

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7636400号
(P7636400)

(45)発行日 令和7年2月26日(2025.2.26)

(24)登録日 令和7年2月17日(2025.2.17)

(51)国際特許分類

F I

G 1 6 B 30/20 (2019.01)

G 1 6 B 30/20

C 1 2 Q 1/6869(2018.01)

C 1 2 Q 1/6869

Z Z N A

C 1 2 N 15/11 (2006.01)

C 1 2 N 15/11

Z

請求項の数 19 (全44頁)

(21)出願番号	特願2022-517384(P2022-517384)	(73)特許権者	518199104
(86)(22)出願日	令和2年9月29日(2020.9.29)		イルミナ シンガポール ピーティーイー
(65)公表番号	特表2022-550013(P2022-550013 A)		リミテッド
(43)公表日	令和4年11月30日(2022.11.30)		I L L U M I N A S I N G A P O R E
(86)国際出願番号	PCT/GB2020/052358		P T E L T D
(87)国際公開番号	WO2021/064365		シンガポール国 7 5 7 7 1 6 , ウッド
(87)国際公開日	令和3年4月8日(2021.4.8)		ランズ インダストリアル パーク イー1
審査請求日	令和5年8月24日(2023.8.24)		2 9 , ノーステック, ナンバー0 2 -
(31)優先権主張番号	1914064.9	(74)代理人	100107489
(32)優先日	令和1年9月30日(2019.9.30)		弁理士 大塩 竹志
(33)優先権主張国・地域又は機関	英国(GB)	(72)発明者	ダーリング, アーロン アール
			オーストラリア国 2 0 0 0 ニューサウ
			スウェールズ, シドニー, ビット ス
			トリート 5 6 , レベル 3 , ロンガス
			最終頁に続く

(54)【発明の名称】 2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定するための方法

(57)【特許請求の範囲】

【請求項1】

コンピュータ実施方法であって、前記方法は、

複数の変異シーケンスリードを受け取ることであって、各変異シーケンスリードは、変異を含むシーケンスのサブシーケンスに対応し、変異を含む前記シーケンスは、変異を含まないシーケンスと比較して変異を含む、ことと、

共通ミニマイザ関数を各変異シーケンスリードに適用し、それによって、変異シーケンスリードごとに1つ以上のそれぞれのミニマイザを決定することであって、前記1つ以上のそれぞれのミニマイザが、それぞれのシーケンスリードに存在するk-merのセットからの1つ以上の代表的なk-merである、ことと、

各変異シーケンスリードにおいて前記1つ以上のそれぞれのミニマイザの位置を決定することと、

各変異シーケンスリードにおいて1つ以上の変異の位置を決定することと、

共通ミニマイザを有する少なくとも2つの変異シーケンスリードについて、前記それぞれのミニマイザがアラインされると、位置が一致する、及び／又は位置が一致しない変異の数をカウントすることと、

前記それぞれのミニマイザがアラインされると、位置が一致する、及び／又は位置が一致しない変異の数に基づいて、前記共通ミニマイザを有する前記少なくとも2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定することと、

前記確率と相関する前記尺度に基づいて、前記複数の変異シーケンスリードをアセンブルすることによって、変異を含む前記シーケンスの少なくとも一部をアセンブルする又は再構築することと、
を含む、方法。

【請求項 2】

複数の非変異シーケンスリードを受け取ることを更に含み、各非変異シーケンスリードは、変異を含まない前記シーケンスのサブシーケンスに対応し、

ここで、共通ミニマイザ関数を適用するステップ、又は各変異シーケンスリードにおいて 1 つ以上の変異の位置を決定するステップは、前記複数の変異シーケンスリードからの k - m e r 及び前記複数の非変異シーケンスリードからの k - m e r を比較することを含む、請求項 1 に記載の方法。

10

【請求項 3】

共通ミニマイザ関数を各変異シーケンスリードに適用するステップは、i) 可能な k - m e r の順序付きリストに最初に列挙されている、前記それぞれの変異シーケンスリードにおける 1 つ以上の k - m e r、又は ii) 可能な k - m e r の所定のセットに存在する、前記それぞれの変異シーケンスリードにおける 1 つ以上の k - m e r を特定することを含み、前記それぞれの変異シーケンスリードについて決定された、前記 1 つ以上のミニマイザは、特定した前記 1 つ以上の k - m e r である、請求項 1 又は 2 に記載の方法。

【請求項 4】

可能な k - m e r の前記順序付きリスト又は可能な k - m e r の前記所定のセットは、前記複数の非変異シーケンスリードよりも前記複数の変異シーケンスリードに存在することが多い k - m e r からなる、請求項 2、又は請求項 2 に従属する請求項 3 に記載の方法。

20

【請求項 5】

可能な k - m e r の前記所定のセットは、前記複数の変異シーケンスリードに n 回以上存在し、前記複数の非変異シーケンスリードに n 回未満存在する k - m e r からなり、n は 1 以上の整数である、
請求項 2、並びに請求項 2 に直接的又は間接的に従属する請求項 3 及び請求項 4 のいずれか一項に記載の方法。

【請求項 6】

前記複数の変異シーケンスリードにおける前記 k - m e r と前記複数の非変異シーケンスリードにおける前記 k - m e r との比較に基づいて、可能な k - m e r の前記順序付きリスト又は可能な k - m e r の前記所定のセットを生成することを更に含む、請求項 2、並びに請求項 2 に直接的又は間接的に従属する請求項 3 ～ 5 のいずれか一項に記載の方法。

30

【請求項 7】

各ミニマイザは、5 超の長さの k - m e r である、請求項 1 ～ 6 のいずれか一項に記載の方法。

【請求項 8】

各ミニマイザビンは、共通ミニマイザを有する変異シーケンスリードを含み、共通ミニマイザを有さない変異シーケンスリードを含まないように、1 つ以上のミニマイザビンにおいて前記変異シーケンスリードをビンングすることを更に含み、

40

位置が一致する、及び／又は位置が一致しない変異の数をカウントするステップは、同一ミニマイザビン内の変異シーケンスリードのみで実行される、請求項 1 ～ 7 のいずれか一項に記載の方法。

【請求項 9】

各変異シーケンスリードにおいて前記 1 つ以上の変異の位置を決定する前記ステップは、
前記複数の非変異シーケンスリードにおいて、1 つ以上のシードパターンのうちのそれぞれ 1 つを k - m e r に適用することによって、シードマスクされた非変異 k - m e r のセットを得ることと、

変異シーケンスリードごとに、前記それぞれの変異シーケンスリードにおいて、前記 1 つ以上のシードパターンを k - m e r に適用して、複数のシードマスクされた変異 k - m

50

e rを得ることと、シードマスクされた非変異k - m e rの前記セットに存在する、前記複数のシードマスクされた変異k - m e rのうちの前記シードマスクされた変異k - m e rに対応する前記シードパターンの全てによってマスクされる前記変異シードシーケンスリードにおいて前記1つ以上の位置を特定することによって、前記1つ以上の変異の前記位置を特定することと、を含む、

請求項2、並びに請求項2に直接的又は間接的に従属する請求項3～8のいずれか一項に記載の方法。

【請求項10】

前記複数の変異シーケンスリードのそれぞれは、複数のサンプルのうちの1つに関連する変異を含むシーケンスのサブシーケンスに対応し、前記複数の非変異シーケンスリードのそれぞれは、前記複数のサンプルのうちの前記1つに関連する変異を含まないシーケンスのサブシーケンスに対応し、変異を含む各シーケンスは、変異を含まないそれぞれのシーケンスと比較して変異を含み、

シードマスクされた非変異k - m e rのセットを得ることは、サンプルごとにシードマスクされた非変異k - m e rのそれぞれのセットを得ることを含み、

非変異サンプルビットベクトルのセットを作成することを更に含み、各非変異サンプルビットベクトルは、シードマスクされた非変異k - m e rの前記セット内のそれぞれのk - m e rについて、前記それぞれのk - m e rが存在する前記複数のサンプルを定義し、

前記1つ以上の変異の位置を決定することは、変異シーケンスリードごとに、並びにシードマスクされた非変異k - m e rのセットごとに及び／又はシードマスクされた非変異k - m e rのセットの組み合わせごとに、シードマスクされた非変異k - m e rの前記それぞれのセット及び／又はシードマスクされた非変異k - m e rのセットの前記それぞれの組み合わせに存在する、前記複数のシードマスクされた変異k - m e rのうちの前記シードマスクされた変異k - m e rに対応する前記シードパターンの全てによってマスクされる前記変異シーケンスリードにおいて前記1つ以上の位置を特定すること、並びに特定した前記1つ以上の位置を、シードマスクされた非変異k - m e rの前記それぞれのセット又はシードマスクされた非変異k - m e rのセットの前記それぞれの組み合わせと関連付けられた前記1つ以上のサンプルと関連付けることを含む、請求項9に記載の方法。

【請求項11】

各変異シーケンスリードにおいて前記1つ以上の変異の位置を決定する前記ステップは、前記変異シーケンスリードのうちの1つ以上について、前記それぞれの変異シーケンスリードをリファレンスアセンブリにアラインすることと、

前記それぞれの変異シーケンスリードと前記リファレンスアセンブリとの間での違いの、前記それぞれの変異シーケンスリードにおける位置を特定することによって、前記それぞれの変異シーケンスリードにおいて前記1つ以上の変異の位置を決定することと、を含む、請求項2～10のいずれか一項に記載の方法。

【請求項12】

各変異シーケンスリードにおいて前記1つ以上の変異の位置を決定する前記ステップは、変異シーケンスリードごとに、

前記それぞれの変異シーケンスリードにおける位置が前記リファレンスアセンブリにアラインされる場合、前記それぞれの変異シーケンスリードにおける前記位置が、前記それぞれの変異シーケンスリードは前記リファレンスアセンブリとは異なる位置であると、前記それぞれの変異シーケンスリードにおける位置は、前記それぞれの変異シーケンスリードにおける変異の位置であると決定することと、

前記それぞれの変異シーケンスリードにおける位置が前記リファレンスアセンブリにアラインされていない場合、前記それぞれの変異シーケンスリードにおける前記位置が、シードマスクされた非変異k - m e rの前記セットに存在する前記複数のシードマスクされた変異k - m e rのうちの前記シードマスクされた変異k - m e rに対応する前記シードパターンの全てによってマスクされる位置である場合、前記それぞれの変異シーケンスリードにおける位置は、前記それぞれの変異シーケンスリードにおける変異の位置であると

10

20

30

40

50

決定することと、を含む、請求項 9 又は請求項 10 に従属する請求項 11 に記載の方法。

【請求項 13】

前記少なくとも 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する前記確率と相関する前記尺度が、位置が一致する、及び／又は位置が一致しない変異の数である、または

前記少なくとも 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する前記確率に相関する前記尺度は、i) 前記少なくとも 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率密度、及び ii) 前記少なくとも 2 つの変異シーケンスリードが同一の変異を含むシーケンスに由来する確率密度と相関するスコア関数のうちの 1 つである、

請求項 1 ～ 12 のいずれか一項に記載の方法。

【請求項 14】

前記複数の変異シーケンスリードから無向加重グラフを作成することを更に含み、

前記無向加重グラフは、前記複数の変異シーケンスリードに対応するノードを含み、前記ノード間のエッジは、それぞれのエッジ重みと関連し、各エッジ重みは、前記それぞれのエッジに関連する前記 2 つのノードに対応する前記 2 つの変異シーケンスリードについて決定された、位置が一致する、及び／又は位置が一致しない変異の数に基づいて決定される、

請求項 1 ～ 13 のいずれか一項に記載の方法。

【請求項 15】

前記無向加重グラフでグラフクラスタリングを実行し、それによって同一の変異を含むシーケンスに由来すると予想される変異シーケンスリードのクラスタを作成することを更に含む、請求項 14 に記載の方法。

【請求項 16】

前記クラスタにおいて前記変異シーケンスリードをアセンブルすることによって変異を含む前記シーケンスの少なくとも一部を再構築することを更に含む、請求項 15 に記載の方法。

【請求項 17】

変異を含む前記シーケンスの前記再構築部分から変異を含まない可能性の高いシーケンスの少なくとも一部を推測することによって、変異を含まない前記シーケンスの少なくとも一部を再構築することを更に含む、請求項 2 に直接的又は間接的に従属する請求項 16 に記載の方法。

【請求項 18】

標的鋳型核酸分子のシーケンスの少なくとも一部を生成するための方法であって、請求項 14 ～ 17 のいずれか一項に記載の方法を含む、方法。

【請求項 19】

少なくとも 1 つの標的鋳型核酸分子のシーケンスの少なくとも一部を決定するための方法であって、前記方法は、

変異を含む、少なくとも 1 つの標的鋳型核酸分子の領域をシーケンスして、複数の変異シーケンスリードを提供することと、

得られた前記複数の変異シーケンスリードで請求項 1 ～ 18 のいずれか一項に記載の方法を実行することと、を含む、方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定するためのコンピュータ実施方法、シーケンスの少なくとも一部を生成するための方法、及び少なくとも 1 つの標的鋳型核酸分子のシーケンスを決定するための方法に関する。

【背景技術】

【0002】

核酸分子のシーケンス能力は、無数の異なる用途において非常に有用なツールである。しかしながら、反復領域を含む核酸分子など問題のある構造を含む核酸分子の正確なシーケンスを決定することは困難であり得る。また、二倍体生物及び多倍数体生物のハプロタイプ構造、並びにこれらの生物のゲノム中の構造バリエーションなど構造的特徴を解決することも困難であり得る。

【0003】

より最新の技術の多く（いわゆる次世代シーケンシング技術）は、短い核酸分子を正確にシーケンスすることしかできない。次世代シーケンシング技術を使用して、より長い核酸シーケンスをシーケンスすることができるが、これは、困難なことが多く、高価である。次世代シーケンシング技術を使用して、核酸分子の一部のシーケンスに対応する短いシーケンスリードを生成することができ、フルシーケンスは、短いシーケンスリードからアセンブルされ得る。核酸分子が反復領域を含む場合、ユーザには、類似のシーケンスを有する2つのシーケンスリードが、より長いシーケンス内の2つの反復のシーケンスに対応するのか、又は同一シーケンスの2つの複製に対応するのかが不明確であり得る。同様に、ユーザは、2つの類似の核酸分子を同時にシーケンスすることを望む場合があり、類似のシーケンスを有する2つのシーケンスリードが同一の元の核酸分子のシーケンスに対応するのか、又は2つの異なる元の核酸分子のシーケンスに対応するかを決定することは困難であり得る。

【0004】

短いシーケンスリードからのシーケンスのアセンブルは、*Sequencing Aided by Mutagenesis* (SAM) 技術によって支援され得る。概して、SAMは、標的鋳型核酸シーケンスに変異を導入することを含む。導入される変異パターンは、この方法のユーザが短いシーケンスリードから核酸分子のシーケンスをアセンブルする際に支援し得る。

【0005】

例えば、鋳型核酸分子が反復領域を含む場合、反復は異なる変異パターンによって互いに区別され得、それによって反復領域が分解され、正しくアセンブルされることを可能にし得る。

【0006】

概して、SAM技術は、標的鋳型核酸分子のコピーを変異させて、変異標的鋳型核酸分子及び／又は変異を含む1つ以上のシーケンスを生成することと、変異を含む1つ以上のシーケンスをシーケンスして、変異シーケンスリードを含むSAMデータを作成することと、次いで変異パターンに基づいて、変異シーケンスリードからシーケンスをアセンブルすることと、を含む。異なる変異コピーは異なる位置に変異を含むため、アセンブルされたシーケンスは、元の鋳型核酸分子を表し得る。

【0007】

しかしながら、SAMデータを処理するためのより信頼性の高い、及び／又はより計算効率の良い方法が依然として必要とされている。

【発明の概要】

【課題を解決するための手段】

【0008】

本発明者らは、変異シーケンスリードを含むSAMデータを処理するための新しい、改善された方法を開発した。したがって、本発明の一態様では、2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定するための方法が提供される。この方法は、複数の変異シーケンスリードを受け取ることを含む。各変異シーケンスリードは、変異を含むシーケンスのサブシーケンスに対応する。変異を含むシーケンスは、変異を含まないシーケンスと比較して変異を含む。この方法は、共通ミニマイザ関数を各変異シーケンスリードに適用し、それによって、変異シーケンスリードごとに1つ以上のそれぞれのミニマイザを決定することを更に含む。この方法は、各変異シー

10

20

30

40

50

ケンスリードにおいて1つ以上のそれぞれのミニマイザの位置を決定することを更に含む。この方法は、各変異シーケンスリードにおいて1つ以上の変異の位置を決定することを更に含む。この方法は、共通ミニマイザを有する少なくとも2つの変異シーケンスリードについて、それぞれのミニマイザがアラインされると、位置が一致する、及び／又は位置が一致しない変異の数をカウントすることを更に含む。

【0009】

本発明の別の態様では、標的鋳型核酸分子のシーケンスの少なくとも一部を生成するための方法が提供される。

【0010】

本発明の別の態様では、少なくとも1つの標的鋳型核酸分子のシーケンスの少なくとも一部を決定するための方法が提供される。

10

【0011】

本発明の更なる態様は、従属請求項及び「発明を実施するための形態」で提供される。

【図面の簡単な説明】

【0012】

【図1】本発明による、少なくとも1つの標的鋳型核酸分子の少なくとも一部を決定するための方法の一実施形態を示す。

【図2】本発明による、2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定するための方法を示す。

【図3】変異シーケンスリードにおいて1つ以上の変異の位置を決定するステップの例を示す。

20

【図4A】本発明の方法を使用しない、2.3Mbpの *Arcobacter butzleri* ゲノムの短リードアセンブリの比較例を示す。

【図4B】本発明の方法を使用して、2.3Mbpの *Arcobacter butzleri* ゲノムをアセンブルする一例を示す。

【図5】本発明の方法の結果における長鋳型ごとの短リードのカバレッジ深度の効果に関する実験データを示す。

【発明を実施するための形態】

【0013】

一般的な定義

30

別途定義されない限り、本開示書で使用する全ての技術用語及び科学用語は、本発明が属する技術分野の当業者に一般的に理解されるものと同じ意味を有する。

【0014】

一般に、用語「含む」は、含むが、限定されないことを意味することを意図する。例えば、語句「[特定のステップ]を含む方法」は、列挙されたステップを含むが、追加ステップが実行され得ることを意味すると解釈されるべきである。

【0015】

本発明のいくつかの実施形態では、単語「含む」は、語句「からなる」で置き換えられる。用語「からなる」は限定的であることが意図される。例えば、語句「[特定のステップ]からなる方法」は、当該方法が列挙されたステップを含み、追加ステップは実行されないことを意味すると理解されるべきである。

40

【0016】

いくつかの態様では、本発明は、少なくとも1つの標的鋳型核酸分子のシーケンスの少なくとも一部を決定する、又は生成するための方法を提供する。この方法は、少なくとも1つの標的鋳型核酸分子の完全なシーケンスを決定する、又は生成するために使用され得る。あるいは、この方法は、部分シーケンス、すなわち、少なくとも1つの標的鋳型核酸分子の一部のシーケンスを決定する、又は生成するために使用され得る。例えば、完全なシーケンスを決定することができないか、又は容易ではない場合、ユーザは、少なくとも1つの標的鋳型核酸分子の一部のシーケンスが、自身の目的に有用である、又は十分であると決定することがある。

