数の従来の方法、例えばフェイスブックＡＩ類似性探索（ＦＡＩＳＳ）、ナビガブルスモールワールド（ＮＳＷ）、階層的ナビガブルスモールワールド（ＨＮＳＷ）、ＰｙＮＮＤｅｓｃｅｎｔ、およびアノイが含まれ得る。最近傍探索のために用いられるＴＣＲ非類似性の尺度には、スミス－ウォーターマン距離、高次元ユークリッド空間への埋め込み、または２つのＴＣＲの共通の抗原特異性を推定するために用いられる他の任意の距離または非類似性の計量の１つまたは複数が含まれ得る。続いてＣＤＲ３プレクラスターを、ｋマーにガイドされる探索テーブルを用いて、一致したＴＲＢＶアレルおよび高いスミス－ウォーターマンアラインメントスコアについてフィルタリングし、最終のＴＣＲクラスターを出力として産生してよい。

【００５０】

ここで図６を参照して、ＧＩＡＮＡワークフローのグラフィックな説明を示す。ステップ６０２において、ＧＩＡＮＡプロセスは、短いＣＤＲ３ペプチド配列を一連のユニタリー変換によって数値ベクトルにエンコードすることによって開始することができる。以下にさらに詳細に記載するように、変換は６次の巡回群の要素を含み得る。ステップ６０４において、エンコードされたそれぞれのＣＤＲ３配列が高次元ユークリッド空間に投影され得る。ステップ６０６において、高速の最近傍探索が実施され得る。ステップ６０８において、反復セントロイドクラスター化が実施され得る。ステップ６１０において、ＴＲＢＶ遺伝子アレルを一致させ、低いアラインメントスコアの対を除去するように、フィルタリングステップが実施され得る。ステップ６１２において、最終のＴＣＲクラスターが出力され得る。

【００５１】

ステップ６０２において実施されるＧ$_6$変換に対してさらにまたはその代わりに、入力ＣＤＲ３ストリングの座標として積み上げＭＤＳベクトルを用いる同様な方法（ＧＩＡＮＡｓｖ）を用いてよい。

【００５２】

ここで図７を参照して、ＧＩＡＮＡｓｖワークフローのグラフィックな説明を示す。ステップ７０２において、入力配列は、ストリング中の全てのアミノ酸の連結されたベクトルとしてエンコードされ得る。入力配列のエンコーディングは別として、他のプロセッシ

ングステップはＧＩＡＮＡと同様でよい。例えば、ステップ７０４において、エンコードされたそれぞれのＣＤＲ３配列が高次元ユークリッド空間に投影され、高速の最近傍探索が実施され得る。ステップ７０６において、反復セントロイドクラスター化が実施され得る。ステップ７０８において、ＴＲＢＶ遺伝子アレルを一致させ、低いアラインメントスコアの対を除去するように、フィルタリングステップが実施され得る。ステップ７１０において、最終のＴＣＲクラスターが出力され得る。ＧＩＡＮＡおよびＧＩＡＮＡｓｖプロセスを、本明細書でさらに詳細に記載する。

【００５３】

一例では、ＧＩＡＮＡプロセスは、参照ＴＣＲ－ｓｅｑデータから抗原特異的ＣＤＲ３配列を同定し分類するために用いられる。ＴＣＲレパートリーシーケンシング試料は、１つまたは複数のデータベース、例えばＡｄａｐｔｉｖｅ　Ｂｉｏｔｅｃｈｎｏｌｏｇｙによって提供されるｉｍｍｕｎｅＡＣＣＥＳＳデータベースからアクセスでき、これは現在のところＴＣＲ－ｓｅｑ試料の最大のデータベースであり、全てｉｍｍｕｎｏＳＥＱプラットフォームを用いてプロファイリングされている。抗原特異的ＴＣＲおよび一致した抗原は、ＶＤＪｄｂ、免疫エピトープデータベースおよび解析リソース（Ｉｍｍｕｎｅ　Ｅｐｉｔｏｐｅ　Ｄａｔａｂａｓｅ　ａｎｄ　Ａｎａｌｙｓｉｓ　Ｒｅｓｏｕｒｃｅ）（ＩＥＤＢ）、および以前の文献からプールしてよい。２つ以上のエピトープに特異的なＴＣＲは、不一致を避けるために参照ＴＣＲ－ｓｅｑデータから除外してよい。

【００５４】

ＣＤＲ３配列の等長埋め込みのための数学的フレームワークを用いて、任意の短いペプチド配列ｓの数値的表現（高次元空間の座標でもある）ｘを見出し、それにより、s_iおよびs_jについて２つの座標x_iおよびx_jの間のユークリッド距離、$||x_i - x_j||$が想定上の進化的置換行列によって測定した配列類似性スコアと完全に相関するようにしてよい。

【００５５】

この問題は、「短い配列の等長埋め込み」と称してよい。この概念は、典型的には１２～１７アミノ酸の範囲の長さを有するＣＤＲ３配列の数値的エンコーディングの課題を解決するために導入される。所与のＣＤＲ３配列の数学的変換は、等長性をほぼ満足するように見出され得る。第１に、以下に記載するように、ＢＬＯＳＵＭ６２行列についてほぼ等長の埋め込みが見出され得る。

【００５６】

アミノ酸A_iを実空間

【００５７】

【数１】

$$\mathbb{R}^r$$

における数値ベクトルβ_iによって表す。実空間ｒの次元は、ユークリッド距離行列（ＥＤＭ）のランクによって決定される。このシナリオでは、ＭはＢＬＯＳＵＭ６２によって誘導される非類似性行列：Ｍ＝４－ＢＬＯＳＵＭ６２を意味するとし、Ｍの対角値を０と設定する。等長性は、

$||\beta_i - \beta_j|| x_i = M_{ij}$（式１）

を示す。

【００５８】

この問題に対する解は、ＥＤＭが平坦である場合かつその場合にのみ、また埋め込み空間がｎより大きくない次元を有する場合に存在し、ここでｎはＥＤＭの次元である。残念ながら、ＢＬＯＳＵＭ６２行列は、三角ルール：

$\forall i, j, k$について$M_{ik} + M_{kj} \geqq M_{ij}$（式２）

を満たさないので、ＥＤＭではない。
【００５９】
したがって、ＢＬＯＳＵＭ６２の正確な等長埋め込みは存在しないかもしれない。しかし、ＭＤＳはＭがＥＤＭでない場合に適用される近似的な解を提供し得る。ＭＤＳは埋め込みベクトルβ_iを誘導するために用いてよい。古典的なＭＤＳの場合には、埋め込み空間のための最大次元は１３である。１３より高い次元性を探索するために、Ｐｙｔｈｏｎにおけるスクラーンパッケージを用いる非計量ＭＤＳ計算を適用してよい。埋め込み等長性を最大化するため、長さ１４の２，３００個の訓練ＴＣＲをＴＣＧＡデータセットから選択してよく、ペアワイズＳＷアラインメントスコアを計算してよい。１３から１９の範囲の様々な次元の等長埋め込みベクトルを得るためにＭＤＳを適用してよい。それぞれの長さについて、ＧＩＡＮＡ法において記載したように、ＣＤＲ３配列のユークリッド座標を計算してよい。ペアワイズ距離をＳＷスコアと比較してよい。最大スコアは次元１６で観察された。これは等長表現における最適の次元であろう。この表現によって、ＢＬＯＳＵＭ行列：
$||\beta_i - \beta_j||$　M_{ij}（式３）
に対してほぼ８７％の類似性が達成され得る。
【００６０】
次に、数値エンコーディングスキームを導入してよく、それにより、ＣＤＲ３配列中のそれぞれのアミノ酸は、量子物理学における概念に例えられる「オペレーター」と考えてよい。一般に、オペレーターＡは、既存の波動関数Φに対する数学的変換であってよい。操作は、ディラックブラケット$A|\Phi\rangle$によって示される波動関数に適用してよい。１つの例は角度モーメントオペレーターL_x、L_y、L_zである。アミノ酸ｉについてのオペレーターはA_iと定義され、これは以下の方式
【００６１】
【数２】

$$A_i x \coloneqq \Omega \times (x + \beta_i) \qquad (式4)$$

で数値ベクトルｘに適用される。
【００６２】
ここでΩは決定する必要がある行列である。オペレーターは非互換性（$i \neq j$であれば$A_i A_j \neq A_j A_i$）であるので、この定義は配列の中の文字の順序付けを強調している。次いで、ＣＤＲ３配列は、ある種の初期ベクトルβ_0における１つまたは複数の連続的線形操作とみなしてよい。計算を簡単にするため、$\beta_0 = 0$とする。したがって、最も右側のアミノ酸における操作の後では、座標は
【００６３】
【数３】

$$A_i\beta_0 \coloneqq \Omega \times (\beta_i + \mathbf{0}) = \Omega \times \beta_i \qquad (式5)$$

となる。
【００６４】
Ωの望ましい性質を説明するいくつかの例を以下に記載する。
【００６５】
第１の例では、２つのアミノ酸配列が１つのアミノ酸によってオフになってよい（例え

ば単一のミスマッチ）。例えば、配列$s_1 = A_k A_i$、かつ配列$s_2 = A_k A_j$である。これらの数値エンコーディングベクトルは、以下のように計算することができる。

$s_1 = A_k A_i$（式６）

$s_2 = A_k A_j$（式７）

$x_1 = \Omega \times (\Omega \times \beta_i + \beta_k)$（式８）

$x_2 = \Omega \times (\Omega \times \beta_j + \beta_k)$（式９）

【００６６】

ここでx_1およびx_2はs_1およびs_2のエンコーディングベクトルである。s_1とs_2の間のユークリッド距離は、

$||x_1 - x_2|| = \delta^T \delta$、ここで$\delta = x_1 - x_2$（式１０）

$= (\Omega^2 (\beta_i - \beta_j))^T (\Omega^2 (\beta_i - \beta_j))$（式１１）

$= (\beta_i - \beta_j)^T \Omega^T \Omega^T \Omega \Omega (\beta_i - \beta_j)$（式１２）

によって計算してよい。

【００６７】

これが唯一のミスマッチであるので、上の値はアミノ酸A_iとアミノ酸A_jとの間の距離に等しくてよい。即ち、

$(\beta_i - \beta_j)^T \Omega^T \Omega^T \Omega \Omega (\beta_i - \beta_j) = (\beta_i - \beta_j)^T (\beta_i - \beta_j) \quad M_{ij}$（式１３）

【００６８】

一般性を失うことなく、$\Omega^T \Omega = I$である。換言すれば、Ωはユニタリー行列であってよい。１つのミスマッチを有するより長い配列は、上と同じパターンに従うことが容易に示される。

【００６９】

第２の例では、２つのアミノ酸配列が連続する２つのアミノ酸によってオフになってよい。変数s_1は$s_1 = A_i A_j$によって定義してよく、$s_2 = A_t A_k$である。埋め込みベクトルx_1とx_2との間の距離は、

$||x_1 - x_2|| = (\Omega^2 (\beta_j - \beta_k) + \Omega (\beta_i - \beta_t))^T (\Omega^2 (\beta_j - \beta_k) + \Omega (\beta_i - \beta_t))$（式１４）

$= (\beta_j - \beta_k)^T (\beta_j - \beta_k) + (\beta_i - \beta_t)^T (\beta_i - \beta_t) - 2 (\beta_i - \beta_t)^T \Omega (\beta_j - \beta_k)$（式１５）

$M_{jk} + M_{it} - 2 (\beta_i - \beta_t)^T \Omega (\beta_j - \beta_k)$（式１６）

であることが示される。

【００７０】

好ましくは、第３項は$\forall i, j, t, k$についてゼロであってよい。１つの解は、最初のｒ次元空間から相補空間への垂直回転を課すことによってΩを

【００７１】

【数４】

$\mathbb{R}^{2r}$

における回転とすることでよい。単純な具現化は、

【００７２】

【数５】

$$\Omega_2 = \begin{pmatrix} \underline{0} & I \\ I & \underline{0} \end{pmatrix} \quad (式\ 16)$$

であってよい。
【００７３】
ここでＩはｒ次元の恒等行列であってよく、０はｒ次元のゼロ行列であってよい。実際上、このようにして定義されるΩは、位数２の巡回群G_2の表現であってよい。G_2はただ２つの要素、即ちｅおよびｇを有してよく、ここで$g^2 = e$である。この表記法は、多数の連続的なミスマッチの例で有用であり得る。したがってβ_iは
【００７４】
【数６】
$\mathbb{R}^{2r}$

に拡張され、最初のｒ次元がＭＤＳ埋め込みから誘導される値で満たされ、残りの次元がゼロのベクトル：
【００７５】
【数７】

$$\beta_i^{2r} = \begin{pmatrix} \beta_i \\ \mathbf{0} \end{pmatrix} \qquad (式\ 17)$$

で満たされる。
【００７６】
ここで０は次元ｒを有するゼロのベクトルであってよい。新たなベクトルは、
【００７７】
【数８】

$$\beta_i^{2r^T} \boldsymbol{\Omega}_2 \beta_j^{2r} = (\beta_i^T, \mathbf{0}^T) \begin{pmatrix} \mathbf{0} & \boldsymbol{I} \\ \boldsymbol{I} & \underline{\mathbf{0}} \end{pmatrix} \begin{pmatrix} \beta_j \\ \mathbf{0} \end{pmatrix} = (\beta_i^T, \mathbf{0}^T) \begin{pmatrix} \mathbf{0} \\ \beta_j \end{pmatrix} = 0 \qquad (式\ 18)$$

を満たし得る。
【００７８】
第３の例では、多数の連続的なミスマッチがあり得る。配列
【００７９】
【数９】
$s_i = \boldsymbol{A}_{i_k} \boldsymbol{A}_{i_{k-1}} \ldots \boldsymbol{A}_{i_1}, k \geq 3$

について、そのエンコーディングベクトルは、
【００８０】
【数１０】

$$x_i = \sum_{l=1}^{k} \boldsymbol{\Omega}^l \beta_{i_l}, l = 1,2,\dots,k \qquad (式\ 19)$$

として記されることが証明され得る。

【００８１】

別の配列、

【００８２】

【数１１】

$$s_j = A_{j_k} A_{j_{k-1}} \dots A_{j_1}$$

：

【００８３】

【数１２】

$$x_j = \sum_{l=1}^{k} \boldsymbol{\Omega}^l \beta_{j_l}, l = 1,2,\dots,k \qquad (式\ 20)$$

を考慮する。

【００８４】

x_i と x_j の距離は、

【００８５】

【数１３】

$$\| x_i - x_j \| = \sum_{l=1}^{k} (\beta_{i_l} - \beta_{j_l})^T (\beta_{i_l} - \beta_{j_l}) - 2 \sum_{u=2}^{k} \sum_{1 \le v < u}^{k} (\beta_{i_u} - \beta_{j_u})^T (\boldsymbol{\Omega}^u)^T (\boldsymbol{\Omega}^v)(\beta_{i_v} - \beta_{j_v}) \qquad (式\ 21)$$

となり得る。

【００８６】

理想的なシナリオでは、二重Σの中の全ての項はゼロであってよく、s_i と s_j の間の距離は

【００８７】

【数１４】

$$\sum_{l=1}^{k} M_{i_l j_l}$$

であってよい。これは

【００８８】

【数１５】

$$(\beta_{i_u} - \beta_{j_u})^T \boldsymbol{\Omega}^{u-v} (\beta_{i_v} - \beta_{j_v}) \qquad (式\ 22)$$

を必要とし、
【００８９】
∀ｕ，ｖ，ｋについて０である。または一般に
【００９０】
【数１６】
$\forall x, y \in \mathbb{R}^r$に対して$x^T \boldsymbol{\Omega}^{u-v} y = 0$ (式23)

である。
【００９１】
【数１７】
$\mathbb{R}^r$

におけるそのようなΩについての解はないかもしれないが、第２の例と同様に、埋め込み空間の次元性はｋｒに増大してよい。このようにして、位数ｎの巡回群G_nからΩを構築することができ、これはアーベル群である。しかし、次元性の増大は０（ｋ）の係数でエンコーディングステップにおける計算の複雑性を増大させる可能性がある。また、正確な解をもってしても、ＭＤＳ埋め込みによって計算される距離は、ＢＬＯＳＵＭ６２スコアと完全には整列しない可能性がある。したがって、次元性の増大にはトレードオフがあり得る。実際上、ｋは６に設定してよく、これはＴ細胞のＣＤＲ３配列のメディアン長さが１４であり、最初の４アミノ酸と最後の１アミノ酸がほとんど不変であることを考慮すれば妥当な数である。対応する行列を構築するため、G_3における要素の表現は、
【００９２】
【数１８】

$$\boldsymbol{\Omega}_3 = \begin{pmatrix} \underline{\boldsymbol{0}} & \underline{\boldsymbol{0}} & \boldsymbol{I} \\ \boldsymbol{I} & \underline{\boldsymbol{0}} & \underline{\boldsymbol{0}} \\ \underline{\boldsymbol{0}} & \boldsymbol{I} & \underline{\boldsymbol{0}} \end{pmatrix} \quad \text{(式24)}$$

によって誘導することができる。
【００９３】
G_2とG_3の両方はG_6の正常なサブグループであってよく、
【００９４】
【数１９】
$G_6 = G_2 \otimes G_3$

である。したがって、Ω_2およびΩ_3からΩ_6を構築することができる。
【００９５】
【数２０】

$$\Omega_6 = \begin{pmatrix} \underline{\underline{0}} & \underline{\underline{0}} & \Omega_2 \\ \Omega_2 & \underline{\underline{0}} & \underline{\underline{0}} \\ \underline{\underline{0}} & \Omega_2 & \underline{\underline{0}} \end{pmatrix} \quad (式25)$$

【００９６】
ここで
【００９７】
【数２１】

$$\underline{\underline{0}}$$

は次元２ｒのゼロ行列である。したがって、ＭＤＳ埋め込みベクトルは、
【００９８】
【数２２】

$$\beta_i^{6r} = (\beta_i^T, \mathbf{0}^T, \mathbf{0}^T, \mathbf{0}^T, \mathbf{0}^T, \mathbf{0}^T)^T \quad (式26)$$

となり得る。
【００９９】
この表現において、二重Σの中の項は、ｕ－ｖ≦６の場合に０となり得る。ｕ－ｖ＞６（即ち、２つのストリングが６を超える連続的ミスマッチを有する）の場合、変換行列としてのΩ$_6$の適用は最終の距離に望ましくない変動を導入する可能性がある。行列のそれぞれの側におけるベクトルに応じて、付加はポジティブまたはネガティブになり得る。しかし、６を超えるミスマッチを有するＣＤＲ３配列を比較する場合、何がそれらの間の正確な距離であるかは通常重要でない。それは、最大の類似性を有する配列のみが抗原特異的ＴＣＲクラスターとして選択され、アラインメントスコアの望ましいカットオフにおいて２つのＣＤＲ３配列の間のミスマッチの数は通常３より小さいからである。
【０１００】
第４の例では、多数の非連続的ミスマッチがあり得る。目的の２つの配列、即ちいずれも長さｋを有するｓ$_i$およびｓ$_j$が最初および最後の位置において異なり、
【０１０１】
【数２３】

$$s_i = A_{i_k} A_{i_{k-1}} \ldots A_{i_2} A_{i_1}, k \geq 3$$

および
【０１０２】
【数２４】

$$s_j = A_{j_k} A_{i_{k-1}} \ldots A_{i_2} A_{j_1}, k \geq 3$$

であると仮定する。変換後の等長座標は、

【０１０３】

【数２５】

$$x_i = \sum_{l=1}^{k} \Omega^l \beta_{i_l}, l = 1,2,\dots,k \qquad (式27)$$

$$x_j = \sum_{l=2}^{k-1} \Omega^l \beta_{j_l} + \Omega^k \beta_{j_k} + \Omega \beta_{j_1} \qquad (式28)$$

である。

【０１０４】

それらの距離は、

【０１０５】

【数２６】

$$\| x_i - x_j \| = (\beta_{i_l} - \beta_{j_l})^T (\beta_{i_l} - \beta_{j_l}) + (\beta_{i_k} - \beta_{j_k})^T (\beta_{i_k} - \beta_{j_k}) - 2(\beta_{i_1} - \beta_{j_1})^T (\Omega)^T (\Omega^k)(\beta_{i_k} - \beta_{j_k})$$

(式29)

によって計算され得る。

【０１０６】

これは、

【０１０７】

【数２７】

$$-2(\beta_{i_1} - \beta_{j_1})^T (\Omega^{k-1})(\beta_{i_k} - \beta_{j_k}) \qquad (式30)$$

と書くことができる交差項の数が少ないことを除いて、第３の例（即ち、多数の連続的なミスマッチ）と同様であり得る。

【０１０８】

第３の例と同様にΩ_6としてΩを選択すれば、ＮＩＣが観察される限り、交差項は０になり得る。しかし、ＮＩＣが乱されれば（即ち２つのミスマッチが正確に６アミノ酸だけ離れれば）交差項はノンゼロになり得る。この項は最終結果に影響を及ぼし得る。第１に、交差項が負のままであれば（その確率は１／２）、推定される等長距離は正確な値より小さくなり得る。高い配列類似性を保証するために厳格なスミス－ウォーターマンアラインメントが適用され得るので、これは結果に影響しないかもしれない。長さ１６で最初の３アミノ酸および最後の２アミノ酸が切り取られているＣＤＲ３について、２つのミスマッチがあると仮定して、正確に６アミノ酸だけ離れた２つのミスマッチを有する可能性は、

【０１０９】

【数２８】

$$\frac{5}{\binom{11}{2}} = 0.091$$

であることが示される。これは全ての長さの中で最大の確率であり得る。

【０１１０】

したがって、ＮＩＣの乱れは、２つのミスマッチを有する比較の多くとも０．０９１／２＝４．６％に影響する可能性がある。これが起こった場合、いくらか同様の配列が大き

な距離を有し、下流のクラスター化から除外される可能性がある。この効果を軽減するため、比較的大きなデフォルト等長距離カットオフ（－ｔ　１０）を包括的であるように適用してよい。パラメーター設定の現在の選択は、クラスター化の正確度と計算速度とのバランスである。

【０１１１】

ＣＤＲ３配列の近似的な等長埋め込みは、高速クラスター化のためのユークリッド空間におけるそれらの最近傍（ＮＮ）の効率的な探索を可能にする。ＮＮ探索を実施するために、機械学習に基づく１つまたは複数の分類手法を用いてよい。

【０１１２】

当業者には理解されるように、機械学習に基づく分類手法は、開示した技術から逸脱することなく、所望の実施に応じて変動してよい。例えば、機械学習分類スキームは、以下の、隠れマルコフモデル、回帰性ニューラルネットワーク、畳み込みニューラルネットワーク、ベイジアン記号法、一般敵対的ネットワーク、サポートベクトルマシン、イメージ登録法、適用可能な規則に基づくシステムの単独または組合せの１つまたは複数を利用することができる。回帰アルゴリズムを用いる場合、これらは、それだけに限らないが、確率的勾配低下リグレッサー、および／または受動攻撃的リグレッサー、その他を含んでよい。

【０１１３】

機械学習分類モデルは、クラスター化アルゴリズム（例えばミニバッチＫ手段クラスター化アルゴリズム）、リコメンデーションアルゴリズム（例えばミニワイズハッシングアルゴリズムまたはユークリッドＬＳＨアルゴリズム）、および／または異常検出アルゴリズム、例えば局所アウトライアー因子に基づいてもよい。さらに、機械学習モデルは、次元縮退アプローチ、例えばミニバッチディクショナリー学習アルゴリズム、インクレメンタルプリンシパルコンポーネント解析（ＰＣＡ）アルゴリズム、潜在的ディリクレ配分アルゴリズム、および／またはミニバッチＫ手段アルゴリズム等の１つまたは複数を採用することができる。

