(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6618929号
(P6618929)
(45)発行日　令和1年12月11日(2019.12.11)　(24)登録日　令和1年11月22日(2019.11.22)

(51)Int.Cl.　F I
G16B 30/00 (2019.01)　G16B 30/00
C12Q 1/6869 (2018.01)　C12Q 1/6869 Z

請求項の数 15　(全 36 頁)

(21)出願番号　特願2016-565058 (P2016-565058)
(86)(22)出願日　平成27年5月12日 (2015.5.12)
(65)公表番号　特表2017-520821 (P2017-520821A)
(43)公表日　平成29年7月27日 (2017.7.27)
(86)国際出願番号　PCT/EP2015/060442
(87)国際公開番号　WO2015/173222
(87)国際公開日　平成27年11月19日 (2015.11.19)
審査請求日　平成30年5月11日 (2018.5.11)
(31)優先権主張番号　61/991,820
(32)優先日　平成26年5月12日 (2014.5.12)
(33)優先権主張国・地域又は機関
米国 (US)

(73)特許権者　591003013
エフ．ホフマン－ラ　ロシュ　アーゲー
F．　HOFFMANN－LA　ROCHE　AKTIENGESELLSCHAFT
スイス・シーエイチ－４０７０バーゼル・グレンツアーヘルストラツセ１２４
(74)代理人　100099759
弁理士　青木　篤
(74)代理人　100077517
弁理士　石田　敬
(74)代理人　100087871
弁理士　福本　積
(74)代理人　100087413
弁理士　古賀　哲次

最終頁に続く

(54)【発明の名称】ウルトラディープシークエンシングにおける希少バリアントコール

(57)【特許請求の範囲】
【請求項１】
第一の試料中の標的領域における低頻度バリアントを検出するための、コンピュータで実行される方法であって、コンピュータシステムにおいて、
－　１又は２以上の試料からのＤＮＡ断片のシークエンシングから得られる複数の配列リードを受け取り、ここで前記１又は２以上の試料は第一の試料を含み、前記シークエンシングは前記ＤＮＡ断片中の標的領域を標的化することを含み；
－　前記複数の配列リードを、参照配列の標的領域にアラインメントし；
－　第一の試料の配列リードに基づいて、標的領域の第一の位置において第一のバリアントアレルを有する第一の候補バリアントを同定し、ここで前記第一のバリアントアレルは、前記参照配列の第一の位置の参照アレルとは異なり；
－　前記参照配列の第一の位置にアラインメントする第一の試料の配列リードに基づいて、第一の位置における第一のバリアントアレルに関する第一のバリアント頻度を決定し、
－　前記参照アレルを有する前記参照配列の標的領域中の１セットの第二の位置を同定し、ここで前記１セットの第二の位置は前記第一の位置を含み；
前記１セットの第二の位置の各々の位置において、かつ前記１又は２以上の試料の各々に関して：
－　前記参照配列の１セットの第二の位置の各々の位置にアラインメントする試料の配列リードに基づいて、第一のバリアントアレルの第二のバリアント頻度を決定し、ここで前記第二のバリアント頻度は統計分布を形成し；

－　前記第一のバリアント頻度を前記統計分布の統計値と比較して、前記統計分布の統計値に対する第一のバリアント頻度の確率値を決定し；そして
－　前記第一の位置における第一のバリアントアレルに関して、第一の試料において第一の候補バリアントが真陽性であるか否かを決定する一部として、前記確率値を閾値と比較すること、ここで前記閾値は前記第一のバリアントアレルに関して偽陽性と真陽性とを区別する、
含む、方法。
【請求項２】
前記参照配列は、正常細胞から決定されるとおりのコンセンサス配列に対応する、請求項１に記載の方法。
【請求項３】
前記１又は２以上の試料は無細胞ＤＮＡ断片由来である、請求項１又は２に記載の方法。
【請求項４】
前記１又は２以上の試料は生物試料のＲＮＡ由来である、請求項１又は２に記載の方法。
【請求項５】
複数の試料は、単一のシークエンシングランにおいてシークエンシングされる、請求項１～４のいずれか１項に記載の方法。
【請求項６】
前記確率値は、ｚスコア、修正されたｚスコア、累積確率、フレッド（Phred）クオリティスコア又は修正されたフレッドクオリティスコアである、請求項１～５のいずれか１項に記載の方法。
【請求項７】
前記統計分布は、前記第二のバリアント頻度の対数変換の統計分布である、請求項１～６のいずれか１項に記載の方法。
【請求項８】
前記閾値は、既知の真陽性と偽陽性とを有するトレーニングデータに基づいて、サポートベクターマシン分類器を用いて決定される、請求項１～７のいずれか１項に記載の方法。
【請求項９】
第一の試料中の標的領域における第一の位置に第一のバリアントアレルを有するバリアントを検出するための、コンピュータで実行される方法であって、コンピュータシステムにおいて、
－　少なくとも２つの試料からのＤＮＡ断片のシークエンシングから得られる複数の配列リードを受け取り、ここで前記少なくとも２つの試料は第一の試料を含み、前記シークエンシングは前記ＤＮＡ断片中の標的領域を標的化することを含み；
－　前記複数の配列リードを、参照配列の標的領域にアラインメントし；
－　第一の位置において各試料のアラインメントされた配列リードに基づいて、前記第一のバリアントアレルが、前記少なくとも２つの試料の各々の試料中の第一の位置に存在するか否かを同定し、ここで前記第一のバリアントアレルは、前記参照配列の第一の位置における参照アレルとは異なり；
－　前記少なくとも２つの試料の各試料に関する第一の位置にアラインメントされた配列リードを用いて、前記第一の位置における第一のバリアントアレルのバリアント数及び前記第一の位置における参照アレルの野生型数を決定し；
－　前記少なくとも２つの試料から、１つ又は２つ以上の試料を１つ又は２つ以上の参照試料として選択し；
－　第一の試料に関する第一の位置における第一のバリアントアレルの第一のバリアント数及び第一の位置における参照アレルの第一の野生型数と、前記１つ又は２つ以上の参照試料に関する第一の位置における第一のバリアントアレルの第二のバリアント数及び第一

の位置における参照アレルの第二の野生型数とを比較して、第一の試料に関する第一の位置に第一のバリアントアレルを有するバリアントが発生する確率値を決定し；そして
－　第一のバリアントアレルに関して第一の試料中の第一の位置における第一のバリアントアレルが真陽性を含むか否かを決定する一部として、前記確率値を閾値と比較すること、ここで前記閾値は第一の位置の第一のバリアントアレルに関して偽陽性と真陽性とを区別する、
含む、方法。
【請求項１０】
前記確率値は、以下の式：
【数１】

$$\chi^2 = n \times (a1 \times w2 - a2 \times w1)^2 / (n1 \times n2 \times a \times w)$$

［式中、
ａ２は第一のバリアント数であり、ｗ２は第一の野生型数であり、ａ１は第二のバリアント数であり、ｗ１は第二の野生型数であり、ａはａ１及びａ２の合計であり、ｗはｗ１及びｗ２の合計であり、ｎ１はａ１及びｗ１の合計であり、ｎ２はａ２及びｗ２の合計であり、ｎはｎ１及びｎ２の合計である］
によって定義されるカイ二乗統計値に基づくカイ二乗累積分布関数を用いて決定される、請求項９に記載の方法。
【請求項１１】
前記確率値は、２つの比率であるｐ１及びｐ２に基づいて決定され、ここでｐ１＝ａ１／ｎ１であり、ｐ２＝ａ２／ｎ２であり、ａ２は第一のバリアント数であり、ａ１は第二のバリアント数であり、ｎ２は第一のバリアント数及び第一の野生型数の合計であり、そしてｎ１は第二のバリアント数及び第二の野生型数の合計である、請求項９に記載の方法。
【請求項１２】
前記確率値は、ｚスコア、修正されたｚスコア、ｐ値、カイ二乗値、累積確率値及び信頼水準を示すクオリティスコアの１つ又は２つ以上である、請求項９に記載の方法。
【請求項１３】
前記クオリティスコアはルックアップテーブルを用いて決定され、ここで前記ルックアップテーブルの入力は、カイ二乗値又は正規クオンタイル値の少なくとも１つである、請求項１２に記載の方法。
【請求項１４】
前記閾値は、既知の真陽性と偽陽性とを有するトレーニングデータに基づいて、サポートベクターマシン分類器を用いて決定される、請求項９～１３のいずれか１項に記載の方法。
【請求項１５】
第一の試料の標的領域における真のバリアントを検出するようにコンピュータシステムを制御するためのコンピュータプログラムであって、以下の指示：
－　１又は２以上の試料からのＤＮＡ断片のシークエンシングから得られる複数の配列リードを受容させ、ここで前記１又は２以上の試料は第一の試料を含み、前記シークエンシングは前記ＤＮＡ断片中の標的領域を標的化することを含み；
－　前記複数の配列リードを、参照配列の標的領域にアラインメントさせ；
－　複数のバリアントクラス中の１つのバリアントクラス中のバリアントの参照アレルを各々が有する参照配列の標的領域における１セットの配列位置を同定させ、ここで、前記複数のバリアントクラスの各々は、１又は２以上のバリアントを含むように定義され、各々のバリアントは対応する参照アレルとは異なるバリアントアレルを有し、そして前記１セットの配列位置は第一の位置を含み、

前記１セットの配列位置の各々の位置において、かつ前記１又は２以上の試料の各々の試料に関して：
－　各試料に関する各位置におけるリード数を決定させ；
－　各試料の配列リードに基づいて、前記バリアントクラス中のバリアントに関するバリアントアレルを有する候補バリアントを同定させ、ここで前記バリアントアレルは、前記参照配列の同一の位置における参照アレルとは異なり、そして各試料中の各位置における候補バリアントの総数は、各試料に関する各位置中のバリアント数であり；
－　前記リード数及びバリアント数に基づいてバリアントクラス中のバリアントのバリアント頻度を決定させ、ここで各試料中の各位置に関するバリアント頻度は、統計分布を形成し、第一の試料に関する前記１セットの配列位置中の第一の位置におけるバリアント頻度は、第一のバリアント頻度であり；
－　前記第一のバリアント頻度を統計分布の値と比較して、前記統計分布の値に対する前記第一のバリアント頻度の確率値を決定させ；そして
－　第一の試料中の候補バリアントが真陽性であるか否かを決定する一部として、前記確率値を閾値と比較させること、ここで前記閾値は、前記バリアントクラス中のバリアントに関して偽陽性と真陽性とを区別する、
を含む、前記コンピュータプログラム。
【発明の詳細な説明】
【背景技術】
【０００１】
核酸シークエンシングは、あるＤＮＡ分子又はＲＮＡ分子中に存在するヌクレオチドの順序を決定する。より安価で迅速なシークエンシング法に対する要求が、次世代シークエンシング（ＮＧＳ）法の開発を推進してきた。ＮＧＳプラットフォームは、大規模な平行シークエンシングを行い、この間に、複数の試料からの数百万のＤＮＡ断片をまとめてシークエンシングすることができ、こうして伝統的なサンガー（Sanger）シークエンシングに対してはるかに安価で高スループットな代替法を提供する。ＮＧＳは全ゲノムシークエンシング又は標的化シークエンシングで使用することができる。標的化シークエンシングでは、ゲノム中の遺伝子又は規定された領域のサブセットがシークエンシングされるか、又は、例えば主に標的領域を増幅することによりシークエンシングされる。
【０００２】
ウルトラディープシークエンシング（ultra-deep sequencing）は、一般的な及び希少な配列の変化を同定することを目的とする、広範囲のアンプリコンのシークエンシングである。充分なカバー率を有するため、ウルトラディープシークエンシングは、希少配列バリアントを１％未満まで完全に特性評価する能力を有する。ウルトラディープシークエンシングは、低頻度（low-frequency)ＨＩＶ薬耐性変異を検出するために、又は複雑な癌試料中の希少体細胞変異を同定するために使用されている。非侵襲性血液検査などの試験に関して、バイオマーカー変異の頻度は１％未満になることがある。しかしＮＧＳはエラーが発生しやすいプロセスであり、シークエンシング深度（sequencing depth)、試料のタイプ、及びシークエンシングプロトコールによっては、ほぼ１％以上のエラー率になることがある。従って、１％未満の頻度を有するバリアントに関しては偽陽性（false positives)が現れる可能性があるため、現在の多くのＮＧＳソフトウェアパッケージは、１％以上の頻度を有するバリアントのみを報告している。それでも、例えば１％未満の低頻度を有するバリアントに関してさえ、真陽性（true positives)が存在し得る。従って、例えば約０．００２５％～約１％という低い頻度を有するバリアントに関して、真陽性を検出する方法とシステムが必要とされる。
【発明の概要】
【０００３】
実施態様は、例えば標的化シークエンシングから得られた試料のシークエンシングリード（sequencing reads)に基づいてより多くの正確なバリアントコール（variant calls）を行うための方法、システム、及び装置を提供することができる。例えば、いったん配列

リード（sequence reads）が受け取られ（received)、参照配列(reference sequence)にアラインメントされる（aligned）と、ある位置にバリアントを有するシークエンシングリードがカウントされる。試料の１つの位置で測定される特定のバリアントの第一のバリアント頻度を、他の位置で測定される及び／又は他の試料からの特定のバリアントの１又は２以上の第二のバリアント頻度と比較することができる。第二のバリアント頻度は、シークエンシングランに関するシークエンシングエラーの予測値に対応することができる。

【０００４】

いくつかの実施態様において、バリアントがある位置で真陽性である信頼水準を示す確率値（probability)は、１又は２以上の試料の標的領域中の複数の位置におけるバリアント数及び総リード数に基づいて計算することができる。確率値は次に閾値レベルと比較されて、検出されたバリアントが真陽性であるか否かを決定することができる。他の実施態様において、試験試料と参照試料（例えば、その位置においてシークエンシングエラーのみを有すると推定される）中の同一の位置におけるバリアント数と総リード数の差を用いて、試験試料中のバリアントが真陽性であるかどうかを決定することができる。

【０００５】

ある実施態様において、ある方法は、試験試料の標的領域における希少バリアントについて真陽性を検出することができる。各試料について、参照配列上で参照アレル（reference allele)が存在する位置における同じバリアントクラスのバリアントに関するバリアント頻度は、バリアント数と総リード数を用いて計算することができる。同一のクラスのバリアントに関するバリアント頻度の分布を使用して、決定されたバリアント頻度を有する試験試料中のある位置におけるバリアントの確率値を決定することができる。この確率値に基づいて、試験試料中の位置におけるバリアントは、真陽性（変異(mutation)）又は偽陽性として分類される。

【０００６】

他の実施態様において、ある方法は、１又は２以上の参照試料との比較を用いて、試験試料の標的領域における希少バリアントに関する真陽性を検出することができる。試験試料中の特定の位置における特定のバリアントに関するバリアント数と野生型数は、アラインメントされた配列リードから決定することができ、１又は２以上の参照試料中の特定の位置における特定のバリアントに関するバリアント数と野生型数と比較して、確率値を決定することができる。この確率値に基づいて、試験試料中の特定の位置における特定のバリアントは、真陽性又は偽陽性として分類される。

【０００７】

ある実施態様において、第一の試料の標的領域における低頻度バリアントを検出するためのコンピュータで実行される方法が提供される。ここで、この方法は（コンピュータシステムにおいて）、

１又は２以上の試料からのＤＮＡ断片のシークエンシングから得られる複数の配列リードを受け取り、ここで前記１又は２以上の試料は第一の試料を含み、前記シークエンシングは前記ＤＮＡ断片中の標的領域を標的化することを含み；

