(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6533011号
(P6533011)

(45)発行日　令和1年6月19日(2019.6.19)　(24)登録日　令和1年5月31日(2019.5.31)

(51)Int.Cl.			F I		
G16B	30/00	(2019.01)	G06F	19/22	
C12Q	1/6806	(2018.01)	C12Q	1/6806	Z
C12N	15/12	(2006.01)	C12N	15/12	

請求項の数 32　(全 19 頁)

(21)出願番号　特願2018-510056(P2018-510056)
(86)(22)出願日　平成28年8月25日(2016.8.25)
(65)公表番号　特表2018-533111(P2018-533111A)
(43)公表日　平成30年11月8日(2018.11.8)
(86)国際出願番号　PCT/US2016/048768
(87)国際公開番号　WO2017/035392
(87)国際公開日　平成29年3月2日(2017.3.2)
審査請求日　平成30年4月26日(2018.4.26)
(31)優先権主張番号　62/209,858
(32)優先日　平成27年8月25日(2015.8.25)
(33)優先権主張国　米国(US)

早期審査対象出願

(73)特許権者　516001591
ナントミクス，エルエルシー
アメリカ合衆国，カリフォルニア州　９０２３２，カルバー　シティ，９９２０　ジェファーソン　ブールバード
(74)代理人　100114775
弁理士　高岡　亮一
(74)代理人　100121511
弁理士　小田　直
(74)代理人　100202751
弁理士　岩堀　明代
(74)代理人　100191086
弁理士　高橋　香元

最終頁に続く

(54)【発明の名称】高正確度変異体コールのためのシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
患者についてＨＬＡ型をインシリコ予測する方法であって、
複数の既知のおよび異なるＨＬＡ対立遺伝子の配列を含む参照配列を提供すること；
複数の患者配列リードを提供することであって、前記患者配列リードの少なくともいくつかが患者特異的ＨＬＡをコードする配列を含む、複数の患者配列リードを提供すること；
前記複数の患者配列リードを複数のｋ－ｍｅｒのそれぞれのセットへと分解すること；
前記参照配列および前記複数のｋ－ｍｅｒのそれぞれのセットを用いてｄｅ　Ｂｒｕｉｊｎグラフを作成すること；
１つの対立遺伝子についてのそれぞれの投票を全て加えることによって前記１つの対立遺伝子に対する前記複数の患者配列リードのそれぞれの投票から算出される複合マッチスコアを使用して前記既知のおよび異なるＨＬＡ対立遺伝子のそれぞれをランク付けることであって、各投票が前記既知のおよび異なるＨＬＡ対立遺伝子中の対応するセグメントにマッチするｋ－ｍｅｒを使用する、前記既知のおよび異なるＨＬＡ対立遺伝子のそれぞれをランク付けること；
最上位ＨＬＡ対立遺伝子を前記患者の第１の対立遺伝子のＨＬＡ型として識別すること；および
調節された最上位ＨＬＡ対立遺伝子を前記患者の第２の対立遺伝子のＨＬＡ型として識別するために、調節された複合マッチスコアを使用して残りの非最上位の既知のおよび異

なるＨＬＡ対立遺伝子を再びランク付けること；
を含み、
前記調節された複合マッチスコアは、第１のＨＬＡ対立遺伝子とマッチするｋ－ｍｅｒの重みを削除するのではなく減少させることによって算出される、
方法。
【請求項２】
前記参照配列が、少なくとも１％の対立遺伝子頻度を有する少なくとも１つのＨＬＡ型についての対立遺伝子を含む、請求項１に記載の方法。
【請求項３】
前記参照配列が、少なくとも１つのＨＬＡ型についての少なくとも１０の異なる対立遺伝子を含む、請求項１に記載の方法。
【請求項４】
前記参照配列が、少なくとも２つの異なるＨＬＡ型についての対立遺伝子を含む、請求項１に記載の方法。
【請求項５】
前記ＨＬＡ型が、ＨＬＡ－Ａ型、ＨＬＡ－Ｂ型、ＨＬＡ－Ｃ型、ＨＬＡ－ＤＲＢ－１型、および／またはＨＬＡ－ＤＱＢ－１型である、請求項１に記載の方法。
【請求項６】
前記複数の患者配列リードが、複数のＤＮＡシーケンシングリードとＲＮＡシーケンシングリードの少なくとも１つを含む、請求項１に記載の方法。
【請求項７】
前記患者配列リードが、染色体６ｐ２１．３に位置する、請求項１に記載の方法。
【請求項８】
前記患者配列リードが、次世代シーケンシングリードであり、且つメタデータをさらに含む、請求項１に記載の方法。
【請求項９】
前記患者配列リードが、５０塩基と２５０塩基の間の長さを有する、請求項１に記載の方法。
【請求項１０】
前記ｋ－ｍｅｒが、１０～２０の長さを有する、請求項１に記載の方法。
【請求項１１】
前記ｋ－ｍｅｒが、前記患者配列リードの長さの５％と１５％の間の長さを有する、請求項１に記載の方法。
【請求項１２】
前記投票が、患者配列リード当たりのｋ－ｍｅｒの合計数に対するマッチングｋ－ｍｅｒの割合を表す値である、請求項１に記載の方法。
【請求項１３】
最上位ＨＬＡ対立遺伝子を前記患者の第１のＨＬＡ型として識別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１４】
前記参照配列が、少なくとも１％の対立遺伝子頻度を有する少なくとも１つのＨＬＡ型についての対立遺伝子を含むか、または前記参照配列が、少なくとも１つのＨＬＡ型についての少なくとも１０の異なる対立遺伝子を含むか、または前記参照配列が、少なくとも２つの異なるＨＬＡ型についての対立遺伝子を含む、請求項１～１３のいずれか１項に記載の方法。
【請求項１５】
前記ｋ－ｍｅｒが、１０～２０の長さを有するか、または前記ｋ－ｍｅｒが、患者配列リードの長さの５％と１５％の間の長さを有する、請求項１～１４のいずれか１項に記載の方法。
【請求項１６】

前記複合マッチスコアが、前記複数の患者配列リードからのすべての投票の合計である、および／または前記投票が、患者配列リード当たりのｋ－ｍｅｒの合計数に対するマッチングｋ－ｍｅｒの割合を表す値である、請求項１～１５のいずれか１項に記載の方法。

【請求項１７】

患者についてＨＬＡ型をインシリコ予測するためのコンピュータシステムであって、

複数の既知のおよび異なるＨＬＡ対立遺伝子の配列を含む参照配列を格納する参照配列データベースと；

複数の患者配列リードを格納するもしくは提供する患者配列データソースであって、前記患者配列リードの少なくともいくつかが、患者特異的ＨＬＡをコードする配列を含む、患者配列データソースと；

（ｉ）前記複数の患者配列リードを複数のｋ－ｍｅｒのそれぞれのセットに分解する；

（ｉｉ）前記参照配列と前記複数のｋ－ｍｅｒのそれぞれのセットとを使用してｄｅＢｒｕｉｊｎグラフを作成する；

（ｉｉｉ）１つの対立遺伝子についてのそれぞれの投票を全て加えることによって前記１つの対立遺伝子に対する前記複数の患者配列リードのそれぞれの投票から算出される複合マッチスコアを使用して前記既知のおよび異なるＨＬＡ対立遺伝子のそれぞれをランク付ける；

（ｉｖ）最上位ＨＬＡ対立遺伝子を前記患者の第１の対立遺伝子のＨＬＡ型として識別する；および

（ｖ）調節された最上位ＨＬＡ対立遺伝子を前記患者の第２の対立遺伝子のＨＬＡ型として識別するために、調節された複合マッチスコアを使用して残りの非最上位の既知のおよび異なるＨＬＡ対立遺伝子を再びランク付ける；

ようにプログラムされた解析エンジンと、

を含み、

各投票が、前記既知のおよび異なるＨＬＡ対立遺伝子内の対応するセグメントとマッチするｋ－ｍｅｒを使用し、

前記調節された複合マッチスコアは、第１のＨＬＡ対立遺伝子とマッチするｋ－ｍｅｒの重みを削除するのではなく減少させることによって算出される、

コンピュータシステム。

【請求項１８】

前記参照配列が、少なくとも１％の対立遺伝子頻度を有する少なくとも１つのＨＬＡ型についての対立遺伝子を含むか、または前記参照配列が、少なくとも１つのＨＬＡ型についての少なくとも１０の異なる対立遺伝子を含むか、または前記参照配列が、少なくとも２つの異なるＨＬＡ型についての対立遺伝子を含む、請求項１７に記載のコンピュータシステム。

【請求項１９】

前記ＨＬＡ型が、ＨＬＡ－Ａ型、ＨＬＡ－Ｂ型、ＨＬＡ－Ｃ型、ＨＬＡ－ＤＲＢ－１型、および／またはＨＬＡ－ＤＱＢ－１型である、請求項１７に記載のコンピュータシステム。