【０１１４】

一例では、ＦＡＩＳＳ等のパイソンパッケージを用いて、高速の索引されたＮＮ探索を実施してよい。

【０１１５】

【数２９】

$$\mathbb{R}^r$$

におけるＮ数値ベクトルの１つの最近傍を見出すため、ＦＡＩＳＳの時間計算量は０（ｒｌｏｇ（Ｎ））であってよい。

【０１１６】

ＣＤＲ３の座標（ｘ）は近傍クラスターに分割してよい。クラスター化の前に、同一のＣＤＲ３をともにグループ化してよい。最初に、それぞれの特有の配列ｘ$_i$、ｉ＝１、２、・・・、Ｎについて、その最近傍ｘ$_j$、ｊ＝１、２、・・・、Ｎ；ｊ≠ｉを位置付けされ得る。ｘ$_i$とｘ$_j$の間の距離がユーザーによって定義されたカットオフ（－ｔオプション、ｔｈｒ）以内であれば、２つの点は重心

【０１１７】

【数３０】

$$\frac{x_i+x_j}{2}$$

を新たな座標とする新しい点として併合してよい。距離がカットオフを超えれば、両方の点を反復から除去してよい。除去された点には少なくとも２つの型、即ち、１）ただ１つのＣＤＲ３配列を含む点、および２）多数のＣＤＲ３の重心としての点があり得る。第２の型の点のそれぞれについて、ＣＤＲ３プレクラスターを記録してよい。点の数がゼロに達するか、さらに減少しなくなるまで、上記のステップを繰り返してよい。異なる長さのＣＤＲ３は別々にクラスター化してよい。全てのプレクラスターは、さらなるフィルタリングのために保存してよい。

【０１１８】

ＴＣＲ可変遺伝子マッチングを伴うＫマーによってガイドされる高速スミス－ウォーターマンアラインメントを、ＣＤＲ３プレクラスターについて実施してよい。プレクラスター由来のＣＤＲ３は高度に類似し得るが、これらは抗原特異的グループとして適格であるとは限らない。それは、１）等長埋め込みが不完全なために配列が十分に類似していない、および／または２）ＴＲＢＶ遺伝子情報が考慮されていないためである。したがって、スミス－ウォーターマンアラインメントおよびＴＲＢＶ遺伝子マッチングに基づいて抗原特異的ＣＤＲ３クラスターを選択するためにフィルタリングステップを実施してよい。

【０１１９】

プレクラスターのサイズ（ｍ）は大きく、従来の直接ペアワイズ比較は二次時間計算量Ｏ（m^2）をもたらすことがある。クラスターサイズを小さくするために、ＴＲＢＶ情報を適用してよい。具体的には、ＴＲＢＶアレルの対の間でアラインメントスコアの事前計算行列を用いてよい。プレクラスター中のＣＤＲ３配列のそれぞれの対について、そのＴＲＢＶアレルを比較してよい。比較スコアがユーザーによる定義（－Ｇオプション、ｔｈｒ＿ｖ）を超えれば、２つの配列の間にエッジを付加してよい。最終グラフについて深度優先探索（ＤＦＳ）を実施して、それぞれのサブグラフが新たなプレクラスターとなる単離されたサブグラフを生成してよい。このステップによって、元のプレクラスターがいくつかの小さなプレクラスターに分割され得る。

【０１２０】

次に、ｋマーアプローチを用いてスミス－ウォーターマンアラインメントを実施してよい。それぞれのＣＤＲ３配列は、連続する５マーに分割され得る。全ての配列を記憶するために、キーを特有の５マーとし、値を所与の５マーを含むＣＤＲ３であるとして、ｋマーディクショナリー（例えばデータベース１０７の中の）を構築してよい。ディクショナリーを構築する際には５マーの中で１つのミスマッチは容認され得る。例えば、配列ＣＡＳＳＧＶＴＥＡＦＦはＳＳＧＶＴとＳＳＶＡＴの両方の下で索引付けされる。このようにして、ＣＤＲ３配列は共有されたｋマーを介してグラフに連結することができる。このグラフの中のそれぞれのエッジに対して、スミス－ウォーターマンアラインメントがＢＬＯＳＵＭ６２置換行列で実施され、アラインメントスコアが計算され得る。スコアがユーザー定義のカットオフ（－Ｓオプション、ｔｈｒ＿ｓ）未満であれば、エッジを除去してよい。このステップの実際の計算量はＯ（ｍ）からＯ（m^2）まで変動し得る。最悪のシナリオは、プレクラスター中のＣＤＲ３のあらゆる対が類似のｋマーモチーフを共有した場合に到達され得る。最終のグラフについてＤＦＳを実施し、最終ＣＤＲ３クラスターを生成して、これらを最終出力として報告してよい。

【０１２１】

一例では、既存の参照ＴＣＲ－ｓｅｑデータの最終のＣＤＲ３クラスターに対して、新たなＴＣＲ－ｓｅｑ試料を探索要求してよい。入力データセットのＴＣＲクラスターの生成の後で、ＧＩＡＮＡはこのデータ（参照）へのさらなるＴＣＲの探索要求（クエリー）を実施することができる。クエリーモードにおいて、ＧＩＡＮＡはクエリーファイル、参照データ、およびクラスター化した参照データの１つまたは複数を解析することができる。

【０１２２】

最初に、参照ＴＣＲ－ｓｅｑデータおよびクエリーＴＣＲ－ｓｅｑデータを、上記のように等長座標に変換してよい。次いで高速の最近傍探索（例えばＦＡＩＳＳによる）を実

行するが、クエリーＴＣＲ－ｓｅｑデータに限定してよい。ユーザー定義のカットオフ（－ｔオプション、ｔｈｒ）より短い距離を有するＴＣＲは、個別のファイル（ｔｍｐ＿ｑｕｅｒｙ．ｔｘｔ）にエクスポートしてよい。このファイルは、クエリー配列とおそらくクラスター化できる全てのＴＣＲを含み得る。このファイルについてＧＩＡＮＡクラスター化を実施し、スミス－ウォーターマンアラインメントのための厳格なカットオフを満足するＴＣＲクラスターを生成してよい。次いでクエリーＴＣＲクラスターを以下のようにして参照クラスターと併合してよい。それぞれのクエリークラスターについて、いずれかの配列が参照データ中の既存のクラスターに由来していれば、２つのクラスターを併合してよい。このステップは、参照データ中に全ての近傍ＴＣＲを含めることを保証するためである。クエリークラスターが参照クラスター中にいずれの配列も含まない２つ以上の条件、即ち、１）クエリークラスター中の全てのＴＣＲが類似していたが、クエリー試料に限定されていた、および／または２）クエリーＴＣＲが、元のクラスター化において他のいずれの参照試料ともクラスター化しない極めて稀ないくつかの参照ＴＣＲと類似していた、があり得る。いずれかの条件に従って、クエリークラスターは最終出力に含まれ得る。

【０１２３】

一例では、クエリーモードの時間コストは、２００Ｋ、１Ｍ、２Ｍ、６Ｍ、および１０ＭのＴＣＲを含む参照データを生成することによって評価してよい。１０Ｋ、２０Ｋ、３０Ｋ、４０Ｋ、および５０ＫのＴＣＲを含む異なるサイズのクエリーデータをスキャンしてよい。それぞれのクエリーファイルは、参照データのそれぞれに対して、例えば汎用コンピューターを用いてクラスター化してよい。経過時間は、パイソンの時間モジュールを用いて推定してよい。

【０１２４】

ＧＩＡＮＡｓｖプロセスでは、２０アミノ酸のＭＤＳ埋め込みの後でＣＤＲ３ストリングの等長表現：$s = A_1A_2 \cdots A_k$、$k \geqq 5$を得る最も容易な方法は、「積み上げベクトル」を構築する（即ち、埋め込みベクトルβ_i、ｉ＝１、２、・・・、ｋを同じ順序で連結する）ことであってよい。積み上げベクトル表現は、

【０１２５】

【数３１】

$$x = (\beta_1^T, \beta_2^T, \ldots, \beta_k^T)^T$$

であってよい。この表現は、上記の３つの例の望ましい品質を満足し得る。６つ以下のミスマッチを有する配列のみに注目する場合には、２つのアプローチは実質的に同一であり得る。ＣＤＲ３が６を超えるミスマッチを有する場合には、ＧＩＡＮＡｓｖはより正確になり得る。しかし、ＧＩＡＮＡｓｖでは、埋め込みベクトルの次元$r_{GIANAsv}$はＧＩＡＮＡの次元（r_{GIANA}）より大きい可能性がある。ＧＩＡＮＡについては、$r_{GIANA} \equiv 6 \times 16 = 96$である。ＧＩＡＮＡｓｖについては、$r_{GIANAsv}$はＣＤＲ３の様々な長さ（典型的には１２～１７アミノ酸）とともに変動し、r_{GIANA}の２～３倍大きくなり得る。次元性の増大は、メモリー負荷の増大および計算時間の長期化をもたらすことがある。

【０１２６】

上記のＧＩＡＮＡおよびＧＩＡＮＡｓｖプロセスは、従来のＴＣＲクラスター化法（例えばｉＳＭＡＲＴ、ＧＬＩＰＨ２、およびＴＣＲｄｉｓｔ）に対していくつかの改善を提供する。例えば、ＧＩＡＮＡおよびＧＩＡＮＡｓｖプロセスは、大きなＴＣＲデータセットを処理でき、より正確なデータを提供できるだけでなく、これらの結果を生成するために必要なコンピューターのリソース量を低減する。本明細書に記載したＧＩＡＮＡおよびＧＩＡＮＡｓｖ法の改善を実証するために、健常ドナーのＴＣＲレパートリーシーケンシ

ングデータを用いる比較を用いてよい。比較において、ＴＣＲクローンを、その存在量に基づいて順序付けてよく、上位１０Ｋ、２０Ｋ、・・・、１００Ｋの配列を選択してよい。５つ全ての方法をサブ試料のそれぞれに適用してよい。ＧＩＡＮＡ、ＧＩＡＮＡｓｖ、ｉＳＭＡＲＴ、およびＧＬＩＰＨ２は、デフォルトパラメーターを用いて実行してよい。ＴＣＲｄｉｓｔはクラスター化を提供せず、ペアワイズ距離のみが計算され得る。

【０１２７】

ここで図８および図９を参照して、従来の方法に対するＧＩＡＮＡおよびＧＩＡＮＡｓｖの性能を説明するチャートを示す。図８は、様々なＴＣＲクラスター化アルゴリズムについての時間計算量の比較を示す。チャートは、解析したＴＣＲ配列の全数（１０ｋの増分で）をｘ軸に、全計算時間（秒）をｙ軸に示している。線８０２はＴＣＲｄｉｓｔの性能を示し、線８０４はｉＳＭＡＲＴの性能を示し、線８０６はＧＬＩＰＨ２の性能を示し、線８０８はＧＩＡＮＡｓｖの性能を示し、線８１０はＧＩＡＮＡの性能を示す。スピードアップは、１００ＫのＴＣＲ試料についての時間コストに基づいて計算され得る。

【０１２８】

図８に示すように、ＧＩＡＮＡ（線８１０）はベンチマークを通して最小の時間コストを有し、１００Ｋの配列を処理するために要した時間は２３．９秒であった一方、ＴＣＲｄｉｓｔ（線８０２）は１４，３３８秒を要した。ＧＩＡＮＡｓｖ（線８０８）はＧＩＡＮＡ（線８１０）より２．２倍遅い。積み上げベクトルエンコーディングは高次元の等長埋め込み空間をもたらし、最近傍探索の間の時間コストを増大させるので、これは予想される。注目すべきことに、ＧＬＩＰＨ２（線８０６）はＧＩＡＮＡ（線８１０）およびＧＩＡＮＡｓｖ（線８０８）に次いで最も速いアルゴリズムである。これは、ＧＬＩＰＨ２がモチーフに導かれる探索を通じたペアワイズアラインメントを回避するからである。下の表１に、ＧＩＡＮＡ、ＧＩＡＮＡｓｖ、ｉＳＭＡＲＴ、ＴＣＲｄｉｓｔ、およびＧＬＩＰＨ２の計算時間とメモリー消費の比較を示す。一例では、３．５ＧＨｚのＤｕａｌ－Ｃｏｒｅ Ｉｎｔｅｌ Ｃｏｒｅ ｉ７プロセッサーおよび１６ＧＢの２１３３ＭＨｚ ＬＰＤＤＲ３メモリーによってｍａｃＯＳ（登録商標） Ｃａｔａｌｉｎａ ｖ１０．１５．２を走らせるシステムで計算を実施してよい。

【０１２９】

【表１】

	TCR	10K	20K	30K	40K	50K	60K	70K	80K	90K	100K
GIANA	時間(s)	1.5	2.9	4.2	6.3	8.4	10.9	13.7	16.6	20.2	23.9
	MB	87	109	120	137	142	149	161	175	182	194
GIANA sv	時間(s)	2.3	5	8.3	12.2	17	23	29	35.4	44	53.3
	MB	134	212	288	362	394	421	508	540	549	599
iSMART	時間(s)	16.5	88.8	212	409	657	984	1380	1833	2323	2850
	MB	90	152	197	284	320	362	401	439	484	622
TCRdist	時間(s)	145.9	580	1300	2330	3668	5411	7371	9093	11695	14338
	MB	19	21	23	26	27	29	31	34	37	38
GLIPH2	時間(s)	16.7	34.9	51.9	75.1	99.6	127.3	156.7	183	224.2	271.4
	MB	63	92	122	165	183	208	247	270	302	334

表 1:計算時間とメモリー消費の比較

【０１３０】

図９は、時間計算量を評価する際の様々なＴＣＲクラスター化アルゴリズムのメモリー使用量を示す。チャートは、解析したＴＣＲシーケンスの全数（１０ｋの増分で）をｘ軸に、ピークメモリー使用量（メガバイトで）をｙ軸に示す。線９０２はＴＣＲｄｉｓｔの性能を示し、線９０４はｉＳＭＡＲＴの性能を示し、線９０６はＧＬＩＰＨ２の性能を示

し、線９０８はＧＩＡＮＡｓｖの性能を示し、線９１０はＧＩＡＮＡの性能を示す。

【０１３１】

ＧＩＡＮＡおよびＧＩＡＮＡｓｖは、抗原特異的ＴＣＲの予測において、従来の方法より高い正確度も達成し得る。抗原特異性は、ＴＣＲクラスター化の最も望ましい特徴であり得る。パブリックドメインからの６１，３６６の非冗長な既知のＴＣＲ／抗原の対を用い、多様な病原体に由来する９００を超える異なるエピトープをカバーする解析を実施した。それぞれの方法を実施し、出力クラスターにおける全てのＴＣＲがただ１つのエピトープに特異的であれば、それぞれの方法からの出力クラスターを「純粋なクラスター」と表現した。クラスターの純度は、所与のクラスターにおける最も一般的なエピトープに対して特異的なＴＣＲのパーセンテージとして定義してよい。「純粋なクラスター」は、１に等しい純度を有すると定義される。

【０１３２】

ここで図１０および図１１を参照して、従来の方法と比較したＧＩＡＮＡのクラスター化の精度と感度の比較を説明するチャートを示す。図１０はクラスター化の精度をｙ軸に示し、これは出力中の純粋なクラスターのパーセンテージと定義してよい。ＧＩＡＮＡ、ｉＳＭＡＲＴ、ＴＣＲｄｉｓｔ、およびＧＬＩＰＨ２は、それぞれバー１００２、１００４、１００６、および１００８で表している。バー１００２で示すように、ＧＩＡＮＡが全ての方法の中で最高の精度（９３％）を有する一方、ＧＬＩＰＨ２は最低の精度（３５％）を有している。図１１はクラスター化感度をｙ軸に示し、これは全ての純粋なクラスター中のＴＣＲの全数を試験した全てのＴＣＲの数で除したものと定義してよい。ＧＩＡＮＡ、ｉＳＭＡＲＴ、ＴＣＲｄｉｓｔ、およびＧＬＩＰＨ２は、それぞれバー１１０２、１１０４、１１０６、および１１０８で表している。バー１１０２で示すように、ＧＩＡＮＡは最高の感度（２９％）も有している。

【０１３３】

ここで図１２を参照して、４つの方法の間の正規化された相互情報量（ＮＭＩ）の比較を示す。ＴＣＲクラスターの間のＮＭＩおよびエピトープ特異性は、同じ訓練データセットを用いて測定した。全ての方法にわたって同様のＮＭＩレベルが観察されたが、ＧＬＩＰＨ２は最低に留まった。

【０１３４】

下の表２に、ＧＩＡＮＡ、ｉＳＭＡＲＴ、ＴＣＲｄｉｓｔ、およびＧＬＩＰＨ２についての純粋なクラスターの感度およびクラスター化の精度の評価を示す。抗原特異性が既知の全部で６１，３６６のＴＣＲをこの解析に用いた。シングルトンＴＣＲ（エピトープあたりの配列がただ１つ）を除外した後、６０，７００が残っていた。

【０１３５】

【表２】

	GIANA	iSMART	TCRdist	GLIPH2
クラスター化された TCR の数	17,250	16,828	18,383	31,563
クラスターの数	7,586	7,649	7,316	11,945
純粋なクラスターの数	7,289	7,387	7,130	4,333
純粋な TCR の数	16,202	16,096	15,595	11.514
特異性(純粋なクラスター/クラスター)	96.1%	96.6%	97.4%	36.3%
感度(純粋な TCR/TCR の全数)	26.7%	26.5%	25.6%	19.0%

表 2:純粋なクラスターの感度およびクラスター化の精度の評価

【０１３６】
ＧＩＡＮＡ（９６％）、ｉＳＭＡＲＴ（９７％）、およびＴＣＲｄｉｓｔ（９７％）についての分率は同様であるが、ＧＬＩＰＨ２（３６％）については大幅に低い。純粋なクラスターの保持率は、全ての純粋なクラスターにおけるＴＣＲの全数を試験した全てのＴＣＲの数で除したものと定義してよい。ＧＩＡＮＡも、ＧＬＩＰＨ２（１９％）を除く他の方法と同様のレベルの保持率（２７％）を有している。スミス－ウォーターマンアラインメントに依拠する３つの方法（ＧＩＡＮＡ、ｉＳＭＡＲＴ、およびＴＣＲｄｉｓｔ）については、アラインメントスコアカットオフ（ＧＩＡＮＡにおける－Ｓオプション）の範囲の影響を探索した。

【０１３７】
ここで図１３を参照して、パラメーター設定の範囲におけるＧＩＡＮＡの性能を測定する適合率－再現率の曲線を説明するチャートを示す。ｙ軸は適合率を説明し、これはコールの全数の中の真にポジティブなコールの分率として定義される。ｘ軸は再現率を説明し、これは真にポジティブなコールの数を集団中のポジティブな例の数で除したものである。この解析は、３つ全てがＳＷアラインメントに基づくので、ＧＩＡＮＡ、ＴＣＲｄｉｓｔ、およびｉＳＭＡＲＴに適用可能であった。

【０１３８】
０．９５を超える適合率で、３つの方法全ては同様の曲線を共有する。曲線の「肘」形状は、再現率の計算において純粋なクラスターＴＣＲを用いたためである。カットオフを低下させると、異なる抗原由来のＴＣＲがクラスター化する可能性があり、それにより、純粋なクラスターの分率が低下する。カットオフは３．５より３．６の方が、再現率を僅かにのみ低下させる（０．２６８から０．２６７へ）が、適合率をほぼ３％増加させる（０．９３２から０．９６１へ）ので好ましい。したがって、ＧＩＡＮＡにおける－Ｓオプションのデフォルトパラメーターは３．６に設定してよい。

【０１３９】
ここで図１４を参照して、ミスマッチのための置換行列としてＢＬＯＳＵＭ６２を用いたＧＩＡＮＡ、ｉＳＭＡＲＴ、ＴＣＲｄｉｓｔ、およびミスマッチのための置換行列としてＢＬＯＳＵＭ５０を用いてＧＩＡＮＡの性能を比較する適合率－再現率の曲線を説明するチャートを示す。図１３と同じく、ｙ軸は適合率を説明し、これはコールの全数の中の真にポジティブなコールの分率として定義される。ｘ軸は再現率を説明し、これは真にポジティブなコールの数を集団中のポジティブな例の数で除したものである。

【０１４０】
「ＧＩＡＮＡ５０」とラベルした曲線は、ＢＬＯＳＵＭ５０行列によるＧＩＡＮＡの曲線である。この曲線は「ＧＩＡＮＡ」とラベルしたＢＬＯＳＵＭ６２行列によるＧＩＡＮＡの元のバージョンの曲線と極めて類似している。これは部分的に、ＢＬＯＳＵＭ５０行列とＢＬＯＳＵＭ６２行列がそのオフダイアゴナル値において類似しているためであろう。ＧＩＡＮＡにおいて距離行列に変換する際には、ダイアゴナル値の相違を排除した。したがって、ＧＩＡＮＡのクラスター化の正確度はタンパク質置換基準の選択に対しては比較的堅固であり、ＢＬＯＳＵＭ６２行列とＢＬＯＳＵＭ５０行列のいずれを用いるかの選択は最終出力の適合率または再現率に実質的に影響しないであろう。

【０１４１】
既知のエピトープに特異的なＴＣＲは、他の供給源の中でも、免疫エピトープデータベースおよび解析リソース（ＩＥＤＢ）ならびにＶＤＪｄｂオンラインブラウザーから収集したことに留意されたい。ＴＣＲβ　ＣＤＲ３配列、ＴＲＢＶ遺伝子、およびそれらの関連抗原のみを保存した。冗長なまたは不完全な配列を除去した後、全部で６１，３６６のＣＤＲ３が得られ、多様な病原体から約９００個のエピトープがカバーされた。それらのデフォルトパラメーターを用いて抗原特異的クラスター化を実施するために、全ての方法をデータセットに適用した。ＴＣＲｄｉｓｔについては、Ｒコードを作成して、１５より小さな距離を有する配列対について深度優先探索を実施した。ＴＣＲｄｉｓｔについての時間計算量の計算は、ＴＣＲクラスターを見出すための深度優先探索を含まない。この１

５のカットオフは、ｉＳＭＡＲＴ４のそれと匹敵するバランスの取れた感度および特異性を有している。大きなカットオフの選択によって、クラスター化されたＴＣＲの全数が増加するが、それぞれのクラスターの特異性が低下するという犠牲を伴い得る。

【０１４２】

上記のように、同じ抗原に特異的な全てのＴＣＲを有するクラスターは、「純粋なクラスター」と定義される。感度は、全ての純粋なクラスターに含まれるＴＣＲの全数を配列の全数（即ち６１，３６６）で除したものと定義される。クラスター化の精度は、純粋なクラスターの数を全クラスター数で除したものと定義される。これらの尺度を用いて、４つ全ての方法の抗原特異的クラスター化の性能を比較した。