前記複数の配列リードを、参照配列の標的領域にアラインメントし；

第一の試料の配列リードに基づいて、標的領域の第一の位置において第一のアレルを有する第一の候補バリアントを同定し、ここで前記第一のアレルは、前記参照配列の第一の位置の参照アレルとは異なり；

前記参照配列の第一の位置にアラインメントする第一の試料の配列リードに基づいて、第一の位置における第一のアレルに関する第一のバリアント頻度を決定し、

複数のバリアントクラスから選択される第一のバリアントクラスに対応する第一の候補バリアントを同定し、ここで前記複数のバリアントクラスの各バリアントクラスは、異なるタイプのバリアントに対応し；

前記参照アレルを有する前記参照配列の標的領域における１セットの第二の位置を同定し、ここで前記１又は２以上の試料中の少なくとも５０％の他の位置は、第一のアレルに関して偽陽性を示し、そして前記１セットの第二の位置は前記第一の位置を含み；

前記１セットの第二の位置の各々において、かつ前記１又は２以上の試料の各々に関して：

前記参照配列の第二の位置にアラインメントする試料の配列リードに基づいて、第一のアレルの第二のバリアント頻度を決定し、ここで前記第二のバリアント頻度は統計分布を形成し；

前記第一のバリアント頻度を前記統計分布の統計値と比較して、前記統計分布の統計値に対する第一のバリアント頻度の確率値を決定し；そして

前記第一のアレルに関して、第一の試料において第一の候補バリアントが真陽性であるか否かを決定する一部として、前記確率値を閾値と比較すること、ここで前記閾値は前記第一のアレルに関して偽陽性と真陽性とを区別する、を含む。

【０００８】

ある実施態様において、前記参照配列は、正常細胞から決定されるとおりのコンセンサス配列に対応する。いくつかの実施態様において、前記１又は２以上の試料は無細胞ＤＮＡ断片由来である。いくつかの実施態様において、前記１又は２以上の試料は生物試料のＲＮＡ由来である。いくつかの実施態様において、複数の試料は、単一のシークエンシングランにおいてシークエンシングされる。他の実施態様において、前記統計分布の統計値は平均値を含む。他の実施態様において、前記確率値は、ｚスコア、修正された(modified)ｚスコア、累積確率、フレッド（Phred）クオリティスコア又は修正された(modified)フレッドクオリティスコアである。他の実施態様において、前記統計分布は、前記第二のバリアント頻度の対数変換の統計分布である。他の実施態様において、前記閾値は、１又は２以上のシークエンシングランから得られるトレーニングデータに基づくサポートベクターマシン分類器（support vector machines classifier）を用いて決定される。他の実施態様において、前記閾値はバリアント頻度の関数である。

【０００９】

別の実施態様において、第一の試料中の標的領域における第一の位置に第一のアレルを有するバリアントを検出するための、コンピュータで実行される方法が提供される。ここでこの方法は（コンピュータシステムにおいて）、

少なくとも２つの試料からのＤＮＡ断片のシークエンシングから得られる複数の配列リードを受け取り、ここで前記少なくとも２つの試料は第一の試料を含み、前記シークエンシングは前記ＤＮＡ断片中の標的領域を標的化することを含み；

前記複数の配列リードを、参照配列の標的領域にアラインメントし；

第一の位置において各試料のアラインメントされた配列リードに基づいて、前記第一のアレルが、前記少なくとも２つの試料の各々の試料中の第一の位置に存在するか否かを同定し、ここで前記第一のアレルは、前記参照配列の第一の位置における参照アレルとは異なり；

前記少なくとも２つの試料の各試料に関する、第一の位置における第一のアレルのバリアント数及び第一の位置の参照アレルの野生型数を決定し；

前記少なくとも２つの試料から、少なくとも１つの試料を参照試料として選択し；

第一の試料に関する第一の位置における第一のアレルの第一のバリアント数及び第一の位置における参照アレルの第一の野生型数と、前記参照試料に関する第一の位置における第一のアレルの第二のバリアント数及び第一の位置における参照アレルの第二の野生型数とを比較して、第一の試料に関する第一の位置に第一のアレルを有するバリアントの確率値を決定し；そして

第一のアレルに関して第一の試料中の第一の位置における第一のアレルが真陽性であるか否かを決定する一部として、前記確率値を閾値と比較すること、ここで前記閾値は第一の位置の第一のアレルに関して偽陽性と真陽性とを区別する、を含む。

【００１０】

ある実施態様において、前記参照試料は、第一の試料以外の前記少なくとも２つの試料のうち、第一の位置における第一のアレルに関して最も低いバリアント頻度を有する２つの試料を含む。いくつかの実施態様において、前記確率値は、カイ二乗累積分布関数（ch

i-squared cumulative distribution function）を用いて決定される。いくつかの実施態様において、前記確率値は、ピアソン比率検定（Pearson proportion test）を用いて決定される。いくつかの実施態様において、前記確率値は、ｚスコア、修正された（modified）ｚスコア、ｐ値、カイ二乗値、累積確率値及びクオリティスコアの１つ又は２つ以上である。いくつかの実施態様において、前記クオリティスコアはルックアップテーブル(look-up table)を用いて決定される。いくつかの実施態様において、前記閾値は、１又は２以上のシークエンシングランから得られるトレーニングデータに基づくサポートベクターマシン分類器を用いて決定される。いくつかの実施態様において、前記閾値はバリアント頻度の関数である。

【００１１】

別の実施態様において、実施されるときにコンピュータシステムを制御して第一の試料の標的領域における真のバリアントを検出する複数の指示を記憶する、非一過性の(non-transitory)コンピュータ可読媒体を含むコンピュータ製品が提供される。ここで前記指示は、

１又は２以上の試料からのＤＮＡ断片のシークエンシングから得られる複数の配列リードを受け取り、ここで前記１又は２以上の試料は第一の試料を含み、前記シークエンシングは前記ＤＮＡ断片中の標的領域を標的化することを含み；

前記複数の配列リードを、参照配列の標的領域にアラインメントし；

バリアントクラス中のバリアントの参照アレルを有する参照配列の標的領域における１セットの配列位置を同定し、ここで、前記１又は２以上の試料中の少なくとも５０％の配列位置は、前記配列リード中のバリアントクラス中のバリアントに関して偽陽性を示し、そして前記１セットの配列位置は第一の位置を含み、

前記１セットの配列位置の各々の位置において、かつ前記１又は２以上の試料の各々の試料に関して：

各試料に関する各位置におけるリード数を決定し；

各試料の配列リードに基づいて、前記バリアントクラス中のバリアントに関するバリアントアレルを有する候補バリアントを同定し、ここで前記バリアントアレルは、参照配列の同一の位置における参照アレルとは異なり、そして各試料中の各位置における候補バリアントの総数は、各試料に関する各位置中のバリアント数であり；

前記リード数及びバリアント数に基づいてバリアントクラス中のバリアントのバリアント頻度を決定し、ここで各試料中の各位置に関するバリアント頻度は、統計分布を形成し、第一の試料に関する前記１セットの配列位置中の第一の位置におけるバリアント頻度は、第一のバリアント頻度であり；

前記第一のバリアント頻度を統計分布の値と比較して、前記統計分布の値に対する前記第一のバリアント頻度の確率値を決定し；そして

第一の試料中の候補バリアントが真陽性であるか否かを決定する一部として、前記確率値を閾値と比較すること、ここで前記閾値は、前記バリアントクラス中のバリアントに関して偽陽性と真陽性とを区別する、を含む。ある実施態様において、前記統計分布は、各試料に関する各位置におけるバリアント頻度の対数変換の統計分布である。

【００１２】

他の実施態様は、本明細書に記載の方法に関連するシステム、装置、及びコンピュータ可読媒体に関する。

【００１３】

以下の定義、詳細な説明、及び添付図面を参照することにより、本発明の本質と利点のより良い理解が得られるであろう。

【００１４】

定義

本明細書において用語「試料（sample)」又は「生物試料（biological sample)」は、核酸を含むか又は含むと推定される任意の組成物を指す。核酸は、動物（例えば哺乳動物、ヒト）、植物、微生物などに由来してもよい。試料という用語は、細胞、組織、又は血

液の、精製されたか又は分離された成分、例えばＤＮＡ、ＲＮＡ、タンパク質、無細胞部分、又は溶解物を含む。試料はまた、他のタイプの生物試料、例えば皮膚、血漿、血清、全血、及び血液成分（バフィーコート）、唾液、尿、涙、精液、膣液、吸引物若しくは洗浄液、組織生検、及び他の体液や組織（パラフィン包埋組織を含む）を指すこともできる。試料はまた、細胞株を含む個体から得られた細胞のインビトロ培養物の成分及び構成要素を含むことができる。「試験試料（test sample)」は、試料中のバリアントを検出するための試験中の試料を指す。

【００１５】

「ゲノムセグメント(genomic segment)」（「ゲノム断片（genomic fragment)」とも呼ばれる）は、生物のゲノム由来である完全に又は部分的にシークエンシングされた核酸分子である。これは、ＤＮＡセグメント（「ＤＮＡ断片」とも呼ばれる）又はＲＮＡセグメント（「ＲＮＡ断片」とも呼ばれる）でもよい。セグメントは、ゲノムの大きな部分を断片化することによって、例えば細胞を音波に供することによって作成することができる。ゲノムセグメントはシークエンシングして、「シークエンシングリード（sequencing read)」（「配列リード(sequence read)」又は単に「リード（read)」とも呼ばれる）を提供することができる。シークエンシングリードは、全ゲノムセグメント又はセグメントの一部であってもよい。

【００１６】

「参照試料（reference sample)」（「対照試料（control sample)」とも呼ばれる）は、試験試料と比較するため、基準、通常既知の基準となる試料を指す。例えば試験試料は、癌又は癌関連変異を有することが疑われる個体から採取することができ、癌のない個体又は癌関連変異のない個体からの参照試料（陰性対照）、又は癌又は癌関連変異を有することがわかっている個体からの参照試料（陽性対照）と比較することができる。対照はまた、多くの試験若しくは結果から集められた平均値又は範囲を表すことができる。

【００１７】

「標的領域（target region)」とは、分析される配列中の領域であって、診断的関連性を有し得る領域のことである。一例として、標的領域を含む断片は、プライマー及び増幅プロセスを用いて増幅するか、又はプローブを用いて濃縮することができる。「参照配列（reference sequence」（単に「参照（reference)」とも呼ばれる）は、配列リードがアラインメントされる任意の既知の配列である。種々の実施態様において、参照配列は、生物のゲノム又はトランスクリプトームのすべて又は一部のみに対応することができる。参照配列はまた、２種以上の生物のゲノムを含むことができる。例えば配列リードはまた、試料中に存在し得るウイルスのデータベースと比較することができる。

【００１８】

バリアント（variant)（変化（variation)又は変異(mutation)とも呼ばれる）は、２つの配列間の差を指す。バリアントは、例えば１つの塩基の１又は２以上の他の塩基への変化、１又は２以上の塩基の挿入、又は１又は２以上の塩基の欠失でもよい。参照配列中の位置の塩基は参照アレルと呼ばれることがあり、一方、試験試料上の同一の位置の異なる塩基（又は挿入もしくは欠失）はバリアントアレルと呼ぶことができる。例えばＡ＞Ｃの単一塩基置換に関して、Ａは参照アレルであり、そしてＣはバリアントアレルである。参照アレルは、天然に存在する生物のための最も一般的な遺伝子型を示す野生型アレルであってもよい。配列リードと参照配列の標的領域との差はカウントすることができ、真の変異が同定される可能性がある（例えば、十分な配列リードが変異を示す場合）。

【００１９】

試料の異なる配列リード上の特定の位置における同一のバリアントアレルの総数（例えばＡ＞Ｃバリアントについてのの数）は、バリアント数（variant count)と呼ばれる。ある試料の特定の位置のリードの総数は、リード数（read count)と呼ばれる。試料の特定の位置におけるバリアントタイプ又はクラス（例えばＡ＞Ｃ)のバリアント頻度は、試料の特定の位置におけるリード数に対する特定の位置におけるバリアントに関するバリアント数の比率として定義される。

【００２０】
本明細書において用語「位置（location)」は、配列中の（例えばゲノムの標的領域中の）１又は２以上の位置に対応する。例えば多塩基挿入が存在する場合、任意の長さのヌクレオチド（又は塩基対）が位置中に存在してもよい。
【００２１】
特に別の指定がなければ、本明細書で使用される技術用語及び科学用語は、一般に、当業者によって理解されるものと同じ意味を有する。例えば、Pfaffl, Methods: The ongoing evolution of qPCR, vol. 50 (2010); van Pelt-Verkuil et al. Principles and Technical Aspects of PCR Amplification, Springer (2010); Lackie, DICTIONARY OF CELL AND MOLECULAR BIOLOGY, Elsevier (4th ed. 2007); Sambrook et al., MOLECULAR CLONING, A LABORATORY MANUAL, Cold Springs Harbor Press (1989) を参照されたい。
【図面の簡単な説明】
【００２２】
【図１】図１は、本発明の実施態様に従う標的化されたウルトラディープシークエンシングのための次世代シークエンシング（ＮＧＳ）を用いた、ゲノムシークエンシング及びバリアントコーリングを示すフローチャートである。
【図２】参照配列と比較された標的領域の配列リードを示し、ここで、異なる配列位置における同一のクラス及び異なるクラスのバリアントは、本発明の実施態様に従って示される。
【図３Ａ】本発明の実施態様に従う１又は２以上の試料中の標的領域内の複数の位置のそれぞれの位置における、バリアントクラスのバリアントに関するバリアントの頻度分布の理想的な統計モデルを示す。
【図３Ｂ】特定の試料上の特定の位置におけるバリアントのバリアント頻度が、本発明の実施態様に従う特定のＺ値を有するであろう確率を示す。
【図３Ｃ】Ｚ値が、本発明の実施態様に従うｚ未満の又はｚに等しい値をとる確率の累積分布関数を示す。
【図３Ｄ】バリアント頻度値又はＺ値を有するバリアントが偽陽性であり、本発明の実施態様に従うバリアントコールを作成するための関連するクオリティスコアである確率を示す。
【図４】本発明の実施態様に従う統計モデルを用いるバリアントコーリングの方法を示すフローチャートである。
【図５】本発明の実施態様に従うサポートベクターマシン（ＳＶＭ）により決定したセパレーターラインを有するエキソン２０のＥＧＦＲ　Ｔ７９０Ｍのトレーニングデータと試験データに関する統計モデルを用いて決定されたバリアントクオリティスコアQ_{AMP}を示す。
【図６】本発明の実施態様に従うＳＶＭにより決定したセパレーターラインを有するエキソン２１のＥＧＦＲ　Ｌ８５８Ｒのトレーニングデータと試験データに関する統計モデルを用いて決定したバリアントクオリティスコアQ_{AMP}を示す。
【図７】本発明の実施態様に従う参照試料と試験試料の配列リード上の特定のゲノム位置における特定のバリアントを示す。
【図８】本発明の実施態様に従う試験試料と参照試料の配列リードデータを比較することにより、特定の配列位置における特定のバリアントに関するバリアントコーリングを示すフローチャートである。
【図９】本発明の実施態様に従うＳＶＭにより決定したセパレーターラインを有するエキソン２０のＥＧＦＲ　Ｔ７９０Ｍのトレーニングデータと試験データに関する２つの試料を比較することにより決定した局所化されたバリアントクオリティスコアQ_{LOC}を示す。
【図１０】本発明の実施態様に従うＳＶＭにより決定したセパレーターラインを有するエキソン２１のＥＧＦＲ　Ｌ８５８Ｒのトレーニングデータと試験データに関する２つの試料を比較することにより決定した局所化されたバリアントクオリティスコアQ_{LOC}を示す。