【請求項２０】

前記複数の患者配列リードが、複数のＤＮＡシーケンシングリードおよびＲＮＡシーケンシングリードの少なくとも１つを含む、請求項１７に記載のコンピュータシステム。

【請求項２１】

前記患者配列リードが、染色体６ｐ２１．３に位置する、請求項１７に記載のコンピュータシステム。

【請求項２２】

前記患者配列リードが、次世代シーケンシングリードであり、且つメタデータをさらに含む、または

前記患者配列リードが、５０塩基と２５０塩基の間の長さを有する、

請求項１７に記載のコンピュータシステム。
【請求項２３】
前記ｋ－ｍｅｒが、１０～２０の長さを有するか、または前記ｋ－ｍｅｒが、前記患者配列リードの長さの５％と１５％の間の長さを有する、請求項１７に記載のコンピュータシステム。
【請求項２４】
前記投票が、患者配列リード当たりのｋ－ｍｅｒの合計数に対するマッチングｋ－ｍｅｒの割合を表す値である、請求項１７に記載のコンピュータシステム。
【請求項２５】
前記解析エンジンが、最上位ＨＬＡ対立遺伝子を前記患者の第１のＨＬＡ型として識別するようにさらにプログラムされている、請求項１７に記載のコンピュータシステム。
【請求項２６】
参照配列データベースおよび患者配列データソースが解析エンジンに情報的に連結されているコンピュータシステムに、
複数の既知のおよび異なるＨＬＡ対立遺伝子の配列を含む参照配列を前記参照配列データベースから前記解析エンジンに提供するステップと；
複数の患者配列リードを患者配列データソースから前記解析エンジンに提供するステップであって、前記患者配列リードの少なくともいくつかが、患者特異的ＨＬＡをコードする配列を含む、ステップと；
前記解析エンジンによって前記複数の患者配列リードを複数のｋ－ｍｅｒのそれぞれのセットへと分解するステップと；
前記参照配列および前記複数のｋ－ｍｅｒのそれぞれのセットを使用してｄｅ　Ｂｒｕｉｊｎグラフを前記解析エンジンによって作成するステップと；
１つの対立遺伝子についてのそれぞれの投票を全て加えることによって前記１つの対立遺伝子に対する前記複数の患者配列リードのそれぞれの投票から算出される複合マッチスコアを使用して前記解析エンジンによって前記既知のおよび異なるＨＬＡ対立遺伝子のそれぞれをランク付けるステップであって、各投票が、前記既知のおよび異なるＨＬＡ対立遺伝子内の対応するセグメントとマッチするｋ－ｍｅｒを使用する、ステップと；
最上位ＨＬＡ対立遺伝子を前記患者の第１の対立遺伝子のＨＬＡ型として識別するステップと；
調節された最上位ＨＬＡ対立遺伝子を前記患者の第２の対立遺伝子のＨＬＡ型として識別するために、調節された複合マッチスコアを使用して残りの非最上位の既知のおよび異なるＨＬＡ対立遺伝子を再びランク付けるステップであって、前記調節された複合マッチスコアは、第１のＨＬＡ対立遺伝子とマッチするｋ－ｍｅｒの重みを削除するのではなく減少させることによって算出される、ステップと；
を含む方法を実施させるためのプログラム命令を含む非一時的なコンピュータ可読媒体。
【請求項２７】
前記参照配列が、少なくとも１％の対立遺伝子頻度を有する少なくとも１つのＨＬＡ型についての対立遺伝子を含むか、または前記参照配列が、少なくとも１つのＨＬＡ型についての少なくとも１０の異なる対立遺伝子を含むか、または前記参照配列が、少なくとも２つの異なるＨＬＡ型についての対立遺伝子を含む、請求項２６に記載のコンピュータ可読媒体。
【請求項２８】
前記ＨＬＡ型がＨＬＡ－Ａ型、ＨＬＡ－Ｂ型、ＨＬＡ－Ｃ型、ＨＬＡ－ＤＲＢ－１型、および／またはＨＬＡ－ＤＱＢ－１型である、請求項２６に記載のコンピュータ可読媒体。
【請求項２９】
前記複数の患者配列リードが、複数のＤＮＡシーケンシングリードおよびＲＮＡシーケンシングリードの少なくとも１つを含む、請求項２６に記載のコンピュータ可読媒体。
【請求項３０】

前記患者配列リードが、染色体６ｐ２１．３に位置する、または
前記患者配列リードが、次世代シーケンシングリードであり、且つメタデータをさらに含む、または
前記患者配列リードが、５０塩基と２５０塩基の間の長さを有する、
請求項２６に記載のコンピュータ可読媒体。
【請求項３１】
前記ｋ－ｍｅｒが、１０～２０の長さを有するか、または前記ｋ－ｍｅｒが、前記患者配列リードの長さの５％と１５％の間の長さを有する、請求項２６に記載のコンピュータ可読媒体。
【請求項３２】
前記投票が、患者配列リード当たりのｋ－ｍｅｒの合計数に対するマッチングｋ－ｍｅｒの割合を表す値である、請求項２６に記載のコンピュータ可読媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、２０１５年８月２５日出願の米国仮出願第６２／２０９，８５８号に対する優先権を主張する。
【０００２】
本発明の分野は、ヌクレオチド配列のインシリコ解析のシステムおよび方法であり、特にＳＮＰ、マルチヌクレオチド変異体、インデル、構造変異体、およびＨＬＡタイピングの高正確度コールに関する。
【背景技術】
【０００３】
本背景技術の記載は、本発明を理解に役立ち得る情報を含む。本明細書で提供する情報のいずれかが従来技術であるもしくは現在主張している発明に関連していること、または具体的もしくは黙示的に参照されたいずれの刊行物が従来技術であることを認めるものではない。
【０００４】
本明細書のすべての刊行物および特許出願は、それぞれ個々の刊行物または特許出願が参照により具体的におよび個別に組み入れられた場合と同程度に、参照により組み入れられる。組み入れられた参考文献における用語の定義または使用が本明細書に示すその用語の定義と一致しないまたは相反する場合、本明細書に示すその用語の定義が適用され、参考文献でのその用語の定義は適用されないものとする。
【０００５】
配列リード中の小さい変化に起因して正しく整列されないことが多く、変異体情報の不正確さまたは消失のいずれかを引き起こす、関連が高いゲノム配列セグメントを正確にアライメントするために、ハイスループットシーケンシングデータについての変異体検出は、ますます重要になってきた。関連が高い配列のアライメントを改善するためにいくつかの試みが行われてきた。例えば、「Ｐｌａｔｙｐｕｓ」（Ｔｈｅ Ｗｅｌｌｃｏｍｅ Ｔｒｕｓｔ Ｃｅｎｔｒｅ ｆｏｒ Ｈｕｍａｎ Ｇｅｎｅｔｉｃｓ）は、ハイスループットシーケンシングデータ中の比較的効率的で正確な変異体検出のために設計されたツールである。リードのローカルリアライメントおよびローカルアセンブリーを用いることにより、Ｐｌａｔｙｐｕｓは、数ｋｂまでのＳＮＰ、ＭＮＰ、短いインデル、置換および欠失の検出のための比較的高い感受性と高い特異性を達成する。Ｐｌａｔｙｐｕｓは従来のアライメントシステムとしてより正確であることが多いが、それにもかかわらず種々の問題点が残されている。特に、全ゲノムをカバーするゲノムデータの処理は問題であなり、類似度が高い複数の配列が存在する場合、所望の正確度に満たないこともある。同様に、ＤＩＳＣＯＶＡＲ（Ｂｒｏａｄ Ｉｎｓｔｉｔｕｔｅ）は、配列を構築し変異体を識別するための比較的正確なツールである。しかし、ＤＩＳＣＯＶＡＲは概して大量のデータ量の処理に適していない。

【０００６】
別の手法において、Ｂｉｇ Ｇｅｎｏｍｉｃｓ Ｉｎｆｅｒｅｎｃｅ Ｅｎｇｉｎｅ（ＢＩＧＧＩＥ；Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，ｖｏｌ．２５，ｐｐ．２０７８－９，２００９）では、最初にゲノムを複雑性が高い領域と低い領域に分類し、続いてそれに応じて情報資源を割り当てることによって、処理速度が上昇する。そのような手法は計算資源に対する要求を減少させる傾向があるが、複雑性が低い領域で変異が起こる場合、変異体コールはそれほど好ましくないことが多い。加えて、次世代のシーケンシングデータのための既知の変異体コーラーの大部分は、変異体を検出しその信頼度を評価するために、確率的フレームワーク（例えば、Ｂａｙｅｓｉａｎ Ｓｔａｔｉｓｔｉｃｓを使用する）を利用する。そのような手法は、通常、十分に機能するが、種々の因子、例えば高度のリード深度、プールサンプル、および混入サンプルまたは不純サンプルは、解析を混乱させる傾向がある。そのような問題を解決するために、ＶａｒＳｃａｎ（Ｇｅｎｏｍｅ Ｒｅｓ．２０１２ ２２：５６８－５７６）は、ヒューリスティク／統計的手法を利用して、リード深度、塩基品質、変異体対立遺伝子頻度、および統計的有意性の所望の閾値を満たす変異体コールを行なう。しかし、そのような手法は、通常、単一リードが及ばないゲノム中のより大きな変化を識別しない。