50

【0017】

本発明の目的のために、「核酸分子」（又は「非変異核酸分子」）は、任意の長さのヌクレオチドのポリマー形態を指す。ヌクレオチドは、デオキシリボヌクレオチド、リボヌクレオチド、又はそれらの類似体であり得る。好ましくは、少なくとも1つの核酸分子は、デオキシリボヌクレオチド又はリボヌクレオチドで構成される。更により好ましくは、少なくとも1つの核酸分子は、デオキシリボヌクレオチドで構成される、すなわち、少なくとも1つの核酸分子は、DNA分子である。

【0018】

「標的鑄型核酸分子」は、ユーザがシーケンスしようとする任意の核酸分子であり得る。少なくとも1つの「標的鑄型核酸分子」は一本鎖であり得る、又は二本鎖複合体の一部であり得る。少なくとも1つの標的鑄型核酸分子がデオキシリボヌクレオチドで構成される場合、これは二本鎖DNA複合体の一部を形成し得る。その場合、一方の鎖（例えば、コード鎖）は、少なくとも1つの標的鑄型核酸分子であるとみなされ、他方の鎖は、少なくとも1つの標的鑄型核酸分子に相補的な核酸分子である。少なくとも1つの標的鑄型核酸分子は、遺伝子に対応するDNA分子であり得、イントロンを含み得、遺伝子間領域であり得る、複数遺伝子にわたる遺伝子内領域であり得る、又は実際には、生物のゲノム全体であり得る。

10

【0019】

本発明の目的のために、「変異核酸分子」又は「変異標的鑄型核酸分子」は、変異が導入された「核酸分子」又は「標的鑄型核酸分子」を指す。変異は、置換変異、任意選択的に遷移変異であり得る。本発明の目的のために、用語「置換変異」は、あるヌクレオチドが異なるヌクレオチドで置換されることを意味すると解釈されるべきである。例えば、シーケンスATCCのシーケンスAGCCへの変換は、単一の置換変異を導入する。本発明の目的のために、用語「遷移変異」は、ヌクレオチドAがヌクレオチドGで置換され、かつその逆も同様である

20

【0020】

【数1】

（すなわち、変異 $A \leftrightarrow G$ ）

こと、又はヌクレオチドCがヌクレオチドTで置換され、かつその逆も同様である

30

【0021】

【数2】

（すなわち、変異 $C \leftrightarrow T$ ）

ことを意味すると解釈されるべきである。

【0022】

語句「少なくとも1つの標的鑄型核酸分子に変異を導入する」は、一对のサンプルの第2のサンプル中の少なくとも1つの標的鑄型核酸分子を、少なくとも1つの標的鑄型核酸分子に変異する条件に曝露することを指す。これは、任意の適切な方法を使用して達成され得る。例えば、変異は、化学的変異誘発及び／又は酵素による変異誘発によって導入され得る。

40

【0023】

本発明の目的のために、「変異を含むシーケンス」は、「変異核酸分子」又は「変異標的鑄型核酸分子」中のヌクレオチドのシーケンスの少なくとも一部に対応する。「変異を含むシーケンス」はまた、「変異シーケンス」と称され得る。「変異を含むシーケンス」は、本明細書では、 μ^i と示され、複数の（すなわち、多数の）「変異を含むシーケンス」はMと示され、 $\mu^1 \dots \mu^n \in M$ である。「変異を含まないシーケンス」は、「核酸分子」又は「標的鑄型核酸分子」中のヌクレオチドのシーケンスの少なくとも一部に対応する。「変異を含まないシーケンス」はまた、「非変異シーケンス」と称され得る。「変異を含まないシーケンス」は、本明細書では、 S^i と示され、複数の（すなわち、多数の

50

「変異を含まないシーケンス」は S と示され、 $S^1 \dots S^n \in S$ である。「変異を含むシーケンス」及び「変異を含まないシーケンス」はしたがって、ヌクレオチド (nt) アデニン (A)、チミン (T)、グアニン (G)、及びシトシン (C) の核酸分子シーケンスの少なくとも一部に対応し得る。そのような染色体シーケンスのサイズは、 $10^3 \sim 10^9$ ヌクレオチド (nt) 超の範囲の長さであり得る。

【0024】

本発明の目的のために、「変異シーケンスリード」は、「変異を含むシーケンス」のサブシーケンスに対応する。すなわち、「変異シーケンスリード」は、少なくとも「変異を含むシーケンス」のサブシーケンスと実質的に同一であり得るが、変異を含むシーケンスと比較して変異を含み、リードエラーに起因した小さな違いを更に含み得る。「変異シーケンスリード」は p^i と示され、複数の（すなわち、多数の）「変異シーケンスリード」は、 P で示され、 $p^1 \dots p^n \in P$ である。「非変異シーケンスリード」は、「変異を含まないシーケンス」のサブシーケンスに対応する。すなわち、「非変異サブシーケンスリード」は、シーケンシング中のリードエラーを除いて、「変異を含まないシーケンス」のサブシーケンスと実質的に同一であり得る。「非変異シーケンスリード」は r^i と示され、複数の（すなわち、多数の）「非変異シーケンスリード」は、 R で示され、 $r^1 \dots r^n \in R$ である。「変異シーケンスリード」は、「変異標的鋳型核酸分子」の領域をシーケンスすることによって得ることができ、「非変異シーケンスリード」は、「標的鋳型核酸分子」の領域をシーケンスすることによって得ることができる。シーケンスリードは、シーケンスよりも小さい長さ、例えば、約150 ntの長さを有し得る。

【0025】

シーケンス分析方法10

図1は、本発明による、少なくとも1つの標的鋳型核酸分子の少なくとも一部を決定する方法10を示す。

【0026】

少なくとも1つの標的鋳型核酸分子の少なくとも一部を決定する方法10は、サンプル調製のステップS110を含み得る。サンプル調製のステップS110は、一对の標的鋳型核酸分子を提供することと、変異標的鋳型核酸分子を生成するために、一对の標的鋳型核酸分子のうちの1つに変異を導入することと、を含み得る。サンプル調製のステップS110は、標的鋳型核酸分子及び変異標的鋳型核酸分子を提供するための任意の既知の技術を含み得る。

【0027】

少なくとも1つの標的鋳型核酸分子の少なくとも一部を決定する方法10は、シーケンシングのステップS120を更に含み得る。シーケンシングのステップS120は、変異を含む少なくとも1つの標的鋳型核酸分子の領域をシーケンスし、それによって複数の変異シーケンスリード P を提供することを含む。加えて、シーケンシングのステップS120は、少なくとも1つの（非変異）標的鋳型核酸分子（変異標的鋳型核酸分子に対応する標的鋳型核酸分子）の領域をシーケンスし、それによって複数の非変異シーケンスリード R を提供することを含み得る。ステップS120は、複数の変異シーケンスリード P を生成するための任意の既知の技術を含み得る。

【0028】

少なくとも1つの標的鋳型核酸分子の少なくとも一部を決定する方法10は、2つの変異シーケンスリード p^i 、 p^j が同一の変異を含むシーケンス μ^i に由来する（又はそれから生じる）かどうかを決定するステップ200又は方法200を含む。2つの変異シーケンスリード p^i 、 p^j が同一の変異を含むシーケンス μ^i に由来する（又はそれから生じる）かどうかを決定することは、2つの変異シーケンスリード p^i 、 p^j が変異を含むシーケンス μ^i の同一部分、又は類似部分、又は重複部分に由来する（又はそれから生じる）かどうか、すなわち、2つの変異シーケンスリード p^i 、 p^j の両方が、変異を含むシーケンス μ^i の同一部分に対応するシーケンスを含むかどうかを決定することを含む。方法200は、コンピュータ実施方法であり、コンピュータのプロセッサによって実行され得

る。方法200は、2つの変異シーケンスリード p^i 、 p^j が、同一の変異を含むシーケンス μ^i に由来する確率と相関する尺度を作成する。

【0029】

少なくとも1つの標的鋳型核酸分子の少なくとも一部を決定する方法10は、シーケンスアセンブリのステップS300を更に含む。シーケンスアセンブリのステップS300は、シーケンス μ^i の少なくとも一部 S^i をアセンブルするか、又は再構築することを含む。変異を含むシーケンス μ^i は、それぞれ2つの変異シーケンスリード p^i 、 p^j が同一の変異を含むシーケンス μ^i に由来する確率と相関する尺度に基づいて、複数の変異シーケンスリードPをアセンブルすることによって得ることができる。これは、例えば、複数の変異シーケンスリードPを、変異を含むシーケンス μ^i に対応するグループに分類し、次いで、各グループを別個にアセンブルして、個々の変異を含むシーケンス μ^i の一部又は全体を別個に再構築することによって達成され得る。変異を含まないシーケンス S^i は、変異を含むシーケンス μ^i のエラー補正によって、例えば、複数の非変異シーケンスリードRを使用して、変異を含むシーケンス μ^i から、最も可能性の高い、変異を含まないシーケンス S^i を推定することによって得ることができる。シーケンスアセンブリのステップS300は、それぞれ2つの変異シーケンスリード p^i 、 p^j が同一の変異を含むシーケンス μ^i に由来する確率と相関する尺度に基づいて、複数の変異シーケンスリードPから変異を含むシーケンス μ^i をアセンブリするための任意の既知の方法を含み得る。

【0030】

図2は、本発明による、2つの変異シーケンスリード p^i 、 p^j が同一の変異を含むシーケンス μ^i に由来するかどうかを決定する方法200を示す。

【0031】

方法200は、複数の変異シーケンスリード $p^1, \dots, p^n \in P$ を受け取るステップS210を含む。各変異シーケンスリード p^i は、変異を含むシーケンス μ^i のサブシーケンスに対応する。変異を含むシーケンス μ^i は、変異を含まないシーケンス S^i と比較して、変異、例えば置換変異、任意選択的に遷移変異を含む。変異を含むシーケンス μ^i は変異標的鋳型核酸分子のシーケンスの少なくとも一部であり得、変異を含まないシーケンスは、(非変異) 標的鋳型核酸分子の少なくとも一部であり得、変異標的鋳型核酸分子は、変異、例えば置換変異、任意選択的に遷移変異を標的鋳型核酸分子に導入することによって得られる。変異を含むシーケンス μ^i の各サブシーケンスは、変異標的鋳型核酸分子のフラグメントのシーケンスの少なくとも一部であり得る。変異を含まないシーケンス S^i の各サブシーケンスは、標的鋳型核酸分子のフラグメントのシーケンスの少なくとも一部であり得る。複数の変異シーケンスリードPを受け取るステップS210は、変異標的鋳型核酸分子のシーケンシングに使用されるシーケンシングマシンから複数の変異シーケンスリードPを直接受け取ること、又は、複数の変異シーケンスリードPを記憶するデータ記憶装置から複数の変異シーケンスリードPを受け取ることを含み得る。

【0032】

方法200は、共通ミニマイザ関数を各変異シーケンスリード p^i に適用するステップS220を更に含む。共通ミニマイザ関数を適用すると、変異シーケンスリード p^i ごとに1つ以上のそれぞれのミニマイザが決定される。方法200は、各変異シーケンスリード p^i において1つ以上のそれぞれのミニマイザの位置を決定するステップS222を更に含む。

【0033】

好ましい実施形態では、方法200は、変異シーケンスリードPをそれぞれのミニマイザによってビンニングするステップS224を含む。2つ以上のミニマイザが決定された変異シーケンスリード p^i は、多数のそれぞれのミニマイザビン内に提供され得る。

【0034】

方法200は、各変異シーケンスリード p^i において1つ以上の変異の位置を決定するステップS230を更に含む。各変異シーケンスリード p^i において1つ以上の変異の位置を決定するステップS230は、共通ミニマイザ関数に関連するステップS220、S

2 2 2、及び S 2 2 4 の前後に、又は同時に実行され得る。

【0035】

方法 2 0 0 は、共通ミニマイザを有する少なくとも 2 つの変異シーケンスリード p^i 、 p^j について、それぞれのミニマイザがアラインされると、すなわち、一方の変異シーケンスリード p^i のミニマイザの位置が他方の変異シーケンスリード p^j の位置と同一であるように、一方の変異シーケンスリード p^i のヌクレオチドの位置が他方の変異シーケンスリード p^j のヌクレオチドの位置に対して変更されると、位置が一致する、及び／又は位置が一致しない変異の数をカウントすることを更に含む。位置が一致する、及び／又は位置が一致しない変異の数は、2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度であり得る。あるいは、方法 2 0 0 は、位置が一致する、及び／又は位置が一致しない変異の数に基づいて、2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定する、更なるステップ 2 4 2 を含む得る。

10

【0036】

複数の変異シーケンスリードを受け取るステップ S 2 1 0

ステップ S 2 1 0 は、複数の変異シーケンスリード $p^1 \dots p^n \in P$ を受け取ることを含む。ステップ S 2 1 0 は、複数の非変異シーケンスリード $r^1 \dots r^m \in R$ を受け取ることを更に含む得る。各変異シーケンスリード p^i は、変異を含むシーケンス μ^i のサブシーケンスに対応し得る。各非変異シーケンスリード r^i は、変異を含まないシーケンス S^i のサブシーケンスに対応し得る。

20

【0037】

変異を含むシーケンス μ^i は、変異を含まないシーケンス S^i に変異を導入することによって得ることができる。各変異シーケンスリード p^i は、したがって、変異を含み得る。すなわち、変異を含む変異標的鋳型核酸分子の領域、すなわち、変異を含むシーケンスのサブシーケンスに対応する。一実施形態では、各変異シーケンスリード p^i は置換変異を含む。すなわち、置換変異を含む変異標的鋳型核酸分子の領域に対応する。好ましい実施形態では、置換変異は遷移変異であり、したがって各変異シーケンスリード p^i は遷移変異を含む。すなわち、遷移変異を含む変異標的鋳型核酸分子の領域に対応する。

【0038】

各シーケンスリード p^i 、 r^i の各ヌクレオチドは、好ましくは、2 ビットを使用してバイナリ形式で符号化される。特に、複数の変異シーケンスリード P が遷移変異

30

【0039】

【数 3】

($A \leftrightarrow G$ 及び $C \leftrightarrow T$)

を含む場合、2 ビットのうちの 1 つ（例えば、第 1 のビット）は、ヌクレオチドがプリン（A 又は G）であるか、又はピリミジン（T 又は C）であるかを定義することは有利である。例えば、ヌクレオチドは、A : 0 0、G : 0 1、C : 1 0 及び T : 1 1 という形式を使用して、バイナリで符号化され得る。この符号化は、本明細書全体を通して使用される。しかしながら、本発明はこの符号化に限定されるものではなく、本発明は、任意の他のヌクレオチドの符号化を使用して容易に実行され得ることは明らかであろう。

40

【0040】

各シーケンスリード p^i 、 r^i は、シーケンスリード p^i 、 r^i におけるホモポリマーエラーを説明するために符号化され得る。ホモポリマーエラーは、同一ヌクレオチドのランが誤った長さにおいて読み間違えられるときに生じ得、例えば、シーケンス T A A A A G C は、多数の A のラン中に A ' の数を区別することが困難であるために T A A G C と読み間違えられる得る。そのようなホモポリマーエラーを説明するために、多数の同一ヌクレオチドのランは、ヌクレオチドの単一インスタンスとして符号化され得る。あるいは、ホモポリマーエラーは、シーケンスリード p^i 、 r^i の後続の処理（すなわち、初期符号化ではない）中に、例えば、方法 2 0 0 に使用される任意の k -mer、及び／又はステップ S

50

230で使用される任意のシードパターンを符号化することにより説明され得、その結果、複数の同一ヌクレオチドのランがヌクレオチドの単一インスタンスとして符号化される。

【0041】

ステップS220及びS222：共通ミニマイザ関数

ミニマイザは、 $k\text{-mer}$ のセットで共通ミニマイザ関数 $\min(\cdot)$ を満たす $k\text{-mer}$ のセットからの $k\text{-mer}$ である。

【0042】

本出願の目的のために、 $k\text{-mer}$ は、長さ k のヌクレオチドサブシーケンスである。長さ n のシーケンス $S = [S_1, S_2, \dots, S_{n-1}, S_n]$ における位置 i で始まる $k\text{-mer}$ は $k(S_i)$ と示され、式中、 $k(S_i) = [S_i, S_{i+1}, \dots, S_{i+k-1}]$ である。 i と j との間に開始位置を有する、シーケンス S 内の $k\text{-mer}$ のセットは、 $k(S_i \dots S_j)$ と示される。シーケンス S の $i \sim j$ の範囲の開始位置を有する全ての $k\text{-mer}$ からのミニマイザは、 $\min(k(S_i \dots S_j))$ と示されるであろう。

【0043】

共通ミニマイザ関数 $\min(\cdot)$ は、シーケンスリード p^i 、 r^i からなる $k\text{-mer}$ のセット、好ましくは全ての又は実質的に全ての $k\text{-mer}$ 、すなわち、シーケンスリード p^i 、 r^i 内に存在する $k\text{-mer}$ 、好ましくは全て又は実質的に全ての $k\text{-mer}$ から1つ以上のミニマイザ（すなわち、1つ以上の代表的な $k\text{-mer}$ ）を決定するために使用される。本発明の目的のために、シーケンスリード p 内に存在する $k\text{-mer}$ のセットは、シーケンスリード p^i 、 r^i の逆補体の $k\text{-mer}$ を含み得る。好ましくは、各ミニマイザは、5以上（すなわち、5mer以上）、好ましくは10以上（すなわち10mer以上）、更に好ましくは15以上（すなわち15mer以上）の長さの $k\text{-mer}$ である。各ミニマイザは、50未満、任意選択的に30未満、更に任意選択的に25未満の長さの $k\text{-mer}$ であり得る。共通ミニマイザ関数 $\min(\cdot)$ がより長いミニマイザを決定するために使用される場合、決定されるミニマイザは、シーケンスの特定の部分を表す可能性が高い。すなわち、ミニマイザが、シーケンスの多数の別個の、無関係な部分で生じる可能性は低い。ミニマイザのサイズに上限を設定すると、ミニマイザがシーケンシングエラーを含むリスクは低減する。

【0044】

共通ミニマイザ関数 $\min(\cdot)$ を適用するステップS220は、それぞれの変異シーケンスリード p^i において、可能な $k\text{-mer}$ の順序付きリストに最初に列挙されている、1つ以上の $k\text{-mer}$ を特定することを含み得る。それぞれの変異シーケンスリード p^i について決定された1つ以上のミニマイザは、特定された1つ以上の $k\text{-mer}$ であり得る。可能な $k\text{-mer}$ の順序付きリストは、所定の順序で、全て又はいくつかの可能な $k\text{-mer}$ を含み得る。ステップS220は、可能な $k\text{-mer}$ の順序付きリストを生成することを含み得るか、又は可能な $k\text{-mer}$ の順序付きリストを生成することを含み得ない（例えば、以下のいくつかの例のように、リストと直接比較してミニマイザを決定することが必要とされない状況において）。

【0045】

例えば、共通ミニマイザ関数 $\min(\cdot)$ は、変異シーケンスリード p^i において全ての $k\text{-mer}$ の2ビットバイナリ符号化の最小整数を有する $k\text{-mer}$ をミニマイザとして決定し得る。換言すると、共通ミニマイザ関数 $\min(\cdot)$ は、2ビットバイナリ符号化の整数値によって順序付けられた $k\text{-mer}$ のリストに最初に列挙されている $k\text{-mer}$ を特定し得る。例えば、A:00、G:01、C:10、及びT:11というバイナリ符号化に基づいて、共通ミニマイザ関数は、例示的な順序付きリストAAAAA, AAAAG, AAAAC, AAAAT, AAAGA, AAAGG, ..., CTTTC, CTTT, TTTTTで最初に列挙されている変異シーケンスリードにおいて5merを特定することができる。例えば、例示的な変異シーケンスリード:

