(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6141310号
(P6141310)
(45)発行日　平成29年6月7日(2017.6.7)　(24)登録日　平成29年5月12日(2017.5.12)

(51)Int.Cl.		F I		
G06F 19/20	(2011.01)	G06F	19/20	
C12Q 1/68	(2006.01)	C12Q	1/68	A
C12N 15/09	(2006.01)	C12N	15/00	A

請求項の数 9　(全 21 頁)

(21)出願番号	特願2014-545409 (P2014-545409)
(86)(22)出願日	平成24年12月3日 (2012.12.3)
(65)公表番号	特表2015-501987 (P2015-501987A)
(43)公表日	平成27年1月19日 (2015.1.19)
(86)国際出願番号	PCT/IB2012/056911
(87)国際公開番号	WO2013/084133
(87)国際公開日	平成25年6月13日 (2013.6.13)
審査請求日	平成27年11月30日 (2015.11.30)
(31)優先権主張番号	61/568,336
(32)優先日	平成23年12月8日 (2011.12.8)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	11193585.4
(32)優先日	平成23年12月14日 (2011.12.14)
(33)優先権主張国	欧州特許庁 (EP)

(73)特許権者　590000248
コーニンクレッカ　フィリップス　エヌヴェ
KONINKLIJKE PHILIPS N. V.
オランダ国　５６５６　アーエー　アインドーフェン　ハイテック　キャンパス　５
High Tech Campus 5, NL-5656 AE Eindhoven
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦

最終頁に続く

(54)【発明の名称】強固な変異体特定および検証

(57)【特許請求の範囲】
【請求項１】
整列された配列決定リードを含む組み立てられた遺伝子配列中の可能性がある変異体を特定するステップと、
前記組み立てられた遺伝子配列の配列決定リードについて複数のリード特性の値を計算するステップと、
受入れ基準を満たす前記可能性がある変異体を含む前記組み立てられた遺伝子配列の配列決定リードについての、前記少なくとも１つのリード特性の前記計算された値を条件として、前記可能性がある変異体をコールするステップと、
を含む方法であって、
前記可能性がある変異体が、前記可能性がある変異体を含む前記リードについての、前記複数のリード特性の計算された値に影響を及ぼす分類子を使用して、受け入れられるもの、または拒絶されるものとして分類され、前記可能性がある変異体が、前記可能性がある変異体が受け入れられるものとして分類される場合にだけコールされ、分類が、Ｆスコアを使用するサポート・ベクター・マシン分類を使用して行われ、
前記特定するステップ、計算するステップ、およびコールするステップが電子データ処理デバイスによって実施され、
前記少なくとも１つのリード特性が、前記配列決定リードの物理的特性を含む、方法。
【請求項２】
前記少なくとも１つのリード特性が少なくとも１つの塩基組成特性を含む、請求項１に

記載の方法。
【請求項３】
前記少なくとも１つのリード特性が少なくとも１つの熱力学的特性を含む、請求項１または２に記載の方法。
【請求項４】
前記少なくとも１つのリード特性が、リード安定性を示す少なくとも１つのエネルギー特性を含む、請求項１～３のいずれか１項に記載の方法。
【請求項５】
前記少なくとも１つのリード特性が少なくとも１つの二次構造特性を含む、請求項１～４のいずれか１項に記載の方法。
【請求項６】
配列決定リードのセットを組み立てて、前記組み立てられた遺伝子配列を生成するステップをさらに含み、
塩基クオリティ・スコアおよびカバー度基準を満たさない配列決定リードが破棄され、前記組み立てられた遺伝子配列に含まれず、
前記組み立てられた遺伝子配列が全ゲノム配列を含む、請求項１～５のいずれか１項に記載の方法。
【請求項７】
請求項１～６のいずれか１項に記載の方法が含む各ステップを、電子データ処理デバイスに実行させるコンピュータプログラム。
【請求項８】
整列された配列決定リードを含む組み立てられた遺伝子配列を記憶する持続性記憶媒体と、
電子処理デバイスであって、
前記組み立てられた遺伝子配列中の可能性がある変異体を特定する動作、
前記組み立てられた遺伝子配列のリードについて複数のリード特性の値を計算する動作、および、
受入れ基準を満たす前記可能性がある変異体を含む前記組み立てられた遺伝子配列の配列決定リードについての、前記少なくとも１つのリード特性の前記計算された値を条件として、前記可能性がある変異体をコールする動作、
を含む動作を行うように構成された電子処理デバイスと、
を備える装置であって、
前記可能性がある変異体が、前記可能性がある変異体を含む前記リードについての、前記複数のリード特性の前記計算された値に影響を及ぼす分類子を使用して、受け入れられるもの、または拒絶されるのとして分類され、前記可能性がある変異体が、前記可能性がある変異体が受け入れられるものとして分類される場合にだけコールされ、分類が、Ｆスコアを使用するサポート・ベクター・マシン分類を使用して行われ、
前記少なくとも１つのリード特性が、前記配列決定リードの物理的特性を含む、装置。
【請求項９】
前記少なくとも１つのリード特性が、塩基組成特性、熱力学的特性、エネルギー特性、および二次構造特性からなる群から選択された少なくとも１つのリード特性を含む、請求項８に記載の装置。
【発明の詳細な説明】
【技術分野】
【０００１】
以下は遺伝子解析技術に関し、また医療技術、獣医学技術、腫瘍学技術などの、その応用分野に関する。
【背景技術】
【０００２】
遺伝子解析には、腫瘍学および様々な遺伝学的に関連する疾病など、多数の医学および

獣医学の分野に用途が見出される。従来、臨床研究は、特定の癌または疾病と相互にその存在が関係する遺伝子変異体を特定するために行われている。このような変異体には、例示的なものとして、一塩基多型（ＳＮＰ）、挿入または欠失（インデル）、コピー数多型（ＣＮＶ）、染色体異常または再配置などが含まれうる。

【０００３】

遺伝子解析は通常、以下の作業を伴う。組織試料を患者から採取および処理し、シークエンサーの中にロードする。この装置は、長さが通常は数ダースから数百塩基程度の遺伝子配列の短い部分を表す配列決定「リード（ｒｅａｄ）」を生成する。配列決定リードをフィルタリングして、重複するリードを破棄し、容認できないほど低いカバー度、および／または、例えばｐｈｒｅｄ塩基クオリティ・スコアで測定された容認できないほど低い塩基クオリティを有するリードがあれば除去する。２０未満のｐｈｒｅｄスコアが一般に、リードを破棄する基準として用いられる。残りのリードを組み立てて、整列された遺伝子配列を生成する。この組立ては、配列決定リードの重なり合う部分を整列することに基づいて新規とすることができ、あるいは、ある一定の割合（例えば、５～１０％）の塩基ミスマッチを許容しながら、配列決定リードを参照配列にマッピングすることに基づくことができる。組み立てられた遺伝子配列を参照配列（場合により、マッピングによる組立てで使用されたものと同じ参照配列）と比較することによって変異体を検出し、また組み立てられた遺伝子配列に注釈を付けて、変異体およびその臨床的重要性（変異体－疾病相関情報が入手可能である場合）を特定する。配列決定または計算のエラーを（少しでもあるとしたら）検出する／特徴付けるために検証を行う。遺伝子解析の出力を記述する医療報告書を生成する。この医療報告書は、遺伝学に精通していない可能性のある治療医師が容易に理解できるという観点で書かれることが好ましく、また検証結果に関連する情報を含まなければならない。

【０００４】

このような解析における重要な作業は、変異体特定すなわち「コーリング（ｃａｌｌｉｎｇ）」である。これは、検出された変異体を評価して、その変異体が患者のゲノム内に実際に存在するのか、それともそうではなく、配列決定および／または組立て処理の際のアーティファクトまたはエラーであるのかを特定する作業である。通常、変異体コーリングは、ヌクレオチド塩基クオリティおよびカバー度に基づいて行われる。塩基クオリティは一般に、ｐｈｒｅｄのようなクオリティ・スコアで測定される。サンガー配列決定の場合、分光写真データからピーク形状および分解能などの配列塩基のパラメータを計算し、かつこれらの値を実験的に開発されたルックアップテーブルと比較することによって、ｐｈｒｅｄクオリティ・スコアを計算する。ｐｈｒｅｄスコアは一般に、塩基が不正確にコールされた確率と対数的関係があると考えられる。例えば、Ｑ＝２０のｐｈｒｅｄスコア（Ｑ）は９９％塩基コール精度に対応し、Ｑ＝３０は９９．９％精度に対応し、Ｑ＝４０は９９．９９％精度に対応する等である。カバー度は、読取りの数の計量であり、乗数の形で表されることが多い。例えば、８×のカバー度は、平均して、あるヌクレオチドが配列決定時に８回読み取られたことを示す。一般に、カバー度は冗長度を示すので、高いカバー度は高いリード信頼度に対応する。