【０００７】
さらなる既知の方法では、ＤｅＢｒｕｉｊｎカラーグラフは、比較的長いｋ－ｍｅｒ（例えばｋは少なくとも５５）と、グラフを暗黙にコードするハッシュテーブルとを使用して（Ｎａｔ Ｇｅｎｅｔ．２０１２；４４（２）：２２６－２３２）シーケンシングデータから作成される。しかし、単離されたＳＮＰ、短いインデル（１～１００ｂｐ）およびＳＮＰとインデル（１～１００ｂｐ）の小複合体の組み合わせの場合、わずか８０％の検出力でヘテロ接合部位を検出し、９０％の検出力でホモ接合変異体部位を検出したことを著者らは報告した。さらに、中等度のサイズ（１００～１０００ｂｐ）のインデルと複合体変異体の場合、ヘテロ接合部位とホモ接合部位に対する検出力はそれぞれ５０％と７５～８０％であり、大きな変異体（１～５０ｋｂ）の場合、わずかな検出力（３５％）でホモ接合変異体部位を検出したことを著者らは報告した。したがって、記載のようにＤｅＢｒｕｉｊｎカラーグラフは、ＳＮＰおよびインデルの解析を少なくともある程度まで容易にするが、正確度と検出力は望ましいものより低い。したがって、その手法の主要な強さは複数のゲノムの同時解析にあり、それは参照ゲノムを必要とせずに変異体検出への強力で正確な手法を可能にする。

【発明の概要】
【発明が解決しようとする課題】
【０００８】
このように、変異体コールのための多数のシステムおよび方法が当技術分野で既知であるが、正確度の高い変異体コールのため、特にインシリコＨＬＡタイピングに関するシステムおよび方法の改善の必要性が依然として存在する。

【課題を解決するための手段】
【０００９】
本発明の主題は、患者の配列データからの正確度の高い変異体コールのための種々のシステム、方法および装置に関し、特にシーケンシング装置からのＤＮＡおよび／またはＲＮＡ配列を使用するＨＬＡタイピングを対象とする。特に好ましい態様において、複数のＨＬＡ対立遺伝子を含む患者配列リードおよび参照配列は、Ｄｅ Ｂｒｕｉｊｎグラフ手法で処理される。各患者配列リードは種々の対立遺伝子に対する重み付き投票を提供し、各対立遺伝子に対する投票総数を次いで使用して対立遺伝子のランクを付ける。ランク付けにおける最上の対立遺伝子が第１のＨＬＡ型であり、第１のＨＬＡ型にマッチするｋ－ｍｅｒに対するバイアスを伴う残りの対立遺伝子の再ランク付けが次いで第２のＨＬＡ型を与える。

【００１０】
本発明の主題の一態様において、本発明者は患者についてＨＬＡ型をインシリコ予測す

る方法を検討し、方法では、既知のおよび異なるＨＬＡ対立遺伝子の複数の配列を含む参照配列を提供し、および複数の患者配列リードを提供し、ここで患者配列リードの少なくともいくつかは患者特異的ＨＬＡをコードする配列を含む。さらなるステップにおいて、患者配列リードは複数のｋ－ｍｅｒのそれぞれのセットへと分解され、参照配列および複数のｋ－ｍｅｒのそれぞれのセットを使用して複合ｄｅ　Ｂｒｕｉｊｎグラフが次いで作成される。既知のおよび異なるＨＬＡ対立遺伝子のそれぞれが、複数の患者配列リードのそれぞれの投票から算出される複合マッチスコアを使用してランク付けられることがさらに考えられ、ここで各投票は、既知のおよび異なるＨＬＡ対立遺伝子中の対応するセグメントにマッチするｋ－ｍｅｒを使用する。

【００１１】

最も一般的に、参照配列は少なくとも１％の対立遺伝子頻度を有する少なくとも１つのＨＬＡ型についての対立遺伝子を含み、または参照配列は少なくとも１つのＨＬＡ型について少なくもとも１０の異なる対立遺伝子、および／もしくは少なくとも２つの異なるＨＬＡ型についての対立遺伝子を含む。ＨＬＡ型に関して、適切なＨＬＡ型はＨＬＡ－Ａ型、ＨＬＡ－Ｂ型、ＨＬＡ－Ｃ型、ＨＬＡ－ＤＲＢ－１型、および／またはＨＬＡ－ＤＱＢ－１型を含むことが考えられる。

【００１２】

患者配列リードは、複数のＤＮＡシーケンシングリードおよびＲＮＡシーケンシングリードの少なくとも１つを一般的に含み、染色体６ｐ２１．３に一般的に位置する。最も一般的には、患者配列リードは次世代シーケンシングリードであり、メタデータをさらに含み、および／または５０塩基と２５０塩基の間の長さである。ｋ－ｍｅｒに関して、好ましいｋ－ｍｅｒは１０～２０の長さであり、および／または患者配列リード長の５％と１５％の間の長さであることが考えられる。本発明の主題に限定されないが、複合マッチスコアは複数の患者配列リードからのすべての投票の合計であることが一般に好ましく、ここで投票は一般的に患者配列リード当たりのｋ－ｍｅｒの合計数に対するマッチングｋ－ｍｅｒの割合を表す値である。

【００１３】

したがって、複合マッチスコアを使用して、意図される方法は、患者の第１のＨＬＡ型として最上位ＨＬＡ対立遺伝子を識別するステップを含み得る。所望される場合、調節された複合マッチスコアを使用して残りの非最上位の既知のおよび異なるＨＬＡ対立遺伝子を再ランク付けする追加のステップを実行して、患者の第２のＨＬＡ型として調節された最上位ＨＬＡ対立遺伝子を識別し得る。最も一般的に、調節された複合マッチスコアは複数の患者配列リードのそれぞれの調節された投票から算出されてよく、および調節された投票は第１のＨＬＡ型にマッチするｋ－ｍｅｒの重みを切り下げることによって算出されてよい。

【００１４】

上記を考慮して、本発明者は、したがって、患者についてＨＬＡ型をインシリコン予測するためのコンピュータシステムも検討する。異なる観点から見て、本発明者はまた、参照配列データベースおよび患者配列データソースが解析エンジンに情報的に連結されるコンピュータシステムに実行させるプログラム命令を含む非一時的なコンピュータ可読媒体も検討する。適切な参照配列、患者配列リード、ＨＬＡ型、ｋ－ｍｅｒ、複合マッチスコア、および追加の再ランク付けステップに関して、上記と同じ考慮が適用される。

本

【００１５】

発明の主題の種々の目的、特徴、態様および利点は、同様の符号は同様の構成成分を表す添付の図面に加えて、以下の発明を実施するための形態からさらに明らかになる。

【図面の簡単な説明】

【００１６】

【図１】本発明の主題による１つの例示的な方法の概念図である。

【図２】本発明の主題による１つの例示的なコンピュータシステムの概念図である。

【発明を実施するための形態】
【００１７】
本発明者は、既知の配列情報を有する参照配列、および統計解析とヒューリスティク解析と組み合わせてｄｅ　Ｂｒｕｉｊｎグラフに基づく方法を使用して配列が処理される手法において、種々の密接に関連している配列の高度に正確なアライメントが容易に達成できることを見いだした。各ＨＬＡ型は多数のしばしば極めて類似した対立遺伝子を有するため、および配列が高い類似度を有する場合に従来のアライメント方法は有意な分別能を有することが一般的にできないので、そのような解析は、ＤＮＡおよび／またはＲＮＡシーケンシング情報からＨＬＡを決定するために、特に有利である。
【００１８】
本発明の主題の１つの例示的な態様において、染色体６ｐ２１．３（またはそこで／あるいはその近くでＨＬＡ対立遺伝子が見いだされるいずれかの他の位置）に位置する比較的多数の患者配列リードは、データベースまたはシーケンシング装置によって提供される。最も一般的に、配列リードは約１００～３００塩基の長さであり、リード品質、アライメント情報、配向、位置などを包含するメタデータを含む。例えば、適切な形式としては、ＳＡＭ、ＢＡＭ、ＦＡＳＴＡ、ＧＡＲなどが挙げられる。本発明の主題に限定されないが、患者配列リードは少なくとも５×、より一般的に少なくとも１０×、より一般的に少なくとも２０×、最も一般的に少なくとも３０×の深度カバレッジを提供することが一般に好ましい。
【００１９】
患者配列リードに加えて、意図される方法は、複数の既知のおよび異なるＨＬＡ対立遺伝子の複数の配列を含む１または複数の参照配列をさらに利用する。例えば、一般的な参照配列は、そのＨＬＡ型の複数のＨＬＡ対立遺伝子を有する少なくとも１つのＨＬＡ型の配列セグメントを含む合成の（ヒトまたは他の哺乳類対応物に対応しない）配列であり得る。例えば、適切な参照配列は、ＨＬＡ－Ａの少なくとも５０の異なる対立遺伝子に関する既知のゲノム配列の一群を含む。または、あるいはさらに、参照配列はＨＬＡ－Ａの少なくとも５０の異なる対立遺伝子に関する既知のＲＮＡ配列の一群も含む。もちろん、以下に詳述するように、参照配列はＨＬＡ－Ａの５０の対立遺伝子に限定されないが、ＨＬＡ型および対立遺伝子の数／組成に関して代替の組成を有することもある。最も一般的に、参照配列はコンピュータ可読形式であり、データベースまたは他のデータ記憶装置から提供される。例えば、適切な参照配列形式としては、ＦＡＳＴＡ、ＦＡＳＴＱ、ＥＭＢＬ、ＧＣＧ、またはＧｅｎＢａｎｋ形式が挙げられ、公開データリポジトリ（例えば、ＩＭＧＴ、Ｉｎｔｅｒｎａｔｉｏｎａｌ　ＩｍＭｕｎｏＧｅｎｅＴｉｃｓ情報システム、またはＴｈｅ　Ａｌｌｅｌｅ　Ｆｒｅｑｕｅｎｃｙ　Ｎｅｔ　Ｄａｔａｂａｓｅ，ＥＵＲＯＳＴＡＭ，ｗｗｗ．ａｌｌｅｌｅｆｒｅｑｕｅｎｃｉｅｓ．ｎｅｔ）のデータから直接取得するまたは構築することができる。または、参照配列は、対立遺伝子頻度、対立遺伝子頻度、民族別対立遺伝子分布、一般的なまたはまれな対立遺伝子型などの１または複数の所定の基準に基づいて、個々の既知のＨＬＡ対立遺伝子から構築されてもよい。
【００２０】
参照配列を使用して、患者配列リードは、現在、ｄｅ　Ｂｒｕｉｊｎグラフを通して、最良の適合で対立遺伝子を識別できる。この文脈において、各個人がＨＬＡ型について２つの対立遺伝子を保有し、これらの対立遺伝子は極めて類似している、または場合によって同一さえあり得ることに留意する必要がある。そのような高類似度は、従来のアライメントスキームに関して重大な問題を提起する。本発明者は、現在、ＨＬＡ対立遺伝子、および極めて密接に関連している対立遺伝子さえ、配列リードを比較的小さいｋ－ｍｅｒ（一般的に１０～２０塩基の長さを有する）へと分解することにより、および各患者配列リードが対立遺伝子の配列にマッチするその配列リードのｋ－ｍｅｒに基づいてそれぞれの対立遺伝子についての投票（「定量的リードサポート」）を提供する重み付き投票処理を実行することによりｄｅ　Ｂｒｕｉｊｎグラフが構築される手法を使用して、解決され得ることを見いだした。対立遺伝子についての累積的に最も高い投票は次いで、最も高い可