A C G G A A A G C G C T A C G A G C G A C T G A T A T T G A G C T A C G T T C

AGAGCC

は、5mer ACGGA、CGGAA、GGAAA、... AGAGC、GAGCCを含む。5mer AAAAGCは、上記の例示的な順序付きリストで最初に列挙されており、共通ミニマイザ関数 $\min(\cdot)$ は、この例示的な変異シーケンスリードのミニマイザとしてAAAAGCを特定する。この共通ミニマイザ関数 $\min(\cdot)$ のために、可能な k -merの順序付きリストを実際に生成して、 k -merセットのミニマイザを決定する必要はないことが理解されよう。

【0046】

変異シーケンスリード p^i において全ての k -merの2ビットバイナリ符号化の整数最小値を決定することは、変異シーケンスリード i に適用されて共通ミニマイザを決定し得る共通ミニマイザ関数 $\min(\cdot)$ の一例に過ぎない。任意の他の共通ミニマイザ関数 $\min(\cdot)$ が使用され得る。例えば、共通ミニマイザ関数 $\min(\cdot)$ にとって、integer minimum関数の順序をランダム化することは有利である。そのようなランダム化を達成するための1つの方法は、最初に、変異シーケンスリード p^i からなる各 k -merに対して、任意のビットベクトルを用いてビット単位の論理XORを適用することであり、その後、integer minimum関数を使用することができる。

【0047】

あるいは、可能な k -merの順序付きリストの代わりに、可能な k -merの所定のセットを使用することができ、共通ミニマイザ関数 $\min(\cdot)$ を適用することは、可能な k -merの所定のセットに存在する1つ以上の k -merを特定することを含む。それぞれの変異シーケンスリード p^i について決定された1つ以上のミニマイザは、特定された1つ以上の k -merであり得る。可能な k -merの所定のセットは、順序付き又は順序無しであり得る。可能な k -merの所定のセットは、ミニマイザとして使用されることが好適であるか、又は意図される k -merのみを含む、 k -merのセットであり得る。共通ミニマイザ関数 $\min(\cdot)$ を適用するステップS220は、可能な k -merの所定のセットを作成することを含み得る。

【0048】

好ましい実施形態では、可能な k -merの順序付きリストにおいて、 k -merは、 k -merが変異を含むシーケンス μ^i に存在し、変異を含まないシーケンス S^i には存在しない確率に基づいて順序付けられる。すなわち、変異を含むシーケンスに存在し、変異を含まないシーケンスに存在しない可能性が比較的高い k -merは、順序付きリストの前の方に列挙され得、変異を含むシーケンスに存在し、変異を含まないシーケンスに存在する可能性が比較的低い k -merは、リストの後の方に列挙され得る。代替の好ましい実施形態では、可能な k -merの所定のセットは、変異を含むシーケンスに存在し、変異を含まないシーケンスには存在しない可能性が比較的高い k -merを含み、任意選択的に、変異を含まない変異を含むシーケンスに存在し、変異を含まないシーケンスに存在しない可能性が比較的低い k -merを含まない。ステップS220は、例えば、複数の変異シーケンスリードPにおける k -merの発生（又は観測）数を、複数の非変異シーケンスリードRにおける k -merの発生（又は観測）数と比較することによって、変異を含むシーケンスに存在し、変異を含まないシーケンスに存在しない傾向が比較的高い、複数の変異シーケンスリードPからなる k -merを決定することを含み得る。これは、複数の変異シーケンスリードPにおける k -merの発生数をカウントすることと、複数の非変異シーケンスリードRの発生数をカウントすることとを含み得る。

【0049】

両方の好ましい実施形態では、共通ミニマイザ関数 $\min(\cdot)$ は、優先的に、非変異シーケンスリード r^i よりも変異シーケンスリード p^i において生じる可能性が高い、1つ以上のミニマイザ、すなわち k -merとして決定するように、選択される。これにより、各ミニマイザが変異を含む可能性が向上する。

【0050】

より好ましい実施形態では、可能な k -merの順序付きリストは、複数の非変異リー

10

20

30

40

50

ドRよりも複数の変異シーケンスリードPに存在することが多い（又は変異を含まないシーケンスより変異を含むシーケンスに存在することが多い） $k\text{-mer}$ 、すなわち、複数の変異シーケンスリードPで生じる数は複数の非変異シーケンスリードRで生じる数よりも多い $k\text{-mer}$ のみを含む。すなわち、そのような $k\text{-mer}$ からなる。代替のより好ましい実施形態では、可能な $k\text{-mer}$ の所定のセットは、複数の非変異リードRよりも複数の変異シーケンスリードPに存在することが多い（又は変異を含まないシーケンスより変異を含むシーケンスに存在することが多い） $k\text{-mer}$ 、すなわち、複数の変異シーケンスリードPで生じる数は非変異シーケンスリードRで生じる数よりも多い $k\text{-mer}$ のみを含む。すなわち、そのような $k\text{-mer}$ からなる。好ましくは、可能な $k\text{-mer}$ の順序付きリスト又は可能な $k\text{-mer}$ の所定のセットは、複数の変異リードに n 倍以上存在し、複数の非変異シーケンスリードに n 倍未満存在する $k\text{-mer}$ 、すなわち、複数の変異シーケンスリードPで生じる数が n 以上であり、複数の非変異シーケンスRで生じる数は n 未満である $k\text{-mer}$ のみを含む。すなわち、そのような $k\text{-mer}$ からなる。 n は、1以上の整数であり得る。 n は、2以上の整数であり得る。好ましくは、 n は2である。更に好ましくは、可能な $k\text{-mer}$ の順序付きリスト又は可能な $k\text{-mer}$ の所定のセットは、複数の非変異シーケンスリードには存在しない $k\text{-mer}$ 、すなわち、複数の非変異シーケンスリードRで生じる数が0である $k\text{-mer}$ のみを含む。すなわち、そのような $k\text{-mer}$ からなる。

【0051】

例えば、可能な $k\text{-mer}$ の順序付きリスト又は可能な $k\text{-mer}$ の所定のセットは、複数の変異シーケンスリードPの $k\text{-mer}$ のセットに少なくとも2回存在するが、複数の非変異シーケンスリードRの $k\text{-mer}$ のセットには決して存在しない（又はほとんど存在しない） $k\text{-mer}$ のみを含み得る。これにより、可能な $k\text{-mer}$ の順序付きリスト又は可能な $k\text{-mer}$ の所定のセットは、複数の変異シーケンスリードPのうちの2つ以上に存在する変異を含むミニマイザを高確率で含むようになる。任意選択的に、複数の非変異シーケンスリードよりも変異シーケンスリードに存在することの多い $k\text{-mer}$ は、変異を含むシーケンスに存在する可能性が比較的高い。任意選択的に、複数の変異リードに n 倍以上存在し、複数の非変異シーケンスリードに n 倍未満存在する $k\text{-mer}$ は、変異を含むシーケンスに存在する可能性が比較的高い。

【0052】

可能な $k\text{-mer}$ の所定のセットは、変異ミニマイザ U_M のセットを構築することによって作成され得、 U_M は $k\text{-mer}$ 、好ましくは全て、又は実質的に全ての $k\text{-mer}$ を含み、複数の変異シーケンスリードPにおける $k\text{-mer}$ の発生数、すなわち観測数は n 以上であり（好ましくは、 $n \geq 2$ 、より好ましくは n は2である）、複数の非変異シーケンスリードPにおける発生数、すなわち観測数は n 未満である（好ましくは、 n は0又は1であり、より好ましくは、 n は0である）。変異ミニマイザ U_M のセットは、複数の非変異シーケンスリードR及び複数の変異シーケンスリードPに各 $k\text{-mer}$ が存在する回数をカウントすることによって作成され得る。変異ミニマイザのセット U_M は、Bloomフィルタ、又は関連するCuckooフィルタ法及びQuotient法など確率的データ構造を使用して、複数の非変異シーケンスリードR及び複数の変異シーケンスリードPから効率的に計算され得る。可能な $k\text{-mer}$ の順序付きリストは、変異ミニマイザ U_M のセット全体から作成され得る。

【0053】

変異ミニマイザ U_M のセットは、可能な $k\text{-mer}$ の所定のセットとして使用され得る。あるいは、変異ミニマイザ U_M のセットは、可能な $k\text{-mer}$ の所定のセットを作成するために更に処理され得る。好ましい実施形態では、変異ミニマイザ U_M のセットのサブセット W_M は、可能な $k\text{-mer}$ の所定のセットとして使用される。サブセット W_M は、各変異リード $p^i \in P$ を2つ以上の（任意選択的に、実質的に等サイズの）非重複セクション、例えば、サイズ L_w の $k\text{-mer}$ 開始位置の非重複セット（例えば $\{1 \dots L_w\}$ 、 $\{L_w + 1 \dots 2L_w\}$ など）に分割することによって構成され得る。長さ150

10

20

30

40

50

の変異シーケンスリードが使用される場合の典型的な L_w の値は 50 であり得、したがって、可能な $k\text{-mer}$ 開始位置を 3 グループに分割する。次いで、開始位置の各セットについて、サブセット W_M は、以下のように示すことができる。

【0054】

【数4】

$$W_M = \bigcup_{\rho \in P} \bigcup_{j=1 \dots \lceil L/L_w \rceil} \min(k(\rho_{(j-1)L_w+1} \dots \rho_{jL_w}) \cap U_M)$$

【0055】

実際には、複数の変異シーケンスリード P のそれぞれは、2 つ以上のセクション（例えば、3 セクション）に分割され得、ミニマイザは各セクションを表すことが見出され得る。ミニマイザは、最初に、それぞれの変異シーケンスの当該セクションにおける $k\text{-mer}$ と変異ミニマイザ U_M のセットとの交点を用いて候補ミニマイザを最初に見出し、次いで、共通ミニマイザ関数を当該セットに適用して、セクションごとに 1 つのミニマイザを特定することによって決定される。

【0056】

したがって、好ましい実施形態では、共通ミニマイザ関数 $\min(\cdot)$ を各変異シーケンスリードに適用するステップ S220 は、

複数の変異リード P において、複数の変異リード P に n 倍以上存在し、複数の非変異シーケンスリード R に n 倍未満存在する $k\text{-mer}$ 、好ましくは全ての、又は実質的に全ての $k\text{-mer}$ からなる変異ミニマイザ U_M のセットを作成することであって、 n は 2 以上の整数である、ことと、

任意選択的に、複数の変異シーケンスリード P のそれぞれを 2 つ以上のセクションに分割することによって、変異ミニマイザ U_M のセットのサブセット W_M を作成し、変異ミニマイザ U_M のセットに存在する複数の変異シーケンスリード P の各セクションで、 $k\text{-mer}$ 、好ましくは全ての又は実質的に全ての $k\text{-mer}$ を特定し、複数の変異シーケンスリード P のそれぞれのセクションごとに特定した $k\text{-mer}$ のうちの 1 つをサブセット W_M に追加することであって、任意選択的に、複数の変異シーケンスリード P のそれぞれのセクションごとに特定した $k\text{-mer}$ のうちの 1 つは、複数の変異シーケンスリード P のそれぞれのセクションごとに特定した $k\text{-mer}$ に共通ミニマイザ関数 $\min(\cdot)$ を適用すること（例えば、integer minimum 又は任意の他の既知のミニマイザ関数を見出すこと）によって選択されることと、

変異ミニマイザ U_M のセット又は変異ミニマイザ U_M のセットのサブセット（サブセット W_M など）を可能な $k\text{-mer}$ の所定のセットとして使用し、複数の変異シーケンスリード P ごとに、可能な $k\text{-mer}$ の所定のセットに存在するそれぞれの変異シーケンスリード μ^1 で $k\text{-mer}$ 、好ましくは全ての $k\text{-mer}$ 又は実質的に全ての $k\text{-mer}$ を特定することであって、それぞれの変異シーケンスについて決定された 1 つ以上のミニマイザは、特定した $k\text{-mer}$ である、ことと、を含む。

【0057】

方法 200 は、各変異シーケンスリード p^i において 1 つ以上のそれぞれのミニマイザの位置 j を決定するステップ S222 を更に含む。それぞれの変異シーケンスリード p^i におけるミニマイザのそれぞれの位置 j は、それぞれのミニマイザと関連した（例えば、同じ場所又はミニマイザのピン）整数ビット値として記憶され得る。

【0058】

ステップ S224：ミニマイザビンニング

好ましい実施形態では、方法 200 は、1 つ以上のミニマイザビンにおいて変異シーケンスリード P をビンニングするステップ S224 を含む。1 つ以上のミニマイザビンにおける変異シーケンスリード P のビンニングは、1 つ以上のミニマイザビンにおいて変異シーケンスリード p^i に特有のインデックス i をビンニングすることを含む。各ミニマイザビンは

、共通ミニマイザを有する変異シーケンスリードPを含み得、共通ミニマイザを有さない変異シーケンスリードを含み得ない。位置が一致する、及び／又は位置が一致しない変異の数をカウントするステップS240は、同一ミニマイザビンの変異シーケンスリードPのみで実行され得る。これにより、ステップS240を実行する計算効率が向上する。

【0059】

換言すると、1つ以上のミニマイザはハッシュキーとして使用されて、例えば、共通ハッシュバケット（本明細書ではミニマイザビンと称される）にミニマイザを含むシーケンスリードでの追加処理（例えば、ステップS240）に備えて、これらのシーケンスリードを収集し得る。

【0060】

共通ミニマイザ関数 $\min(\cdot)$ を変異シーケンスリードPに適用することによって決定される各ミニマイザは、1つ以上のミニマイザビンにおいて変異シーケンスリードPをビンニングする目的で使用され得る。一実施形態では、可能な $k\text{-mer}$ の順序付きリスト内の各ミニマイザ、又は可能な $k\text{-mer}$ の所定のセット内の各ミニマイザ（例えば、変異ミニマイザ U_M のセット又はサブセット W_M などそのサブセット内の各ミニマイザ）は、1つ以上のミニマイザビンにおいて変異シーケンスリードPをビンニングする目的で使用され得る。

【0061】

1つ以上のミニマイザビンにおいて変異シーケンスリードPをビンニングするステップS224は、1つ以上のミニマイザビンを作成することを含み得る。これは、共通ミニマイザ関数 $\min(\cdot)$ によって決定されたミニマイザごとに1つのミニマイザビン、又は可能な $k\text{-mer}$ の所定のセット U_M 内のミニマイザ（又は $k\text{-mer}$ ）ごとに1つのミニマイザビン、又はサブセット W_M 内の $k\text{-mer}$ ごとに1つのミニマイザを作成することを含み得る。各ミニマイザビンは、RAMの連続ブロックとして実施され得る。好ましくは、ミニマイザビンの収集は、コンピュータ記憶媒体（コンピュータディスク、例えば、回転磁気ディスク又は固体ディスクなど）上のファイルとして実施され、各ビンが大量のデータを格納することを可能にする（シーケンス分析に適切なように）。

【0062】

1つ以上のミニマイザビンにおいて変異シーケンスリードPをビンニングするステップS224は、変異シーケンスリード p^i 又は変異シーケンスリード p^i に特有のインデックス i をそれぞれのミニマイザビンに記憶することを含み得る。各変異シーケンスリード p^i において1つ以上のそれぞれのミニマイザの位置 j を決定するステップS222は、それぞれのミニマイザにおいてそれぞれのミニマイザの位置 j を記憶することを含み得る。加えて、各変異シーケンスリードにおいて1つ以上の変異の位置 α を決定するステップS230によって決定されるように、各変異シーケンスリード p^i における1つ以上の変異の位置 $\alpha = \text{morphomuts}(p^i, V_R)$ は、それぞれのミニマイザビンに記憶され得る。任意選択的に、変異シーケンスリード p^i のシーケンス、シーケンスの精度に関する品質情報、又は他の情報など任意の更なる値は、下流処理に有用である場合、ミニマイザビンに記憶され得る。各変異シーケンスリード p^i に関連するこれらの値は、各ミニマイザビンにタプルとして記憶され得る。表記の目的のために、 z 番目のミニマイザビンの y 番目の要素のタプル要素 $b_{z,y}$ は、 $b_{z,y,i}$ 、 $b_{z,y,j}$ 、及び $b_{z,y,\alpha}$ と示される。各変異シーケンスリード p^i は、複数のミニマイザビンに追加され得る。

【0063】

ステップS230：変異の位置

方法200は、各変異シーケンスリード p^i において1つ以上の変異の位置 α を決定するステップS230を含む。各変異シーケンスリード p^i において1つ以上の変異の位置 α を決定するステップS230は、アライメントを行わない方法を使用して実行され得る。

【0064】

各変異シーケンスリード p^i において1つ以上の変異の位置 α を決定するステップS230は、非変異のシードマスクされた $k\text{-mer}$ のセット V_R 、すなわち、1つ以上のシ

10

20

30

40

50

ードパターン ψ が適用済みである、非変異シーケンスリード R の k -merのセットを得ることを含んでよい。非変異のシードマスクされた k -merのセット V_R を得ることは、非変異のシードマスクされた k -merのセット V_R を作成すること、又は生成することを含み得る。シードマスクされた非変異 k -merのセット V_R は、1つ以上のシードパターンの1つ1つを、変異を含まないシーケンスでの各 k -mer、例えば、非変異シーケンスリードでの各 k -merに適用することによって得ることができる、又は作成することができる。シードパターンを k -merに適用することは、シードパターンのビット単位の論理AND及び(2ビット符号化) k -merを決定することを含み得る。シードパターンを k -merに適用することにより、シードマスクされた k -merが作成される。シードマスクされた非変異 k -merのセット V_R は、

$$V_R = \{ \forall r(i) \in R \forall j = 1 - k \forall \psi \in \Psi : \psi(k(r_j^i)) \} \text{ と示され得}$$

すなわち、シードマスクされた非変異 k -merのセット V_R は、各非変異リード r^i における k -merの全て(又は実質的に全て)の位置 j (すなわち $1 \sim -k$)について、複数の非変異シーケンスリード R における全て(又は実質的に全て)の非変異リード r^i について、シードファミリー Ψ の1つ以上のシードパターン ψ のそれぞれを各 k -mer $k(r_j^i)$ に適用することによって作成される。

【0065】

シードパターンを使用して、 k -merを互いに比較するプロセスを修正することができる。シードパターンは、シードマスクされた k -merを一致とみなすために、両方の k -merで同一であることが要求される、2つの k -mer内の位置のセット(すなわち、ヌクレオチド)として定義される。シードパターンは、マスキング位置と、非マスキング位置と、を含み得る。シードパターンを k -merに適用することにより、シードマスクされた k -merが作成され、対応するシードパターンのマスキング位置に対応する、シードマスクされた k -merの位置は、任意の更なる処理(比較など)で無視されるが、対応するシードパターンの非マスキング位置に対応する、シードマスクされた k -merの位置は、任意の更なる処理(比較など)で無視されない。例えば、シードパターン $\{1, 2, 4, 6, 7\}$ は、一致とみなすために、比較対象の2つの k -mer($k(S_i)$ 及び $k(S_j)$)において1番目、2番目、4番目、6番目、及び7番目の位置(又はヌクレオチド)が同一であることを必要とする($k=7$ の場合)。2つの k -merの3番目及び5番目の位置は、任意のヌクレオチドであり得る。これは、2つのシードマスクされた k -merの3番目、及び5番目の位置がシードパターンによってマスクされることを意味する。