【０００５】

いわゆる「次世代」配列決定（ＮＧＳ）手法では、スループットを何倍も向上させる並列処理技法を使用する。従来のｐｈｒｅｄスコア計算は一般に、ＮＧＳ技法には適用できないが、ほとんどのＮＧＳプラットフォームでは、分光写真データから計算された従来のｐｈｒｅｄスコアと比較可能な、またはスケーリングされる「ｐｈｒｅｄのような」塩基クオリティ・スコアを生成する。しかし、ＮＧＳ技法は一般に、サンガー配列決定などの旧来の技法よりも信頼度が低いと考えられている。ＮＧＳの信頼度を向上させるために、その出力は通常、サンガー配列決定、エクソーム捕獲、遺伝子型同定アレイなどの技法を使用して検証される。ＮＧＳとこれら旧来の技法のスループットスケールが異なるので、検証は通常、（ずっと大きい）ＮＧＳ出力のランダムに選択されたいくつかのターゲット部分に対して行われる。

【０００６】
ＮＧＳは、そのスループットが高いことにより臨床応用分野で魅力的なものになっている。しかし、医学的処置の決定が臨床ＮＧＳ結果に基づくので、臨床応用分野ではまた、高い信頼度も必要とされる。したがって、以前の技法と比較してＮＧＳの信頼度が低いことにより、ＮＧＳを臨床環境で応用することが困難になっている。
Ｙ．Ｓｈｅｎ他、Ｇｅｎｏｍｅ　Ｒｅｓｅａｒｃｈ、ｖｏｌ．２０、ｎｏ．２、２０１０年２月１日、２７３～２８０頁は、系統的配列決定エラーを明らかにするのにロジスティック回帰法および標準スコアを使用して、次世代再配列決定データにより変異体対立遺伝子確率を評価するＳＮＰ発見方法に関する。
Ｍａｔｕｋｕｍａｌｌｉ　Ｌ．Ｋ．他、ＢＭＣ　Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ、ＢｉｏＭｅｄ　Ｃｅｎｔｒａｌ、Ｌｏｎｄｏｎ、ＧＢ、Ｖｏｌ．７、Ｎｏ．１、２００６年1月６日、１～９頁は、ＳＮＰ発見における機械学習に関する。ある特定の機械学習プログラムが、決定木に基づいて使用されてきた。ＭＬプログラムでは、配列深さ、変位型および塩基頻度などの特徴のセットに基づいて分類子を作り出すことが報告されている。
【発明の概要】
【発明が解決しようとする課題】
【０００７】
以下では、前述の制限その他を克服する改善された装置および方法を企図する。
【課題を解決するための手段】
【０００８】
一態様によれば、当該方法は、整列された配列決定リードを含む組み立てられた遺伝子配列中の可能性がある変異体を特定するステップと、組み立てられた遺伝子配列の配列決定リードについて複数のリード特性の値を計算するステップと、受入れ基準を満たす可能性がある変異体を含む組み立てられた遺伝子配列の配列決定リードについての、複数のリード特性の計算された値を条件として、可能性がある変異体をコールするステップとを含み、特定するステップ、計算するステップ、およびコールするステップは、電子データ処理デバイスによって実施される。可能性がある変異体は、可能性がある変異体を含むリードについての、複数のリード特性の計算された値に影響を及ぼす分類子を使用して、受け入れられるもの、または拒絶されるものとして分類される。可能性がある変異体は、可能性がある変異体が受け入れられるものとして分類される場合にだけコールされ、分類は、Ｆスコアを使用するサポート・ベクター・マシン分類を使用して行われる。
【００１０】
別の態様によれば、当該装置は、整列された配列決定リードを含む組み立てられた遺伝子配列を記憶する持続性記憶媒体と、電子処理デバイスであって、組み立てられた遺伝子配列中の可能性がある変異体を特定すること、組み立てられた遺伝子配列のリードについて複数のリード特性の値を計算すること、および受入れ基準を満たす可能性がある変異体を含む組み立てられた遺伝子配列の配列決定リードについての、複数のリード特性の計算された値を条件として、可能性がある変異体をコールすることを含む動作を行うように構成された電子処理デバイスとを備える。可能性がある変異体は、可能性がある変異体を含むリードについての、複数のリード特性の計算された値に影響を及ぼす分類子を使用して、受け入れられるもの、または受け入れられないものとして分類される。可能性がある変異体は、可能性がある変異体が拒絶されるものとして分類される場合にだけコールされ、分類は、Ｆスコアを使用するサポート・ベクター・マシン分類を使用して行われる。
【００１１】
利点の１つは、より正確な変異体コーリングにある。
【００１２】
別の利点は、変異体コーリングにおけるより大きいプラットフォーム独立性にある。
【００１３】
別の利点は、遺伝子解析に関連する検証にある。
【００１４】

多数のさらなる利点および利益が、以下の詳細な説明を読むことによって当業者には明らかになろう。

【００１５】

本発明は、様々な構成要素および構成要素の配置、ならびに様々な処理作業および処理作業の配置で実施可能である。図面は、好ましい実施形態を図示することを目的とするにすぎず、本発明を限定するものと解釈されるべきではない。

【図面の簡単な説明】

【００１６】

【図１】本明細書に記載の遺伝子解析システムを示す図である。

【図２】図１のシステムによって適切に実行される遺伝子解析方法を示す図である。

【図３】図２の変異体コーリングと適切に置き換えられる代替変異体コーリング方法を示す図である。

【図４】トリマーで形成された２つのループを有するプローブの一例を示す図である。

【図５】本明細書に開示の実験のトップ機能の数に対する精度のグラフである。

【図６】次世代配列決定（ＮＧＳ）処理時の情報の流れを示す図であり、ＮＧＳ結果の検証のための検証領域を選択する箇所が太線の楕円内に示されている。

【発明を実施するための形態】

【００１７】

塩基クオリティ・スコアとカバー度の組合せを用いる変異体コーリングが一般に有効である。しかし、この手法にはいくつかの不利点があることが本明細書では認識されている。塩基クオリティ・スコアとカバー度の両方が配列決定プラットフォームおよび整列アルゴリズムに依存しており、その結果、この手法に依拠する、１つの研究所のために開発された遺伝子試験は、別の研究所に移すことが困難になる可能性がある。さらに、複数変異体において誤ってコールされた単一の塩基でも、誤った変異体コールを招く可能性があるので、配列決定エラーは変異体コーリングの中に伝搬する傾向がある。したがって、変異体コールエラー率は、ｐｈｒｅｄ（または他の塩基クオリティ）スコアから予想される塩基コールエラー率よりも大幅に高いことがある。例えば、１つの調査では、様々な次世代配列決定（ＮＧＳ）プラットフォームでの９９．９７％を超える配列決定精度が、９９％未満の変異体コーリング精度になり、いくつかの配列決定プラットフォームでは９５％未満になった。Ｈａｒｉｓｍｅｎｄｙ他、「Ｅｖａｌｕａｔｉｏｎ ｏｆ ｎｅｘｔ ｇｅｎｅｒａｔｉｏｎ ｓｅｑｕｅｎｃｉｎｇ ｐｌａｔｆｏｒｍｓ ｆｏｒ ｐｏｐｕｌａｔｉｏｎ ｔａｒｇｅｔｅｄ ｓｅｑｕｅｎｃｉｎｇ ｓｔｕｄｉｅｓ」、Ｇｅｎｏｍｅ Ｂｉｏｌｏｇｙ ｖｏｌ．１０：Ｒ３２（２００９年）。

【００１８】

カバー度を上げると一般に、変異体コールエラーが低減するはずである。しかし、カバー度を上げることは時間がかかり、試薬コストの増大を招くと共に、変異体コールエラー率を許容レベルまで低減しない可能性がある。例えば、１つの調査では、ＮＧＳを使用するＳＮＰコーリングにおけるエラー率は、２０×のカバー度と実質的に互角であった。Ｗａｎｇ他、「Ｔｈｅ ｄｉｐｌｏｉｄ ｇｅｎｏｍｅ ｓｅｑｕｅｎｃｅ ｏｆ ａｎ Ａｓｉａｎ ｉｎｄｉｖｉｄｕａｌ」、Ｎａｔｕｒｅ ｖｏ．４５６、６０～６５頁（２００８年）。いかなる特定の動作理論にも限定されることなく、これは、高いカバー度によって除去されない、および／または、完全な遺伝子配列の報告された（統計的）カバー度と比較してかなり低い実際のカバー度を有するいくつかの配列部分になる遺伝子配列にわたる高いカバー度変動性によって除去されない一貫したエラーの結果でありうることが場合により疑われる。Ｈａｒｉｓｍｅｎｄｙ他を参照されたい。さらに、より高いカバー度が、ＮＧＳの速度優位点を低減する可能性があるより長い配列決定および整列の時間になる。ランダムに選択されたターゲットの検証ではまた、検証ターゲットとして選択されていないＮＧＳ配列の領域内の変異体コーリング問題が検出されない可能性もある。