能性で予測されるＨＬＡ対立遺伝子を示す。加えて、以下でまた詳細に示すように、対立遺伝子にマッチする各フラグメントも用いて全体のカバレッジおよびその対立遺伝子のカバレッジ深度を算出することが一般に好ましい。

【００２１】

同じＨＬＡ型についての第２の対立遺伝子の識別に関して、本発明者は、比較的類似した第２の対立遺伝子でさえ、最上位ＨＬＡ対立遺伝子をさらなる考慮から外し、および残りの対立遺伝子を調節された（「スケーリングされた」）投票を使用して再ランク付ける、よりヒューリスティックな手法で分離できることを見いだした。より具体的には、最上位対立遺伝子とマッチしたｋ－ｍｅｒの投票値が再ランク付け投票で減少するように再ランク付けが行われる。そのような調節された投票は、最上位対立遺伝子に類似する遺伝子型の重み付け投票を減少させ（しかし削除しない）、したがって遺伝的により関連の少ない対立遺伝子により重みを置く。同時に、類似の対立遺伝子は、無視されない。ランク付けは、全体のカバレッジおよびカバレッジ深度を考慮に入れることによって、さらに改善される。例えば、第１の再ランク付け対立遺伝子は、実質的に低い全体的なカバレッジおよびカバレッジ深度で第２の再ランク付け対立遺伝子よりも高いスコアになり得る。このような場合、第２の再ランク付け対立遺伝子が正しい対立遺伝子である可能性が高い。それゆえ、最上位の再ランク付け対立遺伝子は、同じＨＬＡ型の第２の対立遺伝子である。もちろん、上述のように、再ランク付けは全体的なカバレッジおよびカバレッジ深度を考慮に入れることができ、全体的なカバレッジおよび／またはカバレッジ深度が、ユーザが規定する閾値（例えば、９４％未満の全体的なカバレッジ、および／または１０×未満のカバレッジ深度）を下回るような、対立遺伝子の不適合をもたらすこともあり得る。加えて、投票としてマッチするｋ－ｍｅｒを使用することで、特定の投票でのユニークなｋ－ｍｅｒの識別が可能になり、これはその特定の投票が正しい予測でありそうか、そうでないかのさらなるガイダンスとして役立ち得る。下記の表１は、ｄｅ　Ｂｒｕｉｊｎグラフ手法と、１０００　ｔｈｅ　Ｇｅｎｏｍｅｓ　Ｐｒｏｊｅｃｔ　（ＩＧＳＲ：　Ｔｈｅ　Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｇｅｎｏｍｅ　Ｓａｍｐｌｅ　Ｒｅｓｏｕｒｃｅ）からの単一ゲノム（ＹＲＩ）とを使用する、種々のＨＬＡ型（ＨＬＡ－Ａ、ＨＬＡ－Ｂ、ＨＬＡ－Ｃ、ＤＲＢ１、ＤＱＢ１）の対立遺伝子の例示的な予測を示す。

【００２２】

【表１】

		重み付きスコア	非重み付きスコア	%ユニークなk-mer	カバレッジ深度	カバーされる割合	対立遺伝子
NA19238	HLA-A	23721	23721	1.000	21.9	0.925	A*30:01:01
NA19238	HLA-A	15272	22197	0.269	20.5	0.925	A*36:01
NA19238	HLA-A	3595	9609	0.014	11.9	0.938	A*30:18
NA19238	HLA-A	2164	5211	0.031	9.8	0.970	A*30:53
NA19238	HLA-A	1575	5504	0.087	10.3	0.944	A*30:11:01
NA19238	HLA-B	17523	17523	1.000	16.3	0.921	B*53:01:01
NA19238	HLA-B	15938	16485	0.709	15.3	0.913	B*57:03:01
NA19238	HLA-B	5864	12275	0.000	11.4	0.921	B*58:01:01
NA19238	HLA-B	4830	11329	0.000	12.9	0.926	B*57:01:19
NA19238	HLA-B	1913	9496	0.000	8.8	0.921	B*58:01:07
NA19238	HLA-B	762	4835	0.037	9.1	0.925	B*35:27
NA19238	HLA-B	318	4899	0.000	9.2	0.976	B*58:36
NA19238	HLA-C	28463	28463	1.000	26.2	0.924	C*18:02
NA19238	HLA-C	18111	26189	0.317	24.1	0.924	C*04:01:01:01, C*04:01:01:02, C*04:01:01:03, C*04:01:01:04, C*04:01:01:05, C*04:82
NA19238	HLA-C	4320	14117	0.011	17.5	0.927	C*04:41
NA19238	HLA-C	728	3173	0.042	6.0	0.914	C*04:34
NA19238	DRB1	19990	19990	1.000	25.4	0.916	DRB1*16:02:01
NA19238	DRB1	17110	19954	0.599	25.4	0.914	DRB1*11:01:02
NA19238	DRB1	3119	6310	0.000	11.7	0.909	DRB1*11:97
NA19238	DRB1	2411	2790	0.655	10.9	1.000	DRB1*15:96
NA19238	DRB1	1278	4079	0.065	7.6	0.920	DRB1*11:01:08
NA19238	DRB1	893	2165	0.000	8.5	0.934	DRB1*16:23
NA19238	DQB1	17310	17310	1.000	22.4	0.930	DQB1*06:02:01
NA19238	DQB1	16390	16572	0.895	21.5	0.933	DQB1*05:02:01

表１

【００２３】

例示的な解析から容易に分かるように、各型の最上位ＨＬＡ対立遺伝子は容易に区別され、特に重み付けスコアが観察される場合、同じＨＬＡ型において第２のランク付け対立遺伝子は残りの対立遺伝子と実質的に異なっている。ＨＬＡ型の第１と第２のＨＬＡ対立遺伝子の選択もまた、有意に高いカバレッジ深度によって、ある程度のカバレッジまで十分にサポートされている。％ユニークなｋ－ｍｅｒ（最上位と比較して）もまた、本明細書に示すシステムおよび方法の類似性および識別性の良好な指標を提供することも認識すべきである。

【００２４】

もちろん、解析およびＨＬＡ予測が上記の特定のＨＬＡ型に限定される必要はないが、ＨＬＡ－Ｅ、ＨＬＡ－Ｆ、ＨＬＡ－Ｇ、ＨＬＡ－Ｈ、ＨＬＡ－Ｊ、ＨＬＡ－Ｋ、ＨＬＡ－Ｌ、ＨＬＡ－Ｖ、ＨＬＡ－ＤＱＡ１、ＨＬＡ－ＤＭＡ、ＨＬＡ－ＤＭＢ、ＨＬＡ－ＤＯＡ、ＨＬＡ－ＤＯＢ、ＨＬＡ－ＤＰＡ１、ＨＬＡ－ＤＰＢ１、ＨＬＡ－ＤＲＡ、ＨＬＡ－ＤＲＢ３４５、ＨＬＡ－ＭＩＣＡ、ＨＬＡ－ＭＩＣＢ、ＨＬＡ－ＴＡＰ１、ＨＬＡ－ＴＡＰ２、およびさらに新たに発見されるＨＬＡ型ならびにそれらの対応する対立遺伝子を含む、すべてのＨＬＡ型と対立遺伝子変異体が本明細書で検討されることを認識すべきである。さらに、解析が単一ＨＬＡ型に限定される必要はないが、複数のＨＬＡ型が本明細書での使用に適していることを認識すべきである。したがって、それぞれのＨＬＡ型についての対立遺伝子の一群とともに、参照配列は２、３、４、またはより多くのＨＬＡ型を含み得る。各ＨＬＡ型はかなりの数の対立遺伝子を有するので、既知の対立遺伝子のすべてを参照配列に包含する必要はないと考えられる。例えば、参照配列は、特定の閾値を上回る対立遺伝子頻度、例えば、少なくとも０．１％、もしくは少なくとも０．５％、もしくは少なくとも１％、もしくは少なくとも２％、もしくは少なくとも５％の対立遺伝子頻度を有する対立遺伝子を含み得る。したがって、異なる観点から見て、適切な参照配列は、少なくとも１つのＨＬＡ型について少なくとも１０、もしくは少なくとも３０、もしくは少