【0066】

任意選択的に、1つ以上のシードパターンは、1つ以上の遷移シードパターンであり得る。これは、特に変異を含むシーケンス M が、変異を含まないシーケンス S と比較して遷移変異を含むとき、すなわち、複数の変異シーケンスリード P のそれぞれが、1つ以上の遷移変異を含むときに有利である。

【0067】

遷移シードパターンは、位置が2つのクラスのみではなく、3つのクラスに分かれる特殊なタイプのシードパターンであり、各位置は、(1)厳密に一致する必要がある、又は(2)両方ともプリン又はピリミジンであり得る、又は(3)一致するために、4つのヌクレオチドのうちのいずれかであり得る。変異を含むシーケンスが遷移変異を含む場合、遷移シードパターンは特に有利である。上記で導入された2ビットヌクレオチド符号化を使用してコンピュータで実施される場合、厳密に一致するために必要な位置は、ビットマスク11として実施され得、遷移変異のみが許容される位置は10として、任意のヌクレオチドが許容される位置は00として実施される。シードパターン $\{1, 2, 4, 6, 7\}$ は、ビットマスク11110011001111として書き込まれ得る。遷移シードパターン $\{1, 2, 4, 6, 7\}$ は、ビットマスク11111011101111として書き込まれ得る。2つの k -merは、ビットマスク及び k -merの2ビット符号化のビット単位の論理ANDを個別に計算し、次いで、結果として生じる2つのシード

マスクされた $k\text{-mer}$ の同一性を確認することによって、一致について評価することができる。便宜上、ビット単位の論理 AND を用いてシードパターンを $k\text{-mer } k(S_i)$ に適用する関数は、関数 $\psi(k(S_i))$ として示されるであろう。

【0068】

一実施形態では、1つ以上のシードパターンは、複数の変異シーケンスリード P（又は変異を含むシーケンス）における任意の $k\text{-mer}$ 及び複数の非変異シーケンスリード R（又は変異を含まないリード）における対応する $k\text{-mer}$ に1つ以上のシードパターンのうちの少なくとも1つを適用することにより同一のシードマスクされた $k\text{-mer}$ を得る確率が90%超、好ましくは95%超、更に好ましくは98%超、最も好ましくは99%であるように選択される。

【0069】

1つ以上のシードパターンは、シードファミリー Ψ を構成し得る。シードファミリー Ψ は、一緒に使用すると、高確率、例えば90%超、好ましくは95%超、更に好ましくは98%超、最も好ましくは99%超の確率で特定のヌクレオチド相同性における $k\text{-mer}$ の間の一致を特定することができる2つ以上のシードパターンの集合である。シードファミリー Ψ は、シードパターン $\psi_1 \dots \psi_n \in \Psi$ を適用する n 個の異なる関数セットとして示される。シードパターン $w(\psi)$ の重みは、2つの $k\text{-mer}$ が一致とみなされるために同一であることが求められる位置の数として定義され、 $w(\psi) \leq k$ である。

【0070】

各変異シーケンスリード p^i において1つ以上の変異の位置 α を決定するステップ S230 は、変異シーケンスリードごとに、それぞれの変異シーケンスリード p^i において $k\text{-mer}$ （任意選択的に各 $k\text{-mer}$ ）に1つ以上のシードパターン ψ_j の1つ1つを適用して、複数のシードマスクされた変異 $k\text{-mer}$ を得ることを含み得る。1つ以上の変異の位置は、シードマスクされた非変異 $k\text{-mer}$ のセット V_R に存在する、複数のシードマスクされた変異 $k\text{-mer}$ のうちのシードマスクされた変異 $k\text{-mer}$ に対応するシードパターンの全てによってマスクされる変異シーケンスリード p^i において1つ以上の位置を特定することによって決定され得る。これは、変異シーケンス p^i において変異ではない位置は、シードマスクされた非変異 $k\text{-mer}$ のセット V_R に存在する、複数のシードマスクされた変異 $k\text{-mer}$ のうちのシードマスクされた変異 $k\text{-mer}$ に対応するシードパターンのうちの任意の1つによってマスクされていない変異シーケンスリード p^i における1つ以上の位置として特定され得ることを意味する。

【0071】

例えば、各変異シーケンスリード p^i の1つ以上の変異の位置 α は、以下のことによって決定され得る。

- ・ 長さ $2L$ のビットベクトル α を作成し、ビットベクトル α を0に初期化すること
- ・ 長さ $2k$ のビットベクトル b を作成し、ビットベクトル b を全て1に初期化すること
- ・ $\psi \in \Psi$ ごとに、また $1 \sim k$ の位置 j ごとに、 $\psi(k(p_j^i))$ を計算すること

。 $\psi(k(p_j^i)) \in V_R$ である場合は、 $\alpha \leftarrow \alpha \mid (\psi(b) \gg 2j)$ を割り当てること（式中、演算子 $\mid$ は、ビット単位の論理 OR 演算子を示し、演算子 $\gg$ はビット右シフト演算子を示す）。 V_R のセットメンバーシップクエリは、ハッシュテーブルのようなものを厳密に使用するか、又は Bloom フィルタ、Quotient フィルタ、若しくは類似のアプローチなど高性能の確率データ構造を近似的に使用して実施され得る。

- ・ 任意選択的に、更なる処理を容易にするために、奇数位置を除去することによって、ビットベクトル α をバイナリ2ビット変異シーケンスリード符号化の長さから変異シーケンスリード自身の長さに変換すること（例えば、 $\alpha \leftarrow \{\alpha_2, \alpha_4, \alpha_6, \dots, \alpha_{2L}\}$ ）。

- ・ 任意選択的に、更なる処理を容易にするために、ビットを反転させる論理 NOT を適用して、シードの一致が決して見出されない位置を1が表すようにする。

【0072】

上記の手順の結果はビットベクトル α であり、1を含む各位置は、高確率で変異の位置

に対応する。表記の目的のために、変異シーケンスリード p^i のビットベクトル α を計算する関数は、 $\alpha = \text{morphomuts}(p^i, V_R)$ と示される。

【0073】

図3は、単一のシードパターン $\psi = 1110110011$ を使用する例示的な変異シーケンスリード $p = \text{ACGCAAGCGCTACGAGCGACTGATATT}$ について、ビットベクトル α を得ることができる方法を示す。変異シーケンスリード p の4番目、8番目、11番目、12番目、及び16番目の位置は、変異シーケンスリード p における変異に対応する。すなわち、これらの位置のヌクレオチドは、変異を含まないシーケンスとは異なるであろう。実際には、変異シーケンスリード p は、2ビットバイナリ形式で符号化され得、シードパターン ψ の各位置は、2ビットをカバーし得る（すなわち、シードパターン ψ における各1は2つのバイナリ1として実施され、シードパターン ψ の各0は、バイナリ00又はバイナリ10として実施されるであろう）。シードマスクされていない非変異 $k\text{-mer}$ のセット V_R は、この例で以前に生成された。

10

【0074】

図3の例に示すように、シードパターン ψ は、変異シーケンスリード p の各 $k\text{-mer}$ に適用され、それによって、変異シーケンスリード p の $k\text{-mer}$ ごとに1つのシードマスクされた $k\text{-mer}$ を作成する。次いで、シードマスクされた $k\text{-mer}$ がシードマスクされた非変異 $k\text{-mer}$ のセット V_R に存在するかどうかを確認する。図示の例では、1番目、5番目、13番目、17番目、18番目、及び19番目のシードマスクされた $k\text{-mer}$ は全て、シードマスクされた非種変異 $k\text{-mer}$ のセット V_R に存在する。これらのシードマスクされた $k\text{-mer}$ は、シードパターンによってマスクされていない変異位置を含まない。

20

【0075】

次いで、1番目、5番目、13番目、17番目、18番目、及び19番目のシードマスクされた $k\text{-mer}$ を使用して、これらのシードマスクされた $k\text{-mer}$ に対応するシードパターンの全てによってマスクされた位置を特定する。変異シーケンスリード p の4番目の位置はこれらのシードパターンの全てによってマスクされ、13番目、17番目、18番目、及び19番目のシードマスクされた $k\text{-mer}$ の処理時には、変異シーケンスリード p の4番目の位置は無視される、すなわち、変異シーケンスリード p の4番目の位置は、13番目、17番目、18番目、及び19番目のシードマスクされた $k\text{-mer}$ のシードパターンによってマスクされることに留意されたい。これらのシードパターンは、いずれも変異シーケンスリード p の4番目の位置をマスクしない。したがって、変異シーケンスリード p の4番目の位置は、変異の位置として特定される。対照的に、変異シーケンスリード p の7番目の位置は、1番目、13番目、17番目、18番目、及び19番目のシードマスクされた $k\text{-mer}$ に対応するシードパターンの全てでマスクされるが、変異シーケンスリード p のこの7番目の位置は、5番目のシードマスクされた $k\text{-mer}$ に対応するシードパターンではマスクされない。したがって、変異シーケンスリード p の7番目の位置は、変異の位置として特定されない。代わりに、変異シーケンスリード p の7番目の位置は、変異ではない位置として特定される。

30

【0076】

実質的に、1番目、5番目、13番目、17番目、18番目、及び19番目のシードマスクされた $k\text{-mer}$ に対応するシードパターンの全ては、論理ORを使用して組み合わせられる。結果として得られるビットベクトルのビットは、（例えば、論理NOT演算子を使用して）反転されて、変異シーケンスリード p における変異の位置をビットベクトル α として得ることができる。

40

【0077】

リファレンスアセンブリを使用したステップ230の代替的な実施態様

上記の実施形態では、各変異シーケンスリード p^i において1つ以上の変異の位置 α を決定するステップ230は、複数の変異シーケンスリード P 及び複数の非変異シーケンスリード R を使用して、各変異シーケンスリード p^i にシードパターンを適用することに基

50

づいて実行される。

【0078】

ヒトゲノムなど大きく複雑なゲノムでは、ゲノムのかなりの割合が反復シーケンスからなる。例えば、ヒトゲノムの半分以上は、反復シーケンスの一部であると考えられている。これらの反復シーケンスは、類似の反復シーケンスの「ファミリー」に分類される。ヒトゲノムにおいて最も一般的なファミリーは、約300nt長であり、約100万のコピーで存在するSINE（短鎖散在反復配列）のAluファミリーである。別の一般的なファミリーは、1~6.5kbpのサイズの範囲の要素を有し、数万のコピー数を有する、長鎖散在反復配列（LINE）のL1ファミリーである。

【0079】

ゲノム内の反復シーケンスの様々なコピーは、例えば、1塩基の違いを含むなど、同一ではないことがある。変異の生物学のために、これらの違いは遷移の違いであることが多い。場合によっては、これらの違いは、複数の変異シーケンスリードPと複数の非変異シーケンスリードRとの間での変異の導入による違いに類似して現れることがある。これは、遷移変異を導入する場合のように、複数の変異シーケンスリードPを生成する一環として少なくとも1つの標的鋳型核酸分子に変異を導入するために使用される、変異誘発のための特定のポリメラーゼベースのアプローチの場合に特に該当する。

【0080】

結果として、複数の非変異シーケンスリードRは、いくつかの遷移の違いによってのみ互いに異なる多数のk-merを含み得る。したがって、複数の変異シーケンスリードPは、非変異シーケンスリードRと比較して変異が存在するものの、複数の非変異シーケンスリードRにおけるk-merと同一である1つ以上のk-merを含み得る。場合によっては、異なる非変異シーケンスリードRⁱにおける反復シーケンスの異なるコピー間でのこれらの天然起源の違いが、複数の変異シーケンスリードPに導入された変異を部分的に「マスク」することが可能である。これは、SINEのAluファミリーで特に顕著である。

【0081】

したがって、これらのような場合に、反復シーケンスのコピー間で意図的に導入された変異と天然の違いとをより良好に区別することができる方法の実施形態が提供されるのであれば有利であろう。

【0082】

反復シーケンスのコピー間で意図的に導入された変異と天然の違いとを区別する方法の性能を改善する第1のアプローチは、極めて高い重みでシードパターンを使用することであり、その結果、シードマスクされた変異k-merは、反復シーケンスのコピーを区別する違いを含む1つ以上の位置を含む可能性がより高くなる。第1のアプローチを採用する実施形態では、各シードパターン ψ の重み $w(\psi)$ は、50~100、好ましくは70~90の範囲である。ヒトゲノムの場合、第1のアプローチには約80の重みで十分であろう。

【0083】

しかしながら、第1のアプローチは、全ての場合において理想的というわけではない。重み80のシードパターンは非常に長く、変異シーケンスリード p^i の典型的な長さよりも長いことがあり得る。更に、高感度を確保するために必要とされるシードファミリー Ψ のサイズが非常に大きくなり得、したがって、シードパターンの全てを処理するために著しい追加の計算リソースを必要とする。最後に、インデルエラーをカバーするシードパターンの確率が増加し、インデルエラーの可能性に対応するために追加のアルゴリズムの複雑さが必要とされる。したがって、この第1のアプローチは、場合によっては好ましくないことがある。

【0084】

反復シーケンスのコピー間で意図的に導入された変異と天然の違いとを区別する方法の性能を改善する第2のアプローチは、複数の変異シーケンスリードPをリファレンスアセ

10

20

30

40

50

ンブリ（又はリファレンスゲノム）にアライン（又はマップ）することに基づいたアプローチを使用することである。リファレンスアセンブリは、例えば、Genome Reference Consortium（GRC）によって作成されたhg38ヒトゲノムなど独立して生成される、又は複数の非変異シーケンスリードRに基づいたデノボアセンブリであるのいずれかであり得る。第2のアプローチでは、各変異シーケンスリードにおいて1つ以上の変異の位置を決定するステップは、1つ以上の変異シーケンスリードについて、それぞれの変異シーケンスリードをリファレンスアセンブリにアラインすることを含む。

【0085】

このアプローチは、変異シーケンスリード p^i がペアエンドシーケンスリードである場合に特に好適であり得る。ペアエンド変異シーケンスリードをリファレンスアセンブリにアラインすることの利点は、特にSINE反復に関して、ショートリードショットガンライブラリのフラグメントサイズが、典型的には、反復シーケンスの長さよりも大きいことである。ペアエンドシーケンシングの典型的なフラグメントサイズは400~600bpであり、フラグメントの各末端から約150bpがシーケンスされる。したがって、ペアエンドシーケンスリードのペアにおけるペアエンドシーケンスリードのうちの一方が反復シーケンスに着地する場合、ペアエンドシーケンスリードのペアにおけるペアエンドシーケンスリードのうちの他方は、反復シーケンス外の一意的シーケンスに着地する可能性が高い。したがって、標準的なペアエンドアライニングプログラム（例えば、BWA-MEMなどBurrows-Wheelerアライナー）は、ペアエンドシーケンスリードのペアを、反復シーケンスの正しいコピーを含むリファレンスアセンブリ内の正しい位置に自信を持ってアラインすることができる。次いで、アラインした変異シーケンスリード p^i とリファレンスアセンブリとの間での全ての違いの位置を記録し、各変異シーケンスリード p^i にシードパターンを適用することに基づいたアプローチを使用して導出されたものに類似のビットベクトル α にそれらの位置を記憶する。それにより、それぞれの変異シーケンスリードにおいて1つ以上の変異の位置を決定することは、それぞれの変異シーケンスリードとリファレンスアセンブリとの間での違いのそれぞれの変異シーケンスリード内の位置を特定することによって達成される。

【0086】

しかしながら、複数の変異シーケンスリードPをリファレンスアセンブリにアラインすることは、場合によっては理想的ではないことがある。これは、任意の所与の標的鋳型核酸分子が、通常、リファレンスアセンブリで表されていない領域を有するためである。したがって、変異シーケンスリード p^i を、リファレンスアセンブリに表されていない領域にアラインし、アラインした変異シーケンスリード p^i とリファレンスアセンブリとの違いからビットベクトル α を導出することは不可能である。更に、リファレンスアセンブリに表されていない領域は、多くの場合、リファレンスアセンブリに対する構造バリエーションの挿入を表すため、臨床的に興味深いものである。大きな挿入領域に加えて、リファレンスアセンブリに関連する小さな挿入で生じる変異もまた、複数の変異シーケンスリードPをリファレンスアセンブリにアラインすることに基づいたアプローチで欠落し得る。

【0087】

したがって、反復シーケンスのコピー間で意図的に導入された変異と天然の違いとを識別する方法の性能を改善する、第3のハイブリッドアプローチは、複数の変異シーケンスリードPをリファレンスアセンブリにアラインすることに基づいたアプローチを、各変異シーケンスリード p^i にシードパターンを適用することに基づいたアプローチと組み合わせることである。この第3のアプローチは、本方法のステップ230の代替的实施態様として使用され得る。

【0088】

第3のアプローチでは、各変異シーケンスリードにおける1つ以上の変異の位置は、複数の変異シーケンスリードPをリファレンスアセンブリにアラインすることに基づいたアプローチ及びシードパターンを各変異シーケンスリード p^i に適用することに基づいたア

10

20

30

40

50

プローチの両方を使用して決定される。それぞれの変異シーケンスリードにおける位置がリファレンスアセンブリにアラインされる場合、それぞれの変異シーケンスリードにおける位置が、それぞれの変異シーケンスリードがリファレンスアセンブリとは異なる位置であると、それぞれの変異シーケンスリードにおける位置は、それぞれの変異シーケンスリードにおける変異の位置であると決定される。それぞれの変異シーケンスリードにおける位置がリファレンスアセンブリにアラインされない場合、それぞれの変異シーケンスリードにおける位置が、シードマスクされた非変異 $k\text{-mer}$ のセットに存在する複数のシードマスクされた変異 $k\text{-mer}$ のうちのシードマスクされた変異 $k\text{-mer}$ に対応するシードパターンの全てによってマスクされた位置であると、それぞれの変異シーケンスリードにおける位置は、それぞれの変異シーケンスリードにおける変異の位置であると決定される。

10

【0089】

これを達成するために、上記のタイプのビットベクトル α は、アライニングに基づいたアプローチ及びシードパターンを適用することに基づいたアプローチの両方を使用して導出される。シードパターンを各変異シーケンスリード p^i に適用することに基づいたアプローチによるビットベクトルは α_{mmd} と示され、複数の変異シーケンスリード P をリファレンスアセンブリにアラインすることに基づいたアプローチによるビットベクトルは α_{map} と示される。 α_{mask} と示される、更なるアライメントマスクビットベクトルも構築され、リファレンスアセンブリに正常にアラインされた各変異シーケンスリードの位置を記録する。アライメントマスクビットベクトル α_{mask} は、リファレンスアセンブリに正常にアラインされた各位置に 1 を有し、正常にアラインされなかった位置に 0 を有することになる。

20

【0090】

次いで、最終ハイブリッドビットベクトル α_{hybrid} が構築される。これは、各変異シーケンスリード p^i にシードパターンを適用することに基づいたアプローチによるビットベクトル α_{mmd} と、複数の変異シーケンスリード P をリファレンスアセンブリにアラインすることに基づいたアプローチによるビットベクトル α_{map} を以下のように組み合わせたものである。

$$\alpha_{\text{hybrid}} = \alpha_{\text{map}} \mid (\alpha_{\text{mmd}} \& \sim \alpha_{\text{mask}})$$