【００１９】

これらの問題の認識に関し、比較可能なデータを生成するサンガー配列決定（「黄金標

準」）、エクソーム捕獲、遺伝子型同定アレイなどのより確実な（しかし遅い）技法と比較してＮＧＳ出力の検証を使用することが当技術分野で知られている。ＮＧＳとこれら他の技法のスループットスケールが異なるので、より遅い検証技法は通常、ＮＧＳ出力のランダムに選択されたいくつかのターゲット部分に対して行われる。しかし、このようなランダム検証では、ＮＧＳ配列決定において重要なエラーを容易に見逃す可能性があり、それゆえに誤った変異体コールを検出しない。

【００２０】

本明細書では、可能性がある変異体を含む根源的な配列決定リードの特性の評価を使用する、改善された変異体コーリング技法を開示する。この手法は、配列決定リードがシークエンサーの実際の（すなわち、生の）出力であることを活用する。読み取られる物理的なＤＮＡ鎖またはＲＮＡ鎖は、配列決定時に構造的および熱力学的に安定であると期待される。リードはまた、解析されるリードの種類で典型的な値の範囲に入るという特性を有することが期待される。可能性がある変異体を含む遺伝子配列のリードについて計算されたリード特性が、これらのリードが非常に不安定である、または典型的な値の範囲から逸脱していることを示す場合、これらの読取りが誤っている（つまり、エラーを含む）可能性がある。したがって、これらの疑わしいリードに含まれる、可能性がある変異体もまた誤っている可能性がある。開示された手法のさらなる利点は、分類子を使用して複数のリード特性を組み合わせることによって、変異体コーリングを分類子の訓練中に調整できることである。

【００２１】

本明細書ではさらに、検証処理の改善を開示する。検証のためにＮＧＳ出力のランダムな領域選択を用いるのではなく、非ランダム基準を用いて検証領域を選択することを本明細書で開示する。この非ランダム基準は、ＮＧＳ出力中の関連したエラーがもし存在すれば検証により検出される可能性を向上させる。例えば、検証領域は、エラー計量の可能性に基づいて、または機能性領域である（したがって、既知の臨床的重要性の）領域に基づいて、または反復計量が低い領域に基づいて（多くの生物学的遺伝子配列における共通の反復出現を活用して）、または母集団固有の領域に基づいて（これらの領域は通常、誤って変異体と呼ばれる可能性がより高い）、などで選択することができる。

【００２２】

図１を参照すると、患者４が試料抽出検査室６において組織試料抽出を受けて試料が生成され、この試料は、そのＤＮＡおよび／またはＲＮＡを抽出、配列決定、および解析を行うためにゲノム検査室８で処理される。例えば、試料採取検査室６では、患者４から悪性病変の組織試料を、任意選択で患者４の他のところから抽出された正常組織試料と共に、抽出することができる。他のいくつかの説明的な例として、試料抽出には、生検針または他の介入器具を使用する生検手順、濾胞含有ＤＮＡを得るために髪試料を引き抜くこと、皮下針を使用して血液を引き出すこと、などが含まれてよい。図１は、例示的な抽出組織試料１０を示す。この例示的な実施形態は、臨床応用例において患者４に対して実行可能であるが、他の応用例においては、組織試料は、獣医学の対象、生物学研究の対象などの人間でない対象から抽出できることに留意されたい。例示的な図１では試料１０が例示的なバイアルで表されていることにも留意されたい。しかし、試料１０は一般に、試料採取された組織の種類に適した任意の形を取ることができ、またその試料の種類に適した任意の容器または支持体によって収容または支持できることを理解されたい。例えば、試料１０は流体試料、（例えば、口腔スワブによって取得され、滅菌スライド上または他の適切な表面に置かれる）表面試料などでよい。

【００２３】

ゲノム検査室８において、組織試料１０はシークエンサー装置１４で処理されて、配列決定リードが生成される。シークエンサー装置１４は、好ましくは次世代配列決定（ＮＧＳ）装置であり、Ｉｌｌｕｍｉｎａ、Ｓａｎ　Ｄｉｅｇｏ、カリフォルニア州、米国、Ｋｎｏｍｅ、Ｃａｍｂｒｉｄｇｅ、マサチューセッツ州、米国、Ｉｏｎ　Ｔｏｒｒｅｎｔ　Ｉｎｃ．、Ｇｕｉｌｆｏｒｄ、コネチカット州、米国、または他のＮＧＳシステム供給業

者から入手できるものなどの市販の配列決定装置でよいが、市販されていない、または注文構築のシークエンサーもまた企図されている。シークエンサー１４によって生成された配列決定リードは、フィルタリングモジュール１６によってフィルタリングされて、重複リードが除去され、また２０未満のｐｈｒｅｄ（または別の選択された塩基クオリティ・スコア閾値未満の）スコアを有するリードが破棄される。残りの配列決定リードは配列組立てモジュール１８によって組み立てられて遺伝子配列が生成され、この遺伝子配列は、ハード・ディスクもしくは他の磁気媒体などの持続性記憶媒体、光ディスクもしくは他の磁気媒体、ランダム・アクセス・メモリ（ＲＡＭ）、フラッシュメモリ、または他の電子記憶媒体などに記憶されるＳｅｑｕｅｎｃｅ　Ａｌｉｇｎｍｅｎｔ／Ｍａｐ（ＳＡＭ）フォーマット、２進の同等（例えば、ＢＡＭ）フォーマット、または別の適切なフォーマットの整列ファイル２０の形で保存される。ＳＡＭまたはＢＡＭフォーマットは、個々の配列リードを保持し、またそのリードがどのようにして組立て遺伝子配列を形成するかを明記するさらなるデータを含む。組み立てモジュール１８によって行われる組立ては、重なり合う部分の配列リードを新規整列するものでよく、あるいは、ある一定の割合（例えば、５～１０％）の塩基ミスマッチを許容しながら、配列決定リードを参照配列２２にマッピングするものでよい。後者の場合、参照配列２２は、例えば、図示の人間の患者４の場合ではヒトゲノムの標準参照配列でよい。

【００２４】

いくつかの実施形態では、整列ファイル２０に保存される組み立てられた遺伝子配列は、全ゲノム配列（ＷＧＳ）である。本明細書で使用される場合、「全ゲノム配列」、すなわちＷＧＳ（当技術分野では、「フル」、「完全」、または「全体」ゲノム配列とも呼ばれる）、または類似の用語は、患者の実質的な、しかし必ずしも完全ではない、ゲノムを包含すると理解されたい。当技術分野では、「全ゲノム配列」という用語、すなわちＷＧＳは、いくつかの用途では少なくとも９５％が完全であるような、患者のほぼ完全なゲノムを指すのに用いられる。他の実施形態では、遺伝子配列は、ＷＧＳより少ない遺伝子材料を表す（例えば単一の染色体、または染色体の一部分を表す）ことがある。シークエンサー１４は、高いスループットの次世代配列決定（ＮＧＳ）を使用するので、遺伝子配列が通常は長く、例えば、連続した、またはほぼ連続した数百、数千、または数万以上の塩基対（ｂｐ）の配列を含む。したがって、遺伝子配列は、ゲノムの０．１％未満が通常はカバーされる一塩基多型（ＳＮＰ）遺伝子型同定などの遺伝子特異的技法に通常使用される、ターゲット「配列」の型ではない。遺伝子配列は、ＤＮＡ（一本鎖もしくは二本鎖）またはＲＮＡを表すことができる。ＤＮＡとＲＮＡの間の重要な違いは、ＤＮＡが塩基のアデニン（Ａ）、シトシン（Ｃ）、グアニン（Ｇ）およびチミン（Ｔ）からなる配列を含むのに対し、ＲＮＡが塩基のアデニン（Ａ）、シトシン（Ｃ）、グアニン（Ｇ）およびウラシル（Ｕ）からなる配列を含むことである。言い換えると、ＤＮＡのチミン（Ｔ）塩基が、ＲＮＡではウラシル（Ｕ）に置き換えられている。

【００２５】

変異体コーリングおよび注釈付けモジュール２４では、遺伝子配列を、例えばヒトゲノムの文献に報告されている標準参照配列でよい参照配列２２と比較する。組立てでマッピングを使用する場合、組立ておよび変異体コーリングでは通常、図１に示されるのと同じ参照配列２２を使用する。しかし、組立ておよび変異体コーリングそれぞれに異なる参照配列を使用することもまた企図されている。組み立てられた遺伝子配列を参照配列２２と比較することにより、可能性がある変異体（つまり、解析される遺伝子配列が参照配列２２と異なる場所）を検出することが可能になる。可能性がある変異体は、本明細書で開示されるように、可能性がある変異体を含むリードの特性に基づいて「コールされる」（つまり、臨床目的で実際の変異体として受け入れられる）。

【００２６】

精度を確保するために、検証領域選択モジュール２６は、非ランダム選択基準に基づいて検証の領域を選択する。選択された検証領域は、図示のサンガー配列決定設備２８などの適切なシステムで実施される適切な検証技法によって検証される。（サンガー配列決定