なくとも５０、もしくは少なくとも１００、もしくは少なくとも２００もしくは少なくとも５００、またはさらに多くの対立遺伝子を含み得る。

【００２５】

同様に、患者配列リードの性質および型がかなり変化し得ることを認識すべきである。例えば、検討される患者配列リードはＤＮＡ配列とＲＮＡ配列を含み、それぞれの配列は当技術分野で既知のすべての方法を使用して取得できる。さらに、そのような配列リードは、データ記憶装置（例えばデータベース）から、またはシーケンシング装置から提供され得る。例えば、ＤＮＡ配列リードはＮＧＳシーケンシング装置から導き出され、ＲＮＡ配列はｒｔＰＣＲシーケンシング装置から導き出され得る。したがって、患者配列リードの長さは、一般的に２０塩基超、より一般的に５０塩基超、最も一般的に１００塩基超であるが、通常は５，０００塩基未満、もしくは３，０００塩基未満、もしくは１，０００塩基未満である。したがって、検討される患者配列リードは、１００塩基と５００塩基の間または１５０塩基と１，０００塩基の間の長さであり得る。

【００２６】

計算時間とデータ記憶および／または必要メモリを減らすために、患者配列リードをＨＬＡ型遺伝子が位置するゲノム領域にあらかじめ選択しておくことがさらに好ましい。例えば、染色体６ｐ２１．３に位置する患者配列リードが特に検討される。同様に、患者配列リードはまた、ＨＬＡ対立遺伝子座が知られているゲノムに対してありそうな位置を示す１または複数のアノテーションに基づいて選択され得る。代替方法として、アノテーションはまた、ＨＬＡ対立遺伝子であるという配列の可能性を直接参照することもできる。

【００２７】

患者配列リードの長さに関係なく、患者配列リードが比較的短い長さのｋ－ｍｅｒに分解されることが一般に好ましく、特に好ましい長さは一般的に１０と３０の間である。注目すべきことに、そのような短いｋ－ｍｅｒの長さは、特にそのようなｋ－ｍｅｒを含有するフラグメントについての重み付き投票のために、変異体コールにおいてより高度な分解能と正確度を可能にする。したがって、ｋ－ｍｅｒ長は一般的に１０～３０の間、もしくは１５～３５の間、もしくは２０～４０の間である。異なる観点から見て、ｋ－ｍｅｒは、好ましくは６０未満の、より好ましくは５０未満の、最も好ましくは４０未満の、しかし５より長い、より一般的に８より長い、および最も一般的に１０より長い長さを有する。例えば、適切なｋ－ｍｅｒは、したがって、患者配列リードの長さの５％と１５％の間の長さである。

【００２８】

ランク付けおよび複合マッチスコアに関して、最も好ましい態様においてマッチスコアが患者配列リード中に存在するすべてのｋ－ｍｅｒに基づいて作成され、および各投票（すなわち、マッチング）ｋ－ｍｅｒが同じ投票力を有することに留意する必要がある。その結果、患者配列リードは、参照配列中のそれぞれの対立遺伝子に対して特定の定量的リードサポートを有する。さらに、ほとんどの場合、ゲノム中の各位置は＞１のシーケンシング深度を有し、および各患者配列リードは対立遺伝子の全長の一部分だけをカバーするので、各対立遺伝子は複数の患者配列リードから複数の投票を受け取ることができる。最も一般的に、対立遺伝子についての投票のすべては、その対立遺伝子の複合マッチスコアに達するように加えられる。それぞれの対立遺伝子の複合マッチスコアは次いで、ランク付けおよびさらなる解析のために使用される。

【００２９】

しかし、本発明の主題の別の態様において、複合スコアのスコアリングおよび算出は１または複数の特定の目的を達成するように修正されてもよいことに留意する必要がある。例えば、あるフラグメントのマッチスコアは、マッチングｋ－ｍｅｒのすべてから算出される必要はないが、ｋ－ｍｅｒの無作為な数または選択だけを計数できる。一方では、完全なマッチに満たないｋ－ｍｅｒ（例えば１４／１５マッチング）は、おそらく投票重みが低い投票権を与えられる。同様に、特にメタデータが利用できる場合、投票重みは、ｋ－ｍｅｒに対して減少されてよく、および／またはリード品質が特定の閾値を下回る場合

は患者配列リードに対して軽減されてよい。一方では、低いシーケンシング深度が存在する場合、投票は特定のフラグメントに対して多すぎることがあり得る。さらに別の意図される態様において、特にリード深度が比較的高い（例えば、少なくとも１５×、もしくは少なくとも２０×、もしくは少なくとも３０×）場合、同じ位置に対する患者配列リードは投票に基づいて除外され得る、または含まれ得る。したがって、複合マッチスコアは、利用できる投票のすべてに基づいてもよく、または対立遺伝子について利用できる投票の一部分のみに基づいてもよい。

【００３０】

ランク付けは累積のマッチスコアに一般的に依存するが、ランク付けは少なくとも１つの因子を使用して補正され得ることも認識すべきである。そのような補正因子としては、カバーされる割合、シーケンシング深度、ユニークなｋ－ｍｅｒの量、および利用できるフラグメントのメタデータが挙げられる。例えば、投票重みは、対立遺伝子のカバレッジが所定の閾値を下回る（例えば、９６％未満、もしくは９４％未満、もしくは９２％未満など）場合および／またはシーケンシング深度が所定の閾値を下回る（例えば、１５×未満、もしくは１２×未満、もしくは１０×未満など）場合、対立遺伝子について軽減され得る。一方では、投票重みは、例えば、ユニークなｋ－ｍｅｒのパーセンテージが所定の閾値を上回る（例えば、２％超、もしくは５％超、もしくは１０％超）場合、対立遺伝子について増加されてもよい。

【００３１】

最上位対立遺伝子は、一般的に所与のＨＬＡ型の第１の予測対立遺伝子であり、一方第２のランク付け対立遺伝子は、同じＨＬＡ型についての第２の対立遺伝子であり得る。しかし、最上位に続くランクの多くが類似の複合マッチスコアを有する場合（例えば、それのスコアのかなりの部分がｋ－ｍｅｒの高度に共有されるセットに由来する場合）特に、スコアリングは必要に応じてさらに改善または改良されてもよいことに留意する必要がある。好ましい一例において、スコア改良手法が実行されてよく、それは、最上位ｋ－ｍｅｒとマッチした（完全に、または少なくとも９０％、もしくは少なくとも９５％、もしくは少なくとも９７％、もしくは少なくとも９９％の類似度のいずれかで）ｋ－ｍｅｒの重みが補正因子によって軽減される再算出を含む。そのような補正因子は、任意の所定の量によって投票を低減できる。最も一般的に、補正因子は投票を１０％、もしくは２０～４０％、もしくは４０～６０％、もしくはさらに低減させる。これは最上位対立形質と類似している遺伝子型についての重み付き投票を軽減する効果を有し、異なっている遺伝子型を相対的により重要にする。したがって、第１の対立遺伝子は、すべてのシーケンシングデータからの最高のサポートに基づいて識別され、一方第２の対立遺伝子は、第２の対立遺伝子がデータセット中にサポートを有する（例えば、高いスケーリングされた重み付き投票および遺伝子型カバレッジ）かどうか、またはゲノムが第１の遺伝子型についてホモ接合性であるか（例えば、高い未処理の重み付き投票、極めて低いスケーリングされた重み付き投票、適切なカバレッジを有する他の対立遺伝子がない）を決定するために、未処理の重み付き投票、スケーリングされた重み付き投票の両方と、カバレッジを使用する、よりヒューリスティックスに基づく手法で識別されることを認識すべきである。異なる観点から見て、再ランク付けは、最上位対立遺伝子と類似している対立遺伝子の存在下でも、第２の対立遺伝子のより正確な判別を有利に可能にする。さらに、そのような方法は、ホモ接合ＨＬＡ型の迅速な識別も可能にする。加えて、そのような方法はハッシュテーブルの使用を必要とせず、配列リードをＨＬＡ型へと構築することなく適当なＨＬＡ対立遺伝子の識別を可能にすることを認識すべきである。さらに、意図されるシステムおよび方法は、ＤＮＡおよび／またはＲＮＡデータの使用も可能にする。

【００３２】

意図される方法の一代表実施形態を、図１に例示的に示す。ここで、方法１００はステップ１１０を含み、複数の既知のおよび異なるＨＬＡ対立遺伝子の配列を含む参照配列が提供される。ステップ１２０において、複数の患者配列リードが提供され、患者配列の少なくともいくつかは患者特異的ＨＬＡをコードする配列を包含し、一方ステップ１３０に