【０１４３】

さらに、従来の方法と異なり、ＧＩＡＮＡは、既知の抗原特異性を有するＴＣＲを用いて、実際の大きくノイズが多いＴＣＲ－ｓｅｑ試料から抗原特異的ＴＣＲを検索することができる。上記のベンチマーク抗原特異的ＴＣＲから、健常個体では欠けていると予想される３つのエピトープ、即ち最近大流行した重症急性呼吸器症候群コロナウイルス２（ＳＡＲＳ－ＣｏＶ－２）ウイルス由来のＹＡＷおよびＹＬＱエピトープ、ならびにヒト免疫不全ウイルス１（ＨＩＶ－１）由来のＦＲＤエピトープに特異的なＴＣＲを解析した。これらのＴＣＲの２０％を、試験データとして健常ドナー由来の１００，０００個のＴＣＲと混合した。重複しない残りの８０％の抗原特異的ＴＣＲを、試験配列を回収するための訓練データとして用いた。訓練データとクラスター化したいずれの配列も、「陽性」と同定してよい。２０％添加した抗原特異的ＴＣＲが真陽性であり、偽陽性は健常ドナー由来のＴＣＲである。

【０１４４】

ここで図１５Ａ～１５Ｃを参照して、大きくノイズの多いＴＣＲ－ｓｅｑ試料に適用した場合のＧＩＡＮＡの感度および特異性を示すチャートを示す。図１５Ａおよび１５Ｂは、ＳＡＲＳ－ＣｏＶ－２由来のＹＡＷエピトープの特異性および感度を示す。図１５Ｃおよび１５Ｄは、ＳＡＲＳ－ＣｏＶ－２由来のＹＬＣエピトープの特異性および感度を示す。図１５Ｅおよび１５Ｆは、ＨＩＶ－１由来のＦＲＤエピトープの特異性および感度を示す。図１４Ａ～１４Ｃに示すバイオリンプロットは、データの分布を説明している。「バイオリン」の側部における対称的な曲線は、データ点の実際の確率密度である。中央部における典型的なボックスプロットは、データの平均（中点）および四分位間範囲を説明している。それぞれのチャートのｘ軸はスミス－ウォーターマンアラインメントスコアのカットオフである。

【０１４５】

図１４Ａ～１４Ｃに示すバイオリンプロットは、データの分布を説明している。「バイオリン」の側部における対称的な曲線は、データ点の実際の確率密度である。中央部における典型的なボックスプロットは、データの平均（中点）および四分位間範囲を説明している。

【０１４６】

それぞれのバイオリンプロットのｙ軸は特異性または感度である。特異性は、真陰性の数をアルゴリズムで称する陰性の全数で除したものであると定義される。感度は、真陽性の数をアルゴリズムで称する陽性の全数で除したものであると定義される。アルゴリズムで称する陽性および陰性は、ＧＩＡＮＡクラスターを用いて定義される。配列は、それが既知の抗原特異性を有する添加されたＣＤＲ３とともにクラスター化されれば、陽性と称される。

【０１４７】

ｘ軸はスミス－ウォーターマンアラインメントスコアのカットオフ、即ちＧＩＡＮＡにおける重要なパラメーターであり、最大で４．０である。カットオフはＧＩＡＮＡにおける調節可能なパラメーターである。例えば、カットオフを３．７に設定すれば、３．７より大きなスミス－ウォーターマンアラインメントスコア（配列長によって正規化され、したがって最大は４．０）を有するいずれの配列対も、ともにクラスター化される。３．７

より小さなスコアを有する配列対は分離される。高いカットオフは高い特異性をもたらすが、感度の低下という犠牲を伴う。３つ全てのエピトープについて、ＧＩＡＮＡは２０％～５０％の感度で９９．９９％を超える特異性を達成した。

【０１４８】

ここで図１６Ａ～１６Ｃを参照して、大きくノイズの多いＴＣＲ－ｓｅｑ試料によるＧＬＩＰＨ２の性能を説明するチャートを示す。図１６Ａおよび１６Ｂは、ＳＡＲＳ－ＣｏＶ－２由来のＹＡＷエピトープ、ＳＡＲＳ－ＣｏＶ－２由来のＹＬＣエピトープ、およびＨＩＶ－１由来のＦＲＤエピトープを用いたＧＬＩＰＨ２についての感度および特異性の推定をｙ軸に示す。図１６Ｃは、ＧＬＩＰＨ２およびＧＩＡＮＡを用いた、ＳＡＲＳ－ＣｏＶ－２由来のＹＡＷエピトープ、ＳＡＲＳ－ＣｏＶ－２由来のＹＬＣエピトープ、およびＨＩＶ－１由来のＦＲＤエピトープについての陽性予測値（ＰＰＶ）の推測をｙ軸に示す。ＰＰＶは、正しく予測された特有のＴＣＲの全数を訓練データとともにクラスター化された特有のＴＣＲの全数で除したものとして定義され得る。ＧＬＩＰＨ２は多数のクラスターの中に１つのＴＣＲを置くので、この解析には特有のＴＣＲが必要になる場合がある。ＧＬＩＰＨ２は高い感度に達するが、その特異性はＧＩＡＮＡより低い。より重要なことに、ＧＩＡＮＡのＰＰＶは全てのエピトープについて６０％超に達したが、３つのエピトープのうち２つについてＧＬＩＰＨ２のＰＰＶは２０％より低かった。

【０１４９】

既知の抗原に特異的なＴＣＲの発見におけるＧＩＡＮＡの性能を評価するために、インシリコ混合実験を実施したことに留意されたい。健常ドナーに曝露される可能性が低い３つの抗原、即ちＳＡＲＳ－ＣｏＶ－２由来のＹＡＷおよびＹＬＱエピトープ、ならびにＨＩＶ－１ウイルス由来のＦＲＤエピトープを選択した。それぞれのエピトープに特異的なＴＣＲを、冗長を除去して選択した。それぞれの抗原について、ＴＣＲの２０％（試験データ）をランダムにサンプリングし、健常ドナー由来の１００Ｋ個の配列と混合した。抗原特異的ＴＣＲの残りの８０％（訓練データ）と試験データとの間に重複はなかった。混合した試料は、対応する病原体を有する偽患者であるとみなした。混合した試料を訓練データと合わせ、スミス－ウォーターマンアラインメントスコアのカットオフ（ｔｈｒ＿ｓ）の範囲を３．０～４．０（増分０．１）としてＧＩＡＮＡを適用した。それぞれのエピトープおよびパラメーターの設定について、データの変動を捕捉するためにインシリコ混合を２０回実施した。

【０１５０】

得られたデータから、予測性能を評価した。少なくとも１つのＴＣＲを含むＴＣＲクラスターを訓練データから選択した。訓練データを除くこれらのクラスター中の全てのＴＣＲは陽性コールであった。訓練ＴＣＲのいずれとも共クラスター化しなかった全てのＴＣＲは陰性コールであった。真陽性のコールは「試験データ」とラベルした配列であると定義した一方、真陰性のコールは健常ドナーの元の１００Ｋ個のＴＣＲに由来する配列であった。特異性は、真陰性のコールの数を１００Ｋで除したものであると定義した。感度は、真陽性のコールの数を試験ＴＣＲの全数で除したものであると定義した。

【０１５１】

さらに、従来の方法と異なり、ＧＩＡＮＡの高い速度および特異性によって、クエリーモジュールが新たなＴＣＲ試料を既存の参照データセットとクラスター化することが可能になり、これは現存の全てのツールに欠けている機能である。

【０１５２】

ここで図１７を参照して、等長変換に基づく高速ＧＩＡＮＡクエリーを説明するダイアグラムを示す。上記のように、参照およびクエリーのＴＣＲは、線形複雑度を有するユークリッド空間に変換され、それぞれのクエリー配列の最近傍について探索され、ＴＣＲクラスターに処理され、参照データと併合される。図１７に示す点線の矢印は、探索の方向を示す。

【０１５３】

ステップ１７０２において、参照ＴＣＲ－ｓｅｑデータ１７０１は一連のユニタリー変

換を通して数値ベクトルにエンコードされ、エンコードされたそれぞれのＣＤＲ３配列は高次元ユークリッド空間に投影されて参照等長座標１７０３を形成し得る。ステップ１７０４において、クエリーＴＣＲ－ｓｅｑデータ１７１５は一連のユニタリー変換を通して数値ベクトルにエンコードされ、エンコードされたそれぞれのＣＤＲ３配列は高次元ユークリッド空間に投影されてクエリー等長座標１７０５を形成し得る。ステップ１７０６において、参照等長座標１７０３とクエリー等長座標１７０５の間で最近傍探索が実施され得る。ステップ１７０８において、最小クエリークラスター１７０９が形成し得る。ステップ１７１０において、最小クエリークラスター１７０９と参照クラスター１７１１との間で最近傍探索が実施され得る。ステップ１７１２において、併合されたクラスター１７１３が形成され得る。

【０１５４】

ここで図１８を参照して、様々な数のＴＣＲによる参照／クエリーデータを用いるＧＩＡＮＡクエリーモジュールの時間計算量評価を説明するチャートを示す。ｘ軸はクエリーＴＣＲの数を１０ｋの増分で示し、ｙ軸は計算時間のログ表現を秒で示す。線１８０２、１８０４、１８０６、１８０８、および１８１０は、それぞれ２００ｋの参照ＴＣＲ、１Ｍの参照ＴＣＲ、２Ｍの参照ＴＣＲ、６Ｍの参照ＴＣＲ、および１０Ｍの参照ＴＣＲを示す。示したように、ＧＩＡＮＡは極めて効率が高い。これは10^7の参照配列に対して10^4のＴＣＲを探索要求するためにほぼ１７６秒を要した。これは１０００億のペアワイズ比較に等しい計算負荷を有するタスクである。下の表３は、様々なサイズのＴＣＲ試料のＧＩＡＮＡクエリーの計算時間消費を示す。時間は秒で測定した。

【０１５５】

【表３】

		クエリーTCR 数				
		10K	20K	30K	40K	50K
参照 TCR 数	200K	13	22	33	44	60
	1M	21	43	72	107	151
	2M	35	71	121	184	249
	6M	93	200	387	578	719
	10M	176	379	732	1,066	1,438

表 3:様々なサイズの TCR 試料の GIANA クエリーの計算時間消費

【０１５６】

この型のレパートリー分類は、疾患の診断および予後に対する迅速な適用を伴う重要なタスクである。典型的には、このタスクは多数例学習または深層学習によって取り組まれてきた。これらの方法に共通する限界は、大きなＴＣＲ－ｓｅｑデータセットにスケールアップすることを妨げる大きな計算コストである。

【０１５７】

ＧＩＡＮＡクエリーを、ＴＣＲレパートリーを分類するために用いてよい。例えば、２０、１００、または２００のＴＣＲ－ｓｅｑ試料を有する３つの参照データセットを、ＣＯＶＩＤ－１９患者および健常対照（ＨＣ）に均一に分割してよい。さらに１５４のＣＯＶＩＤ－１９および１２０のＨＣ試料を、参照のそれぞれに対して探索要求してよい。

【０１５８】

ここで図１９を参照して、参照データセットに対してクラスター化することによってクエリーＣＯＶＩＤ－１９患者が健常対照から分離される程度を説明するチャートを示す。参照データのＴＣＲの数をｘ軸のラベルとして示す。ｙ軸におけるｔ統計は、ＣＯＶＩＤ－１９とＨＣクエリー試料を分離するためにＣＯＶＩＤ－１９分率を用いる２試料ｔ検定を実施するｔ検定機能を用いて行った。バー１９０２、１９０４、および１９０６は、それぞれ２００ｋの参照ＴＣＲ、１Ｍの参照ＴＣＲ、および２Ｍの参照ＴＣＲを示す。２試

料ｔ検定は、ｔ統計を得るためにクエリーデータから推定したＣＯＶＩＤ－１９分率を用いて実施した。全てのｐ値は２．２×１０$^{-16}$のレベルで有意であった。それぞれのクエリー試料について、ＣＯＶＩＤ－１９参照患者と共クラスター化されたＴＣＲの分率が計算され得る。この分率はクエリー試料中でＣＯＶＩＤ－１９患者について有意に高く、参照データのサイズの増大とともに、クエリーＨＣからの分離が増大する可能性がある。この分率を予測因子として用いれば、大きな試料サイズでの参照について受信者動作曲線下の面積（ＡＵＣ）の増大が観察される。

【０１５９】

ここで図２０Ａ～２０Ｃを参照して、単一予測因子としてＣＯＶＩＤ－１９分率を用いる受信者動作特性（ＲＯＣ）曲線を説明するチャートを示す。ＣＯＶＩＤ－１９およびＨＣの試料の数をそれぞれのチャートの上に示す。それぞれの試料は１０ＫのＴＣＲ配列を含む。

【０１６０】

ＲＯＣ曲線は、所与の方法の予測能を可視化する偏りのない方式である。ここで連続的な予測因子としてＣＯＶＩＤ－１９分率を用いる。この分率の閾値を変更することによって、特異性（ｘ軸）および感度（ｙ軸）が変化する。図２０Ａに、１０個のＨＣ試料および１０個のＣＯＶＩＤ－１９試料を示す。図２０Ｂに、５０個のＨＣ試料および５０個のＣＯＶＩＤ－１９試料を示す。図２０Ｃに、１００個のＨＣ試料および１００個のＣＯＶＩＤ－１９試料を示す。９５％信頼区間は、層化された２，０００のブートストラップから推定した。注目すべきことに、２００万の参照ＴＣＲで、このアプローチの感度（７９％）および特異性（１００％）は、ＣＯＶＩＤ－１９についてのいくつかの現存する検査を凌駕しており、疾患の診断におけるこのアプローチの潜在的な有用性を示唆している。より重要なことに、レパートリー分類の正確度が、より多くの参照試料で改善された。これはおそらく、通常は低頻度で共有される、様々な患者にわたる疾患特異的なＴＣＲの共有によるものであり、したがって、大きな参照データによって高いクラスター化確率、小さな分散、および良好な精度がもたらされる。

【０１６１】

ここで図２１Ａ～２１Ｄを参照して、ＣＯＶＩＤ－１９患者と健常ドナーの両方についての元のＣＯＶＩＤ－１９分率スコアおよび様々な数の参照ＴＣＲによるＣＯＶＩＤ－１９分率の変動係数を説明するチャートを示す。図２１Ａ～２１は、様々な参照データ構成の下でＣＯＶＩＤ－１９参照試料と共クラスター化したＴＣＲ分率の分布を示す。図２１Ａは１０個のＨＣ試料および１０個のＣＯＶＩＤ－１９試料を示す。図２１Ｂは５０個のＨＣ試料および５０個のＣＯＶＩＤ－１９試料を示す。図２１Ｃは１００個のＨＣ試料および１００個のＣＯＶＩＤ－１９試料を示す。

【０１６２】

図２１Ｄにおいて、ｘ軸は参照ＴＣＲの数を示し、ｙ軸は変動係数を示す。変動係数は、標準偏差をクエリー試料中のＣＯＶＩＤ－１９患者のＣＯＶＩＤ－１９分率の平均で除したものと定義してよい。バー２１０２、２１０４、および２１０６は、それぞれ２００ｋの参照ＴＣＲ、１Ｍの参照ＴＣＲ、および２Ｍの参照ＴＣＲを示す。より多くの参照試料で、ＣＯＶＩＤ－１９分率の変動係数の低下が見られる。

【０１６３】

上記の特徴は、より大きな参照データセットでも実証することができる。例えば、がん、ＣＯＶＩＤ－１９、多発性硬化症（ＭＳ）患者、およびＨＣ由来の１，２１３個の試料からなる１０００万のＴＣＲを含むデータセットを用いた。１０Ｍ全てのＴＣＲの抗原特異的クラスター化を実施し、共有されたＴＣＲクラスターのレベルによって測定される様々なレパートリー試料の類似性を検討するために、ＧＩＡＮＡを適用した。表４に、参照データとして用いたＴＣＲ－ｓｅｑ試料コホートを示す。

【０１６４】

【表４】

疾患の型	コホート	疾患	試料サイズ	特有の試料	PMID
健常対照	Emerson et al., 2017	健常対照 (バッチ 1)	100	100	28369038
多発性硬化症	Emerson et al., 2013	多発性硬化症	50	25	23428915
COVID-19	Nolan et al., 2020	COVID-19 (適応性、ISB)	311	311	32793896
がん	Snyder et al., 2017	膀胱がん	117	30	28552987
	Mansfield et al., 2018	肺がんおよび脳転移	40	20	29391594
	Sims et al., 2016	神経膠腫	32	15	27261081
	Page et al., 2019	乳がん	63	16	31831558
	Reuben et al., 2019	肺がん	121	121	32001676
	Emerson et al., 2013	卵巣がん	96	5	24027095
	Stromnes et al., 2017	膵がん	16	16	29066497
	Tumeh et al., 2014	黒色腫	34	23	25428505
	Le et al., 2017	結腸直腸がん	35	3	28596308
	Duhen et al., 2018	頭頚部、卵巣、および黒色腫	33	8	30006565
	Chow et al., 2020	腎細胞癌腫	53	26	32900949
	Riaz et al., 2017	黒色腫	58	29	29033130
	Sherwood et al., 2013	結腸直腸がん	14	14	23771160

表 4:参照データとして用いた TCR-Seq 試料コホート

【０１６５】

いくつかのコホートについては、参照データを創成する際に利用可能な全ての試料を用いてはいない。それぞれの試料について、最も存在量の多い上位１０，０００のＴＣＲを選択し、データが１０，０００より少ない配列を含んでいた場合には、全てを用いた。特有の試料は、研究に参加した独立の患者の数を示していた。試料サイズは、参照で用いたコホート中の全ＴＣＲ－ｓｅｑ試料の数を記録した。Ｅｍｅｒｓｏｎ２０１７コホートはバッチ１に６６６の健常ドナーを含み、これから１００個の試料をランダムに選択した。ＣＯＶＩＤ－１９コホートは多重国際ＣＯＶＩＤ－１９研究からアセンブルした１，４００名を超える患者を含んでいた。それぞれＡｄａｐｔｉｖｅ　Ｂｉｏｔｅｃｈｎｏｌｏｇ

ｙ（Ａｄａｐｔｉｖｅ、ｎ＝１５４）およびＩｎｓｔｉｔｕｔｅ ｆｏｒ Ｓｙｓｔｅｍ Ｂｉｏｌｏｇｙ（ＩＳＢ、ｎ＝１５７）によって２つのコホートを収集した。一例では、ＧＩＡＮＡは８個のＣＰＵおよび１２８Ｇのメモリーによる高性能計算クラスターで参照データをクラスター化するために１９．５時間を要した。

【０１６６】

ここで図２２Ａ～２２Ｄを参照して、ＴＣＲ共クラスター化に基づくＴＣＲ－ｓｅｑ試料の類似性のグラフ表現を示す。図２２Ａ～２２Ｄにおいて、物理的近接は類似性を表し、したがって互いに近接したＴＣＲ－ｓｅｑ試料（ドットとして示す）は類似性が高い。図２２Ａ～２２Ｄを生成するために、上記の１，２１３個の参照試料の元のＴＣＲクラスター化の結果から、試料ごとのカウント共有行列を計算した。スピアマン相関行列も、共クラスター化されたＴＣＲのカウントに基づき、相関値０．４以下の対をゼロと設定して計算した。得られた疎行列を用いてグラフを作成した。接続が２未満のノードを除外して試料群を可視化した。

【０１６７】

図２２Ａは、がん患者の第１のクラスター（図２２Ｂに詳細を示す）、ＨＣおよびＭＳ患者の第２のクラスター（図２２Ｃに詳細を示す）、ならびに肺がんおよびＣＯＶＩＤ－１９の患者の第３のクラスター（図２２Ｄに詳細を示す）の全体図である。図２２Ａは、第１のクラスター中の大部分のがん患者が第２のクラスター中のＨＣおよびＭＳの患者から明らかに分離されることを示している。興味あることに、肺がん患者とＣＯＶＩＤ－１９患者は、分離した第３のクラスターを形成していた。ウイルス感染またはがん等の局所的炎症状態によって組織に存在するＴ細胞が循環系中に放出され、おそらくＴＣＲレパートリーの共有を惹起していることが知られている。これらの知見は、肺組織においてＴ細胞の移動の規模が疾患の型を超越するために十分に大きいことをさらに示唆している。

【０１６８】

ＴＣＲクラスター化を用いるレパートリー分類の実現可能性を試験するため、１０、５０、および１００個のＣＯＶＩＤ－１９試料を１０、５０、および１００個の健常対照と組み合わせて、２０、１００、および２００個の試料を含む３つの参照データセットを生成したことに留意されたい。それぞれの試料は、クローンの存在量を順位付けることによって選択された１０ＫのＴＣＲを含んでいた。クエリー試料は１５４のＣＯＶＩＤ－１９患者および１２０のＨＣを含んでいた。クエリー試料と参照試料との間に重複はなかった。ＧＩＡＮＡを用いてそれぞれのクエリー試料についてＴＣＲクラスターを生成させた。それぞれの試料について、１００を超える試料を有するＴＣＲクラスターは、これらのＴＣＲがスモールワールド接続から生成され、疾患の特異性については情報が得られない可能性があるので除外した。残りのクラスターについては、ＣＯＶＩＤ－１９患者の寄与がある参照ＴＣＲの分率を計算して、予測因子として用いた。

【０１６９】

多数の疾患分類のタスクにおいて、７１２名のがん、３１１名のＣＯＶＩＤ－１９、２５名のＭＳの患者、および１００個のＨＣの試料を併せて、１０ＭのＴＣＲの参照データを産生した。別に６２名のがん、１９３名のＣＯＶＩＤ－１９、１２名のＭＳ、および１５３名のＨＣの試料を収集し、疾患のラベルが未知であることを仮定してクエリーを作成した。それぞれのクエリークラスターファイルについて同様の解析を実施して、ＨＣを含むそれぞれの疾患カテゴリーの分率を推定した。これらの分率を用いて疾患を予測し、ＲＯＣ解析を実施した。具体的には、ＨＣ試料との全ての比較のためにＨＣ分率を用いた。予備的アプローチとして、３つの疾患のペアワイズ分離のため、２つの疾患分率の間の差異を用いた。例えば、ＭＳ患者からがんを予測する場合、がん分率－ＭＳ分率を予測因子として用いた。

【０１７０】

さらに、従来の方法と異なり、ＧＩＡＮＡは、超大スケールのＴＣＲクラスター化および探索要求による新規な多疾患検出プラットフォームとして用いることができる。ＧＩＡＮＡによる超大スケールクラスター化によって、疾患特異的なＴＣＲと組織特異的なＴＣ

Ｒとの検討が可能になる。一例では、肺がんおよびＣＯＶＩＤ－１９の患者のＴＣＲクラスターを３つのカテゴリー、即ちｉ）ＣＯＶＩＤ－１９特異的、ｉｉ）肺がん特異的、ｉｉｉ）２つの疾患で共有、に分割した。

【０１７１】

ここで図２３Ａ～２３Ｂを参照して、様々なカテゴリーのＴＣＲクローン頻度の分布を説明するビースウォームプロットを示す。ビースウォームプロットは、ＴＣＲクローン頻度である実際のデータ点を示すボックスプロットである（ｙ軸）。ＴＣＲクローン頻度は、所与のＴＣＲのシーケンシングリードのパーセンテージを試料中のリードの全数で除したものと定義される。両方のチャートのｘ軸は、２つの試料カテゴリーを並べた比較である。

【０１７２】

図２３Ａに、ＣＯＶＩＤ－１９の患者のみで見出されたＴＣＲ　２３０２と、ＣＯＶＩＤ－１９と肺がんの患者の両方で見出されたＴＣＲ　２３０４の間の比較を示す。図２３Ｂに、肺がんの患者のみで見出されたＴＣＲ　２３０６と、ＣＯＶＩＤ－１９と肺がんの患者の両方で見出されたＴＣＲ　２３０４との比較を示す。ＣＯＶＩＤ－１９と肺がんの患者の両方で見出されたＴＣＲ　２３０４については、疾患特異的ＴＣＲのコホートに一致するようにクローン頻度を選択した。図２３Ａにおけるｐ値（ｐ＜２．２ｅ－１６）は、統計検定の第１種の過誤の尺度である。図２３Ｂにおけるｐ値（ｎ．ｓ．）は有意でないことを意味し、これは０．０５より大きなｐ値である。