設備２８は、図１でゲノム検査室８の一部として示されているが、選択モジュール２６で選択された検証領域の検証のために、試料を異なる検査室に送り出すこともまた企図されている。）組み立てられ、注釈が付けられ、検証された遺伝子配列に基づいて、臨床評価／報告モジュール３０が、患者４の治療を監督する医師が見直すのに適したフォーマットで臨床報告書を生成する。例えば、生成された報告書では、特定の癌の種類と相互に関連しているコールされた変異体を明らかにすることができ、あるいはそれと別に、疾病／変異体の相関関係（または場合により、このような相関関係がないこと）を強調表示することができる。

【００２７】

リードフィルタリングモジュール１６、配列組立てモジュール１８、変異体コーリング／注釈付けモジュール２４、および検証領域選択モジュール２６を含む様々な処理構成要素は、図示のコンピュータＣまたは別の、デスクトップ・コンピュータ、ノート型コンピュータ、ネットワーク・サーバなどの電子データ処理デバイスによって適切に具体化される。臨床評価／報告モジュール３０もまた、コンピュータＣまたは別の電子データ処理デバイスによって適切に具体化されるが、治療医師を対象とする要約報告書を書くことなどの一部の報告の態様は、例えば、ゲノム検査室８で雇用されている人間の検査室技師によって手作業または半手作業で行われてよい。例示的な図１では、図示の単一のコンピュータＣによって具体化された処理構成要素１６、１８、２４、２６、３０を示しているが、これらの処理構成要素を具体化するために異なるコンピュータを使用することも別法として企図されている。例えば、いくつかの実施形態では、フィルタリングモジュール１６および配列組立てモジュール１８は、コンピュータによって、またはシークエンサー１４と一体化されている他の電子データ処理デバイスによって具体化することができ、残りの組立て後の処理構成要素２４、検証構成要素２６、および報告構成要素３０は、シークエンサー装置とデータ通信する独立型コンピュータまたはネットワーク・サーバによって具体化することができる。

【００２８】

さらに、モジュール１６、２２、２６、３０によって行われる処理の様々な実施形態は、開示された処理を実行するための電子データ処理デバイス（例えば、コンピュータＣ）によって実行可能な命令を記憶する持続性記憶媒体（図示せず）として物理的に具体化することができる。このような持続性記憶媒体には、例えば、ハード・ディスクもしくは他の磁気記憶媒体、または光ディスクもしくは他の光記憶媒体、またはフラッシュメモリ、ランダム・アクセス・メモリ（ＲＡＭ）、読み取り専用メモリ（ＲＯＭ）、もしくは他の電子記憶媒体などが含まれてよい。

【００２９】

図１を引き続き参照し、さらに図２を参照して、図１のシステムの動作を説明する。動作４０で、配列決定装置１４から配列決定リードを取得する。動作４２で、リード評価／フィルタリングモジュール１６は、塩基クオリティ・スコアが低すぎる（例えば、いくつかの実施形態では、２０未満のｐｈｒｅｄスコアを有する）リードを破棄する。動作４２では、重複リードを除去するなどの他のフィルタリング動作を任意選択で実行することができる。動作４４で、配列組立てモジュール１８は、残りのリードを組み立てて、整列ファイル２０に保存される整列遺伝子配列を生成する。この組立ては、参照配列にマッピングすることに基づく新規のものでも、（例えば、参照配列がカバー度を提供するマッピングを使用し、また他のところの新規の整列を使用した）これらの組み合わせでもよい。通常、いくつかの配列決定リードが「残される」可能性があり、これら未整列のリードは破棄されるか、または、その未整列状態を示す適切な注釈付きで整列ファイル２０の中に保存される。

【００３０】

動作４６で、整列された遺伝子配列内で対象の領域を特定する。これらの領域は、その領域の既知の機能性、またはその領域における対象の変異体の考えられる存在などに基づく、様々な方法で特定することができる。さらなる解析のための対象の領域を（特に、整

列された遺伝子配列がＷＧＳである場合）選択することで効率を向上させる一方で、動作４６に関して、整列された遺伝子配列全体を対象の領域として選択することもまた企図されている。一方、対象の領域は、疾病と相関関係がある単一ヌクレオチド変異体（ＳＮＶ）の位置である可能性があると臨床研究で特定されている、単一塩基位置ほどに小さくすることもできる。

【００３１】

動作５０で、対象の領域（１つまたは複数）と一致するリードのリード特性を計算する。配列決定リードは、それが対象の領域と重なる、対象の領域を含む、または対象の領域に含まれる場合に、対象の領域と「一致する」。リード特性は、例えば、熱力学的特性、構造上の特性、塩基組成上の特性などでありうる。様々な例については本明細書で後述する。動作５２で、計算されたリード特性が受入れ基準を満たさない配列リードがあれば破棄する。動作５４で、残りのリードに対し（つまり、リード受入れ基準にパスするリードに対し）変異体コーリングを実行する。変異体は、整列された遺伝子配列を塩基対ごとに（二本鎖ＤＮＡの場合に、あるいは一本鎖ＤＮＡまたはＲＮＡの場合には塩基ごとに）参照配列２２と系統的に比較することによって特定することができる。これは、変異体を特定するための包括的な手法であるが、計算的に集約したものにでき、いかなる既知の疾病相関関係もない多くの特定された変異体が結果として得られる。あるいは、いくつかの実施形態では、参照配列２４は、既知の疾病相関変異体の位置（および任意選択で構成物）を示す注釈を含み、動作５４でこのような注釈付けされた既知の変異体をそれぞれ順に調べる。変異体のいくつかの例示的な種類としては、単一ヌクレオチド変異体（ＳＮＶ、一塩基多型すなわちＳＮＰとしても知られている）、挿入または欠失（すなわちインデル）、コピー数多型多形（ＣＮＶ）、構造多形（ＳＶ）、（参照配列２４と比較した）メチル化の存在または不在などが含まれる。この手法は、塩基ごとに比較する手法よりも高速にすることができ、加えて、特定された変異体は通常、既知の疾病相関関係を有する。これらの手法の組合せもまた使用することができる。

【００３２】

注釈付けは、変異体のように参照配列２４から逸脱する塩基のサブ配列（またはｂｐ）を特定することを含むことができ、また任意選択で、型（例えば、ＳＮＶ、ＣＮＶ、ＳＶなど）に関して変異体を標識することもできる。任意選択で、注釈付けはさらに、疾病相関がある変異体を標識することを、このような情報が参照配列２４で入手可能である場合に含むことができる。

【００３３】

図２を引き続き参照すると、動作６０で、検証領域選択モジュール２６は、整列された遺伝子配列の検証のための領域を選択する。動作６０は、ランダム選択基準を使用するのではなく、変異体を含む、または機能する、またはそれとは別に検証の特別の対象である、検証領域を選択するように設計された非ランダム基準を使用する。動作６２で、選択された領域を、例えば図１のサンガー配列決定設備２８によって検証する。

【００３４】

動作４６、５０、５２、５４は、図１の変異体コーリングおよび注釈付けモジュール２４によって適切に実行される。図２の手法では、対象の領域は動作４６で選択され、これらの領域と一致する配列決定リードをまず、動作５０、５２で受入れ基準に対して試験する。フィルタリング５０、５２をパスするリードだけを次に、動作５４で調べて変異体を特定し、コールする。動作５４で特定されたどの変異体も一般に、それが（図１のフィルタ１６をパスすることによって）塩基／カバー度基準にパスしており、かつ（「第２のパス」動作５０、５２をパスすることにより）リード受入れフィルタリングをパスしているので、直ちにコールすることができる。これにより、受入れ基準を満たす、可能性がある変異体を含む組み立てられた遺伝子配列の配列決定リードについての、リード特性の計算された値を条件として、可能性がある変異体をコールするという目標が達成される。しかし、この手法では、いくつかのリードは、可能性がある変異体を何も含まなくても試験される（場合により破棄される）ことがある。

【００３５】
図３を参照すると、図１の変異体コーリングおよび注釈付けモジュール２４の代替的動作では、可能性がある変異体をまず動作７０で特定する。この時点で、（図１のリード評価／フィルタリングモジュール１６の動作によって）可能性がある変異体をフィルタリングするのに塩基クオリティおよびカバー度だけを使用してきた。可能性がある変異体を特定した後、変異体を含む配列決定リードを動作７２で特定する。リード特性は、可能性がある変異体を含むリードについてだけ計算される。例示的な図３で、これらには、プローブ領域のリードについて（つまり、可能性がある変異体を含むリードについて）配列パラメータおよび構成パラメータを計算する動作７４と、プローブ領域のリードについて熱力学的パラメータまたはエネルギー・パラメータを計算する動作７６と、プローブ領域のリードについて生物物理学的パラメータまたは二次構造パラメータを計算する動作７８とが含まれる。動作７４、７６、７８の諸例については本明細書で後述する。決定動作８０で、動作７４、７６、７８で計算された特性を受入れ基準と比較し、これらの基準が満たされる場合、変異体コーリング／注釈付けモジュール２４は、可能性がある変異体を動作８２での医療目的の実際の変異体としてコールし、適切な注釈（１つまたは複数）を作成する。図３の手法は、可能性がある変異体を含むリードだけが動作７４、７６、７８、８０で処理されるので、計算的により効率的でありうる。しかし、「第２のパス」フィルタ動作７４、７６、７８、８０をパスできない、可能性がある変異体が多くある場合には、図２の動作５０、５２、５４の手法の方がより効率的でありうる。

