

(19)日本国特許庁(JP)

## (12)特許公報(B2)

(11)特許番号

特許第7304852号

(P7304852)

(45)発行日 令和5年7月7日(2023.7.7)

(24)登録日 令和5年6月29日(2023.6.29)

(51)国際特許分類

F I

C 1 2 Q 1/6869(2018.01)

C 1 2 Q 1/6869

Z

C 1 2 M 1/00 (2006.01)

C 1 2 M 1/00

A

C 1 2 Q 1/48 (2006.01)

C 1 2 Q 1/48

Z

C 1 2 Q 1/34 (2006.01)

C 1 2 Q 1/34

C 1 2 Q 1/6886(2018.01)

C 1 2 Q 1/6886

Z

請求項の数 16 (全49頁) 最終頁に続く

(21)出願番号 特願2020-524480(P2020-524480)

(86)(22)出願日 平成30年11月2日(2018.11.2)

(65)公表番号 特表2021-502072(P2021-502072  
A)

(43)公表日 令和3年1月28日(2021.1.28)

(86)国際出願番号 PCT/US2018/059056

(87)国際公開番号 WO2019/090147

(87)国際公開日 令和1年5月9日(2019.5.9)

審査請求日 令和3年11月1日(2021.11.1)

(31)優先権主張番号 62/581,609

(32)優先日 平成29年11月3日(2017.11.3)

(33)優先権主張国・地域又は機関  
米国(US)

(73)特許権者 515059083

ガーダント ヘルス, インコーポレイテ  
ッド

アメリカ合衆国 カリフォルニア 940

63, レッドウッド シティ, ペノブ

スコット ドライブ 505

(74)代理人 100078282

弁理士 山本 秀策

(74)代理人 100113413

弁理士 森下 夏樹

(74)代理人 100181674

弁理士 飯田 貴敏

(74)代理人 100181641

弁理士 石川 大輔

(74)代理人 230113332

最終頁に続く

(54)【発明の名称】 脱アミノ化に誘導される配列エラーの補正

## (57)【特許請求の範囲】

## 【請求項1】

無細胞核酸の集団におけるバリエーションヌクレオチドを識別するための方法であって、

(a) 一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子を含む無細胞核酸の集団を、5' - 3' ポリメラーゼ活性および3' - 5' エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、前記タンパク質が、一方または両方の末端において、3' オーバーハングを消化し、5' オーバーハングを相補的な核酸で充填して、二本鎖平滑末端化核酸を生成する、ステップと、

(b) シーケンシングされた核酸を得るために、前記二本鎖平滑末端化核酸の配列を決定するステップと、

(c) 参照配列におけるそれぞれの指定位置について、

(i) 前記指定位置を含む、シーケンシングされた核酸のサブセットを識別し、

(ii) 前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸を識別する、ステップと、

(d) ステップ(c)の(ii)における前記シーケンシングされた核酸が閾値に合致するそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、

(i) 前記バリエーションが、前記参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および

(ii) 前記バリエーションヌクレオチドが、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 前記サブセット内のシーケンシングされた核酸の5'末端から、前記指定位置における前記CからTへの変動の距離、または前記サブセット内のシーケンシングされた核酸の3'末端から、前記指定位置における前記GからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップとを含む、方法。

【請求項2】

前記指定位置が参照ヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸の数を識別することをさらに含む、請求項1に記載の方法。

【請求項3】

ステップ(b)が、

(a) 前記二本鎖平滑末端化核酸の両方の鎖の配列を決定することを含み、ステップ(c)が、少なくとも1つの指定位置について行われ、前記変動を有する前記サブセット内の前記シーケンシングされた核酸が、前記二本鎖平滑末端化核酸のシーケンシングされた核酸の両方の鎖の配列を含み、および／または