【０１７３】

図２３Ａ～２３Ｂに示すように、ＣＯＶＩＤ－１９患者についてはカテゴリーｉｉｉ）に対してカテゴリーｉ）のクローン頻度が有意に高い一方、肺がん患者についてはカテゴリーｉｉ）とｉｉｉ）の間で差がない。同じコホートの中ではＴＣＲ頻度は一致し、バッチの影響は回避されていた。したがって、ＣＯＶＩＤ－１９特異的ＴＣＲの高い存在量は、おそらくＳＡＲＳ－ＣｏＶ－２への免疫応答によって惹起されている。まさに、ウイルス感染の後でＣＯＶＩＤ－１９特異的ＴＣＲのみが動的制御を受け、これは曝露後の最初の２週間以内にピークに達し、その後低下した。

【０１７４】

ここで図２４Ａ～２４Ｂを参照して、ＳＡＲＳ－ＣｏＶ－２感染の経過の間のＴＣＲクローン頻度の動的変化を説明するグラフを示す。両方のチャートで、ｘ軸は診断から試料採取までの日数を示し、ｙ軸はＴＣＲ頻度のログ表現を示す。中央の実線２４０２はデータの平滑化平均である。それぞれの時点（ｘ軸）で、ＴＣＲクローン頻度の多数の観察があり、実線はその平均値である。同様に、上の点線２４０４は、観察されたデータの９５％信頼区間の上限である。下の点線２４０６は信頼区間の下限である。「ｐ値」は、この解析のために実施したスピアマン相関検定の第１種の過誤である。スピアマンの相関値（ｒｈｏ）を、それぞれの図の中の最初の線として表示している。

【０１７５】

共有されたＴＣＲのクローン存在量は、ＳＡＲＳ－ＣｏＶ－２感染の後のタイムラインによって影響されなかった。これらの図は、大きなＴＣＲレパートリー試料についてのクラスター化によって共有された疾患特異的ＴＣＲが大量にあることが明らかになり、これがレパートリー分類に対するより精細な解を提供し得ることを示している。

【０１７６】

クラスター化されたＴＣＲは、例えば一個抜きバリデーションアプローチを実行することによってレパートリー試料を多数の疾患に割り当てるためのマーカーとして用いてよい。具体的には、所与の試料について、参照コホートの中のがん、ＣＯＶＩＤ－１９、ＭＳの患者、または健常対照と共クラスター化されたＴＣＲの分率を、その試料自体を除いて計算してよい。この方法によって、それぞれの試料について合計して１になる４つのクラス分率が得られる。ＨＣ分率を用いて、患者を健常ドナーから分離してよい。

【０１７７】

ここで図２５Ａ～２５Ｆを参照して、共クラスター化されたＴＣＲから計算した疾患分

率についての一個抜きバリデーションアプローチを用いるＲＯＣ曲線を説明するチャートを示す。それぞれのチャートの右下にＡＵＣ値を示す。それぞれのチャートはｘ軸に特異性、ｙ軸に感度を示している。図２５Ａは、がんのＴＣＲクラスターとＨＣのＴＣＲクラスターの間の比較を示す。図２５Ｂは、ＣＯＶＩＤ－１９のＴＣＲクラスターとＨＣのクラスターの間の比較を示す。図２５Ｃは、ＭＳのクラスターとＨＣのクラスターの間の比較を示す。図２５Ｄは、がんのＴＣＲクラスターとＣＯＶＩＤ－１９のクラスターの間の比較を示す。図２５Ｅは、ＣＯＶＩＤ－１９のクラスターとＭＳのクラスターの間の比較を示す。図２５Ｆは、がんのクラスターとＭＳのクラスターの間の比較を示す。９５％信頼区間は、層化された２，０００のブートストラップを用いて計算してよい。３つ全ての疾患についてほぼ完全な正確度が観察された。疾患の対を識別するため、対応する２つの分率の間の差異を予測因子として用いてよく、これは高い（９３％以上）ＡＵＣ値をもたらす。

【０１７８】

ここで図２６Ａ～２６Ｆを参照して、共クラスター化されたＴＣＲから計算した疾患分率についてのより厳格なアプローチを用いるＲＯＣ曲線を説明するチャートを示す。一例では、参照試料の４０％を訓練データとしてランダムに選択し、残りの６０％を試験データとして残した。訓練試料は、「ＣＯＶＩＤ－１９」、「がん」、「ＭＳ」、または「ＨＣ」とラベルする。それぞれの試験データを全ての訓練試料と共クラスター化して、それぞれの試料カテゴリーとともにクラスター化されたＴＣＲの分率を計算した。疾患と健常の個体を区別するために「ＨＣ」の分率を用いる。３つの疾患を区別するために他の分率を用いる。このより厳格な方法によって、一個抜きバリデーションと同様のレベルの予測正確度が達成された。

【０１７９】

それぞれのチャートの右下にＡＵＣ値を示す。それぞれのチャートはｘ軸に特異性、ｙ軸に感度を示している。図２５Ａは、がんのＴＣＲクラスターとＨＣのＴＣＲクラスターの間の比較を示す。図２５Ｂは、ＣＯＶＩＤ－１９のＴＣＲクラスターとＨＣのクラスターの間の比較を示す。図２５Ｃは、ＭＳのクラスターとＨＣのクラスターの間の比較を示す。図２５Ｄは、がんのＴＣＲクラスターとＣＯＶＩＤ－１９のクラスターの間の比較を示す。図２５Ｅは、ＣＯＶＩＤ－１９クラスターとＭＳのクラスターの間の比較を示す。図２５Ｆは、がんのクラスターとＭＳのクラスターの間の比較を示す。

【０１８０】

ここで図２７を参照して、参照ＴＣＲ－ｓｅｑ試料のクロスコホート類似性を説明するチャートを示す。Ｎ個の試料のＴＣＲクラスター化データを用いて、他の試料のそれぞれと共クラスター化したそれぞれの試料のＴＣＲのパーセンテージを計算してよい。自己共クラスター化パーセンテージをゼロと割り当てて、全てのベクトルの長さをＮとしてよい。スピアマン相関行列は、Ｎ×Ｎの共クラスター化分率行列から計算してよい。次いで行列をがんの型に従って折り畳んでよい。上位５つの最大の相関の平均を、ヒートマップとして図２７に表示する。計算に先立ってそれぞれの試料の自己相関を除外してよいことを除いて、同じ疾患相関（対角値）を同様に計算してよい。カラーコーディングによって、ディスプレイ上の値が表される。例えば、赤色は正の相関を表す。

【０１８１】

肺がんとＣＯＶＩＤ－１９を識別する能力は、疾患内の類似性がまだ高いので、２つの疾患の外見上のグループ化と矛盾しなかった。しかし、疾患の大部分はただ１つの研究から誘導されたので、予測性には未知のバッチ効果による寄与があるのではないかという懸念が生じた。

【０１８２】

この可能性を検討するため、ＧＩＡＮＡを用いて、独立したコホートに由来する未知の試料の疾患ラベルを予測した。同じ疾患の２６７の新たなＴＣＲ－ｓｅｑ試料および参照データセットに対する１５３のＨＣ試料を探索要求するためにＧＩＡＮＡを適用した。全ての試料は末梢血由来であった。同じアプローチを用いて、参照のがん、ＣＯＶＩＤ－１

９、ＭＳ、またはＨＣの配列と共クラスター化したＴＣＲの分率を計算した。下の表５に、クエリーデータとして用いたＴＣＲ－ｓｅｑ試料コホートを示す。

【０１８３】

【表５】

疾患の型	コホート	疾患	試料サイズ	特有の試料	PMID
健常対照	Emerson et al., 2017	健常対照(バッチ2)	120	120	28369038
	DeWitt et al., 2018	活動性結核	33	33	29914888
多発性硬化症	Bertoli et al., 2019	多発性硬化症	12	6	31719595
COVID-19	Nolan et al., 2020	COVID-19 (HUniv120)	193	193	32793896
がん	Beshnova et al., 2020	卵巣、膵、および腎がん	25	25	32817363
	Robert et al., 2014	黒色腫	21	21	24583799
	Beausang et al., 2017	乳がん	16	16	29138313

表 5:クエリーデータとして用いた TCR-Seq 試料コホート

【０１８４】

Ｅｍｅｒｓｏｎの２０１７年の研究からの健常ドナーの第２のバッチの１２０全てを対照として用いた。参照との重複を避けるため、ＣＯＶＩＤ－１９患者についてはＮｏｌａｎの２０２０年の研究からのＨｏｓｐｉｔａｌ Ｕｎｉｖｅｒｓｉｔａｒｉｏ １２ ｄｅ Ｏｃｔｕｂｒｅ（ＨＵｎｉｖ１２０、ｎ＝１９３）コホートを用いた。このコホートの患者はスペインのマドリッドから収集した。一例では、３．５ＧＨｚのＤｕａｌ－Ｃｏｒｅ Ｉｎｔｅｌ Ｃｏｒｅ ｉ７プロセッサーおよび１６ＧＢの２１３３ＭＨｚ ＬＰＤＤＲ３メモリーを有するＭａｃＢｏｏｋＰｒｏで、全４２０試料のクエリーを完了するのにＧＩＡＮＡで２０．５時間を要した。

【０１８５】

ここで図２８Ａ～２８Ｄを参照して、がん、ＣＯＶＩＤ－１９、ＭＳ患者、およびＨＣのクラス分率の分布を示す「バイオリンプロット」を示す。図２８Ａ～２８Ｄに示す「バイオリンプロット」のそれぞれはデータの分布を説明している。「バイオリン」の側部の対称的な曲線は、データ点の実際の確率密度である。中央部において、典型的なボックスプロットは、データの平均（中点）および四分位間範囲を説明している。ｙ軸は所与の疾患カテゴリー（サブパネルの標題として示す）のＴＣＲの分率を示す。ｘ軸は疾患のカテゴリーを示す。図２８Ａ～２８Ｄは、所与の疾患のＧＩＡＮＡによって推定されたＴＣＲ分率がその疾患を有する患者で最大であることを示しており、これは多数の疾患の予測因子としてこれを用いることを正当化している。

【０１８６】

クラス分率（例えばがんの分率）は、がん患者からの参照ＴＣＲとともにクラスター化したクエリーＴＣＲの割合として計算してよい。いずれのモデル訓練もなしに、この単純なアプローチは、それぞれの試料カテゴリーを他から区別することができる。ＨＣ分画は、９１％を超える正確度で３つ全ての疾患から識別される。

【０１８７】
ここで図２９Ａ～２９Ｆを参照して、４つの疾患クラスのペアワイズ分離のための単一予測因子として疾患のクラス分率を用いるＲＯＣ曲線を説明するチャートを示す。分率は、参照データセットにおける所与のクラスの試料と共クラスター化されたＴＣＲのパーセンテージであってよい。それぞれのチャートの右下にＡＵＣ値を示す。それぞれのチャートは、ｘ軸に特異性、ｙ軸に感度を示している。図２９Ａは、がんのＴＣＲクラスターとＨＣのＴＣＲクラスターの間の比較を示す。図２９Ｂは、ＣＯＶＩＤ－１９のＴＣＲクラスターとＨＣのクラスターの間の比較を示す。図２９Ｃは、ＭＳのクラスターとＨＣのクラスターの間の比較を示す。図２９Ｄは、がんのＴＣＲクラスターとＣＯＶＩＤ－１９のクラスターの間の比較を示す。図２９Ｅは、ＣＯＶＩＤ－１９のクラスターとＭＳのクラスターの間の比較を示す。図２９Ｆは、がんのクラスターとＭＳのクラスターの間の比較を示す。９５％信頼区間は、層化された２，０００のブートストラップを用いて計算した。疾患の間のペアワイズ分離は、全て８７％を超えるＡＵＣに達した。クエリー試料は参照データに含まれない研究から誘導されたので、大きなＡＵＣは未知のバッチまたはコホートに特異的な効果によって惹起されたのではなく、おそらく３つの疾患についての真の予測可能性を反映している。

【０１８８】
ＧＩＡＮＡクエリー性能を、多数例学習（ＭＩＬ）およびコホート特異的パラメーターのフィッティング（例えばＤｅｅｐＲＣおよびその他）に基づく従来のレパートリー分類法と比較した。ＧＩＡＮＡはパラメーターフィッティングを何ら要求しない一方、従来の方法はクエリー試料と類似の属性を有する好適な参照データ（例えば真のＣＯＶＩＤ－１９患者由来のレパートリー試料および陰性対照）を提供する。ＨＣＭＶ＋およびＨＣＭＶ－の対象を含むコホートを用いて、試料の７５％を参照として適用し（訓練と同様）、残りの２５％を試験データとして適用した。それぞれの試験試料を参照データに対して探索要求した。それぞれのクエリー試料について、ＨＣＭＶ＋参照対象と共クラスター化したＴＣＲの分率を計算し、予測因子として用いた。この単純なアプローチは、下のチャートに示すように、８３．０６％のＡＵＣに達し、ＤｅｅｐＲＣと同じで、他の方法よりも良好であった。したがって、ＧＩＡＮＡはレパートリー分類のための競争力のある方法であろう。

【０１８９】
【表６】

方法	AUC
GIANA	0.831
DeepRC	0.831
SVM (MM)	0.825
SVM (J)	0.546
KNN (MM)	0.679
KNN (J)	0.534
MIL (KMER)	0.582
MIL (TCRB)	0.515

表 6:GIANA と従来法の AUC

【０１９０】
上記のＲＯＣ曲線を生成し、ＡＵＣ値を推定するために、Ｒプログラミング言語のｐＲＯＣパッケージを用い、２，０００の層化されたブートストラップ反復によって９５％の信頼区間を計算し、ｐＲＯＣパッケージのｃｉ．ａｕｃ機能を用いて実行したことに留意されたい。図２２は、ｉｇｒａｐｈパッケージを用いて作成した。注釈付きの値を有する

図２７のヒートマップは、ｇｐｌｏｔｓパッケージのｈｅａｔｍａｐ．２機能を用いて作成した。図示した全てのボックスプロットについて、中央の線はメディアン値を画定し、ボックスの境界線はデータの２５％（Ｑ１）および７５％（Ｑ３）の四分位数を示す。下および上のウィスカーはＱ１－１．５ＩＱＲおよびＱ３＋１．５ＩＱＲに対応し、ここでＩＱＲは四分位数間範囲の省略形である。

【０１９１】

まとめると、ＧＩＡＮＡは、数千万の配列を効率的に取り扱うことができる新規な抗原特異的ＴＣＲクラスター化アルゴリズムである。ＧＩＡＮＡは現存する全ての方法より高い感度および精度を達成し、既知の抗原に特異的なＴＣＲを高い正確度で検索することができる。新規の試料の超大スケールのＴＣＲクラスター化および高速のクエリーによって、参照に基づく新規なレパートリー分類フレームワークも可能になった。ＧＩＡＮＡはＴＣＲ領域が解明された単一細胞のＲＮＡ－ｓｅｑデータを解析することもでき、パブリックドメインにあるＴＣＲレパートリー試料の大きなデータベースに対するｓｃＲＮＡ－ｓｅｑデータにＴＣＲを探索要求し、共有された抗原特異性に対する新たな洞察を提供することが可能である。最小の改変を加えれば、ＧＩＡＮＡは大型Ｂ細胞受容体のシーケンシングデータをクラスター化または探索要求するために適用することもできる。さらに、等長埋め込みを実施するための数学的フレームワークは、将来において古典的な短いＤＮＡまたはタンパク質の配列アラインメントの課題に対する代替の解決を提供し得る。

【０１９２】

ＨＬＡアレルは、そのようなデータが大部分の現在の研究では利用可能でないので、ＧＩＡＮＡにおいて考慮しなかったことに留意されたい。ＨＬＡタイピングを含めると、ＴＣＲクラスター化およびクエリー法の正確度は改善されると期待される。ＧＩＡＮＡはギャップアラインメントを支援しないが、これはこの機能性を有する他の方法より良好な感度を有している。これは、ギャップを許容することがクラスター化の特異性を低減し、予測正確度を犠牲にする可能性があるからである。

【０１９３】

また、上記のように、疾患のクラスを割り当てるために、単純な分率推定が用いられる。より多くのデータがあれば、この努力は予測正確度を最適化する機械学習モデルによって改善され得る。さらに、全てのがん患者を、がんの局在化を区別することなく、ともに他の疾患と比較した。しかし、参照として関連する十分なＴＣＲ－ｓｅｑ試料を用いてがんの型を分離する能力が意図されている。現在のＧＩＡＮＡ法は既にレパートリー分類の高い正確度を達成しているが、このプラットフォームの診断的価値は、予測的に収集された患者試料によって改善されよう。

【０１９４】

自己免疫疾患および感染疾患において実証したように、低頻度で共有される抗原特異的な周知のＴＣＲは、数千名の個体からの大量のＴＣＲを比較することによって検出され得る潜在的に重要なバイオマーカーである。免疫レパートリーを用いてがん、ＣＯＶＩＤ－１９、およびＭＳを個別に検出するための方法が開発されてきたが、いずれも様々な疾患を同時に診断し分離することはできていない。対照的に、ＧＩＡＮＡは、感染疾患、自己免疫障害、およびがんを診断するための統合されたプラットフォームとして用いることができる。

【０１９５】

これは、従来の方法に対していくつかの改善を提供する。疾患の診断は伝統的に主として症状に基づいて行われており、それぞれの疾患には、放射活性造影、液体生検、侵襲的な内視鏡検査、手術、その他の多様な臨床アッセイから得られるシグネチャーの区別できる組が必要である。多数の疾患を指示するための単一のバイオマーカーとして免疫系を用いることの実現可能性は、症状基準から免疫応答基準へのパラダイムシフトを生じ、これにより、多くの免疫関連障害に対する普遍的な解決策が提供され得る。

【０１９６】

さらに、鑑別診断は通常、臨床的に困難であり、プラットフォームに多くの疾患を追加

することは診断の特異性を低下させることが予想される。しかし、ＧＩＡＮＡの断定正確度は、ＴＣＲ－ｓｅｑ試料をより多く含ませることによって実際に増大する。

【０１９７】

さらに、免疫応答は通常、いずれの測定可能な症状より先行するので、ＧＩＡＮＡプラットフォームは、大部分の疾患が治癒可能であるか管理が容易である初期の段階で疾患を検出する可能性を有している。これは、がんの診断で既に示されており、免疫制御の原理もＭＳ等の自己免疫障害に適用される。最後に、このプラットフォームは、標的化Ｖ（Ｄ）Ｊ捕捉を実施するために少量の血液を必要とするのみであるので、これは低コストでの非侵襲的な検査としての役目を果たし得る。併せて、ＧＩＡＮＡは、抗原特異的ＴＣＲクラスターを見出すため、ＳＡＲＳ－ＣｏＶ－２等の既知の病原体に特異的な配列を検索するため、およびがん、免疫学、および臨床的な研究における急速に増大するＴＣＲデータの集まりによる疾患診断を容易にするために、広く用いることができる。

【０１９８】

本開示の目的のため、モジュールは、本明細書に記載したプロセス、特徴、および／または機能を実施しまたは容易にするソフトウェア、ハードウェア、もしくはファームウェア（またはそれらの組合せ）システム、プロセス、もしくは機能性、またはそれらの構成成分である（ヒトとの相互作用またはヒトによる増強を含むか含まない）。モジュールはサブモジュールを含んでよい。モジュールのソフトウェア成分は、プロセッサーによる実行のためのコンピューター可読媒体に記憶することができる。モジュールは１つもしくは複数のサーバーに不可欠であるか、１つもしくは複数のサーバーによって搭載および実行してよい。１つもしくは複数のモジュールはエンジンまたはアプリケーションにグループ化してよい。

【０１９９】

当業者には、本開示の方法およびシステムは多くの方式で実行され、したがって上記の実施例に限定されないことが認識されよう。換言すれば、ハードウェアおよびソフトウェアまたはファームウェアの種々の組合せにおける単一または多数の成分によって実施される機能的要素ならびに個別の機能は、クライアントのレベルもしくはサーバーのレベルまたはその両方において、ソフトウェアアプリケーションの中に分配してよい。これに関して、本明細書に記載した様々な実施例の任意の数の特徴を単一または多数の実施例と組み合わせてよく、本明細書に記載した特徴の全てより少ないか、これより多い特徴を有する代替の実施例が可能である。

【０２００】

機能性は、全体としてまたは部分的に、現在既知であるか将来既知となる方式で、多数の成分の中に分配してもよい。即ち、本明細書に記載した機能、特徴、インターフェース、および優先性を達成するにあたって、多数のソフトウェア／ハードウェア／ファームウェアの組合せが可能である。さらに、本開示の範囲は、記載した特徴および機能およびインターフェース、ならびに現在および将来、当業者によって理解される本明細書に記載したハードウェアまたはソフトウェアまたはファームウェアの成分に加えられる変形および改変を実行するための従来から既知の方式を包含する。

【０２０１】

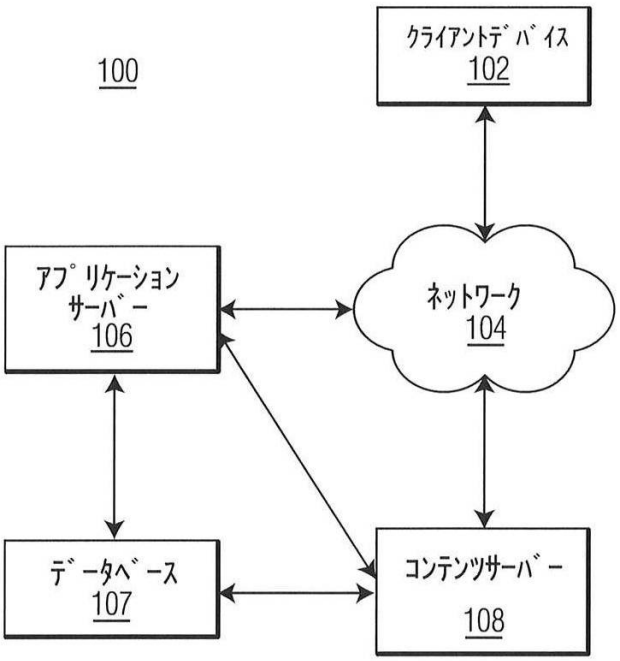
さらに、本開示においてフローチャートとして提示および記載した方法の実施例は、本技術のより完全な理解を提供するために例として提供している。開示した方法は、本明細書で提示した操作および論理的フローに限定されない。その中で種々の操作の順序が変更され、またその中で大きな操作の一部として記載されたサブ操作が独立に実施される代替の実施例も意図している。