【図１１】本発明の実施態様に従うＳＶＭにより決定したセパレーターラインを有するエキソン１９のＥＧＦＲ　１５塩基欠失２２３５＿２２４９ｄｅｌ１５のトレーニングデータと試験データに関する２つの試料を比較することにより決定した局所化されたバリアントクオリティスコアQ_{LOC}を示す。
【図１２】本発明の実施態様に従う効率的なクオリティスコア推定に関するルックアップテーブルを示す。
【図１３】本発明の実施態様に従う低頻度バリアントコーリングに関する例示的コンピュータシステム例のブロック図を示す。
【図１４】シークエンシング装置とコンピュータシステムとの関係を示す一般的なブロック図の例である。
【図１５Ａ】本発明の方法及びシステムを実施するために使用できるソフトウェアとハードウェア資源との関係を示す一般的なブロック図の例である。
【図１５Ｂ】本発明の方法及びシステムを実施するために使用できるソフトウェアとハードウェア資源との関係を示す一般的なブロック図の例である。
【発明を実施するための形態】
【００２３】
シークエンシングは、癌又は他の疾患の突然変異を検出するために使用することができ、またインビトロ診断（ＩＶＤ）検査としても開発することができる。非侵襲的血液検査として、これらの検査を開発することが望ましい。しかし血液試料中のバイオマーカーの変異の頻度は低い。例えば、Kidess and Jeffrey, Circulating tumor cells versus tumor-derived cellfree DNA: rivals or partners in cancer care in the era of single-cell analysis? Genome. Med., 5:70 (2013), Diaz and Bardelli, Liquid biopsies: genotyping circulating tumor DNA, J. Clin. Oncol., 32:579-586 (2014); and Diehl et al., Nat Med., 14:985?990 (2008)を参照されたい。シークエンシングプロセスに関連するエラーが原因で、閾値が１％以下に設定される時、多くのＮＧＳソフトウェアパッケージは、１％以上の頻度を有するバリアントのみを報告する。
【００２４】
本発明の実施態様は、１％未満のバリアント頻度を有する低頻度バリアントに関する真陽性を検出するための解決策を提供する。正確なバリアントコールは、例えば標的化シークエンシングから得られる試料のシークエンシングリードに基づくことができる。例えば、いったん配列リードが受け取られ、参照配列にアラインメントされると、ある位置でバリアントを有する配列リードはカウントされる。試料の１つの位置で測定された特定のバリアントの第一のバリアント頻度は、他の位置で測定された及び／又は他の試料からの特定のバリアントの１又は２以上の第２のバリアント頻度と比較することができる。第２のバリアント頻度は、シークエンシングランのシークエンシングエラーに関する予測値に対応することができる。
【００２５】
いくつかの実施態様において、ある位置でバリアントが真陽性である信頼水準を示す確率値は、１又は２以上の試料中の標的領域内の複数の位置におけるバリアント数及び総リード数に基づいて計算することができる。その後、確率値は閾値と比較され、検出されたバリアントが真陽性であるか否かを決定することができる。他の実施態様において、試験試料と参照試料（例えば、その位置でシークエンシングエラーのみを有すると推定される）中の同一の位置におけるバリアント数と総リード数との差を用いて、試験試料中のバリアントが真陽性であるか否かを決定することができる。
【００２６】
Ｉ．標的化シークエンシングを用いるウルトラディープシークエンシング
ゲノムの特定の領域は、標的化シークエンシングを用いて効率的に分析することができる。例えば生物試料のゲノムセグメントは、標的領域に対応するセグメントをクローニングすることにより（例えば、ポリメラーゼ連鎖反応（ＰＣＲ）などの増幅プロセスにおいてプライマーを用いて）、及び／又は標的領域に対応するセグメントを優先的に捕捉する

プローブを用いることにより、増加又は増幅させることができる。標的増加試料中のゲノムセグメントは、大規模の平行した次世代シークエンシング（ＮＧＳ）を用いてシークエンシングし、標的領域内の可能な変異を調査するために分析することができる。

【００２７】

しかし、このようなプロセスはエラーを発生させることがある。例えば、増幅又は濃縮の前段階を有する高スループットの次世代シークエンシングを用いるバリアント検出では、アンプリコン／濃縮ライブラリ（標的増加試料）は偽陽性リードを含む可能性がある。ＰＣＲは点突然変異及びインデル（indel）を導入することができ、これはまた、組換え配列又はキメラを生成することができる。更に遺伝的バリアントの相対頻度は、ＰＣＲ中の選択的増幅の偏りにより攪乱されることがある。ＰＣＲ中に、追加の単一塩基エラーが起きることがある。シークエンシング自体は、塩基置換エラー及びインデルを導入し得る。これらのエラーは間違った変異報告につながることがあり、疾患の診断のために誤解を招く情報を提供することがある。偽陽性は種々の方法、例えばプライマーの正しい設計及び高忠実度酵素の開発、によって減少させることができる。しかし、それでも偽陽性が残って、多くの場合エラー率が約１％以上に大きくなる可能性がある。

【００２８】

各個別のヌクレオチドについてのシークエンシングの精度は比較的高くなる可能性があるが、ゲノム中の大多数のヌクレオチドは、個々のゲノムが一度だけシークエンシングされた場合、かなりの数のシークエンシングエラーが存在することを意味する。例えば、１塩基対当たり０．２％のエラー率と４００塩基対のリード長さの場合、少なくとも１つのエラーを有するリードの割合は、１－（１－０．００２）^{400}＝０．５５１であり、これは５５％を超える配列リードが少なくとも１つのエラーを有する可能性を意味する。したがって、シークエンシングエラーと希少な真の変異を区別するためには、個々のゲノムに多数回シークエンシングすることによりシークエンシング精度を高めることが望ましい。例えば、たとえ各配列リードが１％のエラー率を含んでいても、バリアントの位置をカバーする８つの同一のリードの組合せは、エラー率が（１０$^{-2}$）^{8}すなわち１０$^{-16}$の強く支持されるバリアント検出を生成するであろう。

【００２９】

ＤＮＡシークエンシングの深度（depth)は、シークエンシングプロセス中にヌクレオチドが読み取られる回数を意味する。ディープシークエンシングは、リードの総数が調査中の配列の長さよりも何倍も大きいことを示している。カバー率は、再構築された配列内のあるヌクレオチドを表すリードの平均数である。「ディープ(deep)」という用語は、７倍超などの広範囲の深度について使用されており、用語「ウルトラディープ(ultra-deep)」は、一般に１００倍超などのより高いカバー率を意味する。シークエンシング深度の要件は、バリアントのタイプ、疾患モデル、及び関心領域の大きさに依存し得る。すなわち、１％以下のバリアント頻度を有する希少バリアントについては、より高いカバー率が所望される。大規模な平行ＮＧＳは、真のバリアント検出のためのそのようなウルトラディープシークエンシングを可能にする。それにもかかわらず、より短いリードの大きい深度を生成することは、必ずしも希少バリアント検出に関する全ての課題を解決しない。

【００３０】

ＩＩ．ウルトラディープシークエンシングにおけるバリアントコール

バリアントコーリングは、試験試料と参照配列の配列リード間の真の差を識別するプロセスである。バリアントコーリングは、試料の特性評価及び疾患の診断において重要である。しかし、非常に低い頻度でしばしば体細胞バリアントが発生するため、バリアントコーリングは本質的に難しい。バリアントコーリングの１つの目標は、謝った偽陽性を最小にするために高い信頼度で体細胞バリアントを同定することである。

【００３１】

図１は、標的化ウルトラディープシークエンシングのための次世代シークエンシング（ＮＧＳ）を用いる、ゲノムシークエンシング及びバリアントコーリングの方法１００を示す。他の方法と同様に、実施態様は、記載された工程のすべて又は一部を含むことができ

、いくつかの工程はコンピュータシステムを用いて行うことができる。方法１００の結果は、生物の診断を決定する際に医師によって使用することができる。

【００３２】

ブロック１１０において、シークエンシングされ、かつ診断されるポリヌクレオチドを含む試料は受け取られ、ここで、前記ポリヌクレオチドは、シークエンシングされるべき標的領域を潜在的に含む。上記で定義したように、用語「試料」は核酸を含むか又は含むと推測される任意の組成物を指す。試料は、そこから試料が得られる生物のゲノムに由来する核酸分子を含む。例えば試料は、染色体中にコードされたゲノムを含有する細胞を含むことができる。試料は、１又は２以上の試験試料を含むことができる。試料はまた、１又は２以上の参照試料又は対照試料を含むことができる。いくつかの試料は、ゲノムの特定の領域における変異について試験されている患者から得ることができる。試料は、癌について試験されている腫瘍の生検から得ることができる。試料は、いくつかの正常細胞、癌進行の初期段階のいくつかの細胞、及び癌の進行の後期ステージのいくつかの細胞を含むことができる。試料は、異なる人や同じ人物（例えば、異なる生検）由来でもよく、異なる実験条件を用いてもよい。

【００３３】

場合により、ブロック１２０においてＲＮＡ又はＤＮＡは、シークエンシング前に試料から分離される。生物試料から核酸を単離するための方法は、例えばSambrookに記載されるように公知であり、いくつかのキットは市販されており、例えば、DNA Isolation Kit for Cells and Tissues, DNA Isolation Kit for Mammalian Blood, High Pure FFPET DNA Isolation Kit, High Pure RNA Isolation Kit, High Pure Viral Nucleic Acid Kit, and MagNA Pure LC Total Nucleic Acid Isolation Kitがあり、全てがRocheから入手可能である。いくつかの実施態様において、単離された核酸はゲノムＤＮＡを含む。いくつかの実施態様において、単離された核酸は、循環遊離ＤＮＡ断片（circulating free DNA fragments）（ｃｆＤＮＡ）を含む。いくつかの実施態様において、単離された核酸は、細胞性ｍＲＮＡ又はｃｆＲＮＡなどのＲＮＡを含む。

【００３４】

ＲＮＡの場合、ブロック１３０において、逆転写反応が行われる。例えばＲＮＡは、逆転写酵素を用いて相補的ＤＮＡ（ｃＤＮＡ）に変換することができる。

【００３５】

場合により、ブロック１４０において、シークエンシングのためにＤＮＡセグメントを調製することができる。これはＤＮＡを、標的領域を含むより小さなＤＮＡセグメントに断片化し、ＤＮＡセグメントの末端にアダプター配列を連結し、そしてＤＮＡ断片が由来する試料を同定する固有のバーコード配列を固定することを含むことができる。標的領域は、例えば任意の癌関連変異があるかどうかを調べるための、診断関連性を有する可能性のあるＤＮＡ中のセグメントである。例として標的領域は、ほぼ数百塩基、例えば１５０～２５０塩基、１５０～４００塩基、又は２００～６００塩基であることができる。別の実施態様において、標的領域に対応するゲノムセグメントを捕捉するためにプローブを使用することができる。例えば、標的領域にハイブリダイズするように設計されたプローブを、表面上に配置することができる。次にゲノムセグメントをその表面の上に配置することができ、標的領域のセグメントが優先的にハイブリダイズされ得る。試料のＤＮＡは、例えば超音波処理又は他の適切な方法によって断片化して、より小さなゲノムセグメントを得ることができる。例えば、２００～５００塩基長さのゲノムセグメントを得ることができる。特定のシークエンシング操作について、ほぼこの長さのゲノムセグメントが好ましい。しかし実施態様は、任意の長さのゲノムセグメントを使用することができる。

【００３６】

ゲノムセグメントは、バーコードやマルチプレックス識別子（ＭＩＤ）配列でマークすることができる。例えば１０塩基の配列を、ゲノムセグメントの末端を（例えば、リガーゼを用いて）加えることができる。このように、種々の試料からのセグメントは、単一のシークエンシングラン中に並行してシークエンシングすることができる。ＭＩＤは配列リ

ードの一部として読み取ることができ、同じＭＩＤを有する配列リードは同じ試料に起因し、一緒に分析することが可能である。ＭＩＤは、異なる試料から配列リードを脱多重化又は区別するために使用することができる。

【００３７】

ブロック１５０においてＤＮＡセグメントは、ＰＣＲ、ＳＤＡ、及びこれらの派生方法などの増幅法により場合により増幅又は増加させて、ＤＮＡセグメントすなわちシークエンシングのための増幅産物を生成することができる。Ｔａｑポリメラーゼ又は他の耐熱性ポリメラーゼなどのＤＮＡポリメラーゼを、ＰＣＲによる増幅のために使用することができる。例えば、増幅法の総説については、Fakruddin et al., J Pharm Bioallied Sci. 5:245 (2013)を参照されたい。これらの増幅産物は、増幅に使用されるプライマーに基づいて規定される。プライマーは、核酸上の標的領域に特異的である。シークエンシングプライマーが増幅産物内の配列に特異的に特異的である（特異的にハイブリダイズする）ように、シークエンシングプライマーは典型的には、増幅プライマーの選択に基づいて設計される。いくつかの実施態様において標的領域は、標的濃縮工程によって濃縮することができる。増幅及び濃縮プロセスの両方を実行することができる。フォワードプライマー及びリバースプライマーは、標的領域を増幅するために使用することができる。これらのフォワードプライマー及びリバースプライマーは種々の長さ、例えば約１５～３０塩基長のものでもよい。

【００３８】

いくつかの実施態様において、試料特異的ＭＩＤの添加は、異なる時点で発生し得る。例えばＭＩＤは、増幅／濃縮後に添加することができ、次に試料は一緒に混合される。こうして、異なる試料は、異なる標的領域について増幅又は濃縮することができるであろう。

【００３９】

ブロック１６０において、１又は２以上の試料からのＤＮＡセグメントは、単一のシークエンシングランで大規模に並列様式でシークエンシングされる。シークエンシングプロセスにおいて、増幅過程で作成された同じセグメントのクローンは、別々に決定された配列を有することができる（及び後にカウントされる）。いくつかの実施態様において単一のシークエンシングランは、１テラ塩基（terabase)を超えるデータを生成することができる。いくつかの実施態様において、１試料当たり約３，０００リード超を得ることができる。リードの数は、試料のサイズ、標的増加の一部としてどの程度の増幅が行われるか、及びシークエンシングプロセスのバンド幅（すなわち、どの程度のシークエンシングに対して装置が設定されるか、例えばいくつのビーズが使用されるか）に依存してもよい。ある実施態様において、リードは約１５０～２５０塩基長である。

【００４０】

シークエンシングプロセスは、Roche 454, Illumina GA, 及び ABI SOLiDなどの種々のＮＧＳプラットフォーム上の種々の技術によって行うことができる。ある実施態様においてＤＮＡセグメントは、シークエンシングの一部として増幅を受けることができる。増幅プロセスが標的増加試料を作成するために使用される実施態様において、この増幅は第二の増幅工程であろう。第二の増幅は、第二の増幅が行われなかった場合よりも、強いシグナル（例えば、特定の塩基：Ａ、Ｃ、Ｇ、又はＴに対応する蛍光シグナル）を提供することができる。

【００４１】

シークエンシング処理の一例において、ブロック１５０からの増幅されたセグメント（例えば、増幅が溶液中で発生した場合）は、それぞれビーズに付着させることができる。付着したセグメントは、次にビーズ上で増幅することができ、各ビーズから１つの配列リードを得ることができる。表面を使用する実施態様において、セグメントを表面に付着させ、次に表面上で単一のクラスタを作成することができる。各クラスタについて単一の配列リードを得ることができる。配列リードは、ゲノムセグメントの全長又はセグメントの一部についてのものであることができる。