(b) 鎖の両方の末端から配列を決定することを含む、請求項1または2に記載の方法。

【請求項4】

(a) 前記バリエーションヌクレオチドが、前記サブセット内のシーケンシングされた核酸において5'末端の規定される近接度内にある前記指定位置における前記CからTへの変動の出現、または前記サブセット内のシーケンシングされた核酸において3'末端の規定される近接度内にある前記指定位置における前記GからAへの変動の出現に基づいて、脱アミノ化エラーとして分類され、前記5'末端へ前記規定される近接度が、前記5'末端まで20ヌクレオチド以内であるか、またはそれよりも少ない数のヌクレオチド以内であるとして規定され、前記3'末端へ前記規定される近接度が、前記3'末端まで20ヌクレオチド以内であるか、またはそれよりも少ない数のヌクレオチド以内であるとして規定され、

(b) 前記CからTまたはGからAへの変動が、少なくとも、前記周囲のコンテキストがTCGからTTGまたはCGAからCAAであることに基づいて、脱アミノ化エラーとして分類され、

(c) 前記バリエーションヌクレオチドは、前記サブセット内のシーケンシングされた核酸の5'末端から、前記指定位置における前記CからTへの変動の平均距離が、前記指定位置における前記参照ヌクレオチドの平均距離よりも小さいこと、または前記サブセット内のシーケンシングされた核酸の3'末端から、前記指定位置における前記GからAへの変動が、前記指定位置における前記参照ヌクレオチドの平均距離よりも小さいことに基づいて、脱アミノ化エラーとして分類され、および／または

(d) 前記指定位置における前記CからTへの変動は、その出現が、前記指定位置が前記5'末端の規定される近接度内にある前記サブセットの第1の比率において少なくとも50%である場合に、脱アミノ化エラーとして分類されるか、または前記指定位置における前記GからAへの変動は、その出現が、前記指定位置が前記3'末端の規定される近接度内にある前記サブセットの第2の比率において少なくとも50%である場合に、脱アミノ化エラーとして分類され、

(i) 前記指定位置における前記CからTへの変動は、前記変動が、前記指定位置が前記5'末端の規定される近接度内にある前記サブセットの第1の比率において、前記サブセット内の他のシーケンシングされた核酸におけるよりも少なくとも2倍の出現を有することに基づいて、脱アミノ化エラーとして分類されるか、もしくは

(ii) 前記指定位置における前記GからAへの変動は、前記変動が、前記指定位置が前記3'末端の規定される近接度内にある前記サブセットの第2の比率において、前記サブセット内の他のシーケンシングされた核酸におけるよりも少なくとも2倍の出現を有することに基づいて、脱アミノ化エラーとして分類される、

10

20

30

40

50

請求項 1 ～ 3 のいずれか一項に記載の方法。

【請求項 5】

前記二本鎖平滑末端化核酸を、バーコードを含むアダプターに連結させること、前記アダプターに結合するプライマー分子からプライミングされる前記核酸を増幅させることをさらに含み、

ステップ (b) が、増幅した核酸分子の配列を決定すること、および前記増幅した核酸分子の配列をファミリーにグループ分けすることであって、ファミリーのメンバーが、前記核酸上の同じ開始点および終止点ならびに同じバーコードを有し、前記ファミリーについて、そのそれぞれのメンバーの配列から、複数の位置のそれぞれにおけるコンセンサヌクレオチドを決定することを含む、

10

請求項 1 ～ 4 のいずれか一項に記載の方法。

【請求項 6】

前記無細胞核酸が、体液に由来し、前記体液が、血液、血漿、唾液、尿、および脳脊髄液からなる群から選択される、請求項 1 ～ 5 のいずれか一項に記載の方法。

【請求項 7】

前記タンパク質が、クレノウである、および／または前記バリエーションヌクレオチドが、単一ヌクレオチドバリエーション (SNV) 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】

前記参照配列が、ヒトゲノム、またはヒト染色体の配列、または前記ヒトゲノムの非連続領域を含む、請求項 1 ～ 7 のいずれか一項に記載の方法。

20

【請求項 9】

前記コールされるバリエーションヌクレオチドのうちの少なくとも 1 つが、がんに関連していることが既知である、請求項 1 ～ 8 のいずれか一項に記載の方法。

【請求項 10】

がんを有するかまたはがんを有することが疑われる対象の集団の試料に由来する核酸集団に行われ、前記個々の対象においてコールされたバリエーションヌクレオチドの数および種類が、前記集団における対象に投与される異なる処置を示す、請求項 1 ～ 9 のいずれか一項に記載の方法。