おいて、複数の患者配列リードは複数のｋ－ｍｅｒのそれぞれのセットに分解される（一般的に、各ｋ－ｍｅｒは１塩基（またはそれほど好ましくないが２塩基、もしくは３塩基、もしくは４塩基）の増分で進む）。ステップ１４０において、ｄｅＢｒｕｉｊｎグラフは、参照配列と、複数のｋ－ｍｅｒのそれぞれのセットとを使用して作成され、およびステップ１５０において、既知のおよび異なるＨＬＡ対立遺伝子のそれぞれは、複数の患者配列リードのそれぞれの投票から算出される複合マッチスコアを使用してランク付けされ、ここで各投票は、既知のおよび異なるＨＬＡ対立遺伝子中の対応するセグメントとマッチするｋ－ｍｅｒを使用する。

【００３３】

そのような方法のための例示的なシステムを図２に示す。ここで、システム２００は、参照配列データベース２０２（例えば、複数の既知のおよび異なるＨＬＡ対立遺伝子の配列を含む参照配列を格納するデータベースまたはファイル）を含み、ならびに患者配列データソース２０４（例えば、複数の患者配列リードを格納もしくは提供する配列データベースまたはシーケンシング装置であって、患者配列リードの少なくともいくつかは患者特異的ＨＬＡをコードする配列を含む）も含み、ここで両者はネットワーク２０６（例えば、ＬＡＮ、ＷＡＮ、イーサネット、インターネット）を介して解析エンジン２０８に情報的に連結されており、解析エンジンは、（ｉ）複数の患者配列リードを複数のｋ－ｍｅｒのそれぞれのセットへと分解する；（ｉｉ）参照配列とｋ－ｍｅｒの複数のそれぞれのセットを使用して複合体ｄｅ　Ｂｒｕｉｊｎグラフを作成する；および（ｉｉｉ）複数の患者配列リードのそれぞれの投票から算出される複合マッチスコアを使用して既知のおよび異なるＨＬＡ対立遺伝子のそれぞれのランク付けを行うようにプログラムされており、ここで各投票は既知のおよび異なるＨＬＡ対立遺伝子中の対応するセグメントとマッチするｋ－ｍｅｒを使用する。

【００３４】

コンピュータに向けられるいずれかの言語は、サーバ、インターフェイス、システム、データベース、エージェント、ピア、エンンジン、コントローラ、または個々にもしくは集合的に作動する他の種類の計算装置を包含する、計算装置の任意の適切な組み合わせを含むように読み取られる必要があることに留意すべきである。計算装置が有形の、非一時的なコンピュータ可読記憶媒体（例えば、ハードドライブ、ソリッドステートドライブ、ＲＡＭ、フラッシュ、ＲＯＭなど）に格納されるソフトウェア命令を実行するように構成されるプロセッサを含むことを認識すべきである。ソフトウェア命令は、開示される装置に関して後述のとおり、役割、責任、または他の機能性を提供するように計算装置を好ましく構成する。特に好ましい実施形態において、種々のサーバ、システム、データベース、またはインターフェイスは、おそらくＨＴＴＰ、ＨＴＴＰＳ、ＡＥＳ、公開鍵／秘密鍵交換、ウェブサービスＡＰＩ、既知の金融取引プロトコル、または他の電子情報交換方法に基づく、標準化プロトコルまたはアルゴリズムを使用してデータを交換する。データ交換は好ましくは、パッケト交換ネットワーク、インターネット、ＬＡＮ、ＷＡＮ、ＶＰＮ、または他の種類のパッケト交換ネットワーク上で行なわれる。

【００３５】

さらに、本明細書に提示されるシステムおよび方法は、従来のデータ形式および処理方式と比較して、ｄｅ　Ｂｒｕｉｊｎグラフエレメントの構築およびランク付け（および重み付け）が正確度および速度を大幅に上昇させるので、コンピュータ機能を改善することに留意すべきである。さらに、本発明者によって解決される問題はバイオインフォマティクス分野に特異的であり、オミクス情報のコンピューティングなしでは存在さえしないことを認識されたい。最後に、解析エンジンによって実行されるタスクは、コンピュータシステムの支援なしに人の一生のうちに合理的に遂行され得ないことを認識すべきである。

【００３６】

上記から容易にわかるように、意図されるシステムおよび方法は、各ＨＬＡ型に、第２位のスコアとして実質的により高く格付け／重み付けされる最上位スコアを提供する。したがって、Ｄｅ　Ｂｒｕｉｊｎグラフ型解析に基づき、ＨＬＡ型は非常に高い正確度で予

測され得ることを認識すべきである。さらに、本明細書に提示するシステムおよび方法は、種々の他のタスク、例えば、病原体変異体が参照配列の一部を形成する場合の病原体（例えば、ＨＰＶなどのウイルス病原体、マイコバクテリアなどの細菌性病原体、または熱帯熱マラリア原虫などの寄生性病原体）のタイピング、または腫瘍の多様性のタイピングなどにも適していることを認識すべきである。

【００３７】

本発明の主題のさらなる態様において、ｄｅ　Ｂｒｕｉｊｎグラフに基づく意図されるシステムおよび方法を利用して、構造変異体を識別し分類することもできる。ここでは、参照および未処理のシーケンシングデータを２つのゲノム領域（例えば、推定上の構造的変異の両側、例えば、ｂｃｒ－ａｂｌ融合）から取得し、これを使用してグラフを構築する。バブルが次いで、境界参照エッジがユーザ定義の最小ゲノム距離を超えて分離されるようなまたは境界参照エッジが異なる染色体上に位置するような、可能な構造的変異として識別される。そのような手法はほとんどの場合、疑われる構造的変異についての先験的な位置の知識（参照エッジの位置は、構造的変異が疑われるゲノム中の正確な位置を提供する）を必要とするが、そのような知識は通常、境界での正確な配列の識別に役立たない。現在Ｄｅ　Ｂｒｕｉｊｎグラフ手法を使用することで、構造的変異のさらに多くの正確な再構築が可能になり、かつ分岐点近くのまたは分岐点内の何らかの新規の配列に役立つ。そのような方法は構造的変異（例えば、挿入、重複など）が同じ鎖上に位置する場合に機能するだけでなく、グラフの構築が算出された逆相補ｋ－ｍｅｒの使用も含む場合に反転を識別するのにもまた同様に有用であることに留意すべきである。すでに前述したように、そのように識別された構造的変異は、続いてｖｃｆ形式または他の適切な形式で報告され得る。

【００３８】

例えば、腫瘍からの収集された配列情報はＤｅ　Ｂｒｕｉｊｎカラーグラフで表わされ、そこではエッジが、ｋ－ｍｅｒが見いだされる入力ソース（例えば、参照、正常サンプル、および／または腫瘍サンプル、様々な時期または年齢で採取されたサンプル、異なる患者または対象群由来のサンプルなど）を識別する「カラー」を有するｋ－ｍｅｒ（例えば、ｋ＝１５）であり、および各エッジが隣接するエッジに連結される。もちろん、配列はＤＮＡ配列ならびにＲＮＡ配列であってよく、このことは発現された体細胞変異、ＲＮＡ編集および選択的スプライシング（例えば、ＤＮＡとＲＮＡが同じ組織に由来する場合）の識別を有利に可能にすることに留意すべきである。最も一般的には、本発明の主題の好ましい一態様において、ゲノム中にｋ－ｍｅｒ位置を保管するために第１のグラフが参照配列から構築される。好ましくは、必要とされる特定のタスクに応じて、ｋ－ｍｅｒは３塩基と３００塩基との間、より好ましくは１０～１００塩基の長さを有する。例えば、インデル解析が所望される場合、ｋ－ｍｅｒ長は２０～５０の間（例えばｋ＝３０）であり得る。したがって、別の観点から見て、ｋ－ｍｅｒ長は、配列リードの平均長の５％から１５％の間であり得る。一旦第１のグラフが確立されると、ゲノムの所与の領域（マップされていないアンカーリードを含む）に位置する腫瘍のおよび正常な未処理シーケンシングデータからのｋ－ｍｅｒが加えられる。必要に応じて、そのための最大サポートがユーザ定義の特定の閾値（例えば、ｋ＝１３の場合、閾値は８である）を下回る弱いエッジをグラフから剪定してリードを除去できる。そのような剪定は、配列予測／アライメントの正確度を一般的に高める。

【００３９】

ｄｅ　Ｂｒｕｉｊｎグラフ（ｋ＝５）内の２つの隣接するエッジについてのデータ構造の例を後述する。

```
Edge0.配列 = ATATC
Edge0.外向き = [TATCG, TATCC]
Edge0.内向き = [TATAT]
Edge0.サポート = {‘参照’: 1, ‘腫瘍’: T0, ‘正常’: N0}
Edge0.quality_sum = {‘tumor’: TQ0, ‘normal’: NQ0}
```

Edge1.配列 = TATCG
Edge1.外向き = [ATCGG]
Edge1.内向き = [ATATC]
Edge1.support = {‘reference’: 0, ‘tumor’: T1, ‘normal’: N1}
Edge1.品質_sum = {‘腫瘍’: TQ1, ‘正常’: NQ1}