【００４２】
ブロック１７０において、場合により配列リードは濾過されて、低クオリティリードと短いリードが除去され、残りの配列リードは参照配列の標的領域にアラインメントされる。いくつかの実施態様において、同一の塩基を有するリードは、単一の配列リードと見なされるように組み合わされる。したがって、唯一のユニークリード (unique read)に関するリード数を記録することができる。平均塩基スコアは、全てのユニークリードに関する全ての塩基位置において計算され得る。塩基スコアは、塩基コールが配列リード上で如何に正確であるかを測定することができる。塩基スコアを使用して、低クオリティリードを除去することができる。いくつかの実施態様において、最小値よりも短いリードも同様に除去される。

【００４３】
アラインメントすることにより、本方法は配列リードを参照配列の標的領域と比較して、配列リードと参照配列との間の変化の数を決定することができる。アラインメントは１又は２以上の標的領域にのみ特異的であることができ、ゲノム全体を検索する必要はないため、アラインメントは高速であることができる。また標的領域に対応するセグメントの割合が増加するにつれて、かなりの数のリードが標的領域に良好に一致するであろう（例えば、比較的少数の変化）。

【００４４】
ある実施態様において、複数の標的領域が使用される場合、配列リードは複数の標的領域のすべてと比較することができ、最良のアラインメントを提供する標的領域を同定することができる。異なる標的領域は、異なる遺伝子又は遺伝子を有する異なるエクソンを有することができる。したがって、最良のアラインメントを有するエクソンが同定され得る。

【００４５】
バーコード又はＭＩＤが使用される場合、それはアラインメント前に除去され得る。特定の試料に関する全てのリードを１つのグループに構成するために、ＭＩＤを使用することができる。このようにして、他の試料からの変異は、特定の試料の分析には影響しないであろう。このグループ化は、脱多重化 (de-multiplexing)と呼ばれる。異なる試料は異なる標的領域を有することができるため、アラインメントについて参照配列のどの標的領域を比較すべきかを決定するためにＭＩＤは使用され得る。

【００４６】
ブロック１８０において、標的領域からのアラインメントされた配列リードは、標的領域における変異を同定するために使用される。この工程の一部として、バリアントアレルの数（又はバリアント数）、参照アレルの数（又は野生型数）、従って各試料に関する配列位置における各バリアントの頻度を決定することができる。例えば、標的領域内の特定の位置について、通常のＡの代わりにＧ変異が現れる回数をカウントすることができる。Ｇ変異が見られる回数の割合は、その位置にアラインメントされる全リードから決定することができる。いくつかの実施態様において、一緒に発生する変化を識別することができ、同じ変異の一部として分類することができる。各試料について、標的領域のシークエンシング深度は、その試料について任意のフィルターを通過するリードの数から決定することができる。

【００４７】
ブロック１９０において、バリアント数、野生型数、及び／又はバリアント頻度に基づいてバリアントコーリングが行われる。ある実施態様において、特定のバリアントのバリアント頻度は、実際の変異と見なすためには、閾値（存在フィルタ (abundance filter)）よりも大きいことが必要とされ得る。表１は、Illumina MSR体細胞変異コーラー（caller）がデフォルト設定で報告するポアソン（Poisson）モデルに基づいて計算される最少バリアント数とバリアント頻度を示す。

【表１】

深度	報告される最小数	報告される最小%
100	5	5
200	7	3.5
500	12	2.4
1000	19	1.9
2000	32	1.6
5000	68	1.36
10000	125	1.25
20000	235	1.175
50000	554	1.108
100000	1075	1.075

表１．種々の深度について報告される最小バリアント数と最小バリアント頻度

【００４８】

いくつかの実施態様において、バリアントが実際に試料中に存在する信頼水準を示すクオリティスコアが提供され、バリアントコールを行うために使用される。いくつかの実施態様において、クオリティスコアは、バリアント数、野生型数、及び／又はバリアント頻度の１又は２以上と組み合わせて使用して、バリアントコールを行うことができる。医師は同定された変異を使用して、癌の素因を診断するか又は癌を有するとして腫瘍を同定することができる。

【００４９】

図２は、参照配列２１０と比較した試験試料中の標的領域２１５の配列リードの例を示し、ここで、種々の配列位置における同一のクラスと異なるクラスのバリアントが示される。図２は、参照配列が塩基Ａの参照アレルを有する標的領域内の４つの位置の例を示す。例示を容易にするために５つの配列リードが明示的に示されるが、実際にはより多くのリードが使用される。参照配列２１０は、位置２０５、２３１、２５５、及び２８１においてＡを有することが示される。

【００５０】

位置２０５について、いくつかの配列リードにおいてＡが検出されるが、いくつかの配列リードではＣが検出される。Ｃの検出は、潜在的なＡ＞Ｃバリアントを示す。バリアントＡ＞Ｃは、特定のバリアントクラスのものである。他の塩基は、示されていない配列リード中で検出され得る。他のアレルの存在は、他のバリアントクラスの他のタイプのバリアントを示す可能性がある。

【００５１】

位置２３１について、いくつかの配列リードにおいてＡが検出される；いくつかの配列リードではＣが検出される；更にいくつかの他のリードにおいて、位置２３１において何も検出されない（「０」）。Ｃの検出は、単一塩基置換Ａ＞Ｃの潜在的なバリアントを示す。「０」の検出は潜在的な欠失のバリアントを示す。

【００５２】

位置２５５における塩基Ａについて、いくつかの配列リードにおいてＡが検出される。いくつかの配列リードではＣが検出される；しかし、いくつかの他のリードではＴが検出される。Ｃの検出は、単一塩基置換Ａ＞Ｃの潜在的なバリアントを示す。Ｔの検出は、異なる単一塩基置換Ａ＞Ｔの潜在的なバリアントを示す。

【００５３】

位置２８１における塩基Ａについて、いくつかの配列リードにおいて、Ａが検出される；いくつかの配列リードにおいて、Ｃが異なる頻度で検出される。異なる頻度でのＣの検出は、異なるバリアント頻度を有する単一塩基置換Ａ＞Ｃの潜在的なバリアントを示す。

【００５４】
試験試料についての配列リードに基づいて、それぞれの位置について、野生型塩基Ａの数、単一塩基置換Ａ＞Ｃの数、単一塩基置換Ａ＞Ｔの数、及びＡの欠失の数をカウントすることができる。図２に示されるバリアントのタイプは、例示のみが目的である。本開示において後述されるように、種々のタイプのバリアント又は変異が存在し得る。
【００５５】
ＩＩＩ．統計分布モデルに基づくバリアントコーリング
本開示のいくつかの実施態様において、ＮＧＳ実験で観察されるすべてのバリアントを報告することができる。低頻度の真陽性を偽陽性から区別するために、ほとんどの観測される低頻度バリアントは偽陽性であってもよいため、偽陽性バリアントの分布を用いて、バリアントコーリングクオリティスコアを確立して、バリアントが真陽性である可能性を決定することができる。
【００５６】
Ａ．統計モデルに基づくバリアントコーリングの数学的理論
図３Ａ～３Ｄは、本発明のいくつかの実施態様に従う統計モデルに基づくバリアントコーリングの基礎となる数学的理論を提供する。バリアントコーリングの偽陽性率は配列状況や位置に依存しているため、全ての試料中の種々の位置におけるＡ＞Ｃなどの同一のクラス又はタイプのバリアントは一緒に比較して、統計分布に基づくバリアントコールを行うことができる。
【００５７】
いくつかの実施態様において、異なる配列位置でのシークエンシングランにおける単純なバリアントは、２０のクラスに分けることができる。すべてのクラスにおいて、バリアントの大部分は偽陽性である。すべてのバリアントクラスの統計分布のパラメータを計算することができる。バリアントクラスは以下のように定義することができる：
（１）Ａ＞Ｃ、Ａ＞Ｇ、Ａ＞Ｔ、Ｃ＞Ａ、Ｃ＞Ｇ、Ｃ＞Ｔ、Ｇ＞Ａ、Ｇ＞Ｃ、Ｇ＞Ｔ、Ｔ＞Ａ、Ｔ＞Ｃ、及びＴ＞Ｇを含む１２の単一塩基置換；
（２）ＡＣ＞ＧＡなどの多塩基置換；
（３）ＡＧＴ＞ＡＴ又はＧＣＡＴ＞ＧＴなどの１～２塩基の欠失；
（４）ＡＴＣＧＡ＞ＡＡなどの３塩基の欠失；
（５）ＧＡＣＣＴＡ＞ＧＡ又はＴＧＣＧＣＧＡ＞ＴＡなどの４～５塩基の欠失；
（６）ＡＴＣＣＴＣＡＧ＞ＡＧなどの６塩基以上の欠失；
（７）ＡＴ＞ＡＡＴ又はＧＣ＞ＧＴＡＣなどの１～２塩基の挿入；
（８）ＧＣ＞ＧＴＡＡＣ又はＡＣ＞ＡＧＡＴＧＣなどの３塩基以上の挿入；そして
（９）単一塩基置換Ａ＞Ｃなどの他の単純な変異に、すぐ続く１塩基の欠失、例えば、元々の参照塩基がＡＴであり、変異塩基がＣである、すなわちＡＴ＞Ｃ。そのような変異ＡＴ＞Ｃはまた、Ａの欠失に単一塩基置換Ｔ＞Ｃが続くと解釈することができる。
【００５８】
本明細書において単純な突然変異は、その中に一致する塩基対無しで、２つの一致する塩基対によって結合された変異である。例えばａＡＴｇやａＣｇにおいて、単純な変異ＡＴ＞Ｃは、一致する対ａ－ａと一致する対ｇ－ｇにより結合され、ここで、小文字は一致する対について使用される。しかしａＡｃＧｇとａＣｃＴｇにおいて、ＡｃＧ＞ＣｃＴは、この中に一致する対ｃ－ｃが存在するため単純な変異ではない。このように、ＡｃＧ＞ＣｃＴは、２つの単純な変異Ａ＞ＣとＧ＞Ｔからなる複合変異である。
【００５９】
いくつかの実施態様において、１又は２以上の試料について参照配列中に参照アレル（例えば、バリアントタイプＡ＞Ｃについての塩基Ａ）が存在する標的領域中の種々の位置における、同一のクラスのバリアント、例えばＡ＞Ｃのバリアント頻度を用いて、バリアントクラスについて統計分布を作成することができる。例えば図２に示すように、位置２０５、位置２３１、位置２５５、位置２８１のそれぞれ、及び試料の参照配列内にＡが存在する標的領域内の他の位置における単一塩基置換Ａ＞Ｃのバリアント頻度は、バリアン

トクラスＡ＞Ｃに関する統計分布のためのデータ点であってもよい。位置２０５、位置２３１、位置２５５、位置２８１のそれぞれ、及び試験試料と同一のシークエンシングランでシークエンシングされる他の試料のそれぞれの参照配列内にＡが存在する標的領域内の他の位置における単一塩基置換Ａ＞Ｃのバリアント頻度は、バリアントクラスＡ＞Ｃに関する統計分布のためのデータ点であってもよい。一方、位置２０５、位置２３１、位置２５５、位置２８１のそれぞれ、及び試験試料と同一のシークエンシングランでシークエンシングされる各試料の参照配列内にＡが存在する標的領域内の他の位置における単一塩基置換Ａ＞Ｔ又は単一塩基欠失Ａ＞０などの異なるバリアントクラスのバリアント頻度は、バリアントクラスＡ＞Ｃについて統計分布のために使用されない。

【００６０】

いくつかの実施態様において、少なくとも３０のデータ点が統計分布に含まれる。少なくとも３０のデータ点は、単一のシークエンシングラン中に２以上の試料からのものであってもよい。３０未満のデータ点の場合は、真の分布はデータ点によって表されない場合がある。

【００６１】

図３Ａは、同一のクラスのバリアントについてのバリアント頻度の理想的な統計分布（正規分布）を示す。図３Ａは例示のみが目的である。バリアントクラスのバリアント頻度の実際の統計分布は試料に依存することがあり、二峰性分布のような他の分布形態であってもよい。いくつかの実施態様において、バリアント頻度の二乗、平方根又は対数のような変換のいくつかの形態は、正規分布に近い分布を形成することができる。

【００６２】

図３Ａにおいて、ｘ軸はバリアントクラスのバリアント頻度値を示し、ｙ軸は特定のバリアント頻度値ｆを有するデータ点の数を示す。図３Ａ中の理想的な正規分布に示されるように、平均値ｍと標準偏差ｓは分布に基づいて決定することができる。

【００６３】

図３Ｂは、特定の試料上の特定の位置におけるバリアントのバリアント頻度が所定のＺ値を有する確率を示し、ここで、この確率とＺ値は、図３Ａに示された統計分布に由来することができる。いくつかの実施態様において図３Ｂは、平均と標準偏差とに基づく図３Ａの正規化された分布であってもよい。いくつかの実施態様において、より複雑な変換又は転換、例えば対数変換を使用することができる。図３Ａ中の斜線部は、ｚに等しいか又はｚより大きい全てのＺ値の累積確率を示す。

【００６４】

図３Ｃは、Ｚ値がｚ未満又はこれに等しい値を取る確率の累積分布関数Ｆを示す。

【００６５】

図３Ｄは、特定のバリアント頻度値又はＺ値を有するバリアントが、左の主軸上で偽陽性である塩基コーリング誤り確率（ｐ値）と、右の２次軸上でバリアントコールを行うための関連するクオリティスコアＱを示す。いくつかの実施態様においてｐ値は、１－Ｆによって計算することができる。いくつかの実施態様においてクオリティスコアＱは、$Q=-10\log_{10}p$によって与えられるフレッド（Phred）クオリティスコア、又はフレッドクオリティスコアの任意の変形であってもよい。

【００６６】

Ｂ．統計モデルに基づくバリアントコーリングの方法

図４は、統計モデルを用いるバリアントコーリングの方法４００を示す。他の方法と同様に実施態様は、記載された操作の全て又は一部を含むことができ、いくつかの操作は追加の操作又はサブ操作を含むことができる。

【００６７】

ブロック４１０において、単一のシークエンシングランで１又は２以上の試料中の標的領域を標的化する配列リードが受け取られる。配列リードデータは受け取られ、読み取り可能なフォーマットで保存され、コンピュータで解析することができる。いくつかの実施態様において、低クオリティリード又はアダプター配列を除去するために、配列リードデ

ータの予備処理を実行することができる。いくつかの実施態様において、バーコード又はＭＩＤを除去することができ、同じ試料からの配列リードは、標識又はグループ化されてもよい。

【００６８】

例えば方法１００のブロック１７０に記載されるように、ブロック４２０において配列リードは参照配列の標的領域にアラインメントされる。

【００６９】

ブロック４３０において、試験試料のアラインメントされた配列リード上の特定の配列位置における同じバリアントクラスのバリアントアレルは、同定され、カウントされて、バリアント数を決定することができる。試験試料のアラインメントされた配列リード上の特定の配列位置についてのリード数も、同様に決定することができる。例えば図２に示すように、試験試料の配列リード中の位置２０５におけるＣの総数は、位置２０５におけるバリアントクラスＡ＞Ｃについてのバリアント数であり、試験試料の配列リード中の位置２０５におけるリードの総数は、位置２０５におけるバリアントクラスＡ＞Ｃについてのリード数である。いくつかの実施態様において、試験試料についての特定の位置のリード数は、別の操作で決定することができる。

【００７０】

ブロック４４０において、特定の位置における同一のクラスのバリアントのバリアント頻度が決定される。ある実施態様においてバリアント頻度は、試験試料中の特定の位置におけるバリアント数をリード数で割ることにより決定することができる。別の実施態様においてバリアント頻度は、試験試料中の特定の位置においてバリアント数を非バリアント数（例えば、リード数－バリアント数）で割ることによって決定することができる。当業者であれば、使用することができるバリアント頻度の種々のタイプの形態を理解し得る。