【請求項 11】

30

脱アミノ化エラーとして分類されるバリエーションヌクレオチドが、前記コールされたバリエーションヌクレオチドのうちの少なくとも 1 % または少なくとも 10 % である、請求項 1 ～ 10 のいずれか一項に記載の方法。

【請求項 12】

バリエーションの存在は、少なくとも 5 個のバリエーションヌクレオチドが脱アミノ化エラーとして分類される場合、コールされない、請求項 1 ～ 11 のいずれか一項に記載の方法。

【請求項 13】

前記脱アミノ化エラーの頻度が、少なくとも 1 % または少なくとも 10 % 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】

40

前記閾値は、前記変動が、前記サブセット内のシーケンシングされた核酸の少なくとも 1 % で存在することである、請求項 1 に記載の方法。

【請求項 15】

核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

前記通信インターフェースと通信するコンピューターであって、1 つまたは複数のコンピュータープロセッサと、前記 1 つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 前記核酸シーケンサーによって生成された前記シーケンシングリードを、前記通信ネットワークを通じて受容するステップであって、前記核酸シーケンサーが、対象に由来

50

する無細胞DNA分子から生成されたシーケンシングライブラリーをシーケンシングし、前記シーケンシングライブラリーが、前記無細胞DNA分子およびバーコードを含むアダプターを含む、ステップと、

(b) 参照配列におけるそれぞれの指定位置について、

(i) 前記指定位置を含む、シーケンシングリードのサブセットを識別し、

(ii) 前記指定位置がバリエーションヌクレオチドによって占有されている、前記サブセット内のシーケンシングリードを識別する、ステップと、

(c) 前記変動を有する前記サブセット内の前記シーケンシングリードが、前記コールをサポートするそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、

(i) 前記バリエーションが、前記参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および

(ii) 前記バリエーションヌクレオチドが、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 前記サブセット内のシーケンシングされた核酸の5'末端から、前記指定位置における前記CからTへの変動の距離、または前記サブセット内のシーケンシングされた核酸の3'末端から、前記指定位置における前記GからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップと

を含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターと

を含む、システム。

【請求項16】

(a) 前記核酸シーケンサーをさらに含む、

(b) 前記核酸シーケンサーが、前記シーケンシングライブラリーにシーケンシングバイインセンシスを行って、前記シーケンシングリードを生成する、

(c) 前記核酸シーケンサーが、前記シーケンシングライブラリーにパイロシーケンシング、単分子シーケンシング、ナノポアシーケンシング、半導体シーケンシング、シーケンシングバイライゲーション、またはシーケンシングバイハイブリダイゼーションを行って、前記シーケンシングリードを生成する、

(d) 前記核酸シーケンサーが、前記シーケンシングライブラリーに由来するクローン単一分子アレイを使用して、前記シーケンシングリードを生成する、

(e) 前記核酸シーケンサーが、前記シーケンシングライブラリーをシーケンシングして前記シーケンシングリードを生成するための、マイクロウェルのアレイを有するチップを含む、

(f) 前記コンピューター可読媒体が、メモリー、ハードドライブ、またはコンピューターサーバーを含む、

(g) 前記通信ネットワークが、遠隔通信ネットワーク、インターネット、エクストラネット、またはイントラネットを含む、

(h) 前記通信ネットワークが、分散コンピューティングの可能な1つまたは複数のコンピューターサーバーを含む、

(i) 前記コンピューターが、前記核酸シーケンサーから遠隔設置されているコンピューターサーバー上に設置されている、

(j) 前記シーケンシングライブラリーが、試料を、1つまたは複数の試料と区別する試料バーコードをさらに含む、および／または

(k) ネットワークを通じて前記コンピューターと通信する、電子ディスプレイであって、(a)～(c)を実施した際の結果を表示するためのグラフィカルもしくはウェブベースのユーザーインターフェース、パーソナルコンピューター、またはインターネット対応コンピューターを含む、電子ディスプレイをさらに含み、前記インターネット対応コンピューターが、前記コンピューターから遠隔した位置に設置されている、