式中、 $\mid$ はビット単位の論理 OR 演算子を示し、 $\&$ はビット単位の論理 AND を示し、 $\sim$ はビット単位の NOT を示す。

30

【0091】

したがって、第3のアプローチは、アライメントが正常に行われた、変異シーケンスリードの位置におけるリファレンスアセンブリへのアライニングから決定された変異の位置及び他の全ての位置におけるシードパターンの適用によって決定された変異の位置を使用する。これは、リファレンスアセンブリに対する全てのタイプの挿入にも対処しつつ、高品質のリファレンスアセンブリを分析に組み込むことを可能にするという利点を有する。ヒトリファレンスゲノムなど独立した高品質のリファレンスアセンブリに対するアライニングは、デノボショートリードアセンブリに対するアライニングよりもはるかに正確であり得る。リファレンスアセンブリへのアライニングから決定された変異の位置を使用することにより、特に反復シーケンス領域において、より高精度の変異位置を推定することができ、一方では、アライメントを行わないシードパターンに基づいた方法は、リファレンスに表されていない領域における変異の位置を特定することができる。後者は、計算的に要求の高いタスクであるアセンブリの計算を必要とせずに行き得る。したがって、ハイブリッドアプローチは、いずれかのアプローチを個別に使用する場合と比較して、変異の位置の特定精度及び計算効率の改善をもたらす。

40

【0092】

バリエーション及び特定の標的鋳型核酸分子からローカルにアセンブルした領域を用いてリファレンスアセンブリを「補強」して、当該標的拡散分子に特有のアセンブリグラフを生成することも可能である。複数の変異シーケンスリード P をリファレンスアセンブリにア

50

ラインすることに基づいたアプローチによるビットベクトル (α_{map} と示される) は、変異シーケンスリードをその補強したアセンブリグラフに対してアラインすることから導出され、次いで、技術的理由又は他の理由によりアラインすることが依然として困難である、標的鋳型核酸分子の任意の領域について、各シーケンスリード p^i にシードパターンを適用することに基づいたアプローチと組み合わせられ得る。

【0093】

ステップ S 2 4 0 及び S 2 4 0 : 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度の決定

方法 2 0 0 は、共通ミニマイザを有する少なくとも 2 つの変異シーケンスリードについて、それぞれのミニマイザがアラインされると、位置が一致する、及び／又は位置が一致しない変異の数をカウントするステップ 2 4 0 を含む。

10

【0094】

これは、最初に、2 つの変異シーケンスリードのそれぞれについて、ステップ S 2 2 2 で決定したミニマイザの位置 j の違いを決定することによって達成され得る。例えば、ミニマイザビン b_z に記憶された 2 つの変異シーケンスリード p^a 及び p^c のそれぞれについてのミニマイザの位置 j の違い $a = b_z, y \cdot i$ 及び $c = b_z, x \cdot i$ は、 $d = b_z, y \cdot j - b_z, x \cdot j$ として決定され得る。

【0095】

位置が一致する変異の数をカウントすることは、2 つの変異シーケンスリードのうちの 1 つについて決定した変異の位置が d だけ右シフトされるときに、ステップ S 2 3 0 で決定した変異の位置の積集合のサイズを決定することを含み得る。例えば、 b_z, y 及び b_z, x として記憶された、2 つの変異シーケンスリード p^x 及び p^y の場合、位置が一致する変異の数は、次のように決定され得る。

20

$$\lambda_{x, y} = | \Omega (b_z, x \cdot \alpha) \cap (\Omega (b_z, y \cdot \alpha) - d) |$$

式中、 $\Omega(\alpha)$ は、非ゼロである α 内の位置インデックスのセット (すなわち、それぞれの変異シーケンスリード p^i における変異の位置のセット) として定義され、 $\Omega(b_z, y \cdot \alpha) - d$ は、 $\Omega(b_z, y \cdot \alpha)$ からの d の要素単位の減算であると理解される。積集合は、ビットシフト CPU 命令及びポップカウント CPU 命令を使用してコンピュータ上で効率的に実施することができる。

【0096】

30

位置が一致しない変異の数をカウントすることは、2 つの変異シーケンスリードのうちの 1 つについて決定した変異の位置が d だけ右シフトされるときに、ステップ S 2 3 0 で決定した変異の位置の対称的な差集合のサイズを決定することを含み得る。例えば、 b_z, y 及び b_z, x として記憶された、2 つの変異シーケンスリード p^x 及び p^y の場合、位置が一致しない変異の数は、次のように決定され得る。

$$\delta_{x, y} = | (\Omega (b_z, x \cdot \alpha) \setminus (\Omega (b_z, y \cdot \alpha) - d)) \cup ((\Omega (b_z, y \cdot \alpha) - d) \setminus \Omega (b_z, x \cdot \alpha)) |$$

【0097】

少なくとも 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定するステップ S 2 4 2 は、位置が一致する変異の数 $\lambda_{x, y}$ 及び／又は位置が一致しない変異の数 $\delta_{x, y}$ に基づき得る。一実施形態では、少なくとも 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度は、位置が一致する変異の数 $\lambda_{x, y}$ に対応する。位置が一致する変異の数 $\lambda_{x, y}$ が高いほど、少なくとも 2 つの変異シーケンスリードが同一の変異を含むシーケンスに由来する確率は高い。代替の一実施形態では、少なくとも 2 つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度は、位置が一致しない変異の数 $\lambda_{x, y}$ に対応する。位置が一致しない変異の数 $\lambda_{x, y}$ が低いほど、少なくとも 2 つの変異シーケンスリードが同一の変異を含むシーケンスに由来する確率は高い。

40

【0098】

好ましい実施形態では、少なくとも 2 つの変異シーケンスリードが、同一の変異を含む

50

シーケンスに由来する確率に相関する尺度は、i) 少なくとも2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率密度、及びii) 少なくとも2つの変異シーケンスリードが同一の変異を含むシーケンスに由来する確率密度と相関するスコア関数のうちの1つである。

【0099】

例えば、位置が一致する変異の数 $\lambda_{x,y}$ 及び位置が一致しない変異の数 $\delta_{x,y}$ は、2つのリードが変異Mを含む同一シーケンス又はそのような確率密度と相関するスコア関数に由来するモデル下で確率密度を計算するために使用され得る。1つのそのようなスコア関数は、 $\omega_{a,c} = \Delta(\lambda_{x,y}) - \Delta(\delta_{x,y})$ であり得、式中、 $\Delta(n) = (0.5n)(n+1)$ であり、 $a = b_{z,x \cdot i}$ 及び $c = b_{z,y \cdot i}$ である。このようにして、 $\omega_{a,c}$ は、2つの変異シーケンスリード p^a と p^c との間のリンクのスコア又は重みを表す。そのようなリンクの集合体は、それぞれのミニマイザビン b_z 内の変異シーケンスリード p^i の全てのペアに対して生成され得る、又はミニマイザビン b_z に多数のエントリが存在する場合、リンク重みの計算若しくは報告は、ミニマイザビン b_z においてランダムに選択されたエントリのペアにサブサンプルされ得る。

10

【0100】

ステップS300：シーケンスのアセンブル又はシーケンスの再構築

方法10は、シーケンス又は少なくともシーケンスの少なくとも一部、例えば、変異を含むシーケンス又は変異を含まないシーケンスをアセンブルする又は再構築するステップS300を更に含み得る。アセンブルされた又は再構築されたシーケンスは、変異を含むシーケンス又は変異を含まないシーケンスであり得る。

20

【0101】

方法200、例えば、シーケンスを再構築する又はアセンブルするステップS300は、複数の変異シーケンスリードから無向加重グラフを作成することを含み得る。無向加重グラフは、複数の変異シーケンスリードに対応するノードを含む。例えば、各ノードは、それぞれの変異シーケンスリードのリードインデックス i によって、又はそれぞれの変異シーケンスリードのシーケンスによって表される、それぞれの変異シーケンスリードに対応し得る。ノード間のエッジは、それぞれのエッジ重みに関連し、各エッジ重みは、それぞれのエッジに関連する2つのノードに対応する2つの変異シーケンスリードについて決定された、位置の一致する変異の数及び／又は位置の一致しない変異の数に基づいて決定され得る。各エッジ重みは、少なくとも2つの変異シーケンスリード（すなわち、エッジ重みに関連するエッジに関連するノードに対応する2つの変異シーケンスリード）が、同一の変異を含むシーケンスに由来する確率に相関する尺度に対応し得る。したがって、2つの変異シーケンスリード（ノード）を結び付けるエッジ重みは、それら2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率、又は当該確率と相関する他の何らかの任意の関数を表す。

30

【0102】

無向加重グラフは、各ミニマイザビンを順次又は並列に処理し、それによって各ミニマイジンビン内の変異シーケンスリード間のエッジを計算することによって作成され得る。エッジ重みは、スコア関数 $a, c \omega$ であり得る。

40

【0103】

エッジ重み $\omega_{a,c}$ を含む無向加重グラフは、次いで、例えば、そのような無向加重グラフを使用してシーケンスをアセンブルするための任意の既知又は未知の技術を使用した、SAMデータ（例えば、変異シーケンスリード）の処理に使用され得る。無向加重グラフからシーケンスをアセンブルすることは、例えば、変異シーケンスリードのクラスタを作成し、各クラスタ内の変異シーケンスリードをアセンブルし、変異を含むシーケンスの少なくとも一部に対応する鋳型を再構築することを含み得る。

【0104】

例えば、シーケンスの少なくとも一部を再構築する又はアセンブルする方法200、又はステップS300は、無向加重グラフでグラフクラスタリングを実行することを含み得

50

、それによって同一の変異を含むシーケンスに由来すると予想される変異シーケンスリードのクラスタを作成することができる。グラフクラスタリングは、Markovクラスタリング (MCL) 又は Infomap など任意の標準的なフローベースのグラフクラスタリングアルゴリズムを使用して実行され得る。あるいは、無向加重グラフのエッジは、いくつかの最小の重み閾値でフィルタリングされ得、次いで、グラフの結び付けられた構成要素が、同一の変異を含むシーケンスに由来する変異シーケンスリードを表すために取得される。

【0105】

シーケンスの少なくとも一部を再構築する、又はアセンブルするステップ S300 は、クラスタ内の変異シーケンスリードをアセンブルすることによって、シーケンスの少なくとも一部を再構築することを更に含み得る。例えば、クラスタ内の変異シーケンスリードは、標準的なデノボアセンブリ法に供されて、変異を含むシーケンスを再構築し得る。そのようなデノボアセンブリ法としては、例えば、「IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth」, Peng Y et al., Bioinformatics. 2012 Jun 1; 28 (11): 1420-8. doi: 10.1093/bioinformatics/bts174. Epub 2012 Apr 11 の IDBA-UD アルゴリズム、又は「SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing」, Benkevich A et al., J Comput Biol. 2012 May; 19 (5): 455-477 の SPAdes 法、又は「A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data」, Coi D et al, Bioinformatics. 2015 Feb 15; 31 (4): 587-9. doi: 10.1093/bioinformatics/btu661. Epub 2014 Oct 22 の A5-miseq 法が挙げられる。

【0106】

シーケンスを再構築する、又はアセンブルするステップ S300 は、変異を含むシーケンスの再構築部分に対するエラー補正を使用して、変異を含まないシーケンスの少なくとも一部分を再構築すること、すなわち、複数の非変異シーケンスリードを使用して、変異を含むシーケンスの再構築部分から最も可能性の高い、変異を含まないシーケンスを推測することを更に含み得る。そのようなエラー補正のための方法としては、例えば、「FMLRC: Hybrid long read error correction using an FM-index」, Jeremy R. Wang et al., BMC Bioinformatics volume 19, Article number: 50 (2018) の FMLRC 法が挙げられる。例えば、変異を含むシーケンスは、非変異シーケンスリードを使用してエラー補正に供されて、導入された変異を除去し、それによって、変異を含まないシーケンスの部分を再構築することができる。エラー補正は、例えば、変異を含むシーケンスを非変異シーケンスリードに適合する、変異を含まないシーケンスに変換することが必要とされる変異を含むシーケンスに対する可能な編集セットを決定し、可能な編集セットから最小サイズ (すなわち、最小限の編集を含む) を有する編集セットを決定し、変異を含むシーケンスに、最小サイズを有する、決定した編集セットを適用することによって、変異を含まないシーケンスの可能性の高い推定値に到達することを含み得る。変異を含まないシーケンスの部分は、Canu、Flye、若しくは Peregrine など標準的なデノボロングリードアセンブルツールを使用して、又は Unicycler 若しくは MaSuRCA などツールを使用した R 内のショートリードとのハイブリッドでアセンブルされ得、それによって、変異を含まないシーケンスをアセンブルし得る。

【0107】

サンプルプールの処理

複数のサンプルを含むサンプルバッチを処理する場合、サンプルバーコードが、サンプルごとに定義されたシーケンスタグの形態で導入され得る。方法200のユーザが、多数のサンプルで当該方法を使用することを望み、各サンプルは1つ以上の変異標的鑄型核酸分子を含む場合、1つの可能性は、実験室で各サンプルを別個に処理し（例えば、変異及び／又は断片化）、次いでシーケンシング前の最終ステップにおいてのみサンプルバーコードを導入することである。別の代替案は、標的鑄型核酸分子の末端にのみサンプルバーコードを導入することであり、この場合、サンプル調製プロセスの初期段階で全てのバーコード付き標的鑄型核酸分子をプールできるようになり、それによって試薬費及び現場に関わる人件費を大幅に低減する（いわゆる初期サンプルプーリングアプローチ）。したがって、サンプル調製は、各サンプル中の標的鑄型核酸分子の末端にそれぞれのサンプルバーコードを導入することを含み得、その結果、各サンプルは、他のサンプル中の標的鑄型核酸分子とは異なるサンプルバーコードを有する標的鑄型核酸分子を含む。サンプル調製は、サンプルをプールしてサンプルプールを作成すること、サンプルプール内の標的鑄型核酸分子を変異させ、任意選択的に断片化すること、及びサンプルプール内の変異標的鑄型核酸分子の一部をシーケンスすることを更に含み得る。

10

【0108】

初期サンプルプーリングアプローチは、結果として生じる複数の変異シーケンスリードPが、多くの異なるサンプルからの変異シーケンスリードPの非標識混合物を含むため、データ処理に更なる課題をもたらす。サンプルは、非変異シーケンスリードRを構築するために別個に処理され得、この場合、非変異シーケンスリードRは、複数セットの非変異シーケンスリード $R^1, \dots, R^{\zeta}$ を含み、式中、 ζ は、バッチで処理されたサンプルの数である。非変異シーケンスリードの各セットは、それぞれのサンプルに関連し得る。方法200は、複数セットの非変異シーケンスリード $R^1, \dots, R^{\zeta}$ 内の非変異シーケンスリードRを受け取ることを含み得、非変異シーケンスリード $R^1, \dots, R^{\zeta}$ の各セットは、それぞれ1つ又は複数のサンプルに関連する。

20

【0109】

したがって、複数の変異シーケンスリードのそれぞれは、複数のサンプルのうちの1つに関連する、変異を含むシーケンスのサブシーケンスであり得る。複数の非変異シーケンスリードのそれぞれは、複数のサンプルのうちの1つに関連する、変異を含まないシーケンスのサブシーケンスであり得る。変異を含む各シーケンスは、それぞれの変異を含まないシーケンスと比較して変異を含み得る。シードマスクされた非変異k-merのセットを得ることは、サンプルごとにシードマスクされた非変異k-merのそれぞれのセットを得ることを含み得る。

30

【0110】

ζ サンプルからデータを処理する単純なアプローチは、 ζ サンプルごとに方法200を1回適用することであろう。代替的なアプローチは、全ての ζ サンプルを同時に処理することができるように、方法200を拡張することである。これは、以下に記載されるように達成され得る。

【0111】

方法200（例えば、ステップS230）は、非変異サンプルビットベクトルのセットを作成することを含み得、各非変異サンプルビットベクトルは、シードマスクされた非変異k-merのセット V_R 内のそれぞれのk-merについて、それぞれのk-merが存在する（又は少なくともx倍存在する（式中、xは1よりも大きい整数））複数のサンプル、及びそれぞれのk-merが存在しない（又はx倍未満存在する）複数のサンプルを定義する。シードマスクされた非変異k-merのセット V_R は、上記の方法で複数の非変異k-merから生成され得る。複数の非変異k-merは、複数のサンプル $R^1, \dots, R^{\zeta}$ のそれぞれにおける全てのk-merの結合として定義され得る。すなわち、複数の非変異k-mer Rは、

40

【0112】

50

【数 5】

$$R = \cup R^1, \dots, R^{\zeta}$$

として定義され得る。

【0113】

例えば、方法200は、各サンプル中にシードマスクされた $k\text{-mer}$ が存在することを示すバイナリインジケータを含むビットベクトルの集合への V_R の全射を定義することを含み得る。各ビットベクトルは、複数のサンプル中の i 番目のサンプルが $k\text{-mer}$ を含有する（又は $k\text{-mer}$ を少なくとも X 倍含む）場合は位置 i に1を有し、そうでない場合は、位置 i に0を有し得る。ソフトウェアによる実施態様では、全射は、ハッシュマ
 ップなど順序無しのマップデータ構造、又は $\text{Counting Quotient Filter}$ など近似メンバーシップクエリ構造を使用して記憶され得る。全射は、 $Z: V_R \rightarrow v$ と示され得、式中、 v は長さ ζ のビットベクトルの空間である。

10

【0114】

各変異シーケンスリードにおいて1つ以上の変異の位置を決定するステップS230は、多数のサンプルについて同時にビットベクトル α を構築するために拡張され得る。1つ以上の変異の位置を決定することは、変異シーケンスリードごとに、並びにシードマスクされた非変異 $k\text{-mer}$ のセットごとに及び／又はシードマスクされた非変異 $k\text{-mer}$ のセットの組み合わせごとに、シードマスクされた非変異 $k\text{-mer}$ のそれぞれのセット及び／又はシードマスクされた非変異 $k\text{-mer}$ のセットのそれぞれの組み合わせに存在
 する、複数のシードマスクされた変異 $k\text{-mer}$ のうちのシードマスクされた変異 $k\text{-mer}$ に対応するシードパターンの全てによってマスクされる、変異シーケンスリード内の1つ以上の位置を特定すること、並びに特定した1つ以上の位置を、シードマスクされた非変異 $k\text{-mer}$ のそれぞれのセット又はシードマスクされた非変異 $k\text{-mer}$ のセットのそれぞれの組み合わせと関連付けることを含み得る。これは、以下のステップを含む、関数 $\text{morphomuts}(p^i, V_R)$ の $\text{MultiSampleVariantMorphomutsMS}(p^i, V_R)$ によって達成され得る。

20

1. 長さ2及び0のみを含む初期ベクトル a_0 でビットベクトル α のセット A を初期化する長さ $2k$ 、及び1のみを含むビットベクトル b を初期化する。長さ ζ 及び1のみを含む、1つの初期メンバー c_0 で、ビットベクトル c のセット C を初期化する。マッピング

30

【0115】

【数 6】

$$\Gamma: C \Leftrightarrow A$$

を初期化する。

2. $1 \sim \zeta - k$ のリード内の位置 j ごとに、

a. シードパターン $\psi \in \Psi$ ごとに、 $\psi(k(p_j^i))$ を決定する。

b. $\psi(k(p_j^i)) \in V_R$ の場合には、以下のステップを実行する。

i. C の要素ごとに（すなわち、 $c \in C$ ごとに）、 $d \leftarrow c \wedge Z(\psi(k(p_j^i)))$ を計算する。