【０２０２】

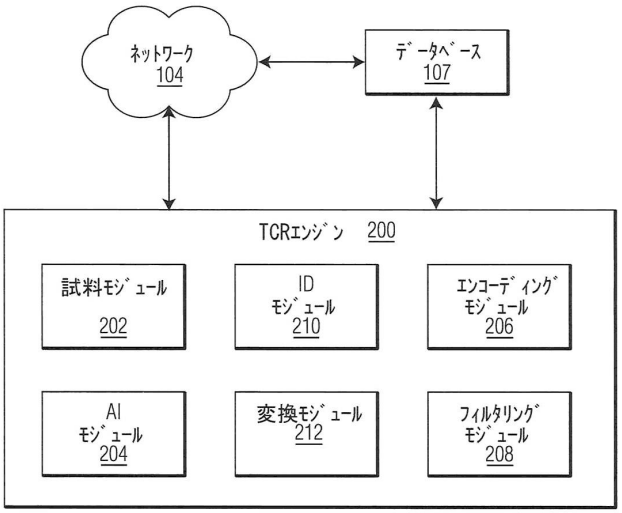
本開示の目的のために種々の実施例を記載したが、これらの実施例は本開示の教示をこれらの実施例に限定するものとみなすべきではない。本開示に記載したシステムおよびプロセスの範囲内に留まる結果を得るために、上記の要素および操作に種々の変更および改変を行ってよい。

【図面】

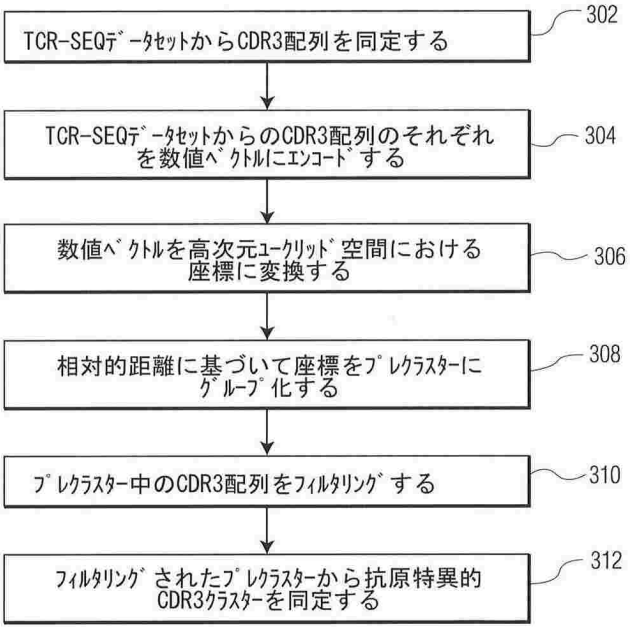
【図１】


100
クライアントデバイス
102
アプリケーション
サーバー
106
ネットワーク
104
データベース
107
コンテンツサーバー
108

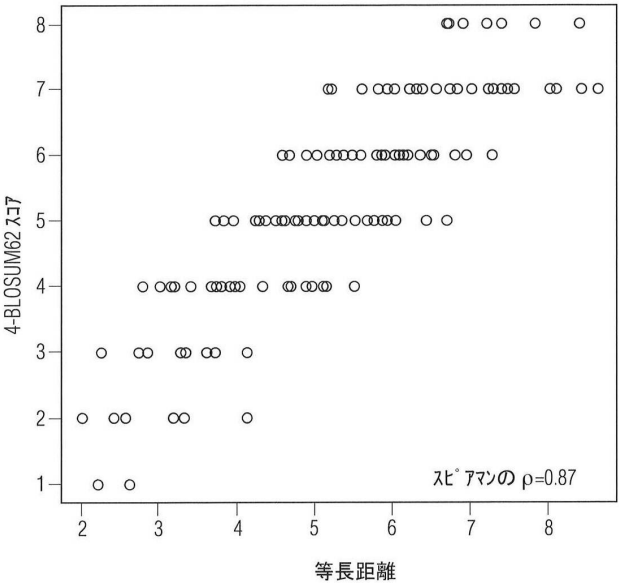
【図２】


ネットワーク
104
データベース
107
TCRエンジン 200
試料モジュール
202
ID
モジュール
210
エンコーディング
モジュール
206
AI
モジュール
204
変換モジュール
212
フィルタリング
モジュール
208

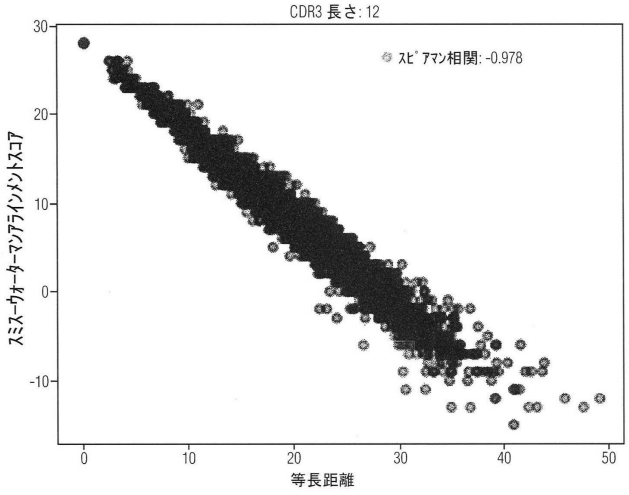
【図３】


TCR-SEQデータセットからCDR3配列を同定する
302
TCR-SEQデータセットからのCDR3配列のそれぞれを数値ベクトルにエンコードする
304
数値ベクトルを高次元ユークリッド空間における座標に変換する
306
相対的距離に基づいて座標をプレクラスターにグループ化する
308
プレクラスター中のCDR3配列をフィルタリングする
310
フィルタリングされたプレクラスターから抗原特異的CDR3クラスターを同定する
312

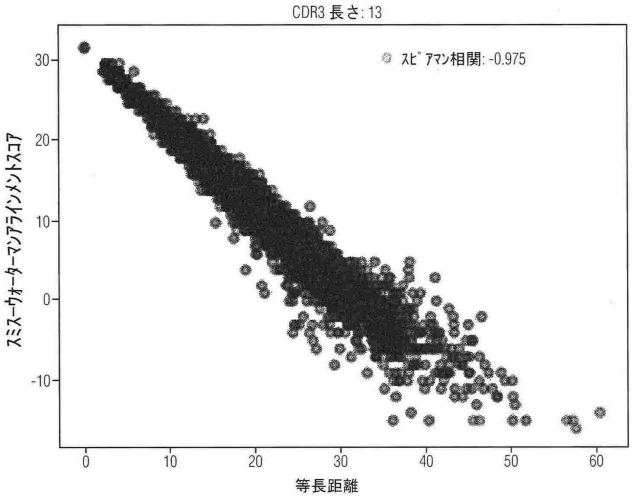
【図４】


4-BLOSUM62 ユニット
等長距離
スピアマンの ρ=0.87
1
2
3
4
5
6
7
8

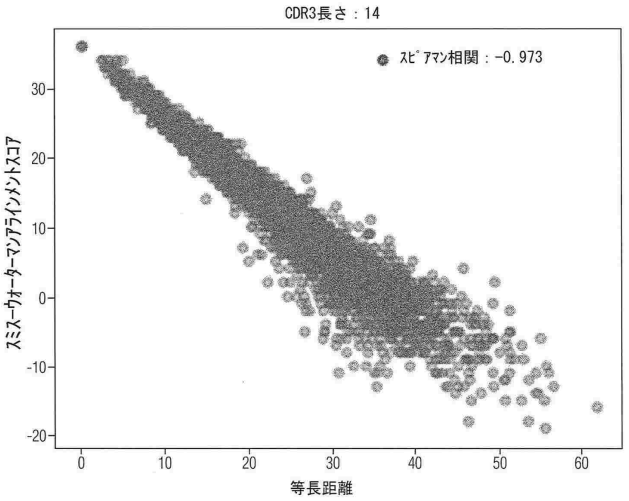
【図５Ａ】



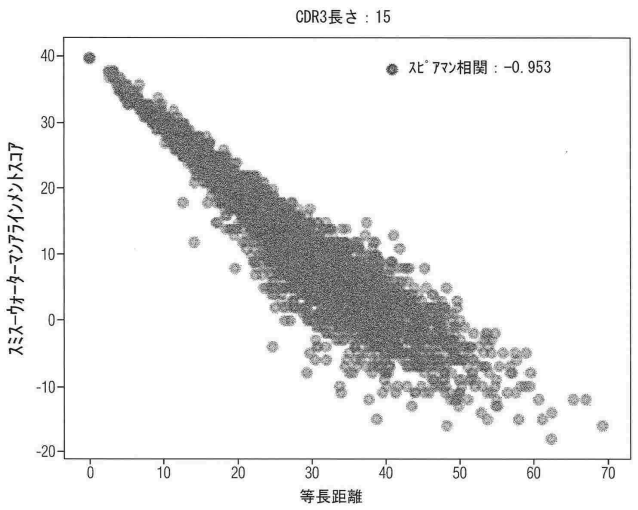
【図５Ｂ】



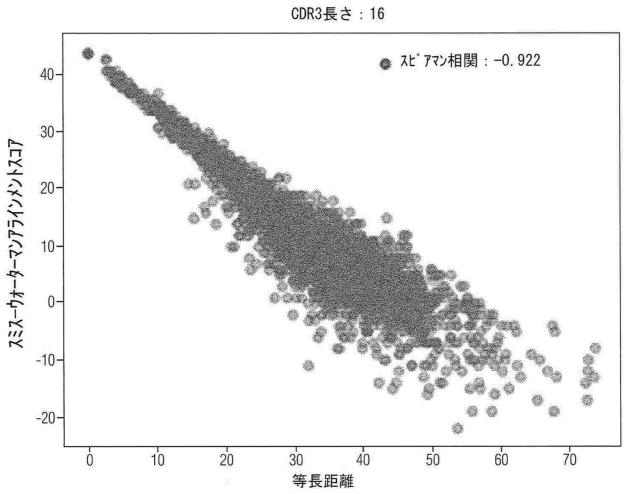
【図５Ｃ】



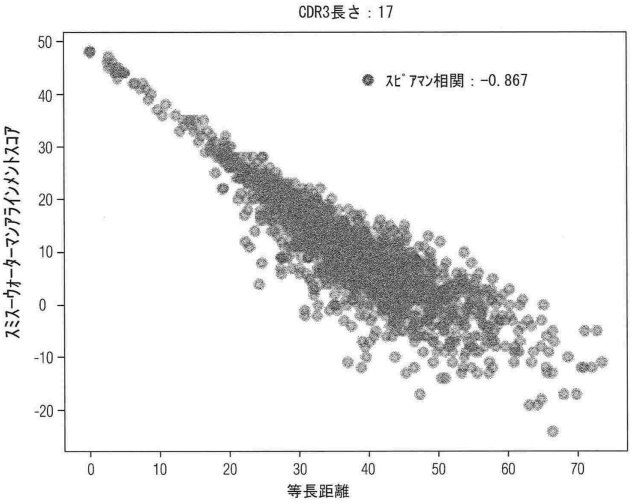
【図５Ｄ】



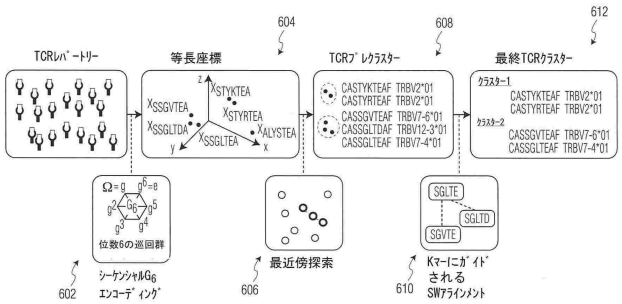
【図５Ｅ】


CDR3長さ：16
スピアマン相関：−0.922
スミスーウォーターマンアラインメントスコア
等長距離

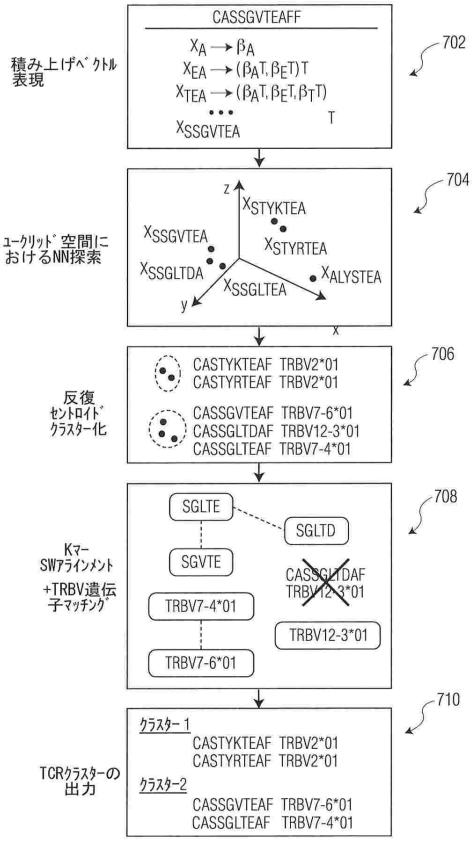
【図５Ｆ】


CDR3長さ：17
スピアマン相関：−0.867
スミスーウォーターマンアラインメントスコア
等長距離

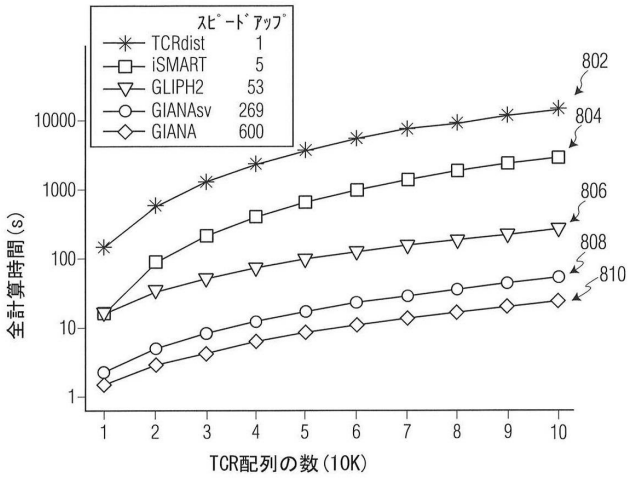
【図６】


TCRレパートリー
等長座標
TCRプレクラスター
最終TCRクラスター
604
608
612
CASTYKTEAF TRBV2*01
CASTYRTEAF TRBV2*01
CASSGVTEAF TRBV7-6*01
CASSGLTDAF TRBV12-3*01
CASSGLTEAF TRBV7-4*01
位数6の巡回群
シーケンシャルG6エンコーディング
最近傍探索
Kマーにガイドされる SWアラインメント
602
606
610

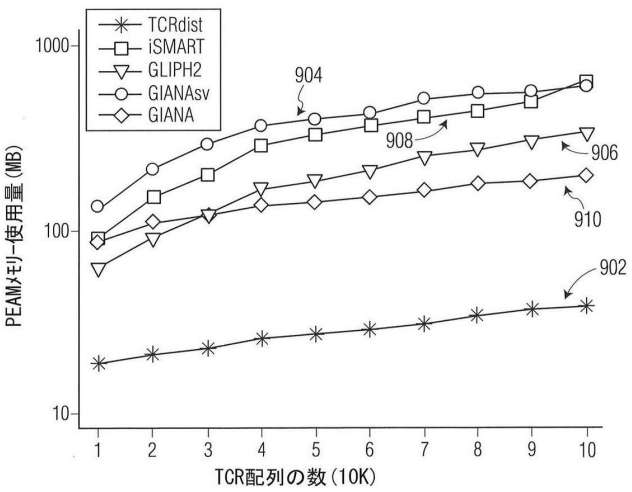
【図７】


積み上げベクトル表現
CASSGVTEAFF
ユークリッド空間におけるNN探索
反復セントロイドクラスター化
CASTYKTEAF TRBV2*01
CASTYRTEAF TRBV2*01
CASSGVTEAF TRBV7-6*01
CASSGLTDAF TRBV12-3*01
CASSGLTEAF TRBV7-4*01
Kマー SWアラインメント +TRBV遺伝子マッチング
CASSGLTDAF TRBV12-3*01
TCRクラスターの出力
クラスター1
クラスター2
702
704
706
708
710

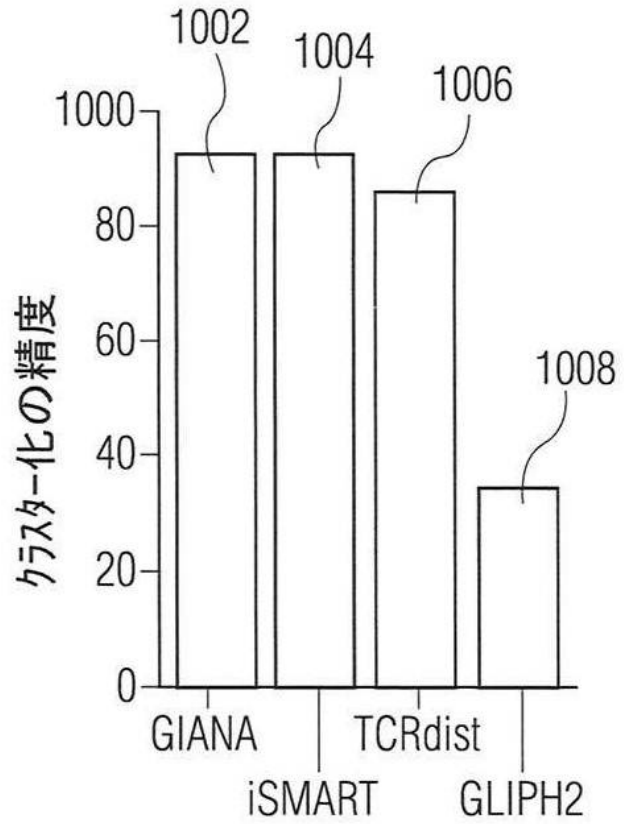
【図８】


スピードアップ
TCRdist 1
iSMART 5
GLIPH2 53
GIANAsv 269
GIANA 600
802
804
806
808
810
全計算時間(s)
TCR配列の数(10K)

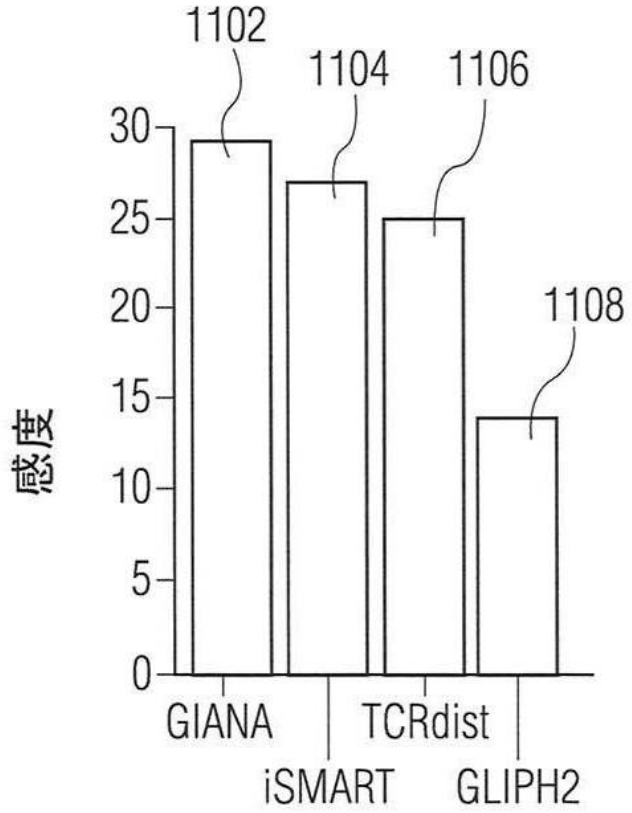
【図９】


TCRdist
iSMART
GLIPH2
GIANAsv
GIANA
904
908
906
910
902
PEAMメモリー使用量(MB)
TCR配列の数(10K)