請求項15に記載のシステム。

## 【発明の詳細な説明】

## 【技術分野】

## 【0001】

## 相互参照

本出願は、2017年11月3日に出願された米国仮出願第62／581,309号に基づく優先権を主張し、そのような仮出願は、参照によりすべての目的で本明細書に全体的に組み込まれる。

## 【背景技術】

## 【0002】

## 背景

腫瘍は、細胞の異常な増殖である。細胞、例えば、腫瘍細胞が死滅すると、断片化DNAが体液中に放出されることが多い。よって、体液中の無細胞DNAの一部は、腫瘍DNAである。腫瘍は、良性である場合も悪性である場合もある。悪性腫瘍は、がんと呼ばれることが多い。

## 【0003】

がんは、全世界で、疾患の主要な原因である。年々、数千万人の人々が、世界中でがんを有すると診断され、半数より多くが、最終的にがんを原因として死亡する。多くの国では、がんは、心血管疾患に続いて2番目の共通死亡原因として位置付けられる。多くのがんでは、早期検出がアウトカムの改善と関連する。

## 【0004】

がんは、個体の正常な細胞内の変異および／またはエピジェネティックな変動の蓄積によって引き起こされ、これらのうちの少なくとも一部は、細胞分裂の不適切な制御をもたらす。そのような変異としては、一般に、コピー数多型(CNV)、コピー数異常(CNA)、一塩基多型(SNV)、遺伝子融合、および挿入欠失が挙げられ、エピジェネティックな変動としては、シトシンの6原子環の5番目の原子に対する改変、ならびにDNAのクロマチンおよび転写因子との会合が挙げられる。

## 【0005】

がんは、腫瘍の生検に続いて、細胞、マーカー、または細胞から抽出されたDNAの分析によって検出されることが多い。しかしながら、最近では、がんを、体液、例えば、血液または尿中の無細胞核酸から検出することもできることが提案されている(例えば、S iravegna et al., Nature Reviews 2017を参照されたい)。そのような検査は、非侵襲的であり、生検によって疑わしいがん細胞を識別することなく行うことができるという利点を有する。しかしながら、そのような検査は、体液中の核酸の量が非常に少なく、何の核酸が存在しているかにより、シーケンシングが行われ得る前に、より均質な形態にプロセッシングする必要があるという事実によって、複雑なものとなっている。

## 【発明の概要】

## 【課題を解決するための手段】

## 【0006】

## 要旨

本開示の一態様は、核酸の集団においてバリエーションヌクレオチドを識別するための方法であって、(a)一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子を含む核酸の集団を、5' - 3' ポリメラーゼ活性および3' - 5' エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、タンパク質が、一方または両方の末端において、3' オーバーハングを消化し、5' オーバーハングを相補的な核酸で充填して、二本鎖平滑末端化核酸を生成する、ステップと、(b)シーケンシングされた核酸を得るために、二本鎖平滑末端化核酸の配列を決定するステップと、(c)参照配列におけるそれぞれの指定位置について、(i)指定位置を含む、シーケンシングされた核酸のサブセットを識別するステップ、そして(ii)指定位置がバリエーションヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸を識別するステップと、および(d)

10

20

30

40

50

変動を有するサブセット内のシーケンシングされた核酸がコールをサポートするそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、(i) バリエーションが、参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および(ii) バリエーションヌクレオチドが、(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または(2) サブセット内のシーケンシングされた核酸の5'末端から、指定位置におけるCからTへの変動の距離、またはサブセット内のシーケンシングされた核酸の3'末端から、指定位置におけるGからAへの変動までの距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップを含む、方法に関する。

【0007】

一部の実施形態では、ステップ(c)の(ii)は、指定位置がバリエーションヌクレオチドによって占有されているサブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、変動を有するサブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在が、コールされる。

【0008】

一部の実施形態では、バリエーションヌクレオチドは、サブセット内のシーケンシングされた核酸において5'末端の規定される近接度内にある指定位置におけるCからTへの変動の出現、またはサブセット内のシーケンシングされた核酸において3'末端の規定される近接度内にある指定位置におけるGからAへの変動の出現に基づいて、脱アミノ化エラーとして分類される。

【0009】

一部の実施形態では、(c)の(ii)は、指定位置が参照ヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸の数を識別することをさらに含む。