【００７１】

ブロック４５０において、試験試料として同一のシークエンシングランでシークエンシングされた各試料について、同一のクラスのバリアント、例えばＡ＞Ｃは、バリアントクラスの参照アレル、例えばＡが、標的領域における参照配列上に存在する複数の位置のそれぞれの位置で、同定されカウントされる。同様に、同一のシークエンシングランにおける各試料について、参照配列上でバリアントクラスの参照アレルが見つかる複数の位置のそれぞれに関するリード数を決定することができる。

【００７２】

ブロック４６０において、試験試料として同一のシークエンシングランでシークエンシングされた各試料について、複数の位置のそれぞれの位置における、同一のバリアントクラス、例えばＡ＞Ｃのバリアント頻度は、各位置についてのバリアント数をその位置のリード数により割ることによって決定することができる。すなわち、もし、例えば３つの試料がシークエンシングランで一緒にシークエンシングされ、標的領域内の参照配列上の３０の位置が、バリアントクラスについての参照アレルを有する場合、各試料上の各位置について１つで最大９０のバリアント頻度を計算することができる。これらのバリアント頻度を用いて、同一のシークエンシングランで同一のクラスのバリアントについてバリアント頻度の統計分布を決定することができる。他のバリアントクラスのバリアント頻度は、統計分布を決定するために含まれていないことに注目されたい。更に、分布モデルの正確性に影響を与える可能性のあるシークエンシングラン間の変動の影響を低減するために、他のシークエンシングランから得られたデータ点は、統計分布を決定するために含まれない。

【００７３】

ブロック４７０において、試験試料中の特定の位置における同一のクラスのバリアントに関するバリアント頻度に対応する確率値は、バリアント頻度をブロック４６０において形成された統計分布のパラメータと比較することにより決定される。いくつかの実施態様において、確率値は実際の確率、累積分布、又はクオリティスコアでもよい。いくつかの実施態様において、統計分布のパラメータは、平均値と標準偏差の１又は２以上でもよい

。

【００７４】

ブロック４８０において、試験試料上の特定の位置におけるバリアントクラスのバリアントが真陽性か否かを決定するために、確率値と閾値に基づいてバリアントコールが行われる。いくつかの実施態様において、閾値は単一の値でもよい。いくつかの実施態様において閾値は、例えばバリアント頻度の関数でもよい。いくつかの実施態様において閾値は、サポートベクターマシン（ＳＶＭ）などのマシン学習アルゴリズムを用いて、トレーニングデータセットに基づいて決定することができる。いくつかの実施態様において、閾値は、異なるシークエンシングランから得られるトレーニングデータに基づいて決定することができる。

【００７５】

上記した方法は、以下の例に照らしてより良く理解することができる。

【００７６】

Ｃ．例

以下の例は、このセクションで上記した方法を示す。以下の例において、バリアント頻度が正規分布ではなく、一方で対数バリアント頻度の分布が後述されるように正規分布に近いため、バリアントクラスに関する対数バリアント頻度の統計分布に基づくモデルが使用される。

【００７７】

表２は、エクソン２０の置換Ｔ７９０Ｍ（２３６９でＣ＞Ｔ）とエクソン２１の置換Ｌ８５８Ｒ（２５７３でＴ＞Ｇ）の偽陽性を有する野生型データに関する元々のバリアント頻度ｆとその対数変換ｘに適用される、Lilliefors検定及びShapiro-Wilk検定などの正規性検定の結果を示す。この結果は、正規分布の仮定が使用される時、ｘが観察された試料の結果を得るためのより大きな確率（Ｐ値）（＞０．０８）を有することを示し、これは、実際の分布と正規分布との間のより小さな相違を示し、ｆがより小さいＰ値（＜０．０１６）を有することを示す。従って、ｘはｆよりも正規分布に近い。

【表２】

偽陽性	変数	Lilliefors検定のＰ値	Shapiro-Wilk検定のＰ値
T790M	f	0.008805	0.001830
	$\log_{10}$(f+1e-06)	0.348639	0.084104
L858R	f	0.014024	0.015862
	$\log_{10}$(f+1e-06)	0.602277	0.520155

表２．ｆ及びｘに関する正規性検査のＰ値

【００７８】

変換されたバリアント頻度は、ほとんどノイズについて元のバリアント頻度よりも正規分布に近いため、統計分析を行うのに通常の近似を使用するために、最初にバリアント頻度の対数変換が行われる。いくつかの実施態様において、ｆ＝０のときの負の無限大値を避けるために、以下の対数変換が使用される。

【数１】

$$x = \mathrm{Log}_{10}(f + e)$$

【００７９】

ここで、ｅは負の無限大値を回避するための調整定数である。調整定数ｅは、任意の適切な値に設定することができる。例えば、いくつかの実施態様において、ｅは10^{-6}に設定することができ、従って最小のｘ値は－６である。

【００８０】
対数変換した後、正規分布近似の平均値ｍと標準偏差ｓを計算することができる。次に正規分布の近似を用いて、配列の位置で検出されるバリアントの確率値を計算することができる。例えば、バリアント頻度ｆ１、対数バリアント頻度ｘ１＝Log_{10}（ｆ１＋ｅ）、十分な深度（総リード数）を有する、ある位置におけるバリアントクラス中の観察されたバリアントに関して、統計的確率値ｚスコアは、以下によって計算することができる：
【数２】

$$z = (x1 - m) / (s / sqrt(n))$$

【００８１】
ここで、ｎは、ｓとｍの推定に用いられる参照データ点の数である。計算結果は、ｚスコアが大きいｎに対して大きいことを示し、これは非常に小さい塩基コーリングエラー確率（ｐ値）を生成し、したがって非常に大きなクオリティスコアを生成することができることを示す。したがっていくつかの実施態様において、ｚスコアはｚ様スコアで置換することができ、これは、上記式中のｎをｍｉｎ（ｎ，Ｎ）で置換することにより計算される。Ｎは、任意の適切な値に設定することができる。いくつかの実施態様において、Ｎは３６に設定される。いくつかの実施態様において、下限ｓ２はまた、ｓが小さすぎる状況では、ｓ／ｓｑｒｔ（ｍｉｎ（ｎ，Ｎ））に設定することができる。ｓ２は、例えばデフォルト値０．０１のような任意の適切な値に設定することができる。したがって、いくつかの実施態様において、ｚ様スコアは以下により表すことができる。
【数３】

$$z' = (x1 - m) / max(s2, s / sqrt(min(n, N)))$$

【００８２】
ｚスコア又はｚ様スコアｚ’を用いて、塩基コーリングエラー確率ｐ値は、ｐ＝１－Ｆ（ｚ）又はｐ＝１－Ｆ（ｚ’）により決定することができ、ここでＦは標準正規分布の累積分布関数である。次にバリアントコーリングクオリティスコアQ_{AMP}は、フレッドスコアを用いて決定することができる。いくつかの実施態様において、Q_{AMP}はフレッド様スコア：
【数４】

$$Q_{AMP} = -10 \log_{10}(max(p, minP))$$

【００８３】
として定義することができる。ここで、ｍｉｎＰは$10^{-maxQ/10}$である。ｍａｘＱは任意の適切な値に設定することができる。例えばいくつかの実施態様において、ｍａｘＱは、８０又は１３０に設定してもよい。
【００８４】
いくつかの実施態様において、クオリティスコアを計算するために、試料平均と試料標準偏差の代わりに、データの中心位置と変動のロバスト推定（robust estimations)を使用することができる。
【００８５】
いくつかの実施態様において、線形カーネル(linear kernel)を有するサポートベクターマシン（ＳＶＭ）などの分類法は、既知の真陽性と偽陽性とを有するトレーニングデー

タセットを用いて、偽陽性から真陽性を分離するために使用することができる。いくつかの実施態様において、閾値はデータを視覚化することによって設定することができる。

【００８６】

図５及び図６は、実際の試料から配列リードデータに適用された上記方法の結果の例を示す。図５は、ＳＶＭによって決定されるセパレーターラインを有するエクソン２０のＥＧＦＲ　Ｔ７９０Ｍ（２３６９でＣ＞Ｔ）の異なるトレーニングデータと試験データに関する、最大対Ｎ＝４を有するバリアントコーリングクオリティスコアQ_{AMP}を示す。図５は、バリアントと野生型データが十分に分離されていないことを示し、従って、０．１％以下のバリアント頻度で真の変異と偽陽性を区別することは困難であるかも知れない。しかし、０．５％以上のバリアント頻度を有するすべての試験データと少なくとも０．２％のバリアント頻度を有するほとんどの試験データに関しては、真の陽性と偽陽性は正確に区別することができる。

【００８７】

図６は、バリアントは、ＳＶＭによって決定されるセパレーターラインを有するエクソン２１のＥＧＦＲ　Ｌ８５８Ｒ（２５７３でＴ＞Ｇ）のトレーニングデータに関するｍａｘＮ＝４を有するバリアントコーリングクオリティスコアQ_{AMP}を示す。図６は、０．１％のバリアント頻度を有するものを含むすべての試験データが、正しく分類することができることを示す。

【００８８】

ＩＶ．１又は２以上の参照試料との比較を用いる、特定の位置における特定のバリアントに関するバリアントコーリング

本発明のいくつかの実施態様において、異なる試料中の同一の位置におけるバリアント及びバリアントの野生型数を比較して、バリアントコールを行うことができる。この方法は、シークエンシングランにおいて陰性対照として野生型（通常は正常）試料が利用可能である場合に特に有用である。

【００８９】

Ａ．特定の位置における特定のバリアントを検出するために試験試料と参照試料とを比較する方法

この方法は、異なる試料について特定の位置における特定のバリアントを比較するために使用することができ、２つという少ないデータ点に適用することができる。

【００９０】

図７は、参照試料と試験試料の配列リード上の特定の位置１１２における特定のバリアントＣ＞Ｔを示す。図７に示されるように、参照配列の位置１１２における参照アレルはＣであり、そして参照試料の位置１１２における配列リードはほとんどＣであるが、シークエンシングエラーに起因するバリアントＣ＞Ｔを有することがある。試験試料について、位置１１２における配列リードは、低いバリアント頻度に起因してＣであり、真の変異に起因していくつかのＴであり、及びシークエンシングエラーに起因していくつかのＴでもよい。

【００９１】

参照試料は、理論的には真の変異は無いが、図７に示すようにシークエンシングエラーが小さいバリアント数を引き起こす可能性がある。特定の位置における特定のバリアントに関する参照試料と試験試料のバリアント数及び参照試料と試験試料の野生型数は、配列リードに基づいて決定することができ、以下の表３に示される表に入れられる。試験試料及び参照試料のカウントデータは、試験試料のバリアント頻度が、同一の位置における参照試料のバリアント頻度よりも有意に大きいかどうかを決定するために使用することができる。

【００９２】

表３において、ａ１は参照試料中の特定の位置における特定のバリアントの数であり、ｎ１は参照試料の配列リードの深度であり、ｗ１＝ｎ１－ａ１は参照試料中の特定の位置における野生型数を示す。ａ２、ｎ２、及びｗ２は、試験試料に関する対応するバリアン

ト数、深度、及び野生型数である。表３はまた、行の合計ａ＝ａ１＋ａ２、ｗ＝ｗ１＋ｗ２、及び総数ｎ＝ｎ１＋ｎ２を列記する。

【００９３】

【表３】

カウントテーブル	参照試料	試験試料	行の合計
バリアント数	ａ１	ａ２	ａ
野生型数	ｗ１	ｗ２	ｗ
総数（深度）	ｎ１	ｎ２	ｎ

表３．参照試料と試験試料のカウントテーブル

【００９４】

（ａ１，ｗ１）と（ａ２，ｗ２）の比に有意差があるかどうかを試験するために、いくつかの方法がある。いくつかの実施態様において、ｎ１及びｎ２はウルトラディープシークエンシングにおいて非常に大きくなる可能性があるため、片側カイ二乗検定を用いることが好ましい。片側カイ二乗検定では、比率ｆ１＝ａ１／ｎ１とｆ２＝ａ２／ｎ２が最初に計算される。もしｆ２≦ｆ１の場合、すなわち、試験試料の割合が参照試料の割合より大きくない場合（偽陽性であることが知られている）、エラー率ｐ＝０．６３に対応する２などの非常に小さいクオリティスコアを設定することができ、更なる分析を必要としない。しかし、ｆ２＞ｆ１の場合は、カイ二乗統計値（chi-squared statistic)は次のように計算することができる：

【数５】

$$\chi^2 = n \times (a1 \times w2 - a2 \times w1)^2 / (n1 \times n2 \times a \times w)$$

【００９５】

片側バリアントコーリングエラー確率ｐ値は、ｐ＝０．５×（１－ｐｃｈｉｓｑ（χ^2，ｄ））として計算することができ、ここでｐｃｈｉｓｑは、自由度ｄを有するカイ二乗累積分布関数である。いくつかの実施態様において、自由度ｄは１である。

【００９６】

（ａ１，ｗ１）及び（ａ２，ｗ２）の比率が有意に異なるかどうかを試験するための別の方法は、大規模な試料のためのピアソン比率検定（Pearson proportion test）である。ピアソン比率検定において、２つの比率、ｐ１＿ｈａｔ＝ａ１／ｎ１とｐ２＿ｈａｔ＝ａ２／ｎ２が最初に計算される。Ｚスコアは、

【数６】

$$Z = (p2\_hat - p1\_hat) / sqrt(V)$$

【００９７】

により計算することができ、ここで、Ｖは次の２つの式のうちの少なくとも一つを用いて計算することができる。

【数７】

$$V=p1\_hat\times(1-p1\_hat)/n1+p2\_hat\times(1-p2\_hat)/n2、及び$$
$$V=p\_hat\times(1-p\_hat)\times((1/n1)+(1/n2))$$

【００９８】
ここで、ｐ＿ｈａｔ＝（ａ１＋ａ２）／（ｎ１＋ｎ２）である。次に片側ｐ値は、ｐ＝１－p_{norm}（Ｚ）として計算することができ、ここでp_{norm}は累積確率分布関数である。
【００９９】
いくつかの実施態様において、フィッシャーの正確確率検定（Fisher's exact test）を用いて、（ａ１，ｗ１）及び（ａ２，ｗ２）の比率が有意に異なるかどうかを決定することができる。フィッシャーの正確確率検定は、低幾何学的（hypogeometric）分布を用いる。フィッシャーの正確確率検定のための計算はより複雑になることがあり、大規模な試料についてオーバーフローを引き起こす可能性がある。
【０１００】
ｐ値が計算された後、対応するクオリティスコアは$Q_{LOC}=-10\times Log_{10}$（ｐ）として定義することができる。片側カイ二乗検定において、ｐは（０，０．５）の範囲内であることに留意されたい。いくつかの実施態様において、ｐが０に近い場合に、数値計算の困難さを回避するために、$Q_{LOC}=-10\times Log_{10}$（ｍａｘ（ｐ，ｍｉｎＰ））が使用され、ここでｍｉｎＰは任意の適切な値、例えば10^{-13}に設定することができ、これは最大クオリティスコアを１３０に設定することと同等である。
【０１０１】
Ｂ．特定の位置における特定のバリアントを検出するための参照試料を選択する方法
複数の試料のシークエンシングラン中の特定の位置における特定のバリアントに関する参照カウントを設定するために、様々な方法を使用することができる。一つの方法は、特定の位置における特定のバリアントの最も低いバリアント頻度と、最小値ｍｉｎＤ以上の深度を有する、同一のシークエンシングランにおいて、２つの試料のバリアント数の合計と深度の合計を使用する。いくつかの実施態様において、ｍｉｎＤは３０００に設定することができる。いくつかの実施態様において、参照割合がｆ０（これは、例えば０．０１又は１％に設定してもよい）より大きい時、全ての試料が特定の位置における特定のバリアントに関して高バリアント頻度を有するまれな可能性を回避するために、ａ１はｆ０×ｎ１に設定される。すなわち、使用されるａ１値は、実際のａ１値か又はｆ０×ｎ１のいずれか小さい方である。この方法では、野生型試料にバリアントが混入している場合、バリアント混入（variant contamination)を有する野生型試料は高バリアント頻度を示し、したがって特定のバリアントのための参照試料として選択されることはないであろう；従って、他の試料のクオリティスコアQ_{LOC}は通常は影響を受けない。いくつかの複雑な変異は、複数の単純な変異で構成されている。このような状況では、Q_{LOC}は、複雑な変異のすべての単純な変異成分のクオリティスコアQ_{LOC}の中央値として定義することができる。
【０１０２】
既知の野生型試料はまた、参照試料として使用することもできる。しかし野生型試料にバリアントが混入している場合は、他の試料のクオリティスコアQ_{LOC}は小さくてもよい。
【０１０３】
Ｃ．特定の位置における特定のバリアントを検出するために試験試料を参照試料と比較することによるデータ解析
図８は、特定の位置における特定のバリアントを分類するために試験試料を１又は２以上の参照試料と比較することによる、バリアントコーリングの方法８００を示す。他の方