40

ii. d が0のみを含む場合、2b.iに戻り、 C の次の要素を処理する（又はそれ以上存在しない場合、次のシードパターン ψ 又は次の位置 j を処理する）。それ以外の場合は、2b.iiiを続行する。

iii. $\alpha \leftarrow \Gamma(c) \mid (\psi(b) \gg 2j)$ を割り当てる（式中、 $\mid$ はビット単位の論理ORを示し、 $\gg$ はビットシフト右演算子を示し、 C から c を除去する）。

iv. C に d を加え、 A に α を加え、 Γ でマッピング $d \rightarrow a$ を定義する。

v. $d' \wedge c$ が非ゼロの場合、 C に $d' \wedge$ を加え、 A に $\Gamma(c)$ を加え、 Γ でマッピング $d' \wedge c \rightarrow \Gamma(c)$ を定義する。

vi. 2b.iに戻り、 C の次の要素 c を処理する。 c が C にこれ以上存在しない場合、2aに戻り、次のシードパターン ψ を処理する。これ以上 Ψ に ψ が存在しない場

50

合、2に戻り、次の位置 j を処理する。それ以外の場合は、3を続行する。

3. 関数 $morphomuts(\cdot)$ で α を作成するために適用された変換を使用して、 A でビットベクトルを変換する。及び

4. 関数の結果として、 C 、 A 、及びマッピング Γ を戻す。

【0116】

任意選択的に、 A のビットベクトルで一致する位置が少なすぎる場合（例えば、所定の y 未満（式中、 y は1以上、好ましくは2、3、4、又は5以上の整数））、 C 及び A の対応するエントリは破棄され得る。これは、そのようなエントリが入力サンプル間のランダムな類似性に起因して生じ得、したがって結果として生じるビットベクトルは、誤ったサンプルとの誤った一致の結果であるために有利である。更なる処理の前にこれらの位置を破棄することにより、不要な計算を回避することができる。方法200は、特定した位置の数を所定の数 y と比較することであって、 y は、1以上、好ましくは2以上の整数であることと、特定した位置の数が所定の数 y 未満である場合に、特定した1つ以上の位置及び1つ以上のサンプルとの特定した1つ以上の位置の関連付けを破棄する（又は更なる処理で無視する）こととを含み得る。

【0117】

ミニマイザビンに記憶されているタプルは、 C にサンプルビットベクトル情報を含むように拡張され得る。具体的には、記憶されたタプルは、リードインデックス i 、変異シーケンスリード内のミニマイザの位置 j 、並びに c 及び α であり得、 c は、サンプルビットベクトルであり、 α は、 $morphomutsMS(p^i, V_R)$ 関数によって計算される変異ベクトルである。

【0118】

その後、ミニマイザビンが処理されてエッジ重みを計算すると、各エッジ重みは、対応するサンプルセットで注釈付けされ得る。一対の変異シーケンスリードに関連するサンプルビットベクトルのビット単位のワイズ論理 AND がゼロである場合、対応するエッジは破棄され得る。エッジスコアが所定の閾値スコア未満である場合、エッジは廃棄され得る。一対の変異シーケンスリード間に多数の可能なエッジが存在する場合、最も高いエッジの重みを保持することが可能になり、このエッジのサンプルビットベクトルの関連セットは、サンプルビットベクトル間でビット単位の論理 AND として計算され得る。

このアプローチは、異なるサンプルのシーケンスで自然に発生する変異が、サンプル処理中に導入された変異と区別され得るという利点を有する。2つの変異シーケンスリードがアラインされると、位置が一致する及び／又は位置が一致しない変異の数をカウントするステップS240は、2つの変異シーケンスリードについて特定された変異の1つ以上の位置のそれぞれの対に関連するサンプル中に重複が存在する場合、すなわち、2つの変異シーケンスリードについて特定された変異の1つ以上の位置の対が少なくとも1つの共通のサンプルに関連する場合に限り、2つのシーケンスリードについて特定された変異の1つ以上の位置の任意の対について実行され得る。

【0119】

変異標的鋳型核酸分子の末端のみがサンプルタグを含む場合、複数の変異シーケンスリード P のうちの一部にこのサンプルタグが付いていることになる。具体的には、変異標的鋳型核酸分子の末端をシーケンスすることから作成された変異シーケンスリードに、サンプルタグが付いていることになる。変異シーケンスリードのクラスタリングに続いて、各クラスタ内のタグ付きサンプルリードの存在を単純に評価することによって、サンプルにリードクラスタを割り当てることが可能になる。単一のサンプルタグのみがクラスタに現れる場合、サンプルへの割り当ては単純であり、明確である。無向加重グラフでグラフクラスタリングを実行することは、変異シーケンスリードの各クラスタに、それぞれのクラスタ内の変異シーケンスリードのうちの少なくとも1つからなるサンプルタグを関連付けることを含み得る。

【0120】

時には、シーケンシング又はデータ分析手順におけるノイズ又はエラーのいずれかのた

10

20

30

40

50

めに、複数のサンプルタグが単一クラスタに現れることがある。この場合、あるサンプルタグが他のタグを大幅に上回って存在する場合、信頼性のあるサンプル割り当てが依然として可能であり得る。単一の割り当てが可能ではない場合、サンプルタグごとに1つのクラスタを用いて、マルチサンプルクラスタを多数の、より小さいクラスタに分解する半教師付きグラフ分解手順を使用してサンプルを明確にすることが依然として可能であり得る。クラスタがサンプルタグ付きリードを含まない場合でも、リードリンクに関連するサンプルマスクの大部分が全て単一のサンプルを示す場合、自信を持って当該クラスタをサンプルに割り当てることが依然として可能であり得る。無向加重グラフでグラフクラスタリングを実行することは、変異シーケンスリードの各クラスタにおいて、変異シーケンスリードのそれぞれのクラスタ内の変異シーケンスリードからなる1つ以上のサンプルタグを特定することを含み得る。変異シーケンスリードの各クラスタは、それぞれのクラスタ内の変異シーケンスリードにおいて最も頻繁に生じるサンプルタグに関連し得る。任意選択的に、2つ以上の異なるサンプルタグが変異シーケンスリードのクラスタで特定される場合、変異シーケンスリードのクラスタは、2つ以上のクラスタに分割され得、2つ以上のクラスタのそれぞれの、2つ以上の異なるサンプルタグのうちのそれぞれ1つに関連し、変異シーケンスリードのうちの異なる1つを含む。

10

【0121】

サンプルの調製及びシーケンシング

少なくとも1つの標的鋳型核酸分子のシーケンスの少なくとも一部を決定するための方法10は、複数の変異シーケンスリードを提供するために、少なくとも1つの変異を含む標的鋳型核酸分子の領域のシーケンシング100を行うことを含み得る。少なくとも1つの標的鋳型核酸分子のシーケンスの少なくとも一部を決定するための方法10は、シーケンシング100によって得た、複数の変異シーケンスリードに基づいて、2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と相関する尺度を決定する方法200を実行することを更に含み得る。

20

【0122】

シーケンシングのステップは、

a) 一对のサンプルを提供することであって、各サンプルは、少なくとも1つの標的鋳型核酸分子を含む、ことと、

(b) 一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、複数の非変異シーケンスリードを提供することと、

(c) 一对のサンプルのうちの第2のサンプル中の少なくとも1つの標的鋳型核酸分子に変異を導入して、少なくとも1つの変異標的鋳型核酸分子を提供することと、

(d) 少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスして、複数の変異シーケンスリードを提供することと、を含み得る。

30

【0123】

好ましい実施形態では、変異を導入するステップは、一对のサンプルのうちの第2のサンプル中の少なくとも1つの標的鋳型核酸分子に遷移変異を導入することを含む。

【0124】

シーケンシングのステップは、

(a) 複数対のサンプルを提供することであって、サンプルの各対は、第1のサンプルと、第2のサンプルと、を含み、各サンプルは、少なくとも1つの標的鋳型核酸分子を含む、ことと、

(b) サンプルの各対がそれぞれのサンプルバーコードに関連するように、サンプルの各対の少なくとも1つの標的鋳型核酸分子にサンプルバーコードを導入することと、

(c) 各第1のサンプル内の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、複数の非変異シーケンスリードを提供することであって、シーケンシングが、第1のサンプルごとに別々に実行され、それによって第1のサンプルごとに非変異シーケンスリードのそれぞれのセットを提供する、ことと、

(d) 第2のサンプルをプールし、それによって第2のサンプルのサンプルプールを作

40

50

成することと、

(e) サンプルプール中の標的鋳型核酸分子に変異を導入して、変異標的鋳型核酸分子を提供することと、

(d) 変異標的鋳型核酸分子の領域をシーケンスして、複数の変異シーケンスリードを提供することと、を含み得る。

【0125】

任意選択的に、シーケンスのステップは、サンプルバーコードを導入した後、かつ少なくとも1つの標的鋳型核酸分子の領域をシーケンスする前に、各第1のサンプル中の少なくとも1つの標的鋳型核酸分子を断片化することを更に含み得る。任意選択的に、シーケンスのステップは、変異標的鋳型核酸分子の領域をシーケンスする前に、サンプルプール中の少なくとも1つの標的鋳型核酸分子又は変異標的鋳型核酸分子を断片化することを更に含み得る。

【0126】

本発明の方法では、任意の数の少なくとも1つの標的鋳型核酸分子を同時にシーケンスすることができる。したがって、本発明の一実施形態では、少なくとも1つの標的鋳型核酸分子は、複数の標的鋳型核酸分子を含む。任意選択的に、少なくとも1つの標的鋳型核酸分子は、少なくとも10個、少なくとも20個、少なくとも50個、少なくとも100個、又は少なくとも250個の標的鋳型核酸分子を含む。任意選択的に、少なくとも1つの標的鋳型核酸分子は、10～1000個、20～500個、又は50～100個の標的鋳型核酸分子を含む。

【0127】

ステップS110：サンプルの調製

一対のサンプルのうちの第1のサンプル及び一対のサンプルのうちの第2のサンプルは、いずれも少なくとも1つの標的鋳型核酸分子を含むため、一対のサンプルは、同一標的生物に由来するか、又は同一の元サンプルから採取され得る。

【0128】

例えば、ユーザがサンプル中の少なくとも1つの標的鋳型核酸分子をシーケンスすることを意図する場合、ユーザは、同一の元サンプルから一対のサンプルを採取し得る。任意選択的に、ユーザは、サンプルの対が元サンプルから採取される前に、元のサンプル中の少なくとも1つの標的鋳型核酸分子を複製し得る。ユーザは、大腸菌など特定生物からの様々な核酸分子をシーケンスすることを意図し得る。この場合には、一対のサンプルのうちの第1のサンプルは、ある供給源からの大腸菌のサンプルであり得、一対のサンプルのうちの第2のサンプルは、第2の供給源からの大腸菌であり得る。

【0129】

一対のサンプルは、少なくとも1つの標的鋳型核酸分子を含むか、又は含むことが疑われる任意の供給源に由来し得る。一対のサンプルは、ヒトに由来する核酸分子のサンプル、例えば、ヒト患者の皮膚スワブから抽出されたサンプルを含み得る。あるいは、一対のサンプルは、給水設備など他の供給源に由来し得る。そのようなサンプルは、数十億の鋳型核酸分子を含み得る。本発明の方法を使用してこれら数十億の標的鋳型核酸分子のそれぞれを同時にシーケンスすることが可能であり、そのため、本発明の方法で使用され得る標的鋳型核酸分子の数に上限はない。

【0130】

一実施形態では、多数のサンプル対が提供され得る。例えば、少なくとも2、3、4、5、6、7、8、9、10、11、15、20、25、50、75、100、500、1000、又は5000のサンプル対が提供され得る。任意選択的に、10000未満、5000未満、1000未満、100未満、75未満、50未満、25未満、20未満、15未満、11未満、10未満、9未満、8未満、7未満、6未満、5未満、又は4未満のサンプルが提供される。任意選択的に、2～100、2～75、2～50、2～25、5～15、又は7～15のサンプル対が提供される。

【0131】

多数のサンプル対が提供される場合、異なるサンプル対中の少なくとも1つの標的鋳型核酸分子は、異なるサンプルタグ（本明細書ではサンプルバーコードとも称される）で標識され得る。例えば、ユーザが2対のサンプルを提供することを意図する場合、第1のサンプル対中の少なくとも1つの標的鋳型核酸分子の全て又は実質的に全ては、サンプルタグAで標識され得、第2のサンプル対中の少なくとも1つの標的鋳型核酸分子の全て又は実質的に全ては、サンプルタグBで標識され得る。

【0132】

少なくとも1つの標的鋳型核酸分子を増幅するための好適な方法は、当該技術分野で既知である。例えば、PCRが一般的に使用される。PCRについては、「少なくとも1つの標的鋳型核酸分子への変異の導入」の表題で以下により詳細に記載する。

【0133】

少なくとも1つの標的鋳型核酸分子への変異の導入

この方法は、一对のサンプルのうちの第2のサンプル中の少なくとも1つの標的鋳型核酸分子に変異を導入して、少なくとも1つの変異標的鋳型核酸分子を提供することを含み得る。

【0134】

変異は、置換変異、挿入変異、又は欠失変異であり得る。本発明の目的のために、用語「置換変異」は、あるヌクレオチドが異なるヌクレオチドで置換されることを意味すると解釈されるべきである。例えば、シーケンスATCCのシーケンスAGCCへの変換は、単一の置換変異を導入する。本発明の目的のために、用語「挿入変異」は、少なくとも1つのヌクレオチドがシーケンスに付加されることを意味すると解釈されるべきである。例えば、シーケンスATCCのシーケンスATTCCへの変換は、挿入変異の一例である（追加のTヌクレオチドが挿入されている）。本発明の目的のために、用語「欠失変異」は、少なくとも1つのヌクレオチドがシーケンスから除去されることを意味すると解釈されるべきである。例えば、シーケンスATTCCのATCCへの転換、欠失変異の一例である（Tヌクレオチドが除去されている）。好ましくは、変異は置換変異である。

【0135】

語句「少なくとも1つの標的鋳型核酸分子に変異を導入する」は、一对のサンプルの第2のサンプル中の少なくとも1つの標的鋳型核酸分子を、少なくとも1つの標的鋳型核酸分子が変異する条件に曝露することを指す。これは、任意の適切な方法を使用して達成され得る。例えば、変異は、化学的変異誘発及び／又は酵素による変異誘発によって導入され得る。

【0136】

任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、少なくとも1つの標的鋳型核酸分子のヌクレオチドの1%～50%、3%～25%、5%～20%、又は約8%を変異させる。任意選択的に、少なくとも1つの変異標的鋳型核酸分子は、1%～50%、3%～25%、5%～20%、又は約8%の変異を含む。

【0137】

ユーザは、既知のシーケンスの核酸分子に変異を導入し、結果として得られた核酸分子をシーケンスし、元のシーケンスと比較して変化したヌクレオチドの総数の割合を決定するステップを実行することによって、少なくとも1つの変異標的鋳型核酸分子内に含まれる変異の数、及び／又は少なくとも1つの標的鋳型核酸分子に変異を導入するステップが少なくとも1つの標的鋳型核酸分子を変異させる程度を決定することができる。

【0138】

任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、少なくとも1つの標的鋳型核酸分子を実質的にランダムに変異させる。任意選択的に、少なくとも1つの変異標的鋳型核酸分子は、実質的にランダムな変異パターンを含む。

【0139】

少なくとも1つの変異標的鋳型核酸分子は、実質的に類似のレベルでその長さ全体にわたって変異を含む場合、実質的にランダムな変異パターンを含む。例えば、ユーザは、既

10

20

30

40

50

知のシーケンスの試験核酸分子を変異させて変異試験核酸分子を提供することによって、少なくとも1つの変異標的鋳型核酸分子が実質的にランダムな変異パターンを含むかどうかを決定することができる。変異試験核酸分子のシーケンスは、試験核酸分子と比較されて、変異のそれぞれの位置を決定することができる。次いで、ユーザは、

(i) 変異のそれぞれの間の距離を計算すること、

(ii) 当該距離の平均を計算すること、

(iii) 500又は1000など、より小さい数に置換することなく、当該距離をサブサンプリングすること、

(iv) 幾何学的分布から500又は1000個の距離のシミュレーションセットを構築すること（平均は、平均はモーメント法によって得られ、観測した距離に関して以前に計算したものと一致する）、及び

(v) 2つの分布に対してKolmogorov-Smirnovを計算すること、によって変異試験核酸分子の長さ全体で実質的に類似のレベルで変異が生じるかどうかを決定し得る。

【0140】

少なくとも1つの変異標的鋳型核酸分子は、非変異リードの長さに応じて、 $D < 0.15$ 、 $D < 0.2$ 、 $D < 0.25$ 、又は $D < 0.3$ の場合には、実質的にランダムな変異パターンを含むと考えられ得る。

【0141】

同様に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、得られた少なくとも1つの変異標的鋳型核酸分子が実質的にランダムな変異パターンを含む場合、少なくとも1つの標的鋳型核酸分子を実質的にランダムに変異させる。少なくとも1つの標的鋳型核酸分子に変異を導入するステップが、少なくとも1つの標的鋳型核酸分子を実質的にランダムに変異させるかどうかは、既知のシーケンスの試験核酸分子で少なくとも1つの標的鋳型核酸分子に変異を導入して変異試験核酸分子を提供するステップを実行することによって決定され得る。次いで、ユーザは、変異試験核酸分子をシーケンスして、導入された変異を特定し、変異試験核酸分子が実質的にランダムな変異パターンを含むかどうかを決定し得る。

【0142】

任意選択的に、少なくとも1つの変異標的鋳型核酸分子は、不偏変異パターンを含む。任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、不偏の方法で変異を導入する。少なくとも1つの変異標的鋳型核酸分子は、導入される変異のタイプがランダムである場合、不偏変異パターンを含む。導入される変異が置換変異である場合、導入される変異は、類似の割合のA（アデノシン）ヌクレオチド、T（チミン）ヌクレオチド、C（シトシン）ヌクレオチド、及びG（グアニン）ヌクレオチドが導入される場合にランダムである。語句「A（アデノシン）ヌクレオチド、T（チミン）ヌクレオチド、C（シトシン）ヌクレオチド、及びG（グアニン）ヌクレオチドの類似の割合が導入される」では、導入されるアデノシンヌクレオチドの数、チミンヌクレオチドの数、シトシンヌクレオチドの数、及びグアニンヌクレオチドの数が、互いに20%以内であることを意味する（例えば、20Aヌクレオチド、18Tヌクレオチド、24Cヌクレオチド、及び22Gヌクレオチドが導入され得る）。

【0143】

少なくとも1つの標的鋳型核酸分子に変異を導入するステップが、少なくとも1つの標的鋳型核酸分子を不偏の方法で変異させるかどうかは、既知のシーケンスの試験核酸分子で少なくとも1つの標的鋳型核酸分子に変異を導入して変異試験核酸分子を提供するステップを実行することによって決定され得る。次いで、ユーザは、変異試験核酸分子をシーケンスして、導入された変異を特定し、変異試験核酸分子が不偏変異パターンを含むかどうかを決定し得る。

【0144】

有利には、少なくとも1つの標的鋳型核酸分子に変異を導入するステップが偏在する変

10

20

30

40

50

異を導入する場合でも、少なくとも1つの標的鋳型核酸分子のシーケンスを生成する方法が使用され得る。したがって、一実施形態では、少なくとも1つの変異標的鋳型核酸分子は、偏在する変異を含む。任意選択的に、変異を少なくとも1つの変異標的鋳型核酸分子に変異を導入するステップは、偏在する変異を導入する。変異が偏った方法で導入される、すなわち、導入されるアデノシンヌクレオチドの数、チミンヌクレオチドの数、シトシンヌクレオチドの数、及びグアニンヌクレオチドの数が互いに20%以内ではない場合、変異は「偏在する」とみなされる。少なくとも1つの変異標的鋳型核酸分子が偏在する変異を含むか、又は少なくとも1つの標的鋳型核酸分子に変異を導入するステップが、偏在する変異を導入するかどうかは、少なくとも1つの標的鋳型核酸分子に変異を導入するステップが不偏の方法で変異を導入するかどうかを決定することについて上述した方法と類似の方法で決定され得る。