【００３６】
動作８０（図３）は、動作５２が潜在的に誤ったリードをフィルタリングして除くのに対し、動作８０では可能性がある変異体をコールするかどうかを決定する（実際には、潜在的に誤った、可能性がある変異体をフィルタリングして除く）点で、動作５２（図２）と異なる。動作８０の１つの適切な手法では、可能性がある変異体を含むリードを動作５２と同じようにフィルタリングし、次に、残りのリードの数（つまり、受入れ基準を満たすリードの数）が、可能性がある変異体をコールするための規定の最小カバー度を超える場合にだけ、可能性がある変異体をコールする。

【００３７】
動作５２（図２）または動作８０（図３）で使用される受入れ基準は、様々なリード特性の値に基づく。例示的な図３で、値が計算される領域の特性には、ヌクレオチド頻度、ＧＣクランプ、連続する塩基の最大長などの配列パラメータおよび構成パラメータ７４と、エンタルピー、エントロピー、スタッキング・エネルギー、ＤＮＡ変性温度またはエネルギー、二本鎖安定破壊エネルギーなどの熱力学的特性またはエネルギー特性７６と、二量体形成、交差二量体、ヘアピン・ループ形成、可屈曲性、曲げ剛性などの生物物理学的パラメータおよび二次構造パラメータ７８とが含まれる。動作８０（または図２の動作５２）で、これらのリード特性の計算値を受入れ基準と比較する。この第２のパス受入れ試験にも（図１のモジュール１６で行われる「第１のパス」塩基クオリティ／カバー度フィルタリングに加えて）パスした場合のみ、処理の流れが動作８２（または図２の動作５４）に到達し、この時点で、可能性がある変異体をコールし注釈付けすることができる。

【００３８】
以下では、いくつかの適切なリード特性のいくつかの説明的な例について説明する。これらは説明的な例として提供されており、追加の、もっと少ない、または他のリード特性を使用することもまた企図されている。

【００３９】
いくつかの適切な熱力学的な、または配列組成のリード特性は以下の通りである。反塩基対組成リード特性値は、（ｉ）リード内のアデニン（Ａ）塩基およびチミン（Ｔ）塩基の総数を数えること、および（ｉｉ）リード内のグアニン（Ｇ）塩基およびシトシン（Ｃ）塩基の総数を数えることに基づいて計算することができる。（ＲＮＡでは、チミンがウラシルすなわち「Ｕ」に置き換えられる）。ＡＴ（またはＡＵ）数に対するＧＣ数の割合は、典型的なヒトＤＮＡまたはＲＮＡ材料では約４０～６０％になるはずである。使用す

ることができる追加の組成リード特性は、リード内のどこでも連続して生じる単一種類の塩基の数である。例えば、いくつかの実施形態では、１つの特性は連続したＡ塩基またはＴ塩基の数であり、この値が例えば６を超える場合、そのリードは受け入れられない。反復塩基の数がこのように多いことは、ヒトＤＮＡ／ＲＮＡ配列では普通ではなく、それが誤った配列決定の結果であることを示唆するものであり、この結果は、その領域（およびそこに含まれる、可能性がある変異体）を信用できないものにして、可能性がある変異体をコールすることが排除される。

【００４０】

融解温度（T_m）均一性は、１つの適切な熱力学的リード特性である。このパラメータの近似的な値は次式の通りに計算することができる。

【００４１】

【数１】

$$T_m = 2(A+T) + 4(G+C) \qquad (1)$$

このパラメータのもっと正確な値は次式で与えられる。

【００４２】

【数２】

$$T_m = \frac{\Delta H}{\Delta S + R\ln\left(C_1 - \frac{C_2}{2}\right)} - 273.15°C \qquad (2)$$

ここで、ΔＨは標準エンタルピー、ΔＳは標準エントロピーであり、C_1は一本鎖の初期濃度、C_2は相補鎖の初期濃度であり、Ｒは一般気体定数である。

【００４３】

別の適切なリード特性はＧＣクランプであり、これは、連続するグアニン（Ｇ）塩基とシトシン（Ｃ）塩基の、両方のプライマーの３’末端における数である。ＧＣクランプは、ターゲットＤＮＡとの複合体を形成するのに重要である。

【００４４】

オリゴヌクレオチド・プローブの安定性を決めるのに役割を果たす、いくつかの適切な熱力学的リード特性は、スタッキング・エネルギー、プロペラ・ツイスト、可屈曲性、二本鎖安定性自由エネルギー、およびＤＮＡ変性である。ジヌクレオチド塩基スタッキング・エネルギーは、ＤＮＡの一部がどれだけ容易にスタック分離するかを表す。高い値は不安定領域を表し、したがって、計算されたスタッキング・エネルギーが高いリードはフィルタリングされて除かれる。二本鎖安定性自由エネルギーが低い領域は、熱力学的エネルギー含量が高い領域よりも安定である。やはり、この値は、それがあまりに高い場合には、そのリードがフィルタリングされて除かれるべきことを示しうる。ＤＮＡ変性値が低いＤＮＡ領域は、値が高い領域よりも変性しやすい。したがって、ここでは、リードをフィルタリングするのに、低すぎるＤＮＡ変性値を使用することができる。

【００４５】

ジヌクレオチド・プロペラ・ツイストは、らせんの可撓性の値である。低い値が、より大きい可撓性を示す。可屈曲性値が高い部分は、値が低い領域よりも可屈曲性が大きい。トリヌクレオチド可屈曲性モデルは、主溝に向かうＤＮＡの可屈曲性をモデル化するものである。これらのパラメータの値は、リードが有効なリードである可能性に関しては、「良い」ものか「悪い」ものかについて簡単に確かめられない。しかし、これらの特性を分類子に入力する特徴として使用することによって、また標識されたデータ（すなわち、「良い」または「悪い」と標識されたリード）の分類子を訓練することによって、分類子をこれらのリード・パラメータについて有用に訓練することができる。

【００４６】

プローブの安定性に影響を及ぼす他のいくつかの要因は、オリゴヌクレオチドの内部安定性に関する要因である。通常、プライマーの安定な５’末端および不安定な３’末端は、未知のターゲットに対する間違ったプライミングを低減させることによって最善の結果をもたらす。ＤＮＡ合成を開始する可能性のある二本鎖形成は、低い３’安定性によって防止することができ、５’末端がまた、安定した二本鎖を形成するために対にならなければならない。最適な末端ΔＧは約８．５ｋｃａｌ／ｍｏｌであり、これが変動するとプライミング効率が低減する。再び、これらのパラメータそれぞれに対する受入れ基準を最適にするのに、分類子訓練を用いることができる。

【００４７】

受入れ基準で使用するためのいくつか適切な生物物理学的特性および二次構造リード特性には、ヘアピン・ループ形成、二本鎖安定性破壊エネルギー、ＤＮＡ曲げ剛性、二量／交差二量体形成、および自己相補性に関する特性が含まれる。これらについて順に述べる。

【００４８】

ヘアピン・ループ形成に関して、ΔＧが－２ｋｃａｌ／ｍｏｌの３’末端ヘアピン、およびΔＧが－３ｋｃａｌ／ｍｏｌの内部ヘアピンは、一般に耐容性がある。これに関する１つの適切なリード特性は、三量体で形成されるループを作らないヌクレオチドの数（以後「ＳＳ特徴１」）、または四量体で形成されるループを作らないヌクレオチドの数（以後「ＳＳ特徴２」）である。別のヘアピン・ループ関連のリード特性は、ループが三量体で形成されている最長配列の長さ（以後「ＳＳ特徴３」）またはループが四量体で形成されている最長配列の長さ（以後「ＳＳ特徴４」）である。