法と同様に、実施態様は、記載された操作の全て又は一部を含むことができ、いくつかの操作は追加の操作又はサブ操作を含むことができる。

【０１０４】

ブロック８１０において、単一のシークエンシングラン中の１又は２以上の試料からのＤＮＡセグメント中の標的領域を標的化する配列リードが受け取られる。配列リードデータは受け取られ、読み取り可能な任意のフォーマットで記憶され、コンピュータにより解析することができる。いくつかの実施態様において、低クオリティリード又はアダプター配列を除去するために、配列リードデータの前処理が行うことができる。いくつかの実施態様において、バーコード又はＭＩＤは除去してもよく、同一の試料からの配列リードは標識又はグループ化することができる。

【０１０５】

ブロック８２０において、配列リードは、方法１００のブロック１７０に記載されたように、参照配列の標的領域にアラインメントされる。

【０１０６】

ブロック８３０において、アラインメントされた配列リード上の特定の配列位置における特定のバリアントのバリアントアレルは、アラインメントされた配列リードを参照配列と比較することにより、すべての試料について同定することができる。当業者に知られているように、任意の適切なアラインメント技術を使用することができる。

【０１０７】

ブロック８４０において、全ての試料について特定の配列位置における特定のバリアントに関するバリアント数とリード数を決定することができる。バリアント数は、試料の異なる配列リード上の特定の位置における同じバリアントアレルの、例えばＡ＞Ｃバリアントの Ｃの、合計数である。リード数は、試料の特定の位置のリードの総数である。

【０１０８】

ブロック８５０において、少なくとも１つの試料が参照試料として選択される。上記したように、いくつかの実施態様において、既知の野生型試料を参照試料として使用することができる。いくつかの実施態様において、シークエンシングラン中で最小バリアント頻度を有する２つの試料を参照試料として使用することができる。このような実施態様において、最小バリアント頻度を有する２つの試料のバリアント数の合計とリード数の合計は、計算において参照試料のバリアント数ａ１及びリード数ｎ１として使用することができる。

【０１０９】

ブロック８６０において、セクションＩＶ（Ａ）で上記した方法を用いて、試験試料と参照試料のための特定の配列位置における特定のバリアントのバリアント数とリード数を比較して、確率値を決定される。確率値は、カイ二乗値、累積確率分布値、ｐ値、Ｚ値、及びクオリティスコアの１又は２以上であってもよい。

【０１１０】

ブロック８７０において、試験試料上の特定の位置における特定のバリアントが真の陽性であるかどうかを決定するために、確率値と閾値とに基づいてバリアントコールが行われる。いくつかの実施態様において、閾値は単一の値であってもよい。いくつかの実施態様において、閾値は例えばバリアント頻度の関数であってもよい。いくつかの実施態様において、閾値はトレーニングデータセットに基づいて、例えばサポートベクターマシン（ＳＶＭ）などのマシン学習アルゴリズムを用いて決定することができる。いくつかの実施態様において、閾値は、異なるシークエンシングランから得られたトレーニングデータに基づいて決定することができる。

【０１１１】

Ｄ．例

以下の例は、特定の位置における特定のバリアントを検出するための、試験試料を参照試料と比較することによるバリアントコーリングの結果を示す。

【０１１２】

図９は、ＳＶＭによって決定されるセパレーターラインを有するエクソン２０のＥＧＦＲ　Ｔ７９０Ｍのトレーニングデータと試験データに関する局所化されたバリアントコーリングクオリティスコアQ_{LOC}を示す。図９から、ＳＶＭにより決定されるセパレータが、真陽性としてＴ７９０Ｍのコーリング０．１％に低下されても、野生型の試験データの誤分類が存在しないことがわかる。更に、例えばｆ≧０．１％又はQ_{LOC}≧１８の単一の閾値は、Ｔ７９０Ｍバリアントの良好な判断ポイントとすることができる。

【０１１３】

図１０は、ＳＶＭによって決定されるセパレーターラインを有するエクソン２１のＥＧＦＲ　Ｌ８５８Ｒのトレーニングデータと試験データに関する局所化されたバリアントコーリングクオリティスコアQ_{LOC}を示す。０．１％のバリアント頻度を有するものを含むすべての試験データが、正しく分類されていることがわかる。更に例えばQ_{LOC}≧１８の単一の閾値は、Ｌ８５８Ｒバリアントの良好な判断ポイントとすることができる。

【０１１４】

図１１は、ＳＶＭによって決定されるセパレーターラインを有するエクソン１９のＥＧＦＲ　１５塩基欠失２２３５＿２２４９ｄｅｌ１５のトレーニングデータと試験データに関するバリアントコーリングクオリティスコアQ_{LOC}を示す。０．１％のバリアント頻度を有するものを含むすべての試験データが、正しく分類されていることがわかる。例えばQ_{LOC}≧１８又は２０の単一の閾値は、ＳＶＭを使用することなく、真の陽性と偽陽性を分離するように設定することができる。

【０１１５】

図９～１１はまた、局所化されたバリアントコーリングスコアQ_{LOC}が、モデルに基づくバリアントコーリングスコアQ_{AMP}より、真陽性と偽陽性の間により広いマージンを有することを示す。

【０１１６】

Ｖ．単純化されたクオリティスコアの推定

いくつかの適用において、すべてのバリアントについてｐ値とクオリティスコアを直接計算することは、時間がかかる。いくつかの実施態様において、クオリティスコアのみを整数として報告する必要があるため、Q_{LOC}とQ_{AMP}の値を離散化することができる。例えばｆ２≦ｆ１である場合、クオリティスコアは２に設定することができる；ｆ２＞ｆ１である場合、クオリティスコアは３、４、・・・、又はｍａｘＱに設定することができ、これは、例えばいくつかの実施態様において１３０に設定することができる。

【０１１７】

いくつかの実施態様において、クオリティスコアは、例えばχ^2値q_{chisq}又は正規クオンタイル値 (normal quantile value)q_{norm}、及び図１２に示されるルックアップテーブルを用いて決定することができる。図１２において．Ｑ＝３．５、４．５、．．．、１２９．５についてχ^2値とq_{norm}値が計算され、ルックアップテーブルに示される。すなわち、二分探索アルゴリズムなどの探索アルゴリズムは、χ^2値又はq_{norm}値に基づいて、３、４、．．．、１３０の最良近似整数値を決定するために使用することができる。

【０１１８】

ＶＩ．ゼロイベントを回避するために必要な試料量

血液検査の一つの実用的な問題は、低頻度変異を検出することができるように、バリアントを検出するために十分なｇＤＮＡの量を決定することである。本発明のいくつかの実施態様において、ゼロイベント検出の確率を用いて、必要な試料の量を推定することができる。

【０１１９】

６．０２２×１０^{23}／モルのアボガドロ定数、塩基対当たり６５０ダルトン（ｇ／ｍｏｌ）の重量平均分子量、及びヒトゲノムあたり３．０９６×１０^{9}塩基対に基づいて、１ナノグラム（ｎｇ）のヒトｇＤＮＡは、６．０２２×１０^{23}／（６５０×３．０９６×１０^{9}×１０^{9}）＝３００分子を含有すると計算される。

【０１２０】

変異を検出するために必要とされるｇＤＮＡの量は、変異頻度に依存し、ゼロイベントを回避する統計的問題を解決することによって決定することができる。例えば、Lachin, Biostatistical Methods: The Assessment of Relative Risks, p.19, Wiley (2000)を参照されたい。血液試料中の変異体コピー数がＢであり、ＤＮＡコピーの総数がＮであり、変異確率がｐ＝Ｂ／Ｎであると仮定する。二項分布に従うと、ランダム試験において変異体コピーが得られない確率は（１－ｐ）であり、Ｎ回のランダム試験において変異体コピーが得られない確率は$(1-p)^N$である。従って以下の不等式を設定することができる。

【数８】

$$(1-p)^N \leqq \alpha$$

【０１２１】

ここで、αは、変異が検出されない最大許容確率（最大許容失敗率）であり、１－αは、上側信頼限界である。したがって、試料サイズＮは、以下の不等式を解くことによって推定することができる。

【数９】

$$N \geqq \ln(\alpha) / \ln(1-p)$$

【０１２２】

ｐ＜＜１である稀な変異の場合、推定はテイラー展開

【数１０】

$$\ln(1-p) \fallingdotseq -p$$

【０１２３】

を用いて単純化することができ、そして試料サイズの推定値は以下の通りとなる。

【数１１】

$$N \geqq -\ln(a) / p$$

【０１２４】

－ｌｎ（０．０５）＝２．９９５７、そして－ｌｎ（０．００５）＝５．２９８３であるため、３／ｐ又は５．３／ｐは、それぞれ０．９５と０．９９５の上側信頼限界を持つ稀な変異の試料サイズＮを推定するために使用することができる。

【０１２５】

表４は、最大許容失敗率αが０．０５と０．００５を有する少なくとも一つの変異体コピーを含むのに必要なｇＤＮＡ分子の推定数を示す。例えば、試料中で少なくとも一つの変異体コピーを得るための９５％の上側信頼水準（α＝０．０５）を有する０．１％（ｐ＝０．００１）の変異を検出するためには、２９９５のｇＤＮＡコピーが必要であり、これは約１０ｎｇのｇＤＮＡ分子と同等である。

【０１２６】

【表４】

p	α=0.05		α=0.005	
	コピー数	ng	コピー数	ng
0.01	300	1.0	530	1.8
0.005	600	2.0	1060	3.5
0.002	1500	5.0	2650	8.8
0.001	2995	10.0	5300	17.7

表４．希少変異検出のためのｇＤＮＡ分子のコピー数と重量の推定

【０１２７】

ＶＩＩ．適用と検証

セクションＩＩＩ及びＩＶにおいて上記した方法は、判定基準として使用されるバリアント頻度の閾値を決定する補助となり得る。この方法は、十分な入力（input）ＤＮＡ量で０．１～０．３％の頻度で、置換をうまく検出することができる。偽陽性率は変異の状態と位置に依存するため、特定の位置における特定の置換について、０．０３％という低いバリアント頻度を有するバリアントを正しく検出することができる。

【０１２８】

適度なサイズの挿入、欠失、及び１５塩基の欠失などの複雑な変異について、シークエンシングにおいてこれらのタイプの変異をランダムに発生させることは困難であり、誤差の主原因は他の試料からのキャリーオーバー混入（carry-over contamination)である。すなわち、ラン間で充分に確立された洗浄プロトコルを用いることにより、０．００２５％と低いバリアント頻度を有するこれらのタイプのバリアントを正確に検出することができる。

【０１２９】

Illumina MiSeq Reporter（ＭＳＲ）は、本開示に記載された方法により検出される低頻度バリアントを確認するための非標準的な方法で使用することができる。ＭＳＲは、組み込みポアソン（Poisson）モデルを用いる体細胞バリアントコーラーを使用して、低頻度バリアントを報告する。ＭＳＲが報告する最低頻度は、深度に依存する。ポアソンモデルに基づくと、ＭＳＲ体細胞バリアントコーラーが報告する最低のバリアント数と頻度を計算することができ、表１に示されるようにデフォルト設定される。例えば深度が１００である場合、最低の報告される頻度は５％である；深度が５００である場合、最低の報告される頻度は１．３６％である；深度が更に大きい場合、最低の報告される頻度は上記の１％に近くなる。

【０１３０】

いくつかの実施態様において、既知のバリアントを含む試料を参照試料として用いるＭＳＲは、ＭＳＲが、野生型アレルを参照試料の「バリアントアレル」として報告し、実際のバリアントアレルを「野生型アレル」として報告するように実行することができる。こうして、本開示に記載された方法を用いるバリアントコーリングを検証することができる。ＭＳＲのこの非標準的な使用は、いくつかの欠点を有する。第一に、これは既知のバリアントを確認するためにのみ使用することができる。第二に、ＭＳＲが報告するバリアントコーリングクオリティスコアは、実際のバリアントのためというより野生型のものである。第三に、複数の重複する既知のバリアントが存在する場合、この方法を使用することが面倒又は困難になる。しかし、上記の欠点を考慮した後、ＭＳＲは、既知のバリアントのための検証ツールとして使用することができる。これは、ゲノム全体が参照配列として使用される場合、ＭＳＲマッピング／アラインメントソフトウェアが、マッピングされていないリードとして報告する適度なサイズのインデルのために特に有用である。

【０１３１】

ＶＩＩＩ．コンピュータシステムとシークエンシングシステム

本明細書に記載の任意のコンピュータシステムは、任意の適切な数のサブシステムを利用することができる。そのようなサブシステムの例は、図１３でコンピュータ装置１３００内に示されている。いくつかの実施態様において、コンピュータシステムは単一のコンピュータ装置を含み、ここでサブシステムはコンピュータ装置の構成要素とすることができる。他の実施態様においてコンピュータシステムは、内部構成要素を含む、それぞれがサブシステムである複数のコンピュータ装置を含むことができる。コンピュータシステムは、デスクトップコンピュータ及びラップトップコンピュータ、タブレット、携帯電話、及び他のモバイルデバイスを含むことができる。

【０１３２】

図１３に示されるサブシステムは、システムバス１３０５を介して相互接続される。プリンタ１３４０、キーボード１３７０、記憶装置１３８０、モニタ１３５２（これはディスプレイアダプタ１３５０に接続されている）などの追加のサブシステムが示されている。周辺機器及び入力／出力（Ｉ／Ｏ）装置（これらは、Ｉ／Ｏコントローラ１３１０に接続されている）は、当技術分野で任意の数の公知の手段、例えば、シリアルポート１３６０などによりコンピュータシステムに接続することができる。例えばシリアルポート１３６０又は外部インタフェース１３９０（例えば、イーサネット（登録商標）、Ｗｉ－Ｆｉなど）は、コンピュータシステム１３００をインターネットなどの広域ネットワーク、マウス入力装置、又はスキャナに接続するために使用することができる。システムバス１３０５を介する相互接続は、中央プロセッサ１３３０が各サブシステムと通信し、システムメモリ１３２０又は記憶装置１３８０（例えば、固定ディスク）からの命令の実行、ならびにサブシステム間の情報の交換を制御することを可能にする。システムメモリ１３２０及び／又は記憶装置１３８０は、コンピュータ読み取り可能媒体を具体化することができる。本明細書に記載の任意の値は、一つの構成要素から別の構成要素に出力することができ、ユーザに出力することができる。