【００４０】

この例において、Ｅｄｇｅ０データ構造は、それらのｋｍｅｒ配列ＴＡＴＣＧとＴＡＴＣＣによって定義される２つの外向きエッジを有し、配列の前者は、後のＥｄｇｅ１データ構造中に記載される。Ｅｄｇｅ１の内向きエッジはＥｄｇｅ０へ戻って連結する。上記のデータ構造に記載されるサポートは、シーケンシングデータ（「腫瘍」または「正常」）または参照ゲノム（「参照」）中にエッジ配列が見られた回数をまとめる。上記のエッジ中のサポートに基づいて、Ｅｄｇｅ０は参照ゲノム中にサポートを有するが、一方Ｅｄｇｅ１に連結される外向きエッジはサポートをもたない。これは、Ｅｄｇｅ１が非参照変異体の始まりであり得ることを示すが、その接続形態が真の変異体（例えば、ＳＮＶに起因する、または参照ゲノム中に存在するエッジに囲まれている小さい挿入／欠失に起因するｄｅ　Ｂｒｕｉｊｎグラフ中の「バブル」）または人為的変異体（例えば、ジャンクまたはランダムなシーケンシングデータに起因することもあり得る、参照ゲノム中のエッジに再連結しないグラフ中の「チップ」）と一致するかどうかを決定するために、後続のエッジのさらなる内観が必要である。「腫瘍」および「正常」シーケンシングデータ（例えばＴ０、Ｎ０、Ｔ１、およびＮ１）中のサポートのレベルに応じて、非参照変異体の体細胞のまたは生殖系列の分類が決定され得る。分類の１つの単純な方法では、変異体は、Ｔ１＞０およびＮ１＞０ならば生殖系列として、Ｔ１＞０およびＮ１＝０ならば体細胞として、またはＴ１＝０およびＮ１＞０ならばＬＯＨとして分類されるが、ほとんどすべての実際の形では、体細胞または生殖系列の状態は、非参照変異体を記述するパス全体の概略分析（すなわち、非参照パス内の平均／最小／最大サポートおよびエッジの塩基品質）を介して決定される。

【００４１】

さらなるステップでは、そのように構築された複合グラフが次いで、腫瘍および参照が分岐する分岐点について解析される。各分岐について、深さ優先探索を使用して、参照に収束する腫瘍をもたらす腫瘍エッジを介するすべてのユニークなパスを識別し、これはｄｅ　Ｂｒｕｉｊｎグラフ中のバブルとして一般的に示される。ブレッドクラムを用いてループを回避できる。複合グラフが次いで、追加の配列で確立される。ここでは、一配列は、同じ患者のマッチする正常組織を表わすことがあり、そこから２つの他の配列、腫瘍ＤＮＡおよび腫瘍ＲＮＡを取得する。そのような例において、腫瘍ＤＮＡおよび腫瘍ＲＮＡは、同一である（これは必ずしもいつもそうとは限らない）。分岐点および収束点は、ｋ－ｍｅｒを使用する配列情報での相違によって決定される。上述の通り、分岐の領域は、グラフ中で「バブル」を生成する。したがって、別の観点から見て、腫瘍配列は分岐点と再収束点の両方を有し得ることを認識すべきである。また留意すべきであるが、腫瘍ＤＮＡおよびＲＮＡグラフは互いに同等であってよく、このことはＤＮＡとその対応する転写物の配列同一性を示す。

【００４２】

各バブル解の終わりから統計解析を次いで利用して、最も可能性の高いアライメントおよび／または配列を識別できる。最も一般的な実施形態において、配列は単なる未処理配列リードではなくアノテーション付きのＳＡＭまたはＢＡＭファイルであるので、統計解析は各リードについてのメタベースに基づくリード特異的パラメータを含み得る。したがって、統計解析は、最大のサポート、ｋ－ｍｅｒのマッピング／塩基品質、マッチした正常でのサポートなどを含み得る。結果として、参照配列を再構築するための参照エッジに沿ったバックトラッキングおよびゲノム中の位置の決定は、一般的にユーザ定義の基準（例えば、最小サポート＞Ｘリード、正常での最大サポート＜Ｙリードなど）を満たすグラフにおけるパスについて実行され得ることを認識すべきである。そのように構築された配

列および／または構造を次いで使用して、特定の変異体を分類できる。好ましくは、変異体分類はｖｃｆ形式で提示されるが、他の形式も考えられる。

【実施例】

【００４３】

ＨＬＡ予測を確証するために、３種の独立した既知の患者記録とサンプルを１０００Ｇｅｎｏｍｅ　ｐｒｏｊｅｃｔ（ＮＡ１９２３８、ＮＡ１９２３９およびＮＡ１９２４０）から取得し、次いで上述のようにＨＬＡ型を予測した。注目すべきことに、かつ予想外に、上述のようにＤｅ　Ｂｒｕｉｊｎグラフ方法を使用するＨＬＡの決定および予測は、以下の表２Ａおよび２Ｂに見られるように、ＨＬＡ－Ｃ（ＮＡ１９２３８について）、ＤＲＢ１（ＮＡ１９２３９について）およびＨＬＡ－Ｃ（ＮＡ１９２４０について）を除いてほぼ完全にマッチした。

【００４４】

【表２】

予測：

NA19238	HLA-A	23721	23721	1.000	21.9	0.925	A*30:01:01
NA19238	HLA-A	15272	22197	0.269	20.5	0.925	A*36:01
NA19238	HLA-B	17523	17523	1.000	16.3	0.921	B*53:01:01
NA19238	HLA-B	15938	16485	0.709	15.3	0.913	B*57:03:01
NA19238	HLA-C	28463	28463	1.000	26.2	0.924	C*18:02
NA19238	HLA-C	18111	26189	0.317	24.1	0.924	C*04:01:01
NA19238	DRB1	19990	19990	1.000	25.4	0.916	DRB1*16:02:01
NA19238	DRB1	17110	19954	0.599	25.4	0.914	DRB1*11:01:02
NA19238	DQB1	17310	17310	1.000	22.4	0.930	DQB1*06:02:01
NA19238	DQB1	16390	16572	0.895	21.5	0.933	DQB1*05:02:01
NA19239	HLA-A	24093	24093	1.000	22.2	0.926	A*02:01:01
NA19239	HLA-A	17596	20701	0.537	19.1	0.927	A*68:02:01
NA19239	HLA-B	21308	21308	1.000	19.8	0.918	B*35:01:01
NA19239	HLA-B	15080	20254	0.286	18.8	0.912	B*52:01:02
NA19239	HLA-C	18529	18529	1.000	17.0	0.920	C*04:01:01
NA19239	HLA-C	17846	18484	0.707	17.0	0.919	C*16:01:01
NA19239	DRB1	26014	26014	1.000	33.1	0.914	DRB1*13:01:01
NA19239	DRB1	16174	24412	0.178	31.0	0.914	DRB1*12:01:01
NA19239	DQB1	18503	18503	1.000	24.0	0.930	DQB1*05:01:01
NA19239	DQB1	13459	13510	0.939	17.5	0.931	DQB1*03:01:01
NA19240	HLA-A	21944	21944	1.000	20.2	0.924	A*30:01:01
NA19240	HLA-A	20059	20512	0.800	18.9	0.929	A*68:02:01
NA19240	HLA-B	18637	18637	1.000	17.3	0.927	B*35:01:01
NA19240	HLA-B	17850	18550	0.682	17.3	0.926	B*57:03:01
NA19240	HLA-C	28054	28054	1.000	25.8	0.923	C*18:02
NA19240	HLA-C	20132	27609	0.390	25.3	0.919	C*04:01:01
NA19240	DRB1	22869	22869	1.000	29.1	0.917	DRB1*16:02:01
NA19240	DRB1	17094	20016	0.591	25.4	0.915	DRB1*12:01:01
NA19240	DQB1	14654	14654	1.000	19.0	0.930	DQB1*05:02:01
NA19240	DQB1	10926	10959	0.951	14.2	0.931	DQB1*03:01:01

表２Ａ

真実：

NA19238	HLA-A	A*30:01
NA19238	HLA-A	A*36:01
NA19238	HLA-B	B*53:01
NA19238	HLA-B	B*57:03
NA19238	HLA-C	C*18:01
NA19238	HLA-C	C*04:01
NA19238	DRB1	DRB1*16:02
NA19238	DRB1	DRB1*11:01
NA19238	DQB1	DQB1*06:02
NA19238	DQB1	DQB1*05:02
NA19239	HLA-A	A*02:01
NA19239	HLA-A	A*68:02
NA19239	HLA-B	B*35:01
NA19239	HLA-B	B*52:01
NA19239	HLA-C	C*04:01
NA19239	HLA-C	C*16:01
NA19239	DRB1	DRB1*13:01
NA19239	DRB1	DRB1*13:01
NA19239	DQB1	DQB1*05:01
NA19239	DQB1	DQB1*03:01
NA19240	HLA-A	A*30:01
NA19240	HLA-A	A*68:02
NA19240	HLA-B	B*35:01
NA19240	HLA-B	B*57:03
NA19240	HLA-C	C*18:01
NA19240	HLA-C	C*04:01
NA19240	DRB1	DRB1*16:02
NA19240	DRB1	DRB1*12:01
NA19240	DQB1	DQB1*05:02
NA19240	DQB1	DQB1*03:01

表２Ｂ

【００４５】

ここでは、不明瞭な数字を上記の対立遺伝子から除いた。例えば、予測がＡ＊０４：０２：０１とＡ＊０４：０２：０２の場合、最後の不明瞭な数字（ここでは０１または０２）を除き、したがって予測Ａ＊０４：０２を得た。

予測されたＨＬＡ型と実験に基づいて決定されたＨＬＡ型（「真実」）との間の相違をさらに調査することにより、以下にさらに詳細に検討するように、ＮＡ１９２３８とＮＡ１９２３９がＮＡ１９２４０の両親であった場合、実験に基づいて決定されたＨＬＡは予想された遺伝パターンと一致しなかったことが驚くべきことに明らかになった。