【００４９】

図４を参照すると、２つのループが三量体で形成されているプローブの一例が示されている。図４のＳＳ特徴１の値はａ＋ｂ＋ｃと計算され、ＳＳ特徴３の値は最大（ａ；ｂ；ｃ）と計算される。高い二本鎖安定性破壊エネルギーの値を有する領域は、低いエネルギー値を有する領域よりも安定である。Ｂｒｅｓｌａｕｅｒ ＫＪ、Ｆｒａｎｋ Ｒ、Ｂｌｏｅｃｋｅｒ Ｈ、Ｍａｒｋｙ ＬＡ、「Ｐｒｅｄｉｃｔｉｎｇ ＤＮＡ ｄｕｐｌｅｘ ｓｔａｂｉｌｉｔｙ ｆｒｏｍ ｔｈｅ ｂａｓｅ ｓｅｑｕｅｎｃｅ」、Ｐｒｏｃ Ｎａｔｌ Ａｃａｄ Ｓｃｉ ＵＳＡ、１９８６年6月、８３（１１）：３７４６～３７５０頁を参照されたい。高いＤＮＡ曲げ剛性の値は、剛性がより大きいＤＮＡ領域に対応し、低い値は、より容易に曲がる領域に対応する。Ｓｉｖｏｌｏｂ ＡＶ、Ｋｈｒａｐｕｎｏｖ ＳＮ、「Ｔｒａｎｓｌａｔｉｏｎａｌ ｐｏｓｉｔｉｏｎｉｎｇ ｏｆ ｎｕｃｌｅｏｓｏｍｅｓ ｏｎ ＤＮＡ：ｔｈｅ ｒｏｌｅ ｏｆ ｓｅｑｕｅｎｃｅ－ｄｅｐｅｎｄｅｎｔ ｉｓｏｔｒｏｐｉｃ ＤＮＡ ｂｅｎｄｉｎｇ ｓｔｉｆｆｎｅｓｓ」Ｊ Ｍｏｌ Ｂｉｏｌ．１９９５年4月１４日；２４７（５）：９１８～９３１頁を参照されたい。

【００５０】

二量体／交差二量体形成は、リード特性として以下のように特徴付けることができる。オリゴヌクレオチドが、ターゲットＤＮＡとハイブリッド形成するよりも容易に分子間二量体を形成する場合、この分子間二量体は製品収量を低減させる。ΔＧが－５ｋｃａｌ／ｍｏｌの３’末端自己二量体、およびΔＧが－６ｋｃａｌ／ｍｏｌの内部自己二量体は一般に耐容性がある。プローブ自己相補性の適切な尺度は、７ｂｐ未満の回文スコアである。

【００５１】

以下では、実際に実施したいくつかの実験について説明する。

【００５２】

全ゲノム配列決定は以下のように実施した。ＷＧＳに対し対末端ライブラリを、Ｉｌｌｕｍｉｎａ標準対末端オリゴと共にＮＥＢＮｅｘｔ ＤＮＡ試料Ｐｒｅｐ Ｍａｓｔｅｒ Ｍｉｘ Ｓｅｔ１を使用して構築した。Ｉｌｌｕｍｉｎａ １１５塩基対対末端リード・データを複数のランにわたって生成した。合計１，２６７，６５１，６３４対末端リー

ドをＷＧＳから生成した。塩基コーリングのためにＩｌｌｕｍｉｎａパイプラインｖ１．５／ｖ１．６を使用してデータを処理した。Ｉｌｌｕｍｉｎａフィルタリングをパスしたリードを、３０塩基種で２つのミスマッチを許容するように構成されたＢｕｒｒｏｗｓ Ｗｈｅｅｌｅｒｓ Ａｌｉｇｎｅｒ（ＢＷＡ）を用いて、ヒト参照ゲノム・アセンブリ（ＵＣＳＣ Ｈｇ１９）に対して整列させた（Ｌｉ他、２００９年）。重複リード対を除去して、試料調製時に生成されたＰＣＲアーティファクトを相殺した。全ゲノム組立て後のマッピングされた塩基の平均カバー度は約２８×であった。

【００５３】

整列後ＳＮＶコーリングをＳＡＭｔｏｏｌｓで、２０の最小ＳＮＶクオリティ・スコアおよび２０×の最小リード深さを使用して実施した（Ｌｉ他、２００９年）。挿入／欠失変異体（インデル）に近接したＳＮＶと、過剰なリード深さの領域内のＳＮＶと、１０塩基対窓内に複数のＳＮＶコールがある領域とを除去した。合計３，２９１，５０１およびＳＮＶを全ゲノムに対しコールした。次に、ＳＮＶのリストを選択し、複数の熱力学的パラメータを抽出した。ＳＮＶはまた、サンガー・ジデオキシ毛細管配列決定法を使用する検証のためにランダムに選択した。

【００５４】

次に、リードを２つのカテゴリー、すなわち（１）ＷＧＳからの遺伝子型コールとサンガー配列決定が一致したカテゴリー１と、（２）ＷＧＳからの遺伝子型コールとサンガー配列決定が一致しなかったカテゴリー２とに分類した。次に、これらのリードを特徴抽出し、特徴選択し、分類した。特徴抽出では、本明細書で前述した塩基組成特徴、熱力学的特徴、および二次構造特徴を抽出した。特徴選択では、サポート・ベクター・マシン・ランク付け特徴抽出（ＳＳＶＭ－ＲＦＥ）および最近接ランク付け特徴（ＮＮ－ＲＦ）を使用した。３’末端および５’末端を含むプローブの様々な領域、およびプローブの中間部で、合計３２９１の特徴を抽出および解析した。その後これらに対し分類法を、良いプローブを悪いプローブから区別するために適用した。

【００５５】

サポート・ベクター・マシン（ＳＶＭ）が様々な分類目的に使用されてきた。サポート・ベクター・マシンはデータをより高い次元空間にマッピングする。この空間では、クラス間のより良い分離が行われると共に、最大余裕が得られる超平面が見出される。$y^k \in \{1、-1\}$であるように訓練ベクトル$x^k \in R^n$、２つのクラス内でｋ＝１，．．．，ｍ、およびラベルのベクトル$y \in R^m$とすると、ＳＶＭで最適化問題が次式の通りに解かれる。

【００５６】

【数３】

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{k=1}^{m} \xi_k, \qquad (3)$$

ただし、

【数４】

$$y_k (w^T \phi (x_k) + b) \geqq 1 - \xi_k \qquad (4)$$

および

【数５】

$$\xi_k \geqq 0, \quad k = 1, \ldots, m \qquad (5)$$

とし、

ここで、訓練データはより高い次元空間に核関数φで表され、Ｃは訓練エラーに対するペ

ナルティ・パラメータである。任意の試験ベクトルｘについて、決定関数は次式となる。
【数６】

$$f(x)=\mathrm{sgn}(W^{T}\phi(x)+b) \qquad (6)$$

【００５７】
特徴選択は、精度尺度として使用して実施した。プローブから計算される特徴は、より良い分類子になるように最適に選択されなければならない。特徴のいくつかは、分類子の働きを劣化させうる可能性がある。実験で用いた戦略は、上位特徴を特定し、それを使用して分類子を訓練および試験することであった。Ｆスコアは、良いプローブと悪いプローブの間の特徴の弁別力に基づいて特徴をランク付けできる場合には、便利な表現を提供する。訓練ベクトル$x^{k}\in R^{n}$、ｋ＝１，．．．，ｍとし、良いプローブおよび悪いプローブの数がそれぞれｎ＋およびｎ－であれば、ｉ番目の特徴のＦスコアは次式の通りに計算される。
【００５８】
【数７】

$$F(i)=\frac{\left(\bar{x}_i^{(+)}-\bar{x}_i\right)^2+\left(\bar{x}_i^{(-)}-\bar{x}_i\right)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)}-\bar{x}_i^{(+)}\right)^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)}-\bar{x}_i^{(-)}\right)^2} \qquad (7)$$

ここで、記号
【００５９】
（外１）

$$\bar{x}_i,\quad \bar{x}_i^{(+)},\quad \bar{x}_i^{(-)}$$

はそれぞれ、ｉ番目の特徴の全体での平均、良いプローブ・データ・セット、および悪いプローブ・データ・セットである。
【００６０】
（外２）

$$x_{k,i}^{(+)}$$

は、ｋ番目の良いプローブのｉ番目の特徴であり、
【００６１】
（外３）

$$x_{k,i}^{(-)}$$

はｋ番目の悪いプローブのｉ番目の特徴である。分子は、良いプローブ・セットと悪いプローブ・セットの間の区別を示し、分母は、２つのセットそれぞれのうちの１つを示す。より高いＦスコアは、その特徴がより特異的であることを示唆する。
【００６２】
合計３２１９の特徴をカテゴリー１および２の両方のリードから計算した。これらの特徴は、これらについてＦスコア値を計算する前に範囲０から１にスケール変更した。実験により、上位特徴がデータ・セット全体にわたって変動することが明らかになった。ほとんどの場合で、（以前の実験に基づいて）重要度の高いものと予想された特徴は、上位特徴リストの中から出てきた。表１は、様々なデータ・セットで特定された上位１０の特徴

を示す。ここで、プローブの長さは、これらのデータ・セット全体にわたって変動したこと、および特徴のいくつかは、短いプローブでは関連していないことに留意されたい。しかし、様々なデータ・セットで特定された上位特徴間でいくらかの重なりがあることが注目されてよい。