【０１３３】

コンピュータシステムは、例えば、外部インタフェース１３９０又は内部インタフェースによって接続される、複数の同一の構成要素又はサブシステムを含むことができる。いくつかの実施態様において、コンピュータシステム、サブシステム、又は装置は、ネットワーク上で通信することができる。このような例では一つのコンピュータはクライアントとして、別のコンピュータはサーバと見なすことができ、それぞれは、同じコンピュータシステムの一部であることができる。クライアントとサーバは、それぞれ複数のシステム、サブシステム、又は構成要素を含むことができる。

【０１３４】

なお、本発明の任意の実施態様は、ハードウェア（例えば、特定用途向け集積回路又はフィールドプログラマブルゲートアレイ）を使用するか、及び／又はコンピュータソフトウェアを使用して、一般にプログラム可能なプロセッサを用いてモジュラー又は統合的方法で、制御ロジックの形態で実施することができることを理解すべきである。本明細書において、プロセッサは、同一の集積チップ上でシングルコアプロセッサ、マルチコアプロセッサ、又は単一の回路基板上の又はネットワーク化された複数の処理ユニットを含む。本明細書で提供される開示及び教示に基づき、ハードウェア及びハードウェアとソフトウェアの組み合わせを用いて、本発明の実施態様を実施する他の手法及び／又は方法を、当業者は周知しており理解しているであろう。

【０１３５】

本出願に記載されている任意のソフトウェア構成要素又は機能は、例えばＪａｖａ（登録商標）、Ｃ、Ｃ＋＋、Ｃ＃、Ｏｂｊｅｃｔｉｖｅ－Ｃ、Ｓｗｉｆｔなどのコンピュータ言語、又はＰｅｒｌやＰｙｔｈｏｎなどのスクリプト言語などの任意の適切なコンピュータ言語を用いて、例えば、従来型又はオブジェクト指向技術を用いて、プロセッサによって実行されるソフトウェアコードとして実行することができる。ソフトウェアコードは、記憶及び／又は送信のためのコンピュータ読み取り可能な媒体上の、一連の支持又は命令

として記憶することができる。適切な非一時的コンピュータ可読媒体としては、ランダムアクセスメモリ（ＲＡＭ）、読み出し専用メモリ（ＲＯＭ）、磁気媒体、例えばハードドライブ又はフロッピー（登録商標）ディスク、又は光学媒体、例えばコンパクトディスク（ＣＤ）若しくはＤＶＤ（デジタル多用途ディスク）、フラッシュメモリなどが挙げることができる。コンピュータ可読媒体は、このような記憶又は送信装置の任意の組み合わせであってもよい。

【０１３６】

このようなプログラムは、コード化され、インターネットを含む種々のプロトコルに適合する有線、光、及び／又は無線ネットワークを介して送信するために適合されたキャリア信号を用いて送信することができる。このように、本発明の実施態様に係るコンピュータ可読媒体は、そのようなプログラムでエンコードされたデータ信号を用いて作成することができる。プログラムコードでコード化されたコンピュータ可読媒体は、互換性のある装置と共にパッケージされるか、又は他の装置とは別に（例えば、インターネットダウンロードを介して）提供されてもよい。任意のそのようなコンピュータ可読媒体は、単一のコンピュータ製品（例えばハードドライブ、ＣＤ、又はコンピュータシステム全体）上に又はその中に常駐することができ、システム又はネットワーク内の異なるコンピュータ製品上又はその中に存在してもよい。コンピュータシステムは、ユーザに本明細書に記載の結果のいずれかを提供するための、モニタ、プリンタ、又は他の適切なディスプレイを含むことができる。

【０１３７】

本明細書に記載される任意の方法は、全体的に又は部分的に、工程を実行するように構成することができる１又は２以上のプロセッサを含むコンピュータシステムを用いて行うことができる。すなわち実施態様は、潜在的に各工程又は工程のそれぞれの群を実行する異なる構成要素を用いて、本明細書に記載された任意の方法の工程を実行するように構成されたコンピュータシステムに関する。番号付きの工程として提示されているが、本明細書の方法の工程は、同時に又は異なる順序で行うことができる。さらに、これらの工程の一部は、他の方法の他の工程の一部とともに使用することができる。また、工程の全て又は一部が任意であってよい。また、任意の方法の任意の工程は、これらの工程を実行するためのモジュール、回路、又は他の手段を用いて行うことができる。

【０１３８】

ある態様において本発明はまた、シークエンシングシステムも提供する。典型的なシークエンシングシステムは図１４に表示される。図１４に示されるシステムは、シークエンシング装置内に位置することができるシークエンシング分析モジュールと、コンピュータシステムの一部であるインテリジェンスモジュールとを含む。データセット（シークエンシングデータセット）は、ネットワーク接続又は直接接続を介して、分析モジュールからインテリジェンスモジュールに、又はその逆に、転送される。データセットは、例えば図４又は８に示すようにフローチャートに従って処理することができる。フローチャートに提供された工程は、コンピュータシステムのハードウェアに格納されたソフトウェアによって、例えば図１５Ａ及び１５Ｂに記載されたフローチャートに従って便利に実行することができる。図１５Ａを参照して、コンピュータシステム（１１００）は、例えば複数の配列リードから得られるデータを受け取るための受け取り手段（１１１０）、前記複数の配列リードを参照配列の標的領域にアラインメントするためのアラインメント手段（１１２０）、第一の試料の配列リードに基づいて、標的領域の第一の位置において、前記参照配列の第一の位置の参照アレルとは異なる第一のアレルを有する第一の候補バリアントを同定するための同定手段（１１３０）、前記参照配列の第一の位置にアラインメントする第一の試料の配列リードに基づいて、第一の位置における第一のアレルに関して第一のバリアント頻度を決定するための決定手段（１１４０）、複数のバリアントクラスから選択される第一のバリアントクラスに対応する第一の候補バリアントを同定するための同定手段（１１５０）であって、ここで前記複数のバリアントクラスの各バリアントクラスは、異なるタイプのバリアントに対応する上記手段、前記参照アレルを有する前記参照配列の

標的領域中の１セットの第二の位置を同定するための同定手段（１１６０）であって、前記１又は２以上の試料中の少なくとも５０％の他の位置は、第一のアレルに関して偽陽性を示し、そして前記１セットの第二の位置は前記第一の位置を含む上記手段を含み、前記１セットの第二の位置の各々において、かつ前記１又は２以上の試料の各々に関して、前記参照配列の第二の位置にアラインメントする試料の配列リードに基づいて、第一のアレルの第二のバリアント頻度を決定するための決定手段（１１７０）であって、ここで前記第二のバリアント頻度は統計分布を形成する上記手段、前記第一のバリアント頻度を前記統計分布の統計値と比較して、前記統計分布の統計値に対する第一のバリアント頻度の確率値を決定するための比較手段（１１８０）、そして、前記第一のアレルに関して、第一の試料において第一の候補バリアントが真陽性であるか否かを決定する一部として、前記確率値を閾値と比較するための比較手段（１１９０）であって、ここで前記閾値は前記第一のアレルに関して偽陽性と真陽性とを区別する上記手段、を含むことができる。図１５Ｂを参照して、コンピュータシステム（２１００）は、例えば複数の配列リードから得られるデータを受け取るための受け取り手段（２１１０）、前記複数の配列リードを参照配列の標的領域にアラインメントするためのアラインメント手段（２１２０）、第一の位置における各試料のアラインメントされた配列リードに基づいて、少なくとも２つの試料の各試料中の第一の位置に、前記参照配列の第一の位置における参照アレルとは異なる第一の位置における第一のアレルが存在するか否かと同定するための同定手段（２１３０）、前記少なくとも２つの試料の各試料の、第一の位置における第一のアレルのバリアント数と、第一の位置における参照アレルの野生型数を決定するための決定手段（２１４０）、前記少なくとも２つの試料から少なくとも１つの試料を参照試料として選択するための選択手段（２１５０）、前記第一の試料に関する第一の位置における第一のアレルの第一のバリアント数と第一の位置における参照アレルの第一の野生型数とを、前記参照試料に関する第一の位置における第一のアレルの第二のバリアント数と第一の位置における参照アレルの第二の野生型数とを比較して、第一の試料に関する第一の位置において第一のアレルを有するバリアントの確率値を決定するための比較手段（２１６０）、そして前記第一の試料中の第一の位置における第一のアレルが、第一のアレルについて真陽性であるかどうかを決定する一部として、確率値を閾値と比較するための比較手段（２１７０）であって、ここで前記閾値は、前記第一の位置における第一のアレルに関して偽陽性と真陽性とを区別する上記手段、をさらに備えることができる。

【０１３９】

ある実施態様において、システムはまた、結果をコンピュータスクリーン上に表示するための表示手段を含むこともできる。図１４は、シークエンシング装置とコンピュータシステムとの間の相互作用を示す。システムは、シークエンシング装置内に位置することができる配列分析モジュールと、コンピュータシステムの一部であるインテリジェンスモジュールとを含む。データセット（シークエンシングデータセット）は、ネットワーク接続又は直接接続を介して、分析モジュールからインテリジェンスモジュールに又はその逆に、転送される。データセットは、プロセッサ上で作動しインテリジェンスモジュールの記憶装置に記憶されるコンピュータコードにより、図１５Ａ又は１５Ｂに従って処理することができ、処理後、分析モジュールの記憶装置に転送されて戻され、ここで修正されたデータは表示装置上に表示することができる。いくつかの実施態様において、インテリジェンスモジュールはまたシークエンシング装置で実行することができる。

【０１４０】

特定の実施態様の具体的な詳細は、本発明の実施態様の精神及び範囲から逸脱することなく、任意の適切な方法で組み合わせることができる。しかし、本発明の他の実施態様は、個々の態様、又はこれらの個々の態様の特定の組み合わせに関連する特定の実施態様に関してもよい。

【０１４１】

本発明の例示的な実施態様の上記記載は、例示と説明のために提示されている。これは網羅的であること又は記載した正確な形態に本発明を限定することを意図するものでもな

く、多くの修正及び変更が上記の教示に照らして可能である。

【０１４２】

「ａ」、「ａｎ」又は「ｔｈｅ」の列挙は、特に別の指定がなければ、「１又は２以上」を意味することが意図される。特に別の指定がなければ、「又は」の使用は「含んで又は」を意味し、「含まなくて又は」ではないことを意味する。