【0010】

一部の実施形態では、(b)は、二本鎖平滑末端化核酸の両方の鎖の配列を決定することを含む。

【0011】

一部の実施形態では、(c)は、少なくとも1つの指定位置について行われ、ここで、変動を有するサブセット内のシーケンシングされた核酸は、二本鎖平滑末端化核酸のシーケンシングされた核酸の両方の鎖の配列を含む。

【0012】

一部の実施形態では、(b)は、鎖の両方の末端から配列を決定することを含む。

【0013】

一部の実施形態では、本方法は、二本鎖平滑末端化核酸を、バーコードを含むアダプターに連結させ、アダプターに結合するプライマー分子からプライミングされる核酸を増幅させるステップをさらに含み、ここで、(b)は、増幅した核酸分子の配列を決定し、増幅した核酸分子の配列をファミリーに分類し、ファミリーのメンバーが、核酸上の同じ開始点および終止点ならびに同じバーコードを有し、ファミリーについて、そのそれぞれのメンバーの配列から、複数の位置のそれぞれにおけるコンセンサスヌクレオチドを決定することを含む。コンセンサス配列は、1つのみのメンバーを有するファミリーについては、決定されない。

【0014】

一部の実施形態では、核酸の集団は、対象の無細胞核酸試料に由来する。無細胞核酸試料は、がんを有するかまたはがんを有することに一致する徴候または症状を有する対象の体液に由来し得る。体液は、血液、血漿、唾液、尿、および脳脊髄液からなる群から選択され得る。血液および血液産物(例えば、血漿および血清)は、本明細書に記載される使用に特に有用な無細胞核酸を含有する。

【0015】

一部の実施形態では、指定位置におけるCからTへの変動は、その出現が、指定位置が

10

20

30

40

50

5'末端の規定される近接度内にあるサブセットの第1の比率において少なくとも50%である場合に、脱アミノ化エラーとして分類されるか、または指定位置におけるGからAへの変動は、その出現が、指定位置が3'末端の規定される近接度内にあるサブセットの第2の比率において少なくとも50%である場合に、脱アミノ化エラーとして分類される。指定位置におけるCからTへの変動は、変動が、指定位置が5'末端の規定される近接度内にあるサブセットの第1の比率において、サブセット内の他のシーケンシングされた核酸におけるものよりも少なくとも2倍の出現を有することに基づいて、脱アミノ化エラーとして分類され得るか、または指定位置におけるGからAへの変動は、変動が、指定位置が3'末端の規定される近接度内にあるサブセットの第2の比率において、サブセット内の他のシーケンシングされた核酸におけるものよりも少なくとも2倍の出現を有することに基づいて、脱アミノ化エラーとして分類される。

10

【0016】

一部の実施形態では、閾値は、変動が、サブセット内のシーケンシングされた核酸の少なくとも1%で存在することである。

【0017】

一部の実施形態では、CからTまたはGからAへの変動は、少なくとも、周囲のコンテキストがTCGからTTGまたはCGAからCAAであることに基づいて、脱アミノ化エラーとして分類される。

【0018】

一部の実施形態では、5'末端への規定される近接度は、5'末端まで20ヌクレオチド以内であるか、またはそれよりも少ない数のヌクレオチド以内であるとして規定され、3'末端への規定される近接度は、3'末端まで20ヌクレオチド以内であるか、またはそれよりも少ない数のヌクレオチド以内であるとして規定される。5'末端への規定される近接度は、5'末端まで20ヌクレオチド以内であるとして規定され得、3'末端への規定される近接度は、3'末端まで20ヌクレオチド以内であるとして規定される。

20

【0019】

一部の実施形態では、タンパク質は、クレノウである。

【0020】

一部の実施形態では、(c)および(d)は、これらのステップを実行するコンピューターにより作動されるシステムなどにおいて行われる。これらの実施形態では、したがって、本開示は、核酸の集団においてバリエーションヌクレオチドを識別するためのコンピューターにより実施される方法に関する。