【００６３】

【表１】

表１　検証肯定および検証否定であったリードの特定された上位特徴

熱力学的		二次構造	
特徴	Ｆスコア	特徴	Ｆスコア
ﾌﾟﾛﾍﾟﾗ(完全)	0.463217	5量体ﾙｰﾌﾟの５'末端で自由(第2)	0.022776
Tm:塩50 核酸50(完全)	0.372434	ﾃﾄﾗ・ﾙｰﾌﾟの３'末端で自由	0.014256
ﾌﾟﾛﾍﾟﾗ(5'末端)	0.327285	ﾃﾄﾗ・ﾙｰﾌﾟの５'末端で自由(第1)	0.014256
含有量(完全)	0.298143	三量体ﾙｰﾌﾟの５'末端で自由	0.014045
ﾄﾘﾇｸﾚｵﾁﾄﾞ(完全)	0.293189	三量体ﾙｰﾌﾟの５'末端で自由(第2)	0.011524
剛性(完全)	0.27979	三量体ﾙｰﾌﾟで最大	0.008427
変性(完全)	0.26369	ﾃﾄﾗ・ﾙｰﾌﾟで最小	0.004824
自由ｴﾈﾙｷﾞｰ(完全)	0.24907	三量体ﾙｰﾌﾟで最小	0.003269
ｽﾀｯｷﾝｸﾞ・ｴﾈﾙｷﾞｰ(完全)	0.247236	五量体ﾙｰﾌﾟで最大	0.002597
含有量(5'末端)	0.225024	五量体ﾙｰﾌﾟの３'末端で自由	0.002251

【００６４】

以下のような実験で、Ｆスコアを使用してＳＶＭ分類を行った。これらの特徴のＦスコア値を降順で並び替え、上位特徴の数をｎ個（通常は１０とする）のステップで増分することによってＦスコア値を使用して、ＳＶＭ分類子を訓練した。上位特徴の数は、試験データについて分類子を評価することによって得られた精度に基づいて選択することができる。特徴の最適なセットは、最高の分類精度が結果として得られるものである。

【００６５】

図５は、このような１つのデータ・セットについて得られた上位特徴の数に対する精度のグラフである。ＳＮＶ形状別カテゴリーは、良好な分離を示し、ＷＧＳデータ・セットからのリード・セットから生成されたＳＮＶの精度を決定するために使用されるべき良好な指標になる。

【００６６】

上記の実験において、ＳＮＶは、サンガー・ジデオキシ毛細管配列決定法を使用する検証のためにランダムに選択した。図１の検証領域選択モジュール２６を参照して論じたように、本明細書では、非ランダム選択基準を使用する検証のための領域を選択することが開示される。

【００６７】

図６を参照すると、検証のためのターゲット領域を抽出するステップおよびチェック点が図示されている。図６は、太線の楕円内に示されたＮＧＳ結果の検証のために検証領域を選択する箇所と共に、情報の流れを示す。検証の帯域幅／リソース（例えば、図１のサンガー配列決定設備２８のスループット）に基づいて、いくつかのターゲット領域をそれぞれのステップまたはチェック点で選択することができる。検証領域を選択するための選択基準は、エラーの可能性が高いいずれの区域も検証するために、領域のエラー計量の可能性に基づいて選択すること、機能性領域である領域（つまり、ゲノム内の機能性領域）に基づいて選択すること、反復領域が実質的に無い領域を検証するために領域の低い塩基反復計量に基づいて選択すること、母集団固有の領域である（つまり、遺伝子変異体とし

て誤って検出される可能性がある母集団固有の情報を有する）領域に基づいて選択すること、変異体を示す少なくとも１つの注釈を含む領域に基づいて選択すること、異なる構造特性を有する領域、などを含むことができる。

【００６８】

【表２】

表２　検証のために選択されたＳＮＰ　（ｓ*＝鎖　Ｂ＝ＷＧＳ塩基コール）

染色体	hg19 開始	hg19 終止	rsid	hg18 開始	s*	試料	参照	Cov	B
染色体 15	28705279	28705280	rs7163702	26500067	+	V16	C	52	C
染色体 17	16287670	16287671	rs41447048	16228395	+	V16	A	37	A
染色体 19	22841788	22841789	rs2957833	22633628	+	V16	t	32	T
染色体 19	54781843	54781844	rs4022330	59473655	+	V16	C	5	C
染色体 6	2547684	2547685	rs10456057	2492683	+	V16	A	68	A
染色体 6	2641546	2641547	rs2596437	2586545	+	V16	A	39	A

【００６９】

表２、表３および表４は、単一のターゲット一塩基多型（ＳＮＰ）と、インデルと、前述の非ランダム基準に基づく検証のために選択された構造変異体（ＳＶ）領域との諸例を提供する。

【００７０】

【表３】

表３　検証のために選択されたインデル

染色体	参照位置	遺伝子	参照配列ｱｸｾｯｼｮﾝ	鎖	参照位置	継続列指標
17	45234406	CDC27	NM_001114091	−	45234406	1
19	45016990	CEACAM20	NM_001102599	−	45016990	2
1	152195728	HRNR	NM_001009931	−	152195728	3
19	41123093	LTBP4	NM_001042544	+	41123093	4
17	34100350	MMP28	NM_024302	−	34100350	5
11	17333663	NUCB2	NM_005013	+	17333663	6
12	1.13E+08	OAS1	NM_016816	+	1.13E+08	7
6	31106500	PSOR1	NM_014068	+	31106500	8
17	7470285	SENP3	NM_015670	+	7470285	9
11	1.19E+08	SLC34	NM_001467	−	1.19E+08	10

継続列指標	試料の対立遺伝子	ﾘｰﾄﾞ1/ﾘｰﾄﾞ2	同種/異種	dB SNP ID	ｴｸｿｰﾑ･ﾃﾞｰﾀ	効果
1	*/-T	670/123	異種	−	合致	ﾌﾚｰﾑｼﾌﾄ
2	-A/-A	21/1	同種	−	合致	ﾌﾚｰﾑｼﾌﾄ
3	-A/-A	25/6	同種	−	NC	ﾌﾚｰﾑｼﾌﾄ
4	+G/+G	24/4	同種	−	合致	ﾌﾚｰﾑｼﾌﾄ
5	+G/+G	31/1	異種	−	合致	ﾌﾚｰﾑｼﾌﾄ
6	A	33//0	同種	rs2521998	NA	ｽﾌﾟﾗｲｽ･ｼｸﾞﾅﾙ
7	A	83//0	同種	rs10774671	NA	ｽﾌﾟﾗｲｽ･ｼｸﾞﾅﾙ
8	+C/+C	25/3	異種	−	合致	ﾌﾚｰﾑｼﾌﾄ
9	-A/-A	22/3	異種	−	合致	ﾌﾚｰﾑｼﾌﾄ
10	-G/-G	54/5	同種	−	合致	ﾌﾚｰﾑｼﾌﾄ

【００７１】

【表４】

表４　検証のために選択されたＳＶ

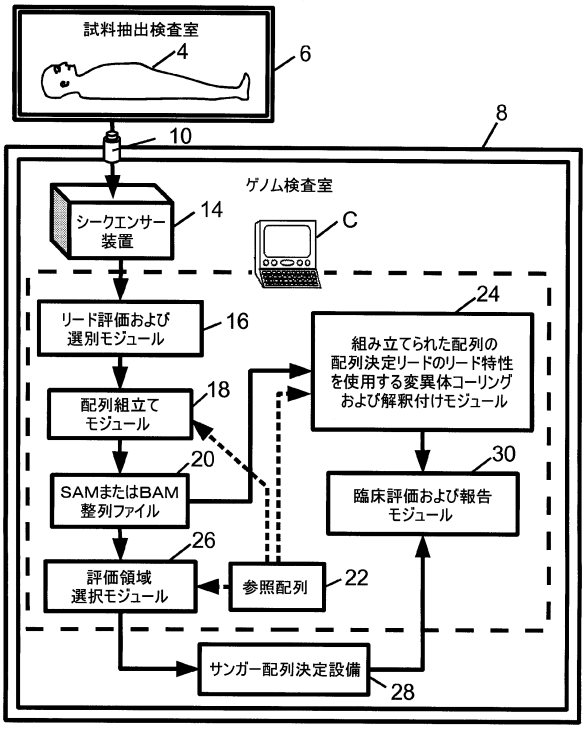
染色体	開始	終止	SV型	長さ	同種/異種	異常ﾘｰﾄﾞ	継続列指標
1	152555546	1.53E+08	Del	32189	同種	25	1
1	152555546	1.53E+08	Del	32189	同種	25	2
20	62596234	62597524	Del	116	同種	15	3
22	24274145	24311299	Del	37154	同種	25	4
22	24274145	24311299	Del	37154	同種	25	5