【図１】

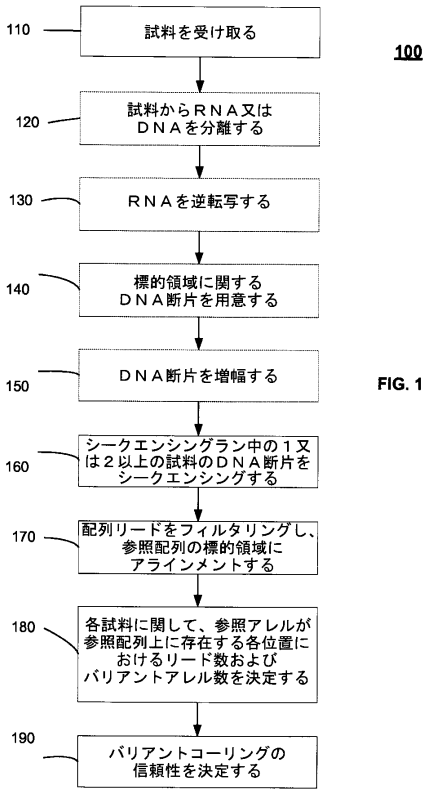


FIG. 1

【図２】

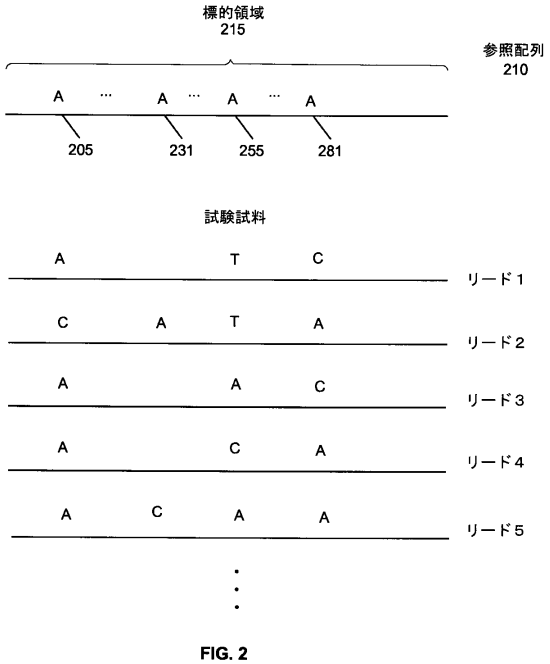


FIG. 2

【図３Ａ】

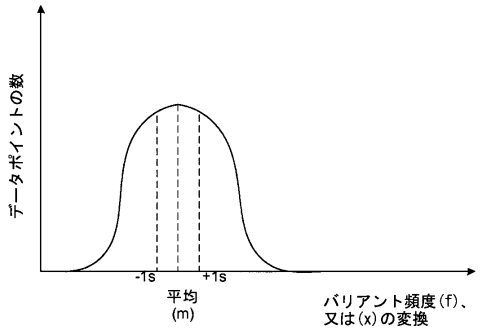


FIG. 3A

【図３Ｂ】

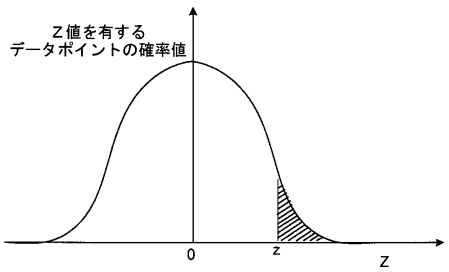


FIG. 3B

【図３Ｃ】

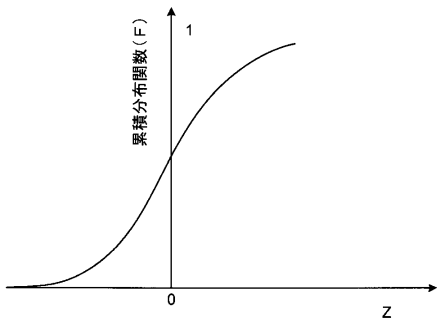


FIG. 3C

【図３Ｄ】

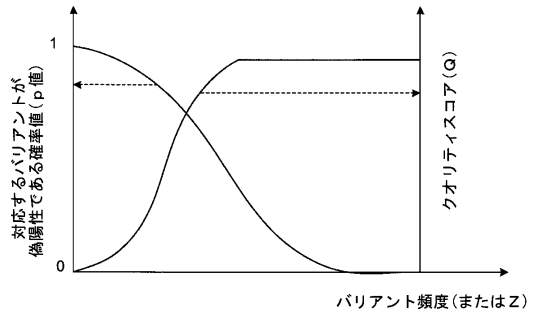


FIG. 3D

【図４】

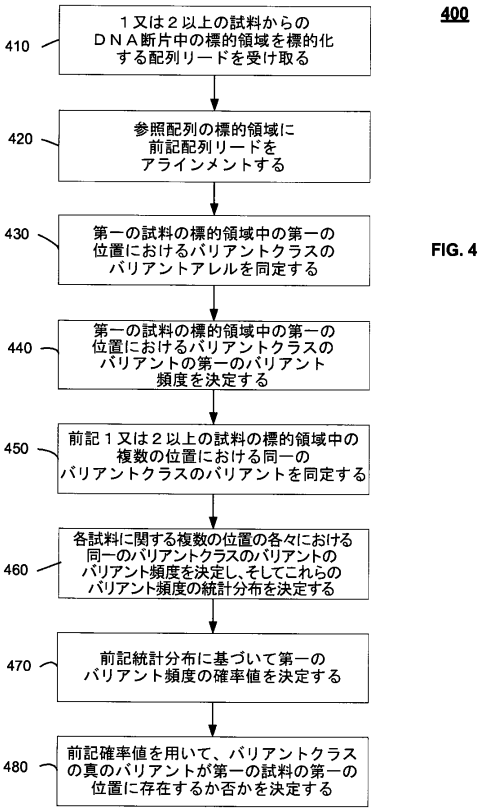


FIG. 4

【図５】

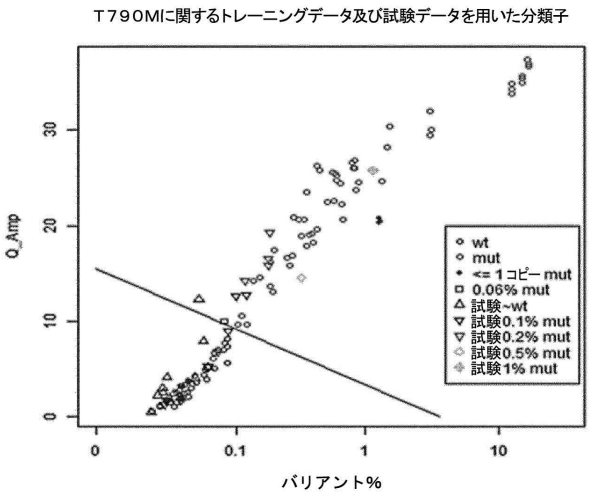


FIG. 5

【図６】

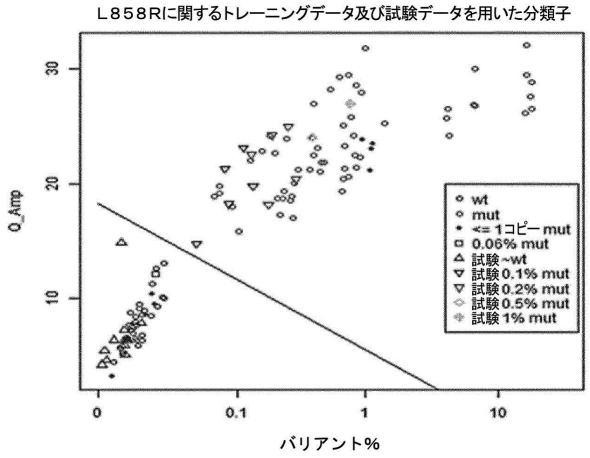


FIG. 6

【図７】

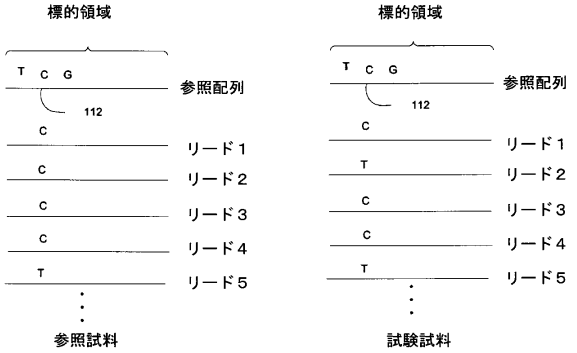


FIG. 7

【図８】

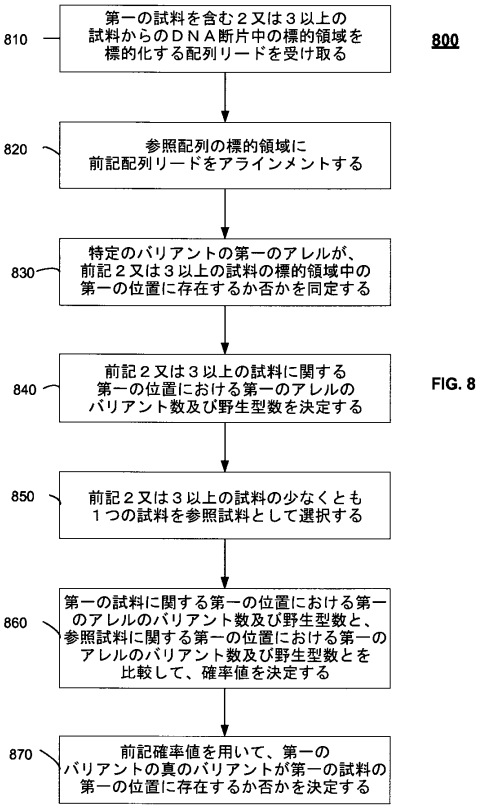


FIG. 8

【図９】

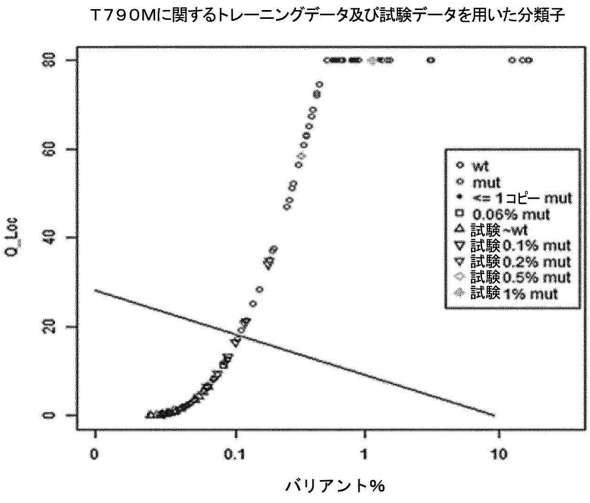


FIG. 9

【図１０】

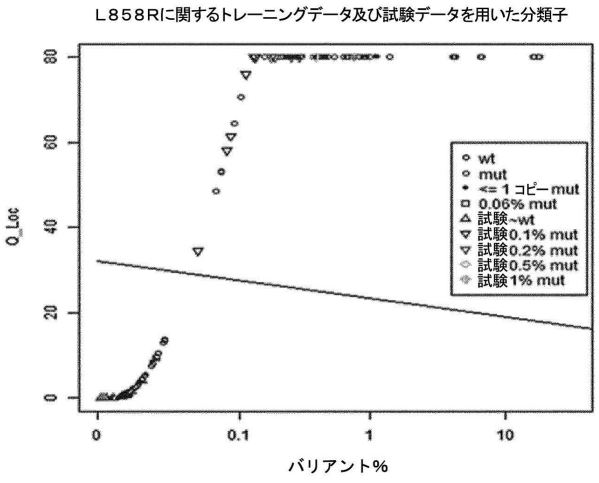


FIG. 10

【図１１】

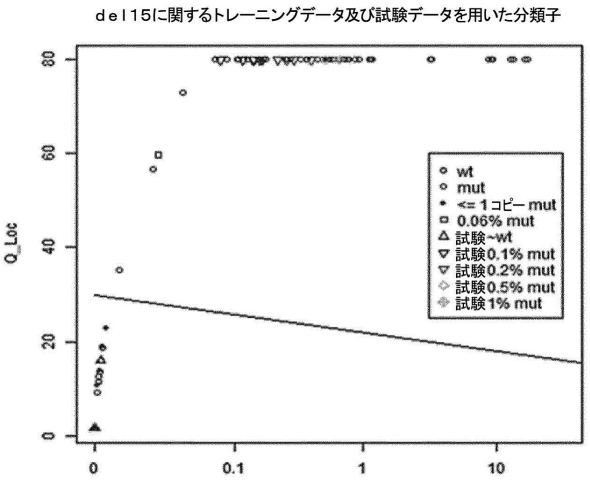


FIG. 11

【図１２】

FIG.12

Q	qnorm	qchisq	Q	qnorm	qchisq	Q	qnorm	qchisq	Q	qnorm	qchisq
3.5	0.1340	0.0180	35.5	3.4485	11.8923	67.5	5.0913	25.9211	99.5	6.3436	40.2417
4.5	0.3724	0.1387	36.5	3.5102	12.3216	68.5	5.1348	26.3657	100.5	6.3790	40.6916
5.5	0.5774	0.3334	37.5	3.5710	12.7519	69.5	5.1779	26.8106	101.5	6.4142	41.1416
6.5	0.7592	0.5764	38.5	3.6308	13.1830	70.5	5.2207	27.2557	102.5	6.4492	41.5917
7.5	0.9237	0.8532	39.5	3.6898	13.6148	71.5	5.2632	27.7010	103.5	6.4840	42.0419
8.5	1.0747	1.1550	40.5	3.7480	14.0475	72.5	5.3053	28.1465	104.5	6.5186	42.4923
9.5	1.2149	1.4760	41.5	3.8054	14.4809	73.5	5.3472	28.5923	105.5	6.5531	42.9427
10.5	1.3462	1.8122	42.5	3.8620	14.9149	74.5	5.3887	29.0383	106.5	6.5874	43.3932
11.5	1.4699	2.1606	43.5	3.9179	15.3497	75.5	5.4300	29.4844	107.5	6.6215	43.8438
12.5	1.5872	2.5192	44.5	3.9730	15.7850	76.5	5.4709	29.9308	108.5	6.6554	44.2945
13.5	1.6989	2.8863	45.5	4.0275	16.2209	77.5	5.5116	30.3774	109.5	6.6892	44.7453
14.5	1.8057	3.2606	46.5	4.0814	16.6574	78.5	5.5519	30.8241	110.5	6.7228	45.1962
15.5	1.9082	3.6412	47.5	4.1345	17.0945	79.5	5.5921	31.2710	111.5	6.7563	45.6472
16.5	2.0068	4.0271	48.5	4.1871	17.5320	80.5	5.6319	31.7181	112.5	6.7896	46.0983
17.5	2.1019	4.4178	49.5	4.2391	17.9701	81.5	5.6715	32.1654	113.5	6.8227	46.5494
18.5	2.1938	4.8127	50.5	4.2905	18.4086	82.5	5.7108	32.6129	114.5	6.8557	47.0007
19.5	2.2828	5.2113	51.5	4.3414	18.8476	83.5	5.7498	33.0605	115.5	6.8885	47.4520
20.5	2.3692	5.6133	52.5	4.3917	19.2870	84.5	5.7886	33.5082	116.5	6.9212	47.9034
21.5	2.4532	6.0182	53.5	4.4415	19.7269	85.5	5.8272	33.9562	117.5	6.9538	48.3549
22.5	2.5349	6.4259	54.5	4.4908	20.1671	86.5	5.8655	34.4043	118.5	6.9862	48.8065
23.5	2.6146	6.8360	55.5	4.5396	20.6077	87.5	5.9036	34.8525	119.5	7.0184	49.2582
24.5	2.6923	7.2484	56.5	4.5879	21.0488	88.5	5.9415	35.3009	120.5	7.0505	49.7099
25.5	2.7682	7.6629	57.5	4.6357	21.4901	89.5	5.9791	35.7494	121.5	7.0825	50.1617
26.5	2.8424	8.0793	58.5	4.6831	21.9318	90.5	6.0165	36.1980	122.5	7.1143	50.6136
27.5	2.9150	8.4974	59.5	4.7301	22.3739	91.5	6.0537	36.6468	123.5	7.1460	51.0654
28.5	2.9862	8.9171	60.5	4.7766	22.8163	92.5	6.0906	37.0958	124.5	7.1776	51.5174
29.5	3.0559	9.3384	61.5	4.8228	23.2590	93.5	6.1274	37.5448	125.5	7.2090	51.9696
30.5	3.1243	9.7610	62.5	4.8685	23.7020	94.5	6.1639	37.9940	126.5	7.2403	52.4217
31.5	3.1914	10.1850	63.5	4.9138	24.1452	95.5	6.2003	38.4433	127.5	7.2714	52.8742
32.5	3.2573	10.6102	64.5	4.9587	24.5888	96.5	6.2364	38.8927	128.5	7.3025	53.3259
33.5	3.3221	11.0365	65.5	5.0033	25.0327	97.5	6.2723	39.3423	129.5	7.3333	53.7788
34.5	3.3858	11.4639	66.5	5.0475	25.4768	98.5	6.3081	39.7919			

【図１３】

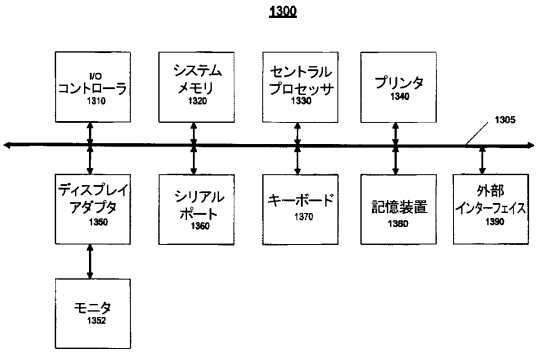


FIG. 13

【図１４】

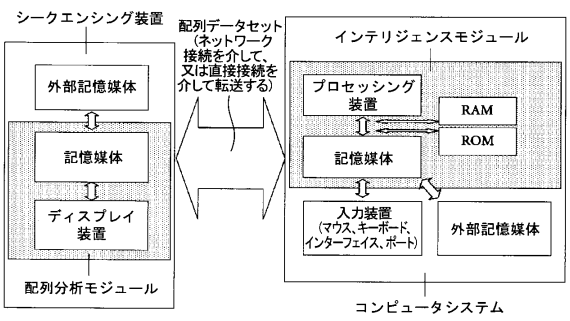
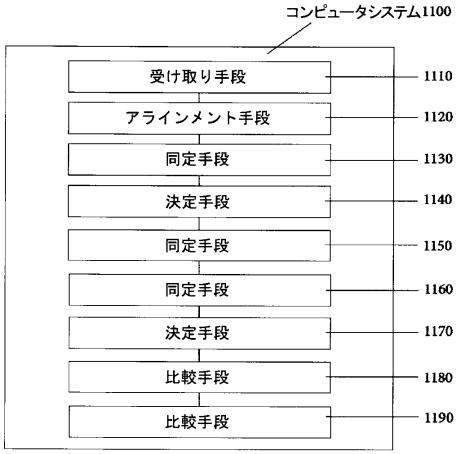


FIG. 14

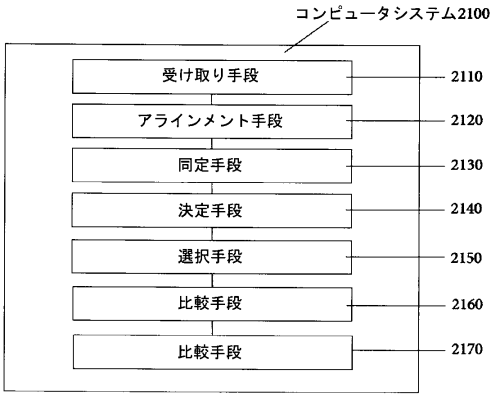
【図１５Ａ】

FIG. 15A



【図１５Ｂ】

FIG. 15B



フロントページの続き

(74)代理人 100117019
弁理士 渡辺 陽一
(74)代理人 100150810
弁理士 武居 良太郎
(74)代理人 100164563
弁理士 佐々木 貴英
(72)発明者 ウエイーミン リウ
アメリカ合衆国, カリフォルニア ９４５６８, ダブリン, シェルトン ストリート ４９２９

審査官 田付 徳雄

(56)参考文献 国際公開第２０１４／０８３０２３（WO, Ａ１）
米国特許出願公開第２０１４／０１４３１８８（ＵＳ, Ａ１）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ１６Ｂ ５／００ － ９９／００
Ｃ１２Ｑ １／６８６９