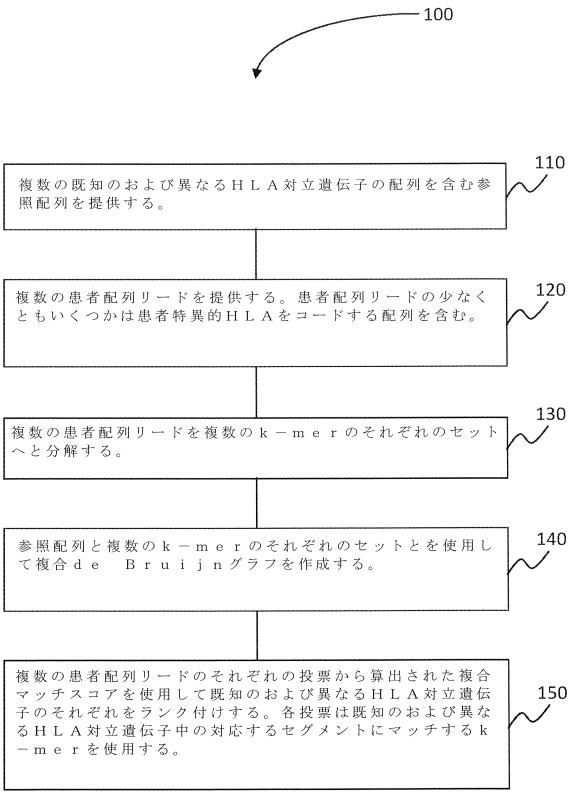
【００４６】

Ｃ＊１８：０１と決定される「真実」および予測されるＣ＊１８：０２に関して、これらの２つの対立遺伝子形態間にわずか一塩基の変化があることが注目される。特に、Ｃ＊１８：０１は、ＷＧＳデータ中にリードサポートがゼロであるＣＴＧＧＴＴＧＴＣ（関連する配列部分のみ）の配列を有するが、Ｃ＊１８：０２はＷＧＳデータ中にそれをサポートする３３のリードがあるＣＴＧＧＣＴＧＴＣ（関連する配列部分のみ）の配列を有する。このデータによれば、「真実」Ｃ＊１８：０１に対するサポートはないが、予測されるＣ＊１８：０２に対しては多数のサポートがある。

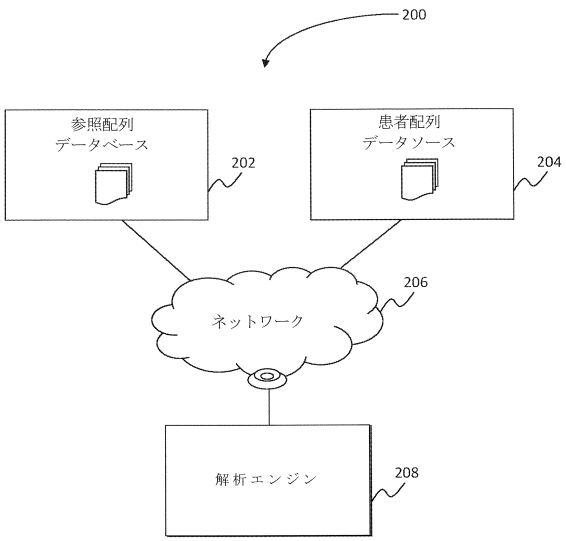
【００４７】
ＤＲＢ１＊１３：０１と決定される「真実」および予測されるＤＲＢ１＊１２：０１に関して：ＮＡ１９２４０は両親ＮＡ１９２３８とＮＡ１９２３９の子供であることが注目される。子供は各親から各ＨＬＡ型についての対立遺伝子を１つだけ受け継ぐので、真の対立遺伝子は単純な基本的メンデル遺伝から決定できる：
親１（ＮＡ１９２３８）：　１６：０２、　１１：０１
親２（ＮＡ１９２３９）：　１３：０１、　？　←問題の対立遺伝子
子供（ＮＡ１９２４０）：　１６：０２、　１２：０１
【００４８】
上記からわかるように、子供は親１から１６：０２を受け継がなければならず、このことは対立遺伝子１２：０１が親２から来なければならないことを意味する。特に、「真実」は１３：０１として親２についての第２の対立遺伝子を記載するが、これは遺伝に基づくと不可能である。親２の予測される対立遺伝子は１２：０１である。しかし、これは、まさに遺伝に基づいて予想するものである。したがって、上記の例に基づいて、「不正確な」予測は、実際のところ「真実」における誤りに起因していた。このように、本明細書に示すＨＬＡ予測方法は、３つの個々のデータセットにおいて５つのＨＬＡのそれぞれ異なるパネルにわたり１００％の正確度を示した。上記の予測が平均的カバレッジのＷＧＳサンプルを用いて行われたことを、さらに認識すべきである。本方法の正確度は、腫瘍によって発現される対立遺伝子の識別を可能にするＲＮＳ配列データを用いてさらにかなり改善され得る。これは、時には、ＤＮＡに存在する２つの対立遺伝子の１つだけであり得る。意図されるシステムおよび方法のさらなる有利な態様において、ＤＮＡもしくはＲＮＡ、またはＤＮＡとＲＮＡの両方の組み合わせを処理して高度に正確であるＨＬＡ予測を行うことができ、かつそれらを腫瘍または血液のＤＮＡもしくはＲＮＡから得ることができる。さらに、意図される方法は、２６のすべてのＨＬＡ型についての予測を極めて迅速に（実行時間は一般的に５分未満）取得し、かつ新たに発見された、または極めてまれなＨＬＡ対立遺伝子が自明な方法で追加できる。最後に、集団に基づくヒューリスティクスは正確な結果を出すために必要とされないことに留意する必要がある。
【００４９】
したがって、本明細書に示すシステムおよび方法を用いて、ゲノム解析で明らかにされる異なるオブジェクトを確証または確認できることを認識すべきである。さらに、同じグラフでＲＮＡ情報を使用する場合、変異対立遺伝子発現を直ちに識別できる。さらに、上記の結果と考察に基づいて、システムおよび方法が、ＲＮＡ－Ｓｅｑを使用して遺伝子融合を、特に「実施可能な融合」（例えばＢＣＲ－ＡＢＬ）または発癌遺伝子のアイソフォーム（例えばＥＧＦＲｖＩＩＩ）をコールできるであろうとも考えられる。
【００５０】
本明細書で用いる場合、文脈が明らかに指示しない限り、「に連結される」という用語は、直接連結（互いに連結される２つのエレメントが互いに接触する）および間接連結（少なくとも１つの追加のエレメントが２つのエレメント間に位置する）の両方を含むと意図される。したがって、「に連結される」および「と連結される」という用語は同義的に使用される。さらに、本明細書に開示される本発明の代替エレメントまたは実施形態のグループ化は、限定として解釈されるべきではない。各グループメンバーは、個別に、または本明細書に見られるグループの他のメンバーもしくは他のエレメントとの任意の組み合わせで参照され、または主張され得る。グループの１または複数のメンバーは、利便性および／または特許性の理由から、グループに包含され、もしくはグループから削除され得る。そのような包含または削除が行なわれた場合、本明細書は、修正されたグループを含み、したがって、添付の特許請求の範囲に使用されるすべてのマーカッシュグループの記載を満たすと本明細書ではみなされる。
【００５１】
すでに記述されているものの他にさらに多くの修正が本明細書の発明概念を逸脱しない範囲で可能であることは当業者にとって明らかである。したがって、本発明の主題は、添

付の特許請求の範囲を除いて限定されるべきではない。さらに、本明細書および特許請求の範囲の両方を解釈する際に、すべての用語は、文脈と一致する最も広い可能な方法で解釈されなければならない。特に、「含む」および「含んでいる」という用語は、エレメント、成分、またはステップを参照して、非排他的方法で解釈すべきであり、言及するエレメント、成分、またはステップが、明白に参照していない他のエレメント、成分、またはステップとともに存在し、もしくは利用され、もしくは組み合わされてもよいことを指示している。本明細書、特許請求の範囲がＡ、Ｂ、Ｃ．．．．、およびＮからなる群から選択されるもののうちの少なくとも１つを指す場合、本文は、Ａ＋ＮまたはＢ＋Ｎなどではないその群からの唯一のエレメントを要求していると解釈すべきである。

【図１】



【図２】



フロントページの続き

(72)発明者　サンボーン，ジョン　ザキャリー
アメリカ合衆国，カリフォルニア州　９５０６５，サンタ　クルス，１９５　ケニー　アベニュー

審査官　松野　広一

(56)参考文献　米国特許出願公開第２０１５／０１１０７５４（ＵＳ，Ａ１）
米国特許出願公開第２０１４／０１１４５８４（ＵＳ，Ａ１）
特開２０１５－０３５２１２（ＪＰ，Ａ）
米国特許出願公開第２０１３／０２６７４２９（ＵＳ，Ａ１）
Alexander DILTHEY et al., Improved genome inference in the MHC using a population reference graph, NATURE GENETICS, ２０１５年　6月, Vol.47 No.6, pp.682-688

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ　５／００－９９／００
Ｃ１２Ｎ　１５／１２
Ｃ１２Ｑ　１／６８０６
ＰｕｂＭｅｄ