継続列指標	合計ﾘｰﾄﾞ	遺伝子	ｺｰﾃﾞｨﾝｸﾞ	鎖	内容
1	25	LCE3B	ﾀﾝﾊﾟｸ質ｺｰﾃﾞｨﾝｸﾞ	+	遅れて確認されたｴﾝﾍﾞﾛｰﾌﾟ3B
2	25	LCE3C	ﾀﾝﾊﾟｸ質ｺｰﾃﾞｨﾝｸﾞ	+	遅れて確認されたｴﾝﾍﾞﾛｰﾌﾟ3C
3	16	ZNF5B	ﾀﾝﾊﾟｸ質ｺｰﾃﾞｨﾝｸﾞ	−	ｼﾞﾝｸ・ﾌｨﾝｶﾞｰ・ﾀﾝﾊﾟｸ質512B
4	25	GST2B	ﾀﾝﾊﾟｸ質ｺｰﾃﾞｨﾝｸﾞ	−	ｸﾞﾙﾀﾁｵﾝS-ﾄﾗﾝｽﾌｪﾗｰｾﾞθ2B (遺伝子/偽遺伝子)
5	25	DDTL	ﾀﾝﾊﾟｸ質ｺｰﾃﾞｨﾝｸﾞ	+	D-ﾄﾞﾊﾟｸﾛﾑ・ﾄｰﾄﾒﾗｰｾﾞ状

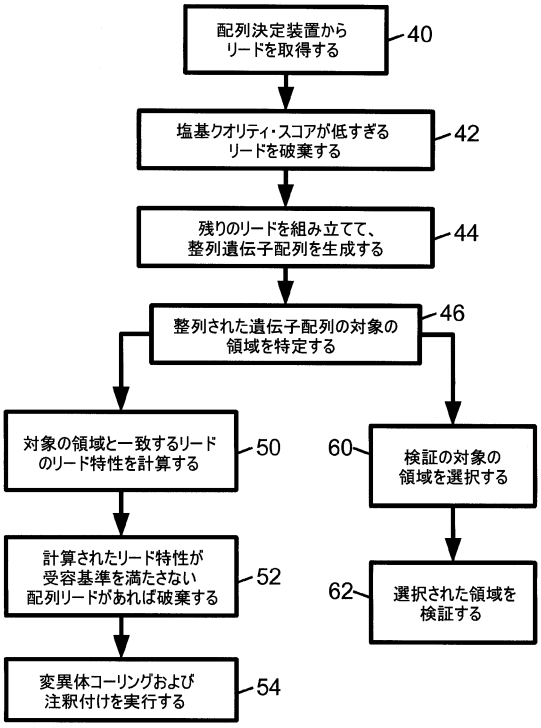
【００７２】

本発明を好ましい諸実施形態を参照して説明してきた。明らかに、前述の詳細な説明を読み理解することによって、修正形態および代替形態が他の人に想起されよう。本発明は、すべてのこのような修正形態および代替形態を、これらが添付の特許請求の範囲またはその同等物の範囲内に入る限り含むと解釈されるものである。

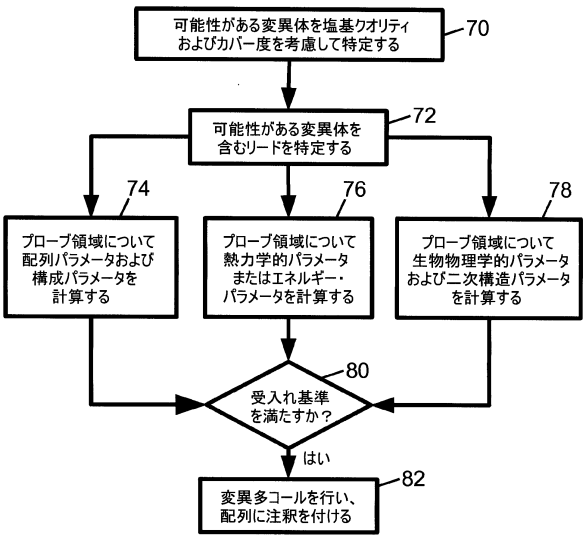
【図１】



【図２】



【図３】



【図４】

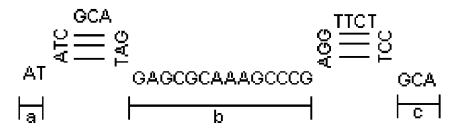
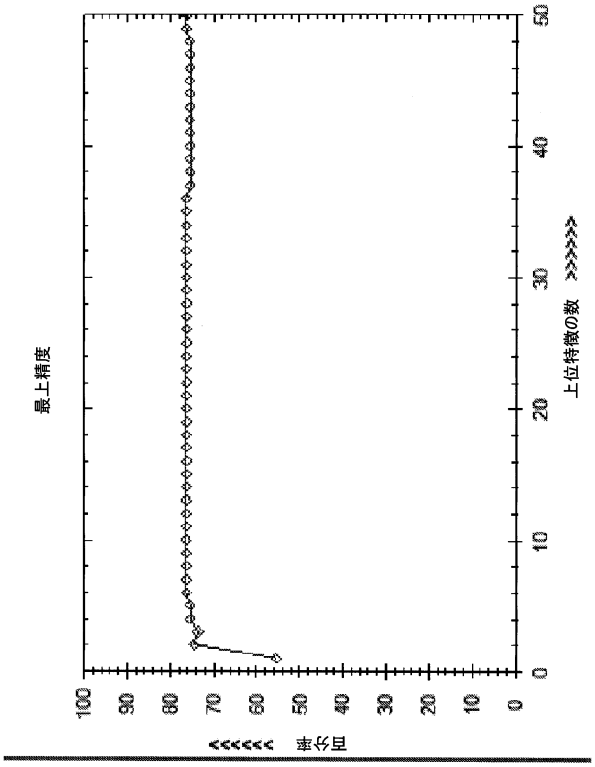
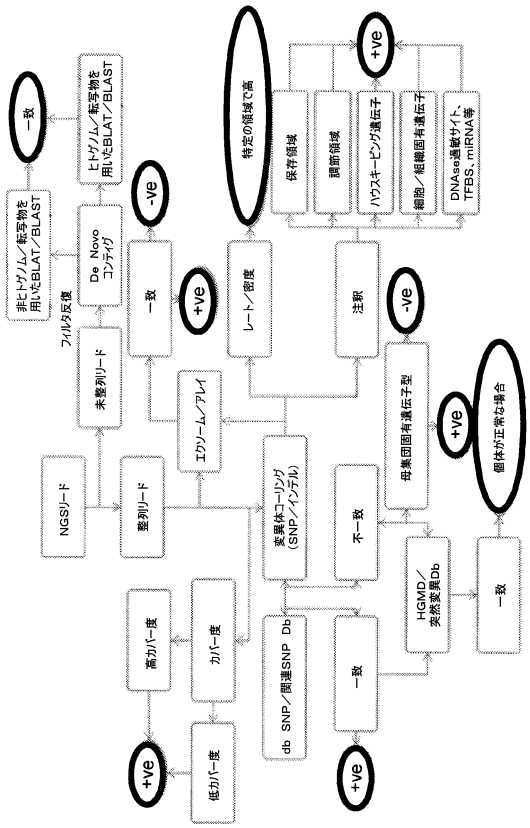


Fig. 4

【図５】



【図６】


NGSリード
整列リード
未整列リード
フィルタ反復
非ヒトゲノム／転写物を用いたBLAT／BLAST
De Novo コンティグ
ヒトゲノム／転写物を用いたBLAT／BLAST
一致
エクソーム／アレイ
一致
-ve
+ve
高カバー度
カバー度
低カバー度
+ve
変異体コーリング（SNP／インデル）
db SNP／関連SNP Db
レート／密度
特定の領域で高
保存領域
調節領域
ハウスキーピング遺伝子
細胞／組織固有遺伝子
DNaseⅠ感受性サイト、TFBS、miRNA等
+ve
注釈
不一致
一致
+ve
-ve
母集団固有遺伝子型
HGMD／突然変異Db
+ve
一致
個体が正常な場合

フロントページの続き

(74)代理人 100091214
弁理士 大貫 進介
(72)発明者 クマル, スニル
オランダ国, ５６５６ アーエー アインドーフェン, ハイ・テク・キャンパス・ビルディング４４
(72)発明者 シン, ランディープ
オランダ国, ５６５６ アーエー アインドーフェン, ハイ・テク・キャンパス・ビルディング４４
(72)発明者 チャクラバルティ, ビスワループ
オランダ国, ５６５６ アーエー アインドーフェン, ハイ・テク・キャンパス・ビルディング４４
(72)発明者 クマル, スボード
オランダ国, ５６５６ アーエー アインドーフェン, ハイ・テク・キャンパス・ビルディング４４

審査官 宮久保 博幸

(56)参考文献 国際公開第２０１１／０１８１９３（WO, Ａ１）
特表２０１３－５３１９８０（ＪＰ, Ａ）
特表２０１３－５２２７４３（ＪＰ, Ａ）
米国特許第６７１４８７４（ＵＳ, Ｂ１）
Shen, Y., A SNP discovery method to assess variant allele probability from next-generation resequencing data, Genome Research, ２０１０年 2月, Vol.20, No.2, p.273-280
Akay, M.F., Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with applications, ２００９年 3月, Vol.36, No.2, p.3240-3247
Matukumalli, L.K., Application of machine learning in SNP discovery, BMC Bioinformatics, ２００６年 1月, Vol.7, No.1, p.4

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １９／２０
Ｃ１２Ｑ １／６８
Ｃ１２Ｎ １５／０９