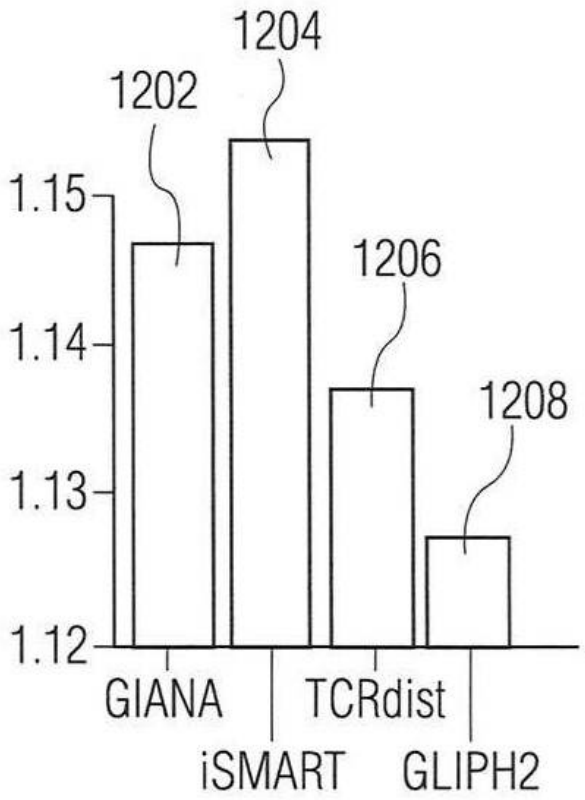
【図１０】


1002
1004
1006
1008
クラスター化の精度
GIANA
iSMART
TCRdist
GLIPH2

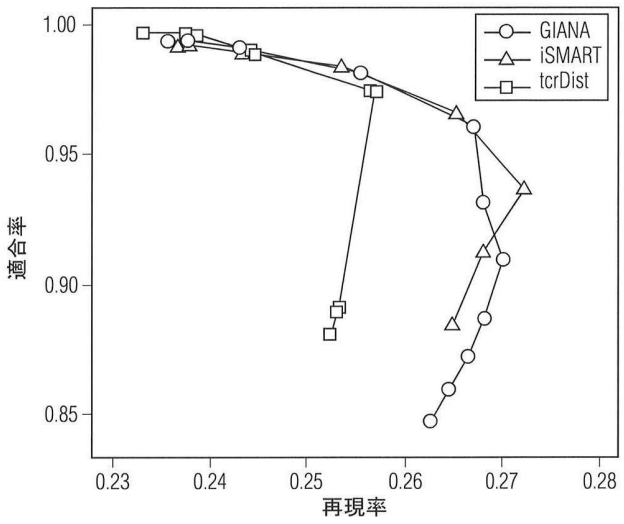
【図１１】


1102
1104
1106
1108
感度
GIANA
iSMART
TCRdist
GLIPH2

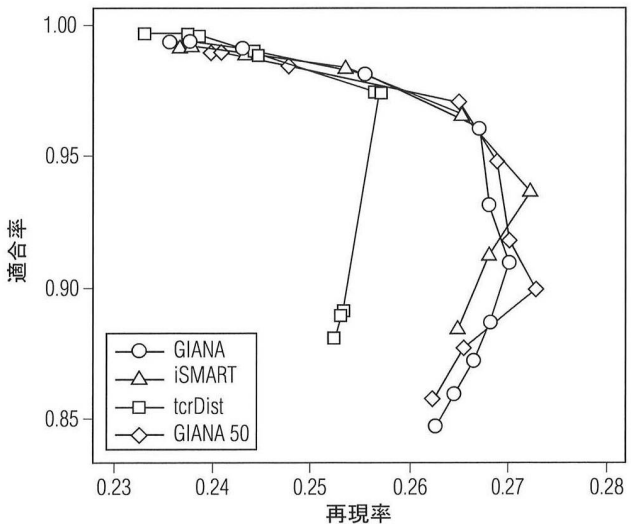
【図１２】



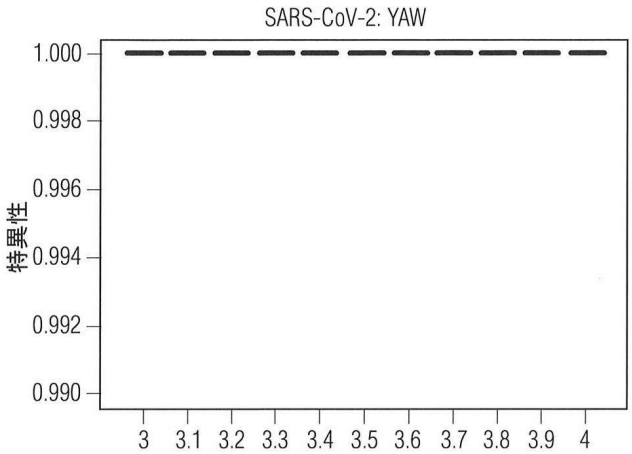
【図１３】



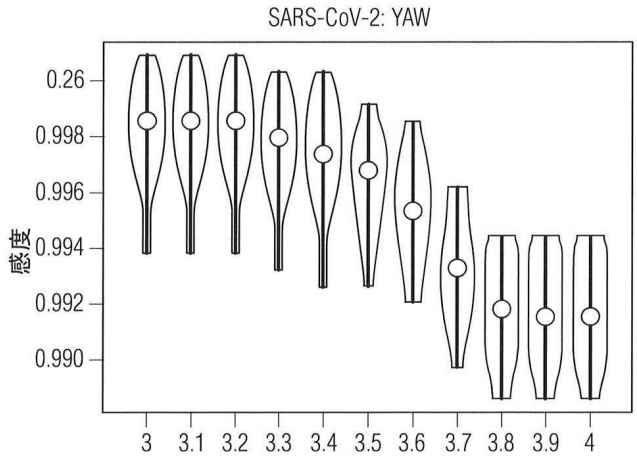
【図１４】



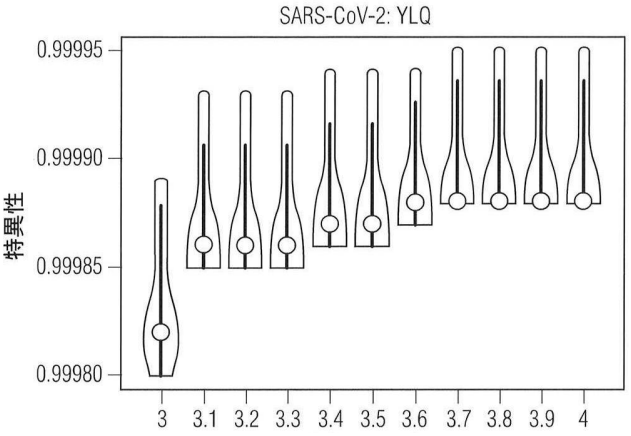
【図１５Ａ】



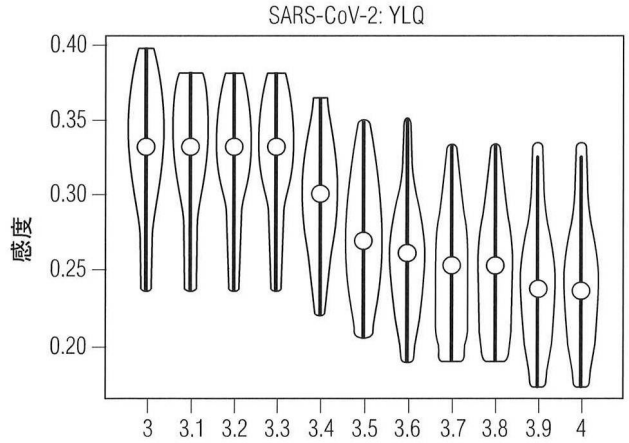
【図１５Ｂ】



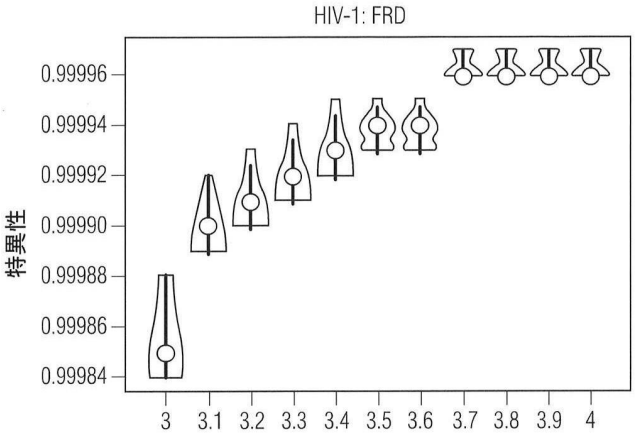
【図１５Ｃ】



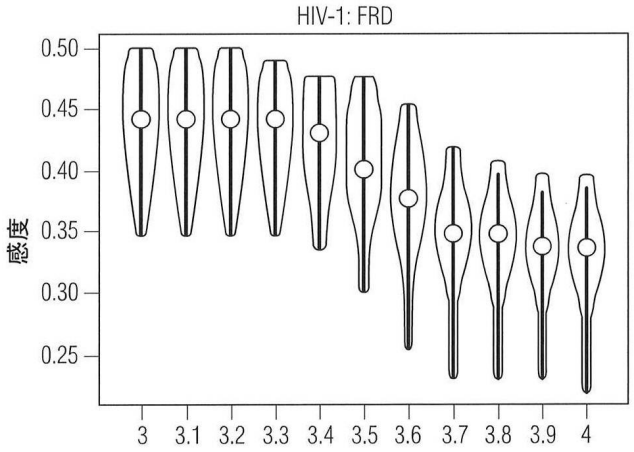
【図１５Ｄ】



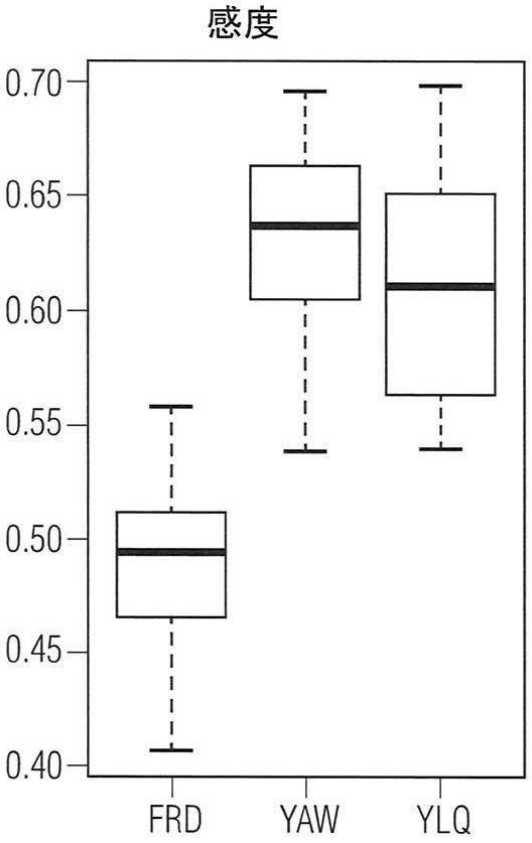
【図１５Ｅ】



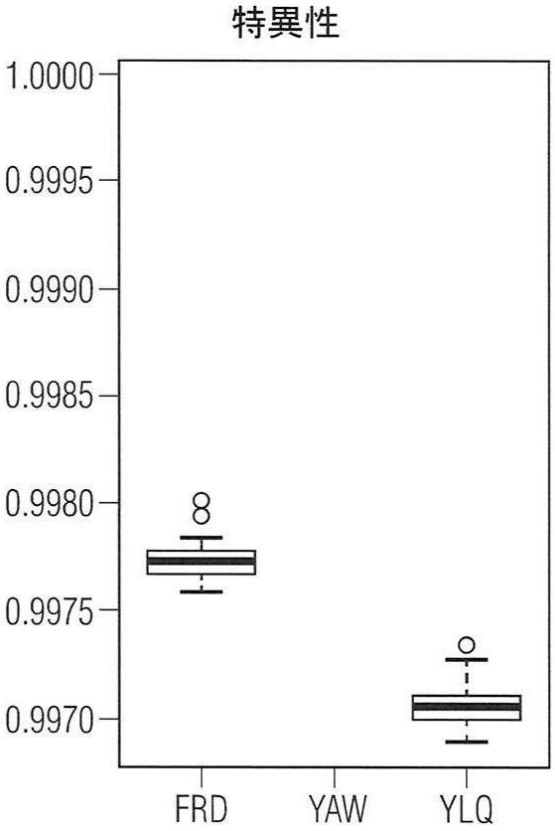
【図１５Ｆ】



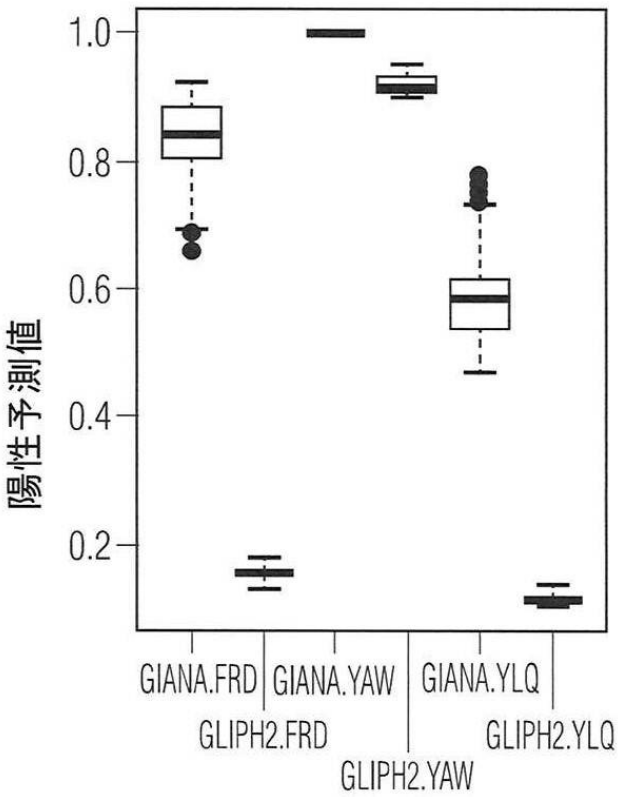
【図１６Ａ】



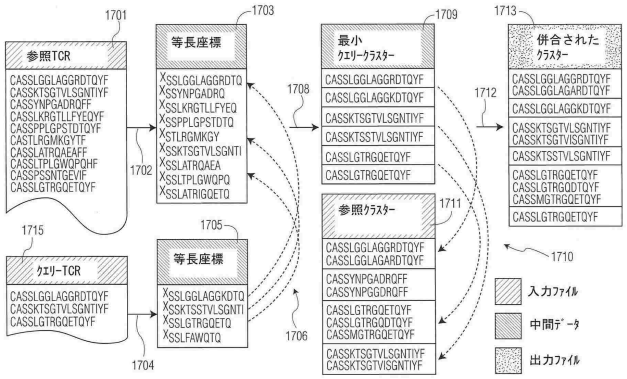
【図１６Ｂ】



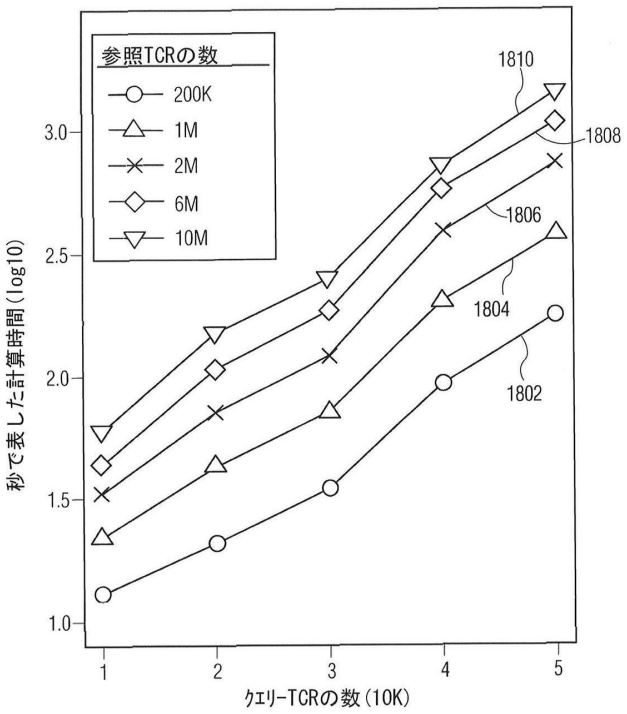
【図１６Ｃ】



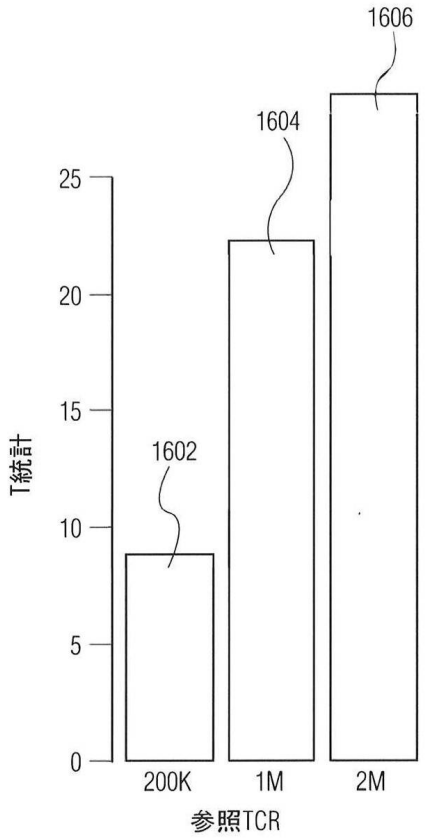
【図１７】



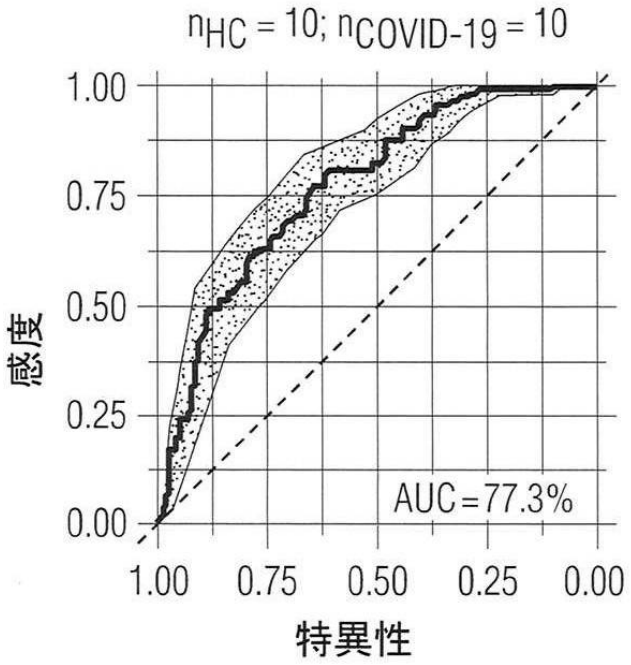
【図１８】



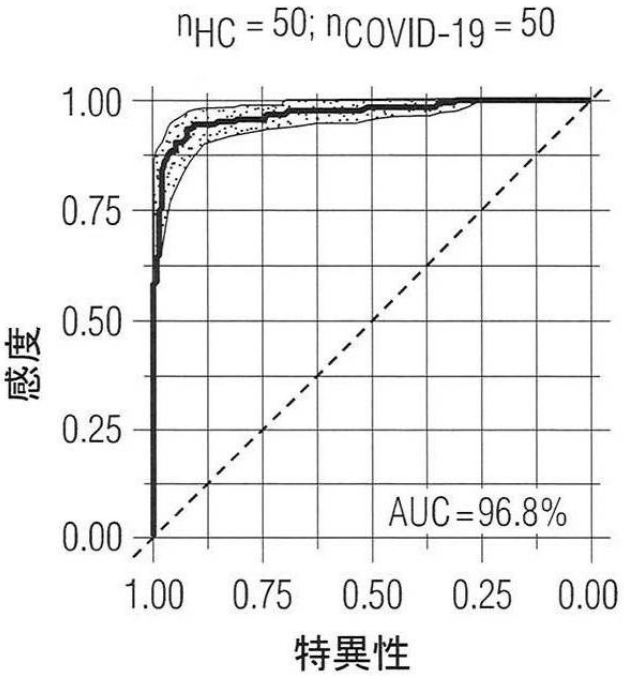
【図１９】



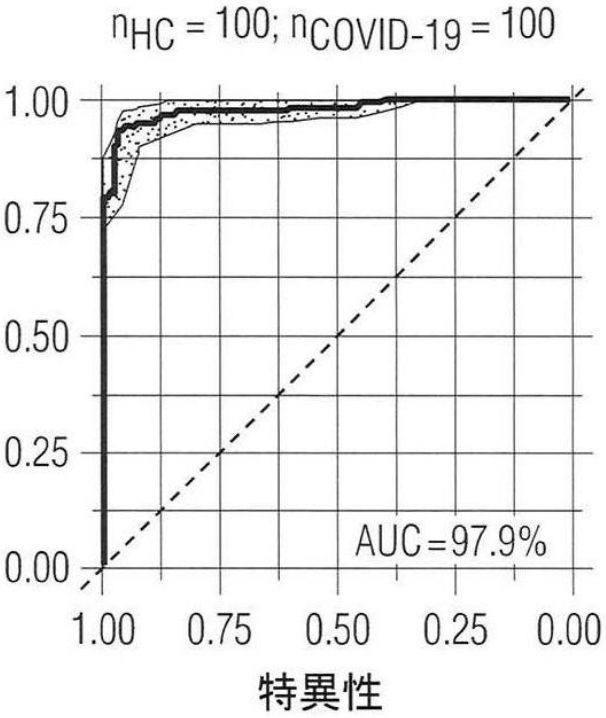
【図２０Ａ】



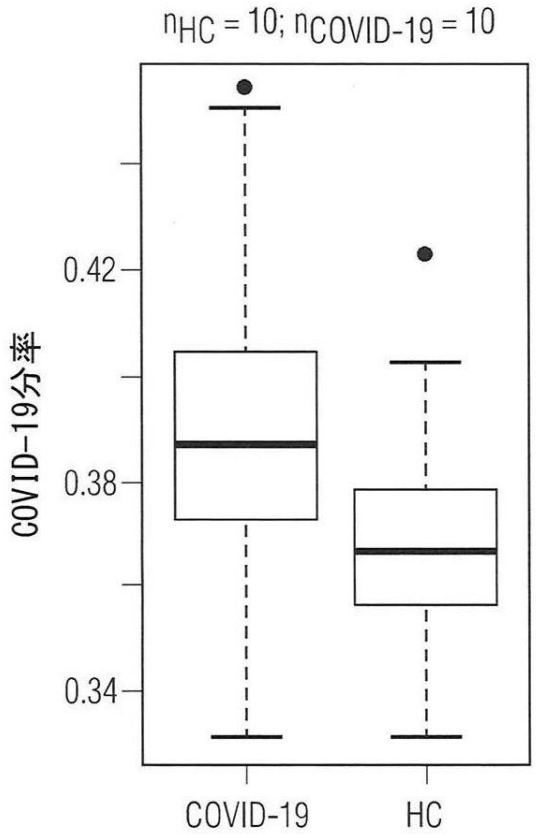
【図２０Ｂ】



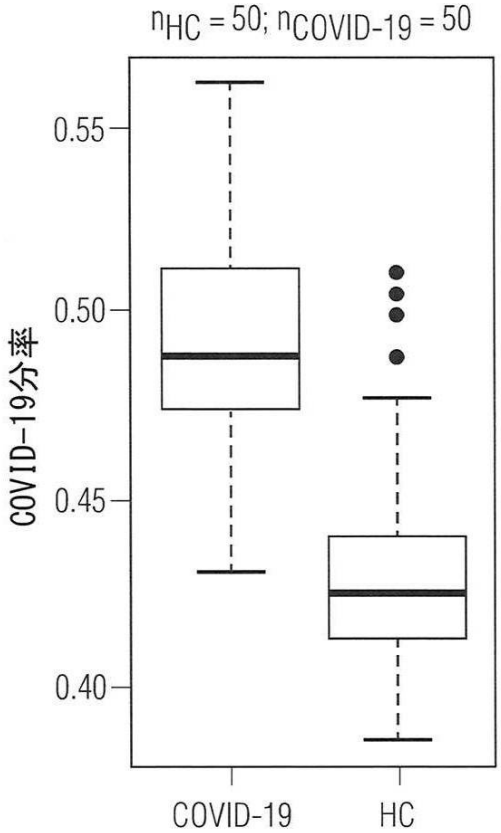
【図２０Ｃ】



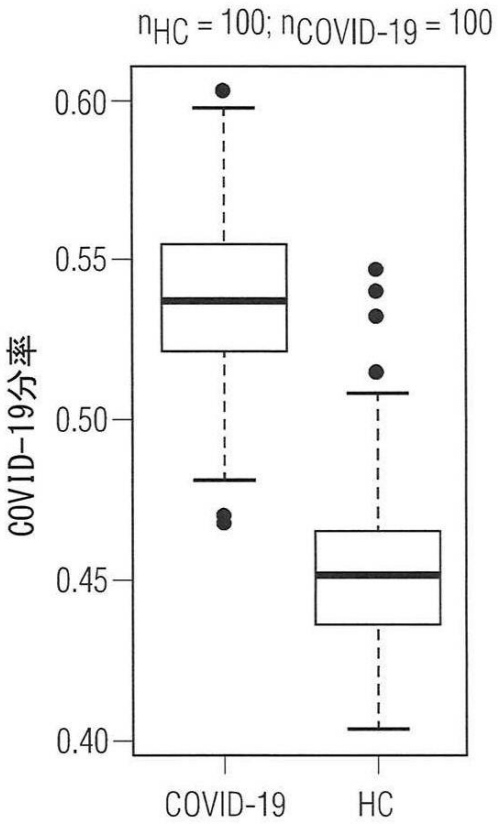
【図２１Ａ】



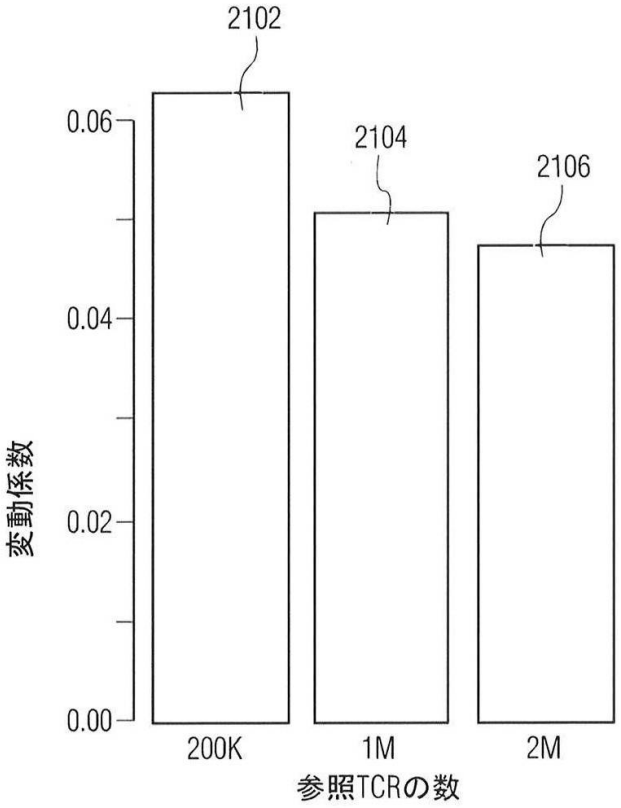
【図２１Ｂ】



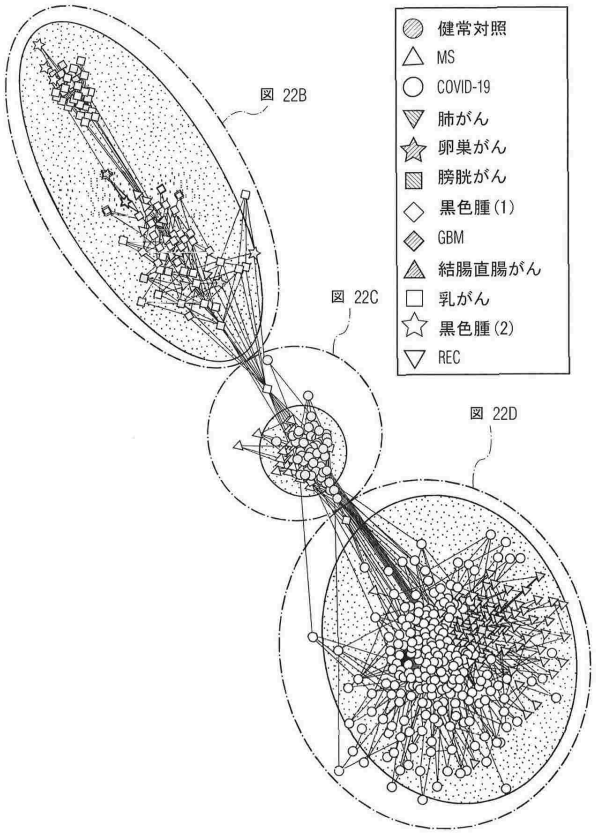
【図２１Ｃ】



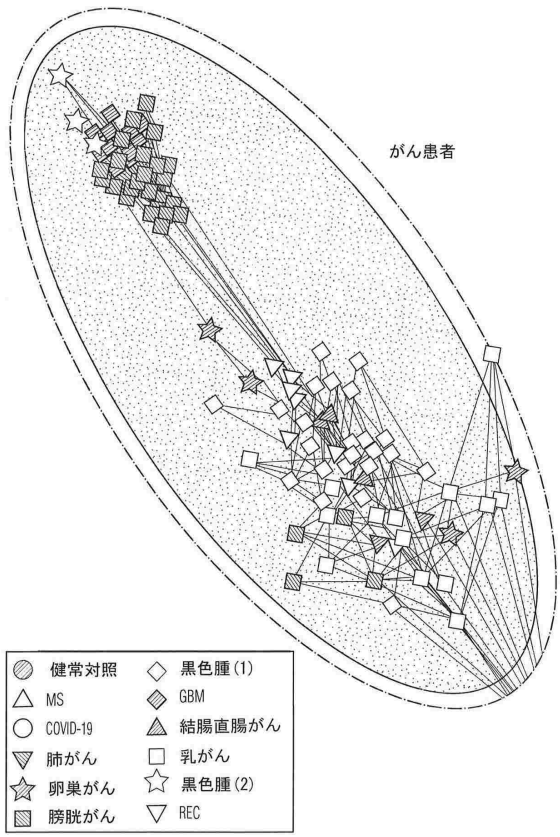
【図２１Ｄ】



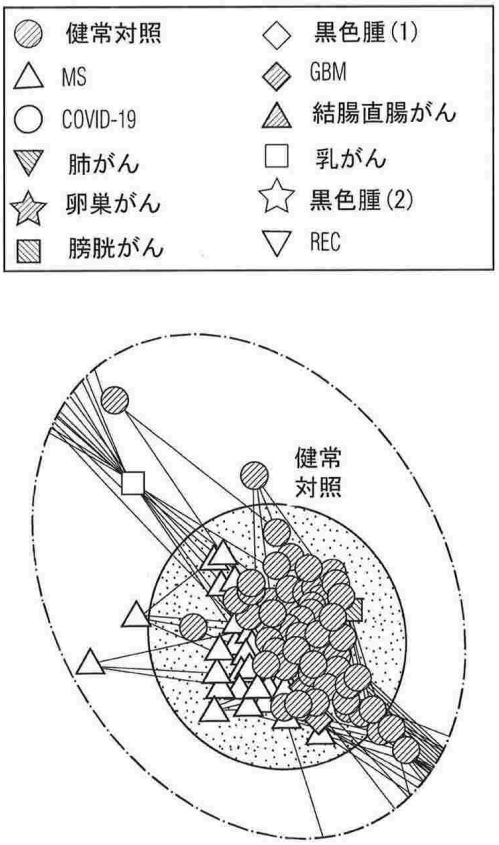
【図２２Ａ】



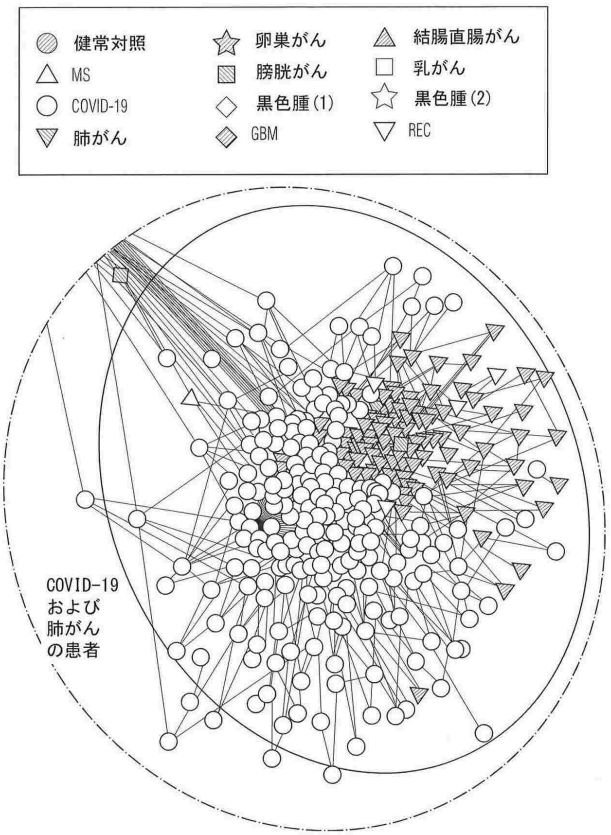
【図２２Ｂ】



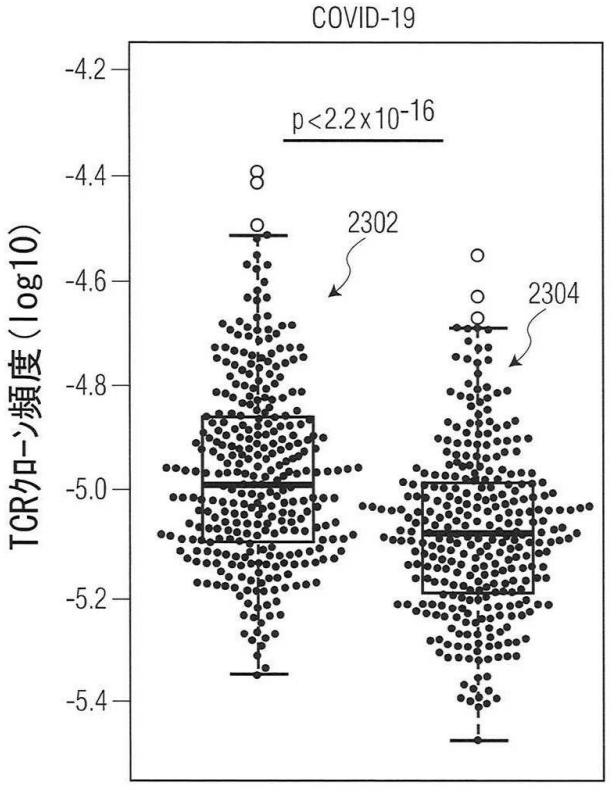
【図２２Ｃ】



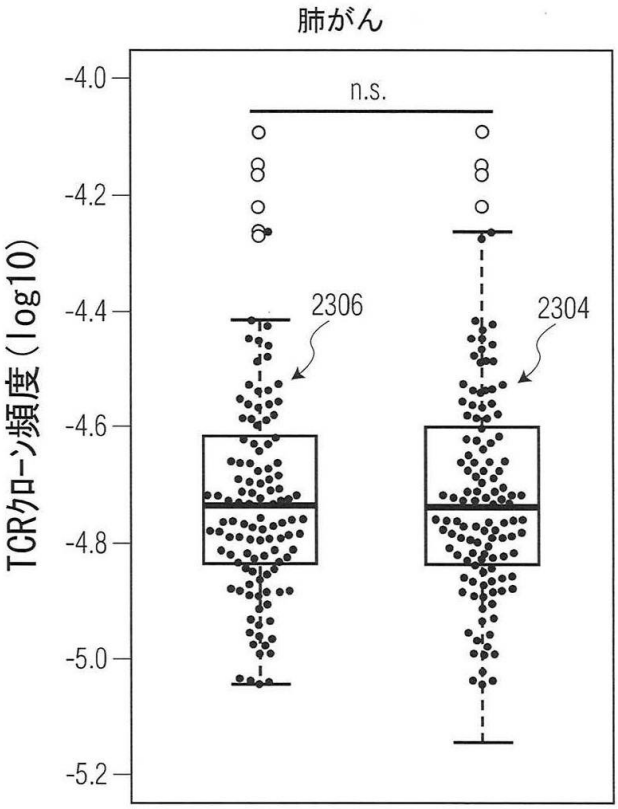
【図２２Ｄ】



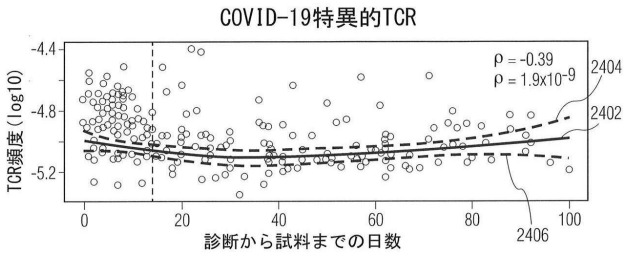
【図２３Ａ】



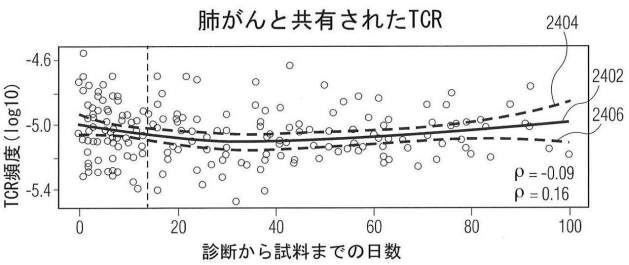
【図２３Ｂ】



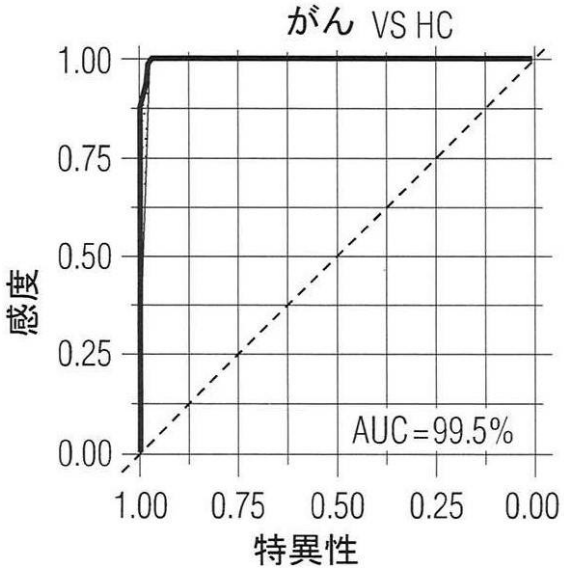
【図２４Ａ】



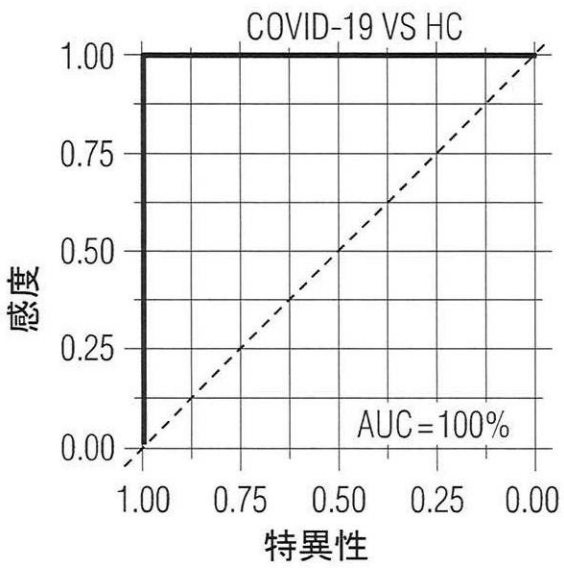
【図２４Ｂ】



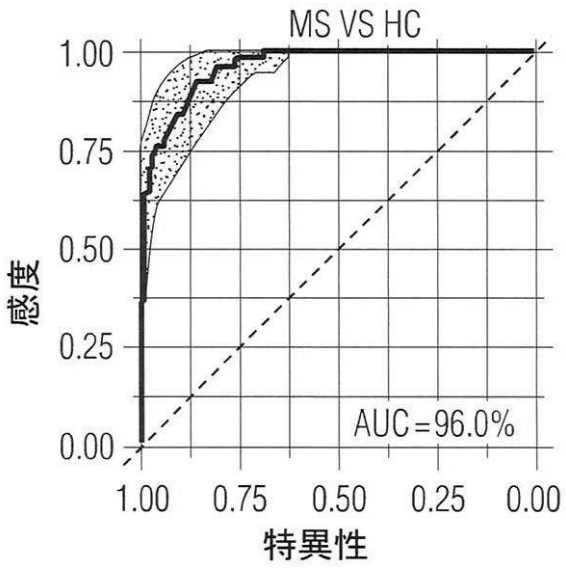
【図２５Ａ】



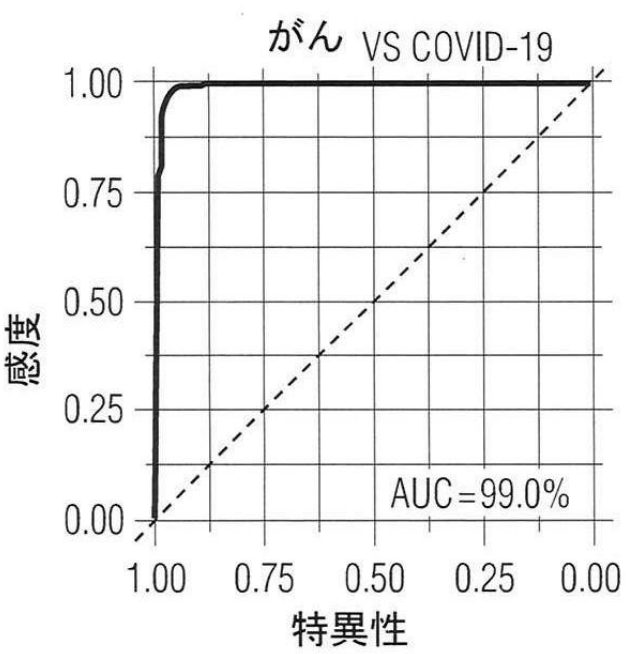
【図２５Ｂ】



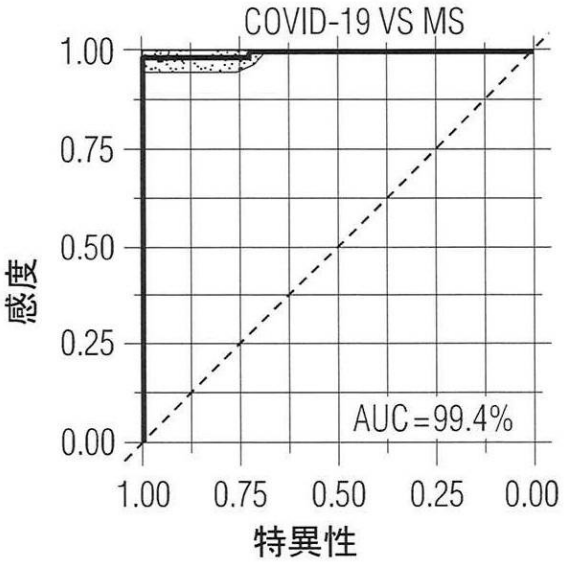
【図２５Ｃ】



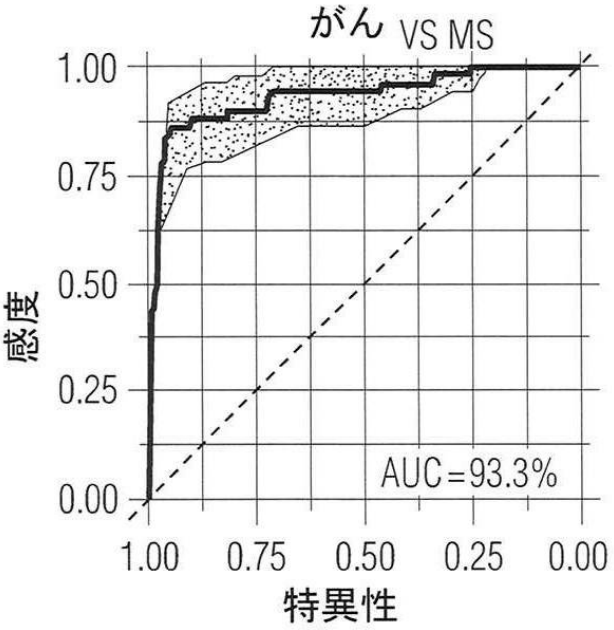
【図２５Ｄ】



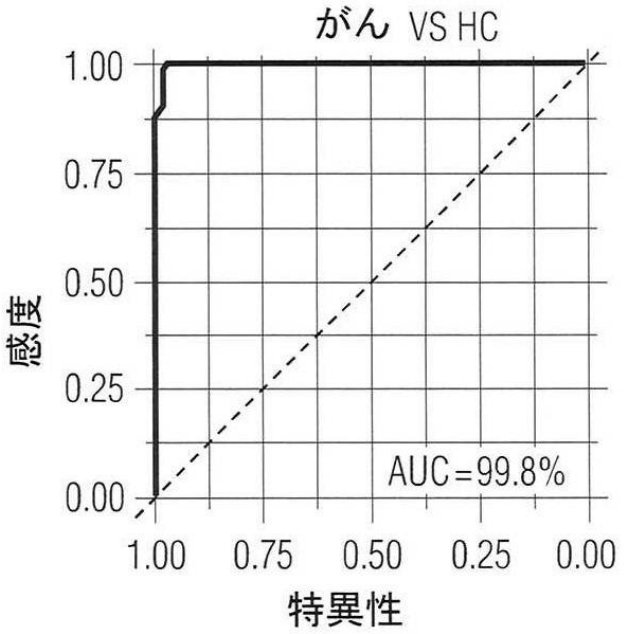
【図２５Ｅ】



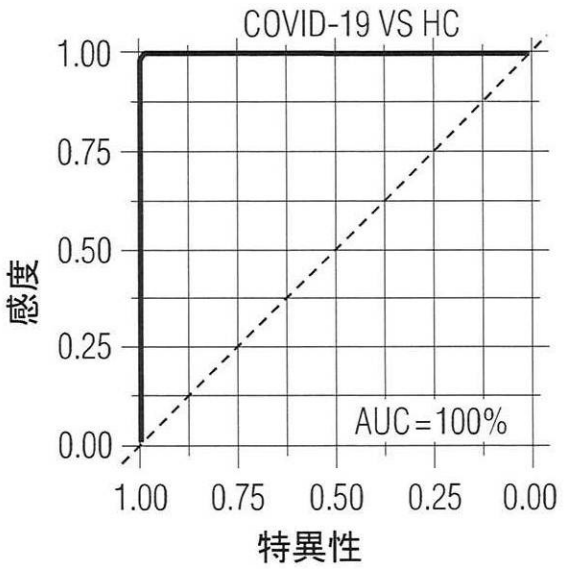
【図２５Ｆ】



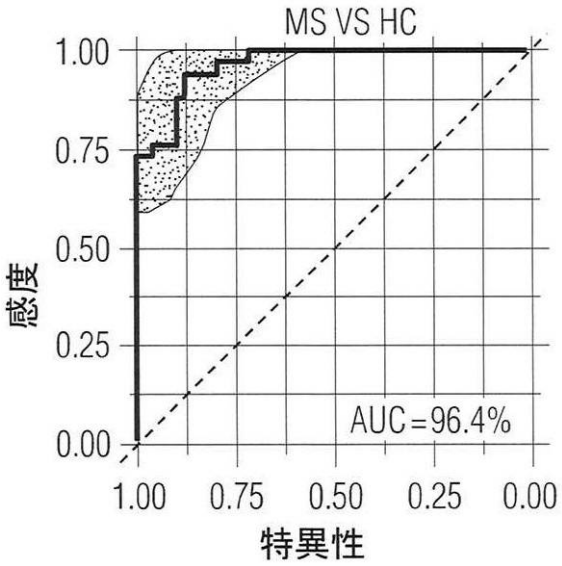
【図２６Ａ】



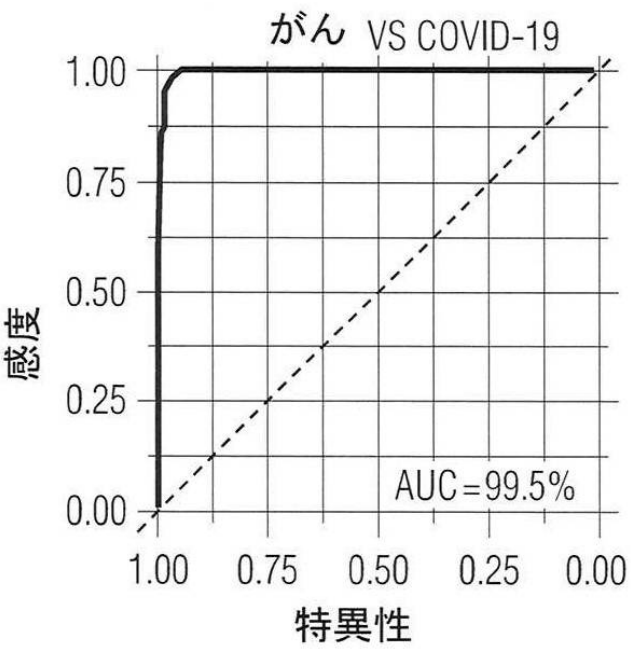
【図２６Ｂ】



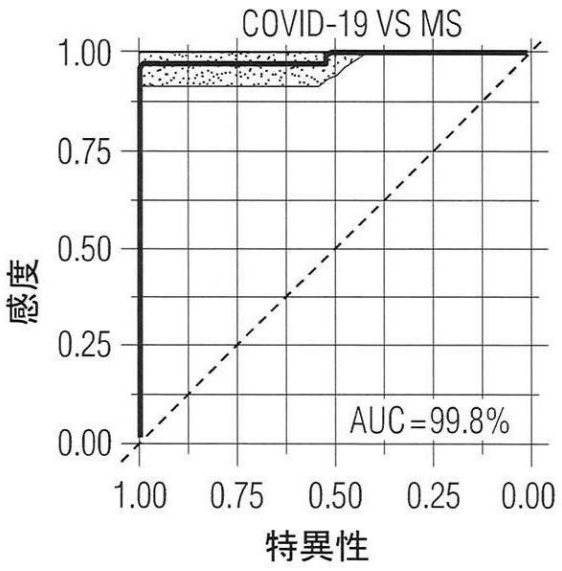
【図２６Ｃ】



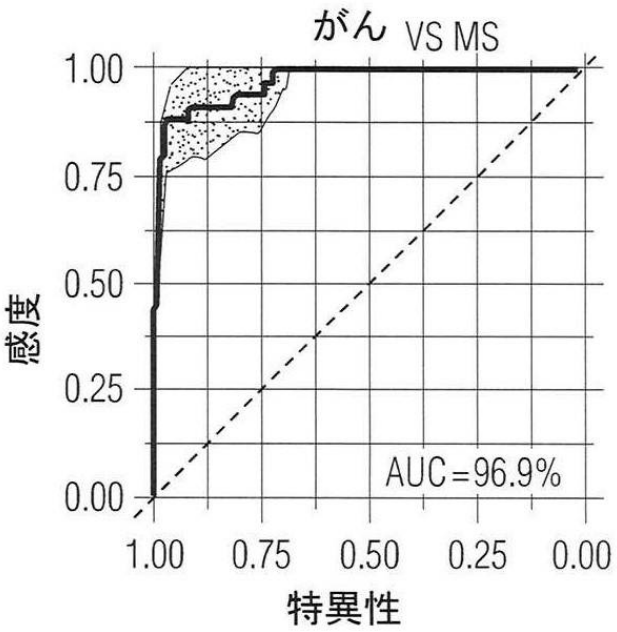
【図２６Ｄ】



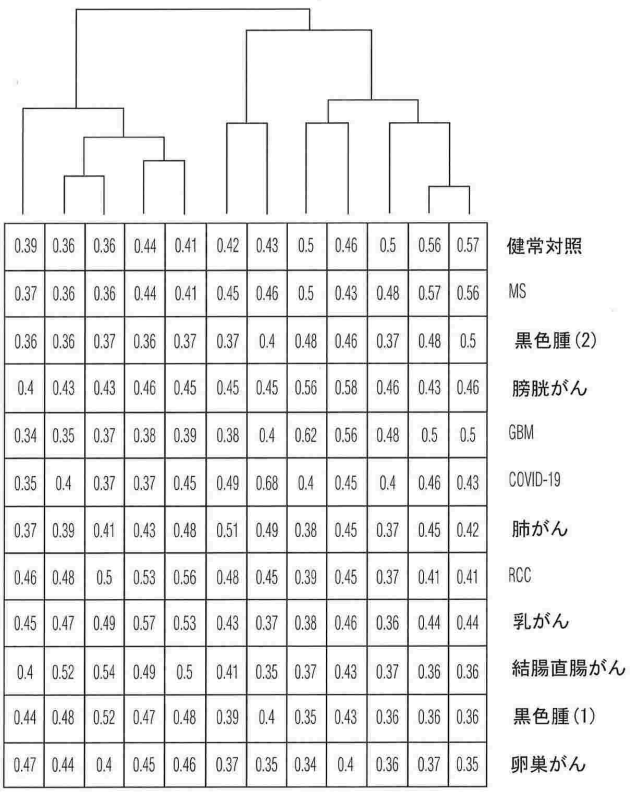
【図２６Ｅ】



【図２６Ｆ】

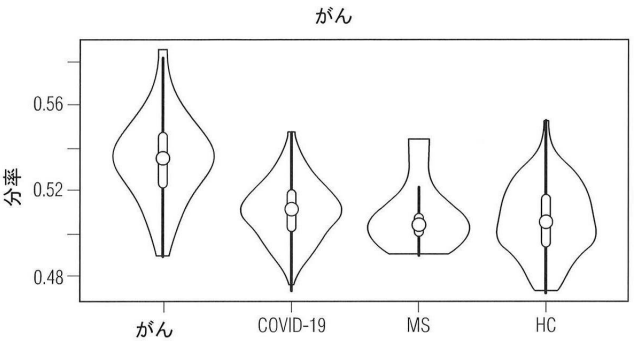


【図２７】

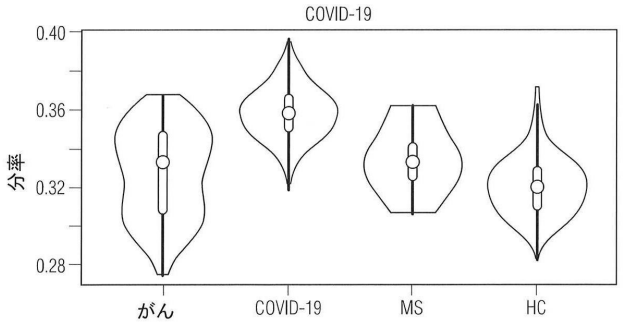


0.39	0.36	0.36	0.44	0.41	0.42	0.43	0.5	0.46	0.5	0.56	0.57	健常対照
0.37	0.36	0.36	0.44	0.41	0.45	0.46	0.5	0.43	0.48	0.57	0.56	MS
0.36	0.36	0.37	0.36	0.37	0.37	0.4	0.48	0.46	0.37	0.48	0.5	黒色腫(2)
0.4	0.43	0.43	0.46	0.45	0.45	0.45	0.56	0.58	0.46	0.43	0.46	膀胱がん
0.34	0.35	0.37	0.38	0.39	0.38	0.4	0.62	0.56	0.48	0.5	0.5	GBM
0.35	0.4	0.37	0.37	0.45	0.49	0.68	0.4	0.45	0.4	0.46	0.43	COVID-19
0.37	0.39	0.41	0.43	0.48	0.51	0.49	0.38	0.45	0.37	0.45	0.42	肺がん
0.46	0.48	0.5	0.53	0.56	0.48	0.45	0.39	0.45	0.37	0.41	0.41	RCC
0.45	0.47	0.49	0.57	0.53	0.43	0.37	0.38	0.46	0.36	0.44	0.44	乳がん
0.4	0.52	0.54	0.49	0.5	0.41	0.35	0.37	0.43	0.37	0.36	0.36	結腸直腸がん
0.44	0.48	0.52	0.47	0.48	0.39	0.4	0.35	0.43	0.36	0.36	0.36	黒色腫(1)
0.47	0.44	0.4	0.45	0.46	0.37	0.35	0.34	0.4	0.36	0.37	0.35	卵巣がん

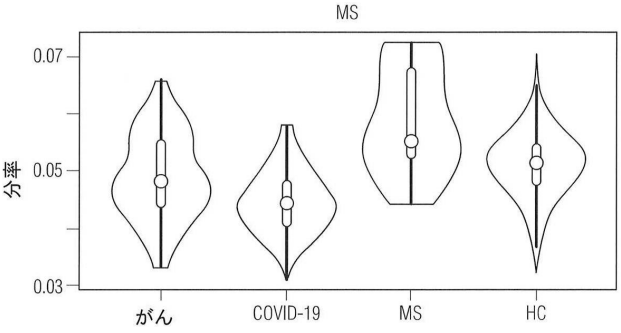
【図２８Ａ】



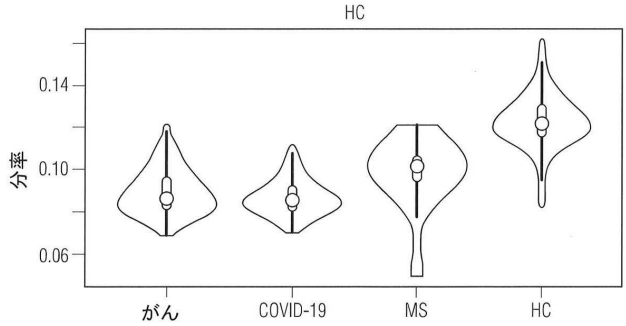
【図２８Ｂ】



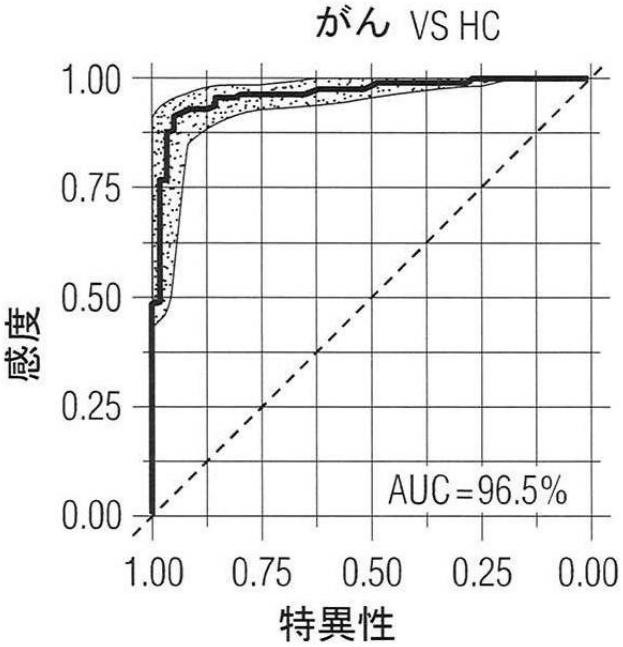
【図２８Ｃ】



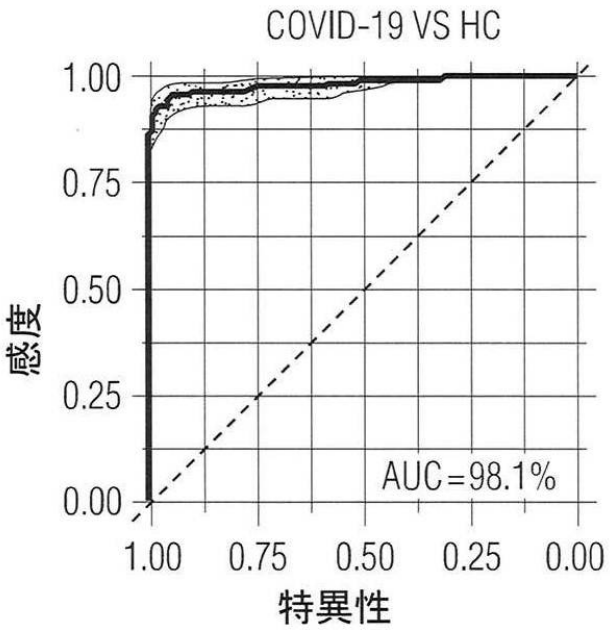
【図２８Ｄ】



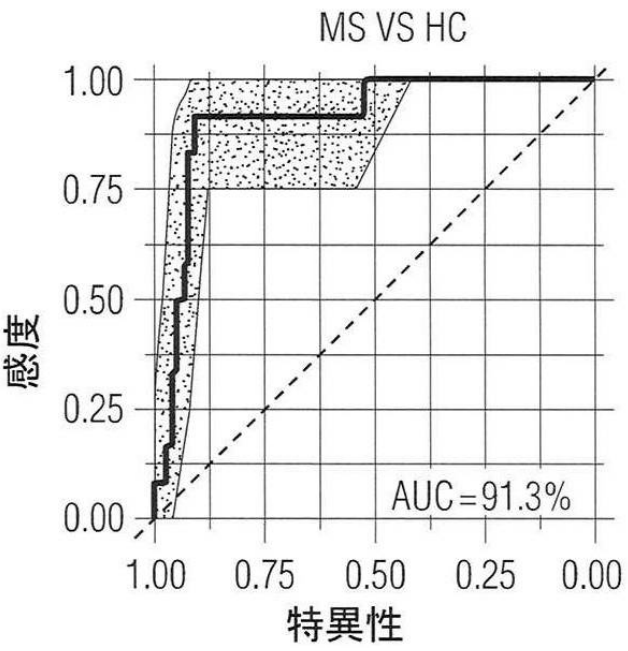
【図２９Ａ】



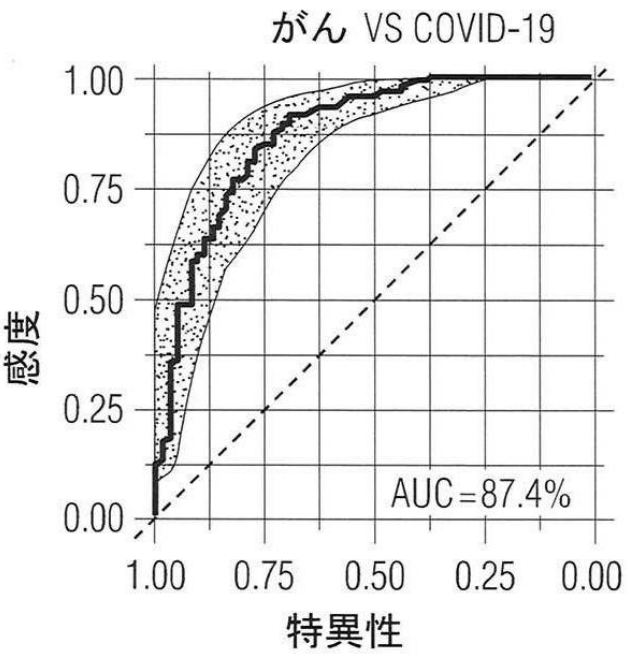
【図２９Ｂ】



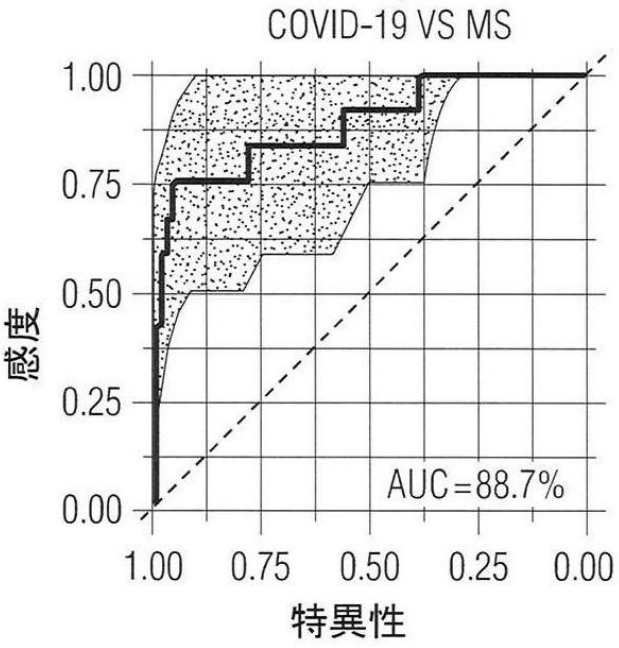
【図２９Ｃ】



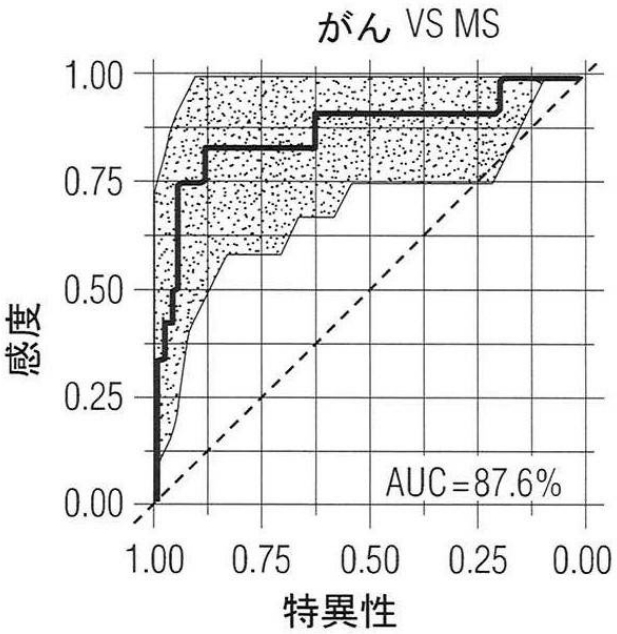
【図２９Ｄ】



【図２９Ｅ】



【図２９Ｆ】



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2022/034068

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - INV. - G06N 3/12; G06N 20/20; C40B 30/04; G06F 19/18; G16B 20/00 (2022.01)
ADD.

CPC - INV. - G06N 3/123; G06N 20/20; G16B 20/00; G16B 30/00; G16B 40/20 (2022.08)
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X — Y	WO 2020/197820 A1 (BOARD OF REGENTS OF THE UNIVERSITY OF TEXAS SYSTEM) 01 October 2020 (01.10.2020) entire document	1, 3, 4 — 5, 6
Y	US 2021/0002722 A1 (THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY) 07 January 2021 (07.01.2021) entire document	5
Y	JOSHI et al. "Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer," Nature Medicine, 01 October 2019 (01.10.2019), Vol. 25, Iss. 10, Pgs. 1549-1559. entire document	6
A	WO 2020/191079 A1 (THE BROAD INSTITUTE INC. et al) 24 September 2020 (24.09.2020) entire document	1-6