10

【0145】

同様に、変異シーケンスリード及び／又は非変異シーケンスリードが偏在するシーケンシングエラーを含む場合でも、少なくとも1つの標的鋳型核酸分子のシーケンスを生成する方法が使用され得る。したがって、一実施形態では、変異シーケンスリード及び／又は非変異シーケンスリードは、偏在するシーケンシングエラーを含む。同様に、一実施形態では、少なくとも1つの標的鋳型核酸分子の領域をシーケンスする、及び／又は少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスするステップは、偏在するシーケンスエラーを導入する。

【0146】

20

少なくとも1つの標的鋳型核酸分子の領域をシーケンスする、及び／又は少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスする特定のステップが偏在するシーケンスエラーを導入するかどうかは、シーケンシング機器の精度に依存し、ユーザには既知である可能性が高い。しかしながら、ユーザは、既知の核酸分子でシーケンシング法を実行し、生成されたシーケンスリードを既知のシーケンスの元の核酸分子のシーケンスリードと比較することによって、少なくとも1つの標的鋳型核酸分子の領域及び／又は少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスするステップが、偏在するシーケンスエラーを導入するかどうかを調査し得る。次いで、ユーザは、実施例6で論じる確率関数を適用し、M及びEの値を決定し得る。E及びマトリックスモデルの値が等しくないか、又は実質的に等しくない（互いに10%以内）場合、少なくとも1つの標的鋳型核酸分子の領域をシーケンスするステップは、偏在するシーケンスエラーを導入する。

30

【0147】

化学的変異誘発により少なくとも1つの標的鋳型核酸分子に変異を導入することは、少なくとも1つの標的鋳型核酸を化学変異源に曝露することによって達成され得る。好適な化学変異源としては、マイトマイシンC (MMC)、N-メチル-N-ニトロソウレア (MNU)、亜硝酸 (NA)、ジエポキシブタン (DEB)、1, 2, 7, 8, -ジエポキシオクタン (DEO)、メタンスルホン酸エチル (EMS)、メタンスルホン酸メチル (MMS)、N-メチル-N'-ニトロ-N-ニトロソグアニジン (MNNG)、4-ニトロキノリン1-オキシド (4-NQO)、2-メチルオキシ-6-クロロ-9-[エチル-2-クロロエチル]-アミノプロピルアミノ)-アクリジン二塩酸塩 (ICR-170)、2-アミノプリン (2A)、重亜硫酸塩、及びヒドロキシルアミン (HA) が挙げられる。例えば、核酸分子が重亜硫酸塩に曝露されると、重亜硫酸塩はシトシンを脱アミノ化してウラシルを形成し、C-T置換変異を実質的に導入する。

40

【0148】

上記のように、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、酵素的変異誘発によって実行され得る。任意選択的に、酵素的変異誘発は、DNAポリメラーゼを使用して実行される。例えば、いくつかのDNAポリメラーゼは、変異性であり（低忠実度ポリメラーゼである）、変異性DNAポリメラーゼを使用して少なくとも1つの標的鋳型核酸分子を複製すると、変異を導入するであろう。Taqポリメラーゼは、低忠実度ポリメラーゼの例であり、少なくとも1つの標的鋳型核酸分子に変異を導入するステッ

50

プは、Taqポリメラーゼを使用して、例えばPCRによって、少なくとも1つの標的鋳型核酸分子を複製することによって実施され得る。

【0149】

DNAポリメラーゼは、低バイアスDNAポリメラーゼであり得る。

【0150】

少なくとも1つの標的鋳型核酸分子に変異を導入するステップがDNAポリメラーゼを使用して実行される場合、少なくとも1つの標的鋳型核酸分子は、DNAポリメラーゼが少なくとも1つの変異標的鋳型核酸分子の生成を触媒するのに好適な条件下で、DNAポリメラーゼ及び好適なプライマーを使用してインキュベートされ得る。

【0151】

好適なプライマーは、少なくとも1つの標的鋳型核酸分子に隣接する領域、又は少なくとも1つの標的鋳型核酸分子に相補的である核酸分子に隣接する領域に相補的である、短い核酸分子を含む。例えば、少なくとも1つの標的鋳型核酸分子が染色体の一部である場合、プライマーは、少なくとも1つの標的鋳型核酸分子の3'末端の直ぐ3'側及び少なくとも1つの標的鋳型核酸分子の5'末端の直ぐ5'側の染色体の領域に相補的である、又はプライマーは、少なくとも1つの標的鋳型核酸分子に相補的な核酸分子の3'末端の直ぐ3'側及び少なくとも1つの標的鋳型核酸分子に相補的な核酸分子の5'末端の直ぐ5'側の染色体の領域に相補的である。

【0152】

好適な条件としては、DNAポリメラーゼが少なくとも1つの標的鋳型核酸分子を複製することができる温度が挙げられる。例えば、40℃～90℃、50℃～80℃、60℃～70℃、又は約68℃の温度である。

【0153】

少なくとも1つの鋳型核酸分子に変異を導入するステップは、複数ラウンドの複製を含み得る。例えば、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、好ましくは、

i) 少なくとも1つの標的鋳型核酸分子を複製して、少なくとも1つの標的鋳型核酸分子に相補的である、少なくとも1つの核酸分子を提供するラウンドと、

ii) 少なくとも1つの標的鋳型核酸分子を複製して、少なくとも1つの標的鋳型核酸分子の複製を提供するラウンドと、を含む。

【0154】

任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、少なくとも2ラウンド、少なくとも4ラウンド、少なくとも6ラウンド、少なくとも8ラウンド、少なくとも10ラウンド、10ラウンド未満、8ラウンド未満、約6ラウンド、2～8ラウンド、又は1～7ラウンドの、少なくとも1つの標的鋳型核酸分子の複製を含む。ユーザは、少数の複製ラウンドを用いて、増幅バイアスを導入する可能性を低減することができる。

【0155】

任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、60℃～80℃の温度での少なくとも2ラウンド、少なくとも4ラウンド、少なくとも6ラウンド、少なくとも8ラウンド、少なくとも10ラウンド、10ラウンド未満、8ラウンド未満、約6ラウンド、2～8ラウンド、又は1～7ラウンドの複製を含む。

【0156】

任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、ポリメラーゼ連鎖反応(PCR)を使用して実行される。PCRは、核酸分子を複製するために、

a) 溶融、

b) アニール、及び

c) 伸長(extension and elongation)の各ステップの複数ラウンドを含むプロセスである。

【0157】

核酸分子（少なくとも1つの標的鋳型核酸分子など）は、好適なプライマー及びポリメラーゼと混合する。溶融ステップにおいて、二本鎖核酸分子を変性させる（2本の鎖へと分離させる）ように、核酸分子を90℃超の温度に加熱する。アニーリングステップでは、75℃未満、例えば、55℃～70℃、約55℃、又は約68℃の温度に核酸分子を冷却して、プライマーが核酸分子にアニールすることを可能にする。伸長ステップでは、核酸分子を60℃超の温度に加熱して、DNAポリメラーゼがプライマー伸長を触媒することを可能にし、鋳型鎖に相補的なヌクレオチドを付加する。

【0158】

任意選択的に、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、変異性反応条件下でTaqポリメラーゼを使用して、少なくとも1つの標的鋳型核酸分子を複製することを含む。例えば、少なくとも1つの標的鋳型核酸分子に変異を導入するステップは、 Mn^{2+} 、 Mg^{2+} 、又は不等dNTP濃度（例えば、過剰のシトシン、グアニン、アデニン、又はチミン）の存在下でTaqポリメラーゼを使用するPCRを含み得る。

【0159】

ステップS120：シーケンシング

非変異シーケンスリード及び変異シーケンスリードを含むデータの取得

本発明の方法は、変異シーケンスリードを受け取るステップ、任意選択的に非変異シーケンスリードを受け取るステップを含み得る。非変異シーケンスリード及び変異シーケンスリードは、任意の供給源から得ることができる。

【0160】

任意選択的に、非変異シーケンスリードは、一対のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスすることによって得られる。任意選択的に、変異シーケンスリードは、一対のサンプルのうちの第2のサンプル中の少なくとも1つの標的鋳型核酸分子に変異を導入して、少なくとも1つの変異標的鋳型核酸分子を提供し、少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスすることによって得られる。

【0161】

任意選択的に、非変異シーケンスリードは、一対のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域のシーケンスを含み、変異シーケンスリードは、一対のサンプルのうちの第2のサンプル中の少なくとも1つの変異標的鋳型核酸分子の領域のシーケンスを含み、一対のサンプルは、同一の元サンプルから採取されたか、同一生物に由来する。

【0162】

少なくとも1つの標的鋳型核酸分子又は少なくとも1つの変異標的鋳型核酸分子の領域のシーケンシング

少なくとも1つの標的鋳型核酸分子のシーケンスを決定するための方法は、一対のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、非変異シーケンスリードを提供するステップ、及び／又は少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスして、変異シーケンスリードを提供するステップを含み得る。

【0163】

シーケンシングステップは、任意のシーケンシング法を使用して実行され得る。可能なシーケンシング法の例としては、Maxam Gilbert Sequencing、Sanger Sequencing、ブリッジ増幅（ブリッジPCRなど）を含むシーケンシング、又はMaxam AM, Gilbert W (February 1977), 「A new method for sequencing DNA」, Proc. Natl. Acad. Sci. U. S. A. 74 (2): 560-4、Sanger F, Coulson AR (May 1975), 「A rapid method for determining sequences in DNA by primed synt

10

20

30

40

50

hesis with DNA polymerase」, J. Mol. Biol. 94 (3): 441-8及びBentley DR, Balasubramanian S, et al. (2008), 「Accurate whole human genome sequencing using reversible terminator chemistry」, Nature, 456 (7218): 53-59に記載される方法など任意の高スループットシーケンシング (HTS) 法が挙げられる。

【0164】

典型的な実施形態では、シーケンシングステップのうちの少なくとも1つ、又は好ましくは両方は、ブリッジ増幅を伴う。任意選択的に、ブリッジ増幅ステップは、5秒超、10秒超、15秒超、又は20秒超の延長時間を使用して実行される。ブリッジ増幅の使用の例は、Illumina Genome Analyzer Sequencerにおけるものである。好ましくは、ペアエンドシーケンシングが使用される。

10

【0165】

任意選択的に、(i) 一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、非変異シーケンスリードを提供するステップ、及び(ii) 少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスして、変異シーケンスリードを提供するステップは、同一のシーケンシング法を使用して実行される。

【0166】

任意選択的に、(i) 一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、非変異シーケンスリードを提供するステップ、及び(ii) 少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスして、変異シーケンスリードを提供するステップは、異なるシーケンシング法を使用して実行される。

20

【0167】

任意選択的に、(i) 一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、非変異シーケンスリードを提供するステップ、及び(ii) 少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスして、変異シーケンスリードを提供するステップは、2つ以上のシーケンシング法を使用して実行され得る。例えば、一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の一部は、第1のシーケンシング法を使用してシーケンスされ得、一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の一部は、第2のシーケンシング法を使用してシーケンスされ得る。同様に、少なくとも1つの変異標的鋳型核酸分子の一部は、第1のシーケンシング法を使用してシーケンスされ得、少なくとも1つの変異標的鋳型核酸分子の一部は、第2のシーケンシング法を使用してシーケンスされ得る。

30

【0168】

任意選択的に、(i) 一对のサンプルのうちの第1のサンプル中の少なくとも1つの標的鋳型核酸分子の領域をシーケンスして、非変異シーケンスリードを提供するステップ、及び(ii) 少なくとも1つの変異標的鋳型核酸分子の領域をシーケンスして、変異シーケンスリードを提供するステップは、異なる時間に実行される。あるいは、ステップ(i) 及び(ii) は、互いに1年以内など、ほぼ同時期に実行され得る。一对のサンプルのうちの第1のサンプル及び一对のサンプルのうちの第2のサンプルは、互いに同時に採取される必要はない。2つのサンプルが同一生物に由来する場合、これらのサンプルは実質的に異なる時間、数年も隔てて提供され得、したがって、2つのシーケンシングステップには、数年の隔たりがあり得る。更に、一对のサンプルのうちの第1のサンプル及び一对のサンプルのうちの第2のサンプルが同一の元サンプルに由来する場合であっても、生体サンプルは一定時間保存することができ、そのため、シーケンシングステップが同時に行われる必要はない。

40

【0169】

変異シーケンスリード及び／又は非変異シーケンスリードは、シングルエンドシーケンスリード又はペアエンドシーケンスリードであり得る。

【0170】

50

任意選択的に、変異シーケンスリード及び／又は非変異シーケンスリードは、50 nt 超、100 nt 超、500 nt 超、200,000 nt 未満、15,000 nt 未満、1,000 nt 未満、50~200,000 nt、50~15,000 nt、又は50~1,000 ntである。

【0171】

任意選択的に、シーケンシングステップは、少なくとも1つの標的鋳型核酸分子ごとのヌクレオチド当たり0.1~500リード、0.2~300リード、又は0.5~150リードのシーケンシング深度を使用して実行される。シーケンシング深度が大きいほど、決定される／生成されるシーケンスの精度が高くなるが、アセンブリはより困難となり得る。

【0172】

パラメータの選択

好ましくは、方法200で使用されるパラメータは、以下に記載するように選択される。

【0173】

好ましい実施形態では、各シードパターン ψ の重み $w(\psi)$ は、5~50、好ましくは10~30、更に好ましくは13~23の範囲である。このため、各シードパターン ψ は、各シードパターン ψ によってマスクされた各 k -merが高確率で確実に固有であるように、確実に十分に大きくなる。例えば、典型的な長さが500万ヌクレオチドである細菌ゲノムの場合、各シードパターン ψ の重み $w(\psi)$ は、好ましくは13~19の範囲内にあり、 $4^{13} > 500$ 万であることを留意されたい。典型的な長さが30億ヌクレオチドであるヒトサイズのゲノムの場合、各シードパターン ψ の重み $w(\psi)$ は、好ましくは19~23の範囲内にあり、 $4^{19} > 3 \times 10^9$ であることを留意されたい。

【0174】

好ましい実施形態では、各変異シーケンスリードにおいて1つ以上の変異の位置を決定するステップS230で使用される各 k -merのサイズ k は、各シードパターン ψ の重み $w(\psi)$ よりも大きい。各 k -merのサイズ k は、各シードパターン ψ の重み $w(\psi)$ の5倍未満、4倍未満、3倍未満、又は2倍未満であり得る。各変異シーケンスリードにおいて1つ以上の変異の位置を決定するステップS230で使用される各 k -merのサイズ k は、10~250、好ましくは13~100、更に好ましくは15~50、最も好ましくは20~40の範囲であり得る。これにより、サイズ k は、任意の k -merが、方法200の文脈において不利である、挿入又は欠失シーケンシングエラーを含む確率が確実に低くなるように、確実に十分に小さくなる。

【0175】

重み $w(\psi) = 16$ 及び $k = 27$ のシードパターン ψ を含む、例示的なシードファミリー Ψ を以下に示す。

$\psi_1 = \{0, 1, 2, 3, 5, 6, 9, 12, 13, 14, 16, 18, 20, 21, 22, 23\}$ 、

$\psi_2 = \{0, 1, 2, 4, 5, 9, 10, 11, 13, 18, 19, 21, 23, 24, 25, 26\}$ 、

$\psi_3 = \{0, 1, 2, 3, 4, 5, 7, 8, 9, 10, 13, 15, 16, 18, 19, 20\}$ 、

$\psi_4 = \{0, 1, 2, 4, 6, 7, 12, 14, 16, 17, 20, 21, 23, 24, 25, 26\}$ 、

【0176】

一実施形態では、共通ミニマイザ関数を適用するステップS220で使用される k -mer、すなわち、変異シーケンスリードごとに決定された1つ以上のミニマイザは、各変異シーケンスリードにおいて1つ以上の変異の位置を決定するステップS230で 사용되는 k -merとは異なるサイズ k を有する。各ミニマイザのサイズ k は、5~50、好ましくは10~30、更に好ましくは13~23の範囲であり得る。

各ミニマイザのサイズ k は、シードパターンの重み $w(\psi)$ の選択と同じ考慮事項に基

10

20

30

40

50

づいて選択され得る。各ミニマイザのサイズ k は、細菌では 13 ～ 19 の範囲、ヒトサイズのゲノムでは 19 ～ 23 の範囲であり得る。

【0177】

方法200の実施態様

方法200は、様々な方法で実施することができる。好ましいアプローチは、変異シーケンスリードP及び非変異シーケンスリードRの一部又は全てまでを含む初回のパスでセット U_M を最初に計算し、次いで、変異シーケンスリードP及び非変異シーケンスリードRまでを含む2回目のパスで W_M を計算する。 W_M を入手したら、変異シーケンスリードPまでを含む3回目のパスで、1つ以上の変異の位置と共にミニマイザの位置を計算することができ、これらは、RAM又は固定ストレージ（例えば、ディスク）のいずれかのミニマイザビンに記憶され得る。任意選択的に、多数のミニマイザビンが、ソートされて、又は順序無しで単一ファイルに記憶され得る。次いで、各ミニマイザビン（又は各ファイル）は順次又は並列に読み出され得、ミニマイザビンは、エッジ重みを計算するように処理される。各変異シーケンスリードは、多数のミニマイザビンに現れ得るため、一対の変異シーケンスリードが、計算された重みの複数の推定値を有することが可能である。これが生じる場合、好ましい重み（通常は最大）を選択するために、いくつかの尺度を使用する必要がある。最後に、シーケンシング化学によってペアエンドリードが生成され、一対のペアエンドリード内のそれぞれが共通ミニマイザを共有する場合、2つの端部のスコアが合計されて、一対のペアエンドリードの単一スコアに到達することができる。

【実施例】

【0178】

実験データ

方法200を使用して、いくつかの実際のSAMデータセットを処理した（各SAMデータセットは、非変異シーケンスリード及び変異シーケンスリードを含む）。

【0179】

Aerobacter butzleri JV22のSAMデータセットを処理した。この生物は、単一の環状染色体として存在する2.3Mbpのゲノムを有する。方法200のAC++実施態様を、Amazon AWSインスタンスで実行した。SAMデータセットは、956133のリファレンスリード対（非変異）、及び約8000の変異長鋳型由来の2154909の変異リード対からなる。変異リードの2087506（96.9%）は変異長鋳型の内部部分に由来し、67403（3.1%）は長鋳型の端部に由来し、サンプルバーコードを含む。個々のリードは、150nt以下の長さである。リード対は、以前に品質及びアダプタのトリミングが行われている。方法200は、データセットの処理に12分のCPU時間及び1.2GBのRAMを要し、リード間で3003393の候補リンクを生成した。次いで、これらのリンクをMarkovクラスタリング（mcl）を使用してグラフクラスタリングに供し、結果として得られた6779グループのリードは、MEGAHIT（Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, Oxford, England, 31(10):1674-1676, May 2015を参照）を使用してデノボアセンブリを行って、変異長鋳型を再構築した。最後に、変異長鋳型を、Unicyclerソフトウェアによって計算されたハイブリッドゲノムアセンブリにおいて未変異リードと共に使用した。得られたアセンブリは、ショートリードのみから得られたアセンブリ（図4Aに示す）と比較して図4Bに示す。図4Aは、方法200の実行前に、Shovillアセンブリパイプラインを使用した2.3Mbpの*Aerobacter butzleri* ゲノムのショートリードアセンブリを示す。これにより、78個の足場のアセンブリが得られ、最大の足場は342kbpに及び、足場N50は約127kbpであった。図4Bは、方法200を使用した2.3