A	US 2018/0201991 A1 (REGENERON PHARMACEUTICALS INC.) 19 July 2018 (19.07.2018) entire document	1-6
P, X	WO 2022/072722 A1 (THE BOARD OF REGENTS OF THE UNIVERSITY OF TEXAS SYSTEM) 07 April 2022 (07.04.2022) entire document	1-6

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 October 2022	NOV 04 2022
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Talna Matos Telephone No. PCT Helpdesk: 571-272-4300

Form PCT/ISA/210 (second sheet) (July 2019)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2022/034068

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

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

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

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

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

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

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

See extra sheet(s).

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

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

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

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-6

Remark on Protest

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

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

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2022/034068

Continued from Box No. III Observations where unity of invention is lacking

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

Group I: claims 1-6 are drawn to a method of improving computational efficiency for T-cell receptor (TCR) comparisons.

Group II: claims 7-12 are drawn to computing devices.

Group III: claims 13-18 are drawn to a non-transitory computer-readable storage medium.

Group IV: claims 19 and 20 are drawn to a method of organizing and querying a T-cell receptor (TCR) database.

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

The special technical features of Group I, a method of improving computational efficiency for T-cell receptor (TCR) comparisons, are not present in Groups II-IV; the special technical features of Group II, computing device, are not present in Groups I, III, or IV; the special technical features of Group III, a non-transitory computer-readable storage medium, are not present in Groups I, II, or IV; and the special technical features of Group IV, a method of organizing and querying a T-cell receptor (TCR) database, are not present in Groups I-III.

Additionally, even if Groups I-IV were considered to share the technical features of identifying, by a computing device, complementary determining region 3 (CDR3) sequences from a reference TCR sequence (TCR-seq) dataset, the reference TCR-seq dataset consisting of TCRs specific to only one epitope; encoding, by the computing device, each of the CDR3 sequences from the reference TCR-seq dataset into numeric vectors, the numeric vectors corresponding to a sequence of amino acids in each of the CDR3 sequences; converting, by the computing device, the numeric vectors to coordinates in a high-dimensional Euclidean space; generating, by the computing device, a predictive model using a neural network by: learning, by the neural network, to generate a tree data structure of the numeric vectors based on relative distances of the coordinates, and grouping, by the neural network, the coordinates into pre-clusters based on the relative distances; filtering, by the computing device, the CDR3 sequences in the pre-clusters; and identifying, by the computing device, antigen-specific CDR3 clusters from the filtered pre-clusters. However, these shared technical features do not represent a contribution over the prior art.

Specifically, WO 2020/197820 to Board Of Regents Of The University Of Texas System discloses identifying, by a computing device, complementary determining region 3 (CDR3) sequences from a reference TCR sequence (TCR-seq) dataset (rapid accumulation of TCR-seq data, more computational tools are required to bridge the gap between basic immunogenomics research and clinical applications beneficial to cancer patients, Para. [0058]; an open source algorithm for calling the TCR transcript hypervariable CDR3 regions, Para. [0060]; a valid training dataset for cancer-associated TCRs, Para. [0067]), the reference TCR-seq dataset consisting of TCRs specific to only one epitope (, In the TCR repertoire sequencing, one read is sufficient to cover one CDR3 region. Therefore, sequencing read counts as CDR3 counts for each clonotype can be used, Para. [0128]); encoding, by the computing device, each of the CDR3 sequences from the reference TCR-seq dataset into numeric vectors, the numeric vectors corresponding to a sequence of amino acids in each of the CDR3 sequences (here the data and metadata relationship can be stored as the n-dimensional vector. Such storage can be realized through any known or to be known vector or array storage, including but not limited to, a hash tree, queue, stack, VList, or any other type of known or to be known dynamic memory allocation technique or technology. It should be understood that any known or to be known computational analysis technique or algorithm, such as, but not limited to, cluster analysis, data mining, Bayesian network analysis, Hidden Markov models, artificial neural network analysis, logical model and/or tree analysis, and the like, and be applied to determine, derive or otherwise identify vector information for patients and/or health care providers, Para. [0053]); converting, by the computing device, the numeric vectors to coordinates in a high-dimensional Euclidean space (Euclidean distances between origin and each of the CDR3 Ca positions were calculated across all the structures. Standard deviation of the distance for each of the positions was then calculated and displayed. Visualization of selected PDB structures for the -6 position of the b chain CDR3 region was performed using Chimera and PyMol, Para. [0121]); generating, by the computing device, a predictive model using a neural network by: learning, by the neural network, to generate a tree data structure of the numeric vectors based on relative distances of the coordinates (and ensemble tree classifiers are then trained for CDR3s with different lengths using the AdaBoost algorithm (or other supervised machine learning methods, including the deep neural network models), as discussed above and in more detail below. Steps 314-316. Each TCR-seq sample was pre-processed (Step 318), and clustered by immuno- similarly measurement (iSMART) (Step 320) to identify antigen-specific groups (Step 322). Then trained tree classifiers (e.g., trained from Step 314) are applied to the grouped CDR3s to evaluate a cancer score, related to the probability of an immune repertoire being cancer- associated, Para. [0094]), and grouping, by the neural network, the coordinates into pre-clusters based on the relative distances (and ensemble tree classifiers are then trained for CDR3s with different lengths using the AdaBoost algorithm (or other supervised machine learning methods, including the deep neural network models), as discussed above and in more detail below. Steps 314-316. Each TCR-seq sample was pre-processed (Step 318), and clustered by immuno- similarly measurement (iSMART) (Step 320) to identify antigen-specific groups (Step 322). Then trained tree classifiers (e.g., trained from Step 314) are applied to the grouped CDR3s to evaluate a cancer score, related to the probability of an immune repertoire being cancer- associated, Para. [0094]); filtering, by the computing device, the CDR3 sequences in the pre-clusters; and identifying, by the computing device, antigen-specific CDR3 clusters from the filtered pre-clusters (the TCGA-derived CDR3s can be filtered in for complete sequence starting with the last cysteine (C) from the variable gene, and the phenylalanine (F) in the FGXG motif in the joining gene. The non-productive sequences containing stop codon between C and F can be excluded. To remove public TCRs that are also found in non-cancer individuals, the top most abundant CDR3s from a cohort of PBMC repertoire samples (e.g., the CDR3s satisfying a threshold - for example, the top 5,000 from 666 healthy or HCMV infected patients) can be collected and filtered out from the set. The resulting CDR3 sequences (e.g., 43,000 CDR3) are expected to be non-public and cancer associated, Para. [0101]).

The inventions listed in Groups I-IV therefore lack unity under Rule 13 because they do not share a same or corresponding special technical features.

フロントページの続き

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

（特許庁注：以下のものは登録商標）

１．ＰＹＴＨＯＮ

弁理士　森田 裕
(74)代理人　100162503
弁理士　今野 智介
(74)代理人　100144794
弁理士　大木 信人
(74)代理人　100204582
弁理士　大栗 由美
(72)発明者　リー，ボー
アメリカ合衆国 ７８７０１ テキサス州，オースティン，ウエスト セブンス ストリート ２１０