Mbpの *Arcobacter butzleri* ゲノムのアセンブリを示す。円形染色体は、単一のコンティグに概ね分解されており、200nt未満の小片のコピー数は未分解のままである。

【0180】

方法200で実施したアプローチの拡張性及び分解力は、シミュレーションデータを使用して測定した。CFTR遺伝子からの50kbpシーケンスを使用して、変異長鋳型カバレッジ及びこれらの鋳型からの対応する変異ショートリードの増加量をシミュレートした。シミュレーションは、最初に変異長鋳型を生成する、新開発のスク립トを使用して実行し、次いで、周知のIlluminaリードシミュレータartsimを呼び出して、変異鋳型からショートリードシーケンシングをシミュレートした。変異データに加えて、非変異シーケンスの30倍のカバレッジをartsimでシミュレートした。10¹～10⁶の規模の増加範囲で変異長鋳型カバレッジをシミュレートした。変異率を6%に固定した。長鋳型ごとに、10倍のショートリードカバレッジをシミュレートした。シミュレーションデータの結果は、方法200で偽陽性リンクが報告された割合を測定することによって評価した。

【0181】

図5は、長鋳型当たりのショートリードカバレッジの深度の影響を示す。長鋳型当たりの生成されたショートリードデータの量をx軸に示し、y軸は、方法200からの結果に関する様々な性能測定基準を示す。鋳型当たりのショートリードカバレッジが低い場合、例えば、<4倍である場合、元の長鋳型の再構築が不十分かつ不完全であることが分かる。しかしながら、変異鋳型当たりのカバレッジが5～10倍の範囲である場合、良好な再構築を得ることができる。

- ・ links_num: 方法200で報告された変異リード間のリンクの数。links_fp: 報告された偽陽性リンクの数。
- ・ links_fprate: 全ての報告されたリンクの中で偽陽性であるリンクの割合。
- ・ mcl_num: mmdremingによって報告されたグラフでのMarkovクラスタリングによって生成されたクラスタの数。
- ・ idba_scaf_num: 変異ショートリードのクラスタをアセンブルすることによって再構築された足場シーケンスの数。
- ・ idba_scaf_bp: アセンブルされた全ての足場の総長さ。

本発明は、例えば、以下の項目を提供する。

(項目1)

2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率と関連する尺度を決定するためのコンピュータ実施方法であって、前記方法は、

複数の変異シーケンスリードを受け取ることであって、各変異シーケンスリードは、変異を含むシーケンスのサブシーケンスに対応し、変異を含む前記シーケンスは、変異を含まないシーケンスと比較して変異を含む、ことと、

共通ミニマイザ関数を各変異シーケンスリードに適用し、それによって、変異シーケンスリードごとに1つ以上のそれぞれのミニマイザを決定することと、

各変異シーケンスリードにおいて前記1つ以上のそれぞれのミニマイザの位置を決定することと、

各変異シーケンスリードにおいて1つ以上の変異の位置を決定することと、

共通ミニマイザを有する少なくとも2つの変異シーケンスリードについて、前記それぞれのミニマイザがアラインされると、位置が一致する、及び／又は位置が一致しない変異の数をカウントすることと、を含む、方法。

(項目2)

複数の非変異シーケンスリードを受け取ることを更に含み、各非変異シーケンスリードは、変異を含まない前記シーケンスのサブシーケンスに対応する、項目1に記載の方法。

(項目3)

共通ミニマイザ関数を各変異シーケンスリードに適用するステップは、i) 可能な $k\text{-mer}$ の順序付きリストに最初に列挙されている、前記それぞれの変異シーケンスにおける前記 1 つ以上の $k\text{-mer}$ 、又は ii) 可能な $k\text{-mer}$ の所定のセットに存在する前記 1 つ以上の $k\text{-mer}$ を特定することを含み、前記それぞれの変異シーケンスについて決定された、前記 1 つ以上のミニマイザは、特定した前記 1 つ以上の $k\text{-mer}$ である、項目 1 又は 2 に記載の方法。

(項目 4)

i) 可能な $k\text{-mer}$ の前記順序付きリストにおいて、前記 $k\text{-mer}$ は、前記 $k\text{-mer}$ が変異を含む前記シーケンスに存在し、変異を含まない前記シーケンスには存在しない確率に基づいて順序付けられる、又は ii) 可能な $k\text{-mer}$ の前記所定のセットは、変異を含む前記シーケンスに存在し、変異を含まない前記シーケンスには存在しない可能性が比較的高い $k\text{-mer}$ を含み、任意選択的に、可能な $k\text{-mer}$ の前記所定のセットは、変異を含む前記シーケンスに存在する可能性が比較的低い $k\text{-mer}$ を含まない、項目 3 に記載の方法。

10

(項目 5)

可能な $k\text{-mer}$ の前記順序付きリスト又は可能な $k\text{-mer}$ の前記所定のセットは、前記複数の非変異シーケンスリードよりも前記複数の変異シーケンスリードに存在することが多い $k\text{-mer}$ からなり、任意選択的に、前記複数の非変異シーケンスリードよりも前記複数の変異シーケンスリードに存在することが多い $k\text{-mer}$ は、変異を含む前記シーケンスに存在する可能性が比較的高い、項目 2 及び項目 3 又は 4 に記載の方法。

20

(項目 6)

可能な $k\text{-mer}$ の前記所定のセットは、前記複数の変異シーケンスリードに n 倍以上存在し、前記複数の非変異シーケンスリードに n 倍未満存在する $k\text{-mer}$ からなり、 n は 1 以上の整数であり、任意選択的に、前記複数の変異シーケンスリードに n 倍以上存在し、前記複数の非変異シーケンスリードに n 倍未満存在する $k\text{-mer}$ は、変異を含む前記シーケンスに存在する可能性が比較的高い、項目 2 及び項目 3 ～ 5 のいずれか一項に記載の方法。

(項目 7)

可能な $k\text{-mer}$ の前記所定のセットは、前記複数の非変異シーケンスリードに存在しない $k\text{-mer}$ からなる、項目 6 に記載の方法。

30

(項目 8)

n は 2 である、項目 6 又は 7 に記載の方法。

(項目 9)

前記複数の変異シーケンスリードにおける前記 $k\text{-mer}$ と前記複数の非変異シーケンスリードにおける前記 $k\text{-mer}$ との比較に基づいて、可能な $k\text{-mer}$ の前記順序付きリスト又は可能な $k\text{-mer}$ の前記所定のセットを生成することを更に含む、項目 2 及び項目 3 ～ 8 のいずれか一項に記載の方法。

(項目 10)

各ミニマイザは、5 超、好ましくは 10 超の長さの $k\text{-mer}$ である、項目 1 ～ 9 のいずれか一項に記載の方法。

40

(項目 11)

各ミニマイザピンは、共通ミニマイザを有する変異シーケンスリードを含み、共通ミニマイザを有さない変異シーケンスリードを含まないように、1 つ以上のミニマイザピンにおいて前記変異シーケンスリードをビンニングすることを更に含む、

位置が一致する、及び／又は位置が一致しない変異の数をカウントするステップは、同一ミニマイザピン内の変異シーケンスリードのみで実行される、項目 1 ～ 10 のいずれか一項に記載の方法。

(項目 12)

各変異シーケンスリードにおいて前記 1 つ以上の変異の位置を決定する前記ステップは、前記複数の非変異シーケンスリードにおいて、1 つ以上のシードパターンのうちのそれ

50

それぞれ1つをk-merに適用することによって、シードマスクされた非変異k-merのセットを得ることと、

変異シーケンスリードごとに、前記それぞれの変異シーケンスリードにおいて、前記1つ以上のシードパターンをk-merに適用して、複数のシードマスクされた変異k-merを得ることと、シードマスクされた非変異k-merの前記セットに存在する、前記複数のシードマスクされた変異k-merのうちの前記シードマスクされた変異k-merに対応する前記シードパターンの全てによってマスクされる前記変異シードシーケンスリードにおいて前記1つ以上の位置を特定することによって、前記1つ以上の変異の前記位置を特定することと、を含む、項目2及び項目2～11のいずれか一項に記載の方法。

(項目13)

前記1つ以上のシードパターンは、前記複数の変異シーケンスリードにおける任意のk-mer及び前記複数の非変異シーケンスリードにおける、対応するk-merに前記1つ以上のシードパターンのうちの少なくとも1つを適用することによって同一のシードマスクされたk-merを得る確率が90%超、好ましくは99%超であるように選択される、項目12に記載の方法。

(項目14)

変異を含む前記シーケンスは、変異を含まない前記シーケンスと比較して遷移変異を含み、

前記1つ以上のシードパターンは、1つ以上の遷移シードパターンからなる、項目12又は13に記載の方法。

(項目15)

前記複数の変異シーケンスリードのそれぞれは、複数のサンプルのうちの1つに関連する変異を含むシーケンスのサブシーケンスに対応し、前記複数の非変異シーケンスリードのそれぞれは、前記複数のサンプルのうちの前記1つに関連する変異を含まないシーケンスのサブシーケンスに対応し、変異を含む各シーケンスは、変異を含まないそれぞれのシーケンスと比較して変異を含み、

シードマスクされた非変異k-merのセットを得ることは、サンプルごとにシードマスクされた非変異k-merのそれぞれのセットを得ることを含み、

非変異サンプルビットベクトルのセットを作成することを更に含み、各非変異サンプルビットベクトルは、シードマスクされた非変異k-merの前記セット内のそれぞれのk-merについて、前記それぞれのk-merが存在する前記複数のサンプルを定義し、

前記1つ以上の変異の位置を決定することは、変異シーケンスリードごとに、並びにシードマスクされた非変異k-merのセットごとに及び／又はシードマスクされた非変異k-merのセットの組み合わせごとに、シードマスクされた非変異k-merの前記それぞれのセット及び／又はシードマスクされた非変異k-merのセットの前記それぞれの組み合わせに存在する、前記複数のシードマスクされた変異k-merのうちの前記シードマスクされた変異k-merに対応する前記シードパターンの全てによってマスクされる前記変異シーケンスリードにおいて前記1つ以上の位置を特定すること、並びに特定した前記1つ以上の位置を、シードマスクされた非変異k-merの前記それぞれのセット又はシードマスクされた非変異k-merのセットの前記それぞれの組み合わせと関連付けることを含む、項目12～14のいずれか一項に記載の方法。

(項目16)

各変異シーケンスリードにおいて前記1つ以上の変異の位置を決定する前記ステップは、前記変異シーケンスリードのうちの1つ以上について、前記それぞれの変異シーケンスリードをリファレンスアセンブリにアラインすることと、

前記それぞれの変異シーケンスリードと前記リファレンスアセンブリとの間での違いの、前記それぞれの変異シーケンスリードにおける位置を特定することによって、前記それぞれの変異シーケンスリードにおいて前記1つ以上の変異の位置を決定することと、を含む、項目2～15のいずれか一項に記載の方法。

(項目17)

10

20

30

40

50

各変異シーケンスリードにおいて前記1つ以上の変異の位置を決定する前記ステップは、変異シーケンスリードごとに、

前記それぞれの変異シーケンスリードにおける位置が前記リファレンスアセンブリにアラインされる場合、前記それぞれの変異シーケンスリードにおける前記位置が、前記それぞれの変異シーケンスリードは前記リファレンスアセンブリとは異なる位置であると、前記それぞれの変異シーケンスリードにおける位置は、前記それぞれの変異シーケンスリードにおける変異の位置であると決定することと、

前記それぞれの変異シーケンスリードにおける位置が前記リファレンスアセンブリにアラインされていない場合、前記それぞれの変異シーケンスリードにおける前記位置が、シードマスクされた非変異 $k - m e r$ の前記セットに存在する前記複数のシードマスクされた変異 $k - m e r$ のうちの前記シードマスクされた変異 $k - m e r$ に対応する前記シードパターンの全てによってマスクされる場合、前記それぞれの変異シーケンスにおける位置は、前記それぞれの変異シーケンスリードにおける変異の位置であると決定することと、を含む、項目12～15のいずれか一項に従属する場合の項目16に記載の方法。

(項目18)

位置が一致する、及び／又は位置が一致しない変異の数に基づいて、前記少なくとも2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する前記確率と相關する前記尺度を決定する、項目1～17のいずれか一項に記載の方法。

(項目19)

前記少なくとも2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する前記確率に相關する前記尺度は、i) 前記少なくとも2つの変異シーケンスリードが、同一の変異を含むシーケンスに由来する確率密度、及びii) 前記少なくとも2つの変異シーケンスリードが同一の変異を含むシーケンスに由来する確率密度と相關するスコア関数のうちの1つである、項目18に記載の方法。

(項目20)

前記複数の変異シーケンスリードから無向加重グラフを作成することを更に含み、前記無向加重グラフは、前記複数の変異シーケンスリードに対応するノードを含み、前記ノード間のエッジは、それぞれのエッジ重みと関連し、各エッジ重みは、前記それぞれのエッジに関連する前記2つのノードに対応する前記2つの変異シーケンスリードについて決定された、位置が一致する、及び／又は位置が一致しない変異の数に基づいて決定される、項目1～19のいずれか一項に記載の方法。

(項目21)

前記エッジ重みは、前記それぞれのエッジに関連する前記2つのノードに対応する前記少なくとも2つの変異シーケンスリードが同一の変異を含むシーケンスに由来する前記確率と相關する前記尺度に対応する、項目20に記載の方法。

(項目22)

前記無向加重グラフでグラフクラスタリングを実行し、それによって同一の変異を含むシーケンスに由来すると予想される変異シーケンスリードのクラスタを作成することを更に含む。項目20又は21に記載の方法。

(項目23)

前記グラフクラスタリングは、Markovクラスタリング又はInfomapを含む、項目22に記載の方法。

(項目24)

前記クラスタにおいて前記変異シーケンスリードをアセンブルすることによって変異を含む前記シーケンスの少なくとも一部を再構築することを更に含む、項目22又は23に記載の方法。

(項目25)

任意選択的に前記複数の非変異シーケンスリードを使用して、変異を含む前記シーケンスの前記再構築部分から変異を含まない可能性の高いシーケンスの少なくとも一部を推測することによって、変異を含まない前記シーケンスの少なくとも一部を再構築することを

10

20

30

40

50

更に含む、項目 2 及び 2 4 に記載の方法。

(項目 2 6)

標的鑄型核酸分子のシーケンスの少なくとも一部を生成するための方法であって、項目 2 0 ～ 2 5 のいずれか一項に記載の方法を含む、方法。

(項目 2 7)

少なくとも 1 つの標的鑄型核酸分子のシーケンスの少なくとも一部を決定するための方法であって、前記方法は、

変異を含む、少なくとも 1 つの標的鑄型核酸分子の領域をシーケンスして、複数の変異シーケンスリードを提供することと、

得られた前記複数の変異シーケンスリードで項目 1 ～ 2 6 のいずれか一項に記載の方法を実行することと、を含む、方法。

(項目 2 8)

シーケンスする前記ステップは、

(a) 一対のサンプルを提供することであって、各サンプルは、少なくとも 1 つの標的鑄型核酸分子を含む、ことと、

(b) 前記一対のサンプルのうちの第 1 のサンプル中の前記少なくとも 1 つの標的鑄型核酸分子の領域をシーケンスして、前記複数の非変異シーケンスリードを提供することと、

(c) 前記一対のサンプルのうちの第 2 のサンプル中の前記少なくとも 1 つの標的鑄型核酸分子に変異を導入して、少なくとも 1 つの変異標的鑄型核酸分子を提供することと、

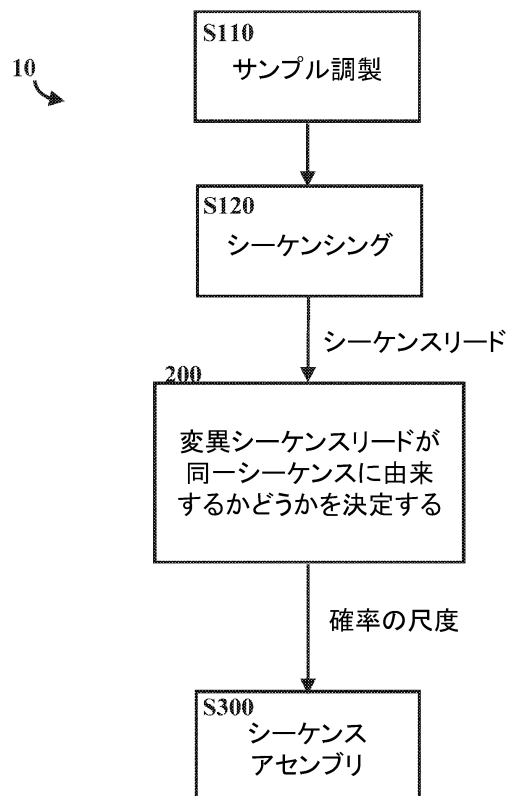
(d) 前記少なくとも 1 つの変異標的鑄型核酸分子の領域をシーケンスして、前記複数の変異シーケンスリードを提供することと、を含む、項目 2 7 に記載の方法。

(項目 2 9)

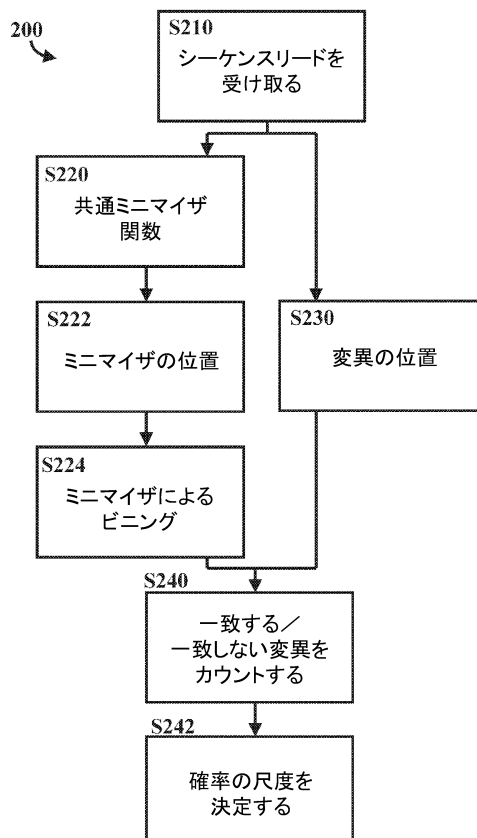
変異を導入する前記ステップは、前記一対のサンプルのうちの前記第 2 のサンプル中の前記少なくとも 1 つの標的鑄型核酸分子に遷移変異を導入することを含む、項目 2 8 に記載の方法。

【図面】

【図 1】



【図 2】



10

20

30

40

50

フロントページの続き

テクノロジーズ ピーティーワイ リミテッド

審査官 田中 寛人

- (56)参考文献 米国特許出願公開第2018/0365375 (US, A1)
米国特許出願公開第2019/0164627 (US, A1)
国際公開第2002/079502 (WO, A1)
国際公開第2019/022018 (WO, A1)
米国特許出願公開第2004/0152108 (US, A1)
米国特許出願公開第2013/0110407 (US, A1)
- (58)調査した分野 (Int.Cl., DB名)
G16B 5/00-99/00
C12Q 1/6869
C12N 15/11