30

【0021】

一部の実施形態では、参照配列は、ヒトゲノムの配列である。参照配列は、ヒト染色体の配列であってもよい。参照配列は、ヒトゲノムの非連続領域を含み得る。

【0022】

一部の実施形態では、コールされるバリエーションヌクレオチドのうちの少なくとも1つは、がんに関連していることが既知である。

【0023】

一部の実施形態では、本方法は、がんを有するかまたはがんを有することが疑われる対象の集団から得られた試料に由来する核酸集団に行うことができ、ここで、集団内の対象は、その後、個々の対象においてどのバリエーションヌクレオチドがコールされたかに応じて、異なる処置を受容する。

40

【0024】

一部の実施形態では、脱アミノ化エラーとして分類されるバリエーションヌクレオチドは、コールされたバリエーションヌクレオチドのうちの少なくとも1%である。

【0025】

一部の実施形態では、脱アミノ化エラーとして分類されるバリエーションヌクレオチドは、コールされたバリエーションヌクレオチドのうちの少なくとも10%である。

【0026】

50

一部の実施形態では、バリエーションの存在は、少なくとも5個のバリエーションヌクレオチドが脱アミノ化エラーとして分類される場合、コールされない。

【0027】

一部の実施形態では、核酸の集団は、固形組織に由来する。

【0028】

一部の実施形態では、体液は、血漿である。

【0029】

一部の実施形態では、5'末端に連結されるバーコードを含むアダプターは、3'末端に連結されるバーコードを含むアダプターとは異なる。

【0030】

一部の実施形態では、脱アミノ化エラーの頻度は、少なくとも1%である。

【0031】

一部の実施形態では、脱アミノ化エラーの頻度は、少なくとも10%である。

【0032】

一部の実施形態では、バリエーションヌクレオチドは、サブセット内のシーケンシングされた核酸の5'末端から、指定位置におけるCからTへの変動の平均距離が、指定位置における参照ヌクレオチドの平均距離よりも小さいこと、またはサブセット内のシーケンシングされた核酸の3'末端から、指定位置におけるGからAへの変動が、指定位置における参照ヌクレオチドの平均距離よりも小さいことに基づいて、脱アミノ化エラーとして分類される。

【0033】

一部の実施形態では、バリエーションヌクレオチドは、単一ヌクレオチドバリエーション(SNV)である。

【0034】

本開示の一態様は、核酸においてバリエーションヌクレオチドを識別する方法であって、(a)一本鎖オーバーハングを有する二本鎖核酸を、5'-3'ポリメラーゼ活性および3'-5'エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、それによって、二本鎖平滑末端化核酸を産生するステップと、(b)二本鎖平滑末端化核酸の配列を決定するステップと、(c)決定した配列を、参照配列と比較するステップであって、決定した配列が、決定した配列の5'末端の20ヌクレオチドまたはそれ未満以内の少なくとも1つの指定位置における少なくとも1つのCからTへの変動、または決定した配列の3'末端の20ヌクレオチドまたはそれ未満以内における少なくとも1つのGからAへの変動を含む、ステップと、(d)核酸の配列を、決定した配列としてコールするステップであって、ただし、CからTへの変動が決定した配列の5'末端の20ヌクレオチドまたはそれ未満以内に存在するか、またはGからAへの変動が決定した配列の3'末端の20ヌクレオチドまたはそれ未満以内に存在する位置のうちの少なくとも1つにおいては、参照配列を占有しているヌクレオチドが、指定位置でコールされることを除く、ステップとを含む、方法に関する。

【0035】

一部の実施形態では、CからTまたはGからAへの変動は、TCGからTTGまたはCGAからCAAの周囲コンテキストにおいて生じる。

【0036】

本開示の一態様は、核酸の集団においてバリエーションヌクレオチドを識別する方法であって、(a)少なくとも1つが一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子である、オーバーラップする配列の核酸の集団を、5'-3'ポリメラーゼ活性および3'-5'エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、タンパク質が、3'オーバーハングを消化し、5'オーバーハングを充填して、二本鎖平滑末端化核酸を生成する、ステップと、(b)二本鎖平滑末端化核酸を、バーコードを含むアダプターに連結させ、アダプターに結合するプライマー分子からプライミングされる核酸を増幅させるステップと、(c)増幅した核酸分子の配列を決定し、および増幅した核

10

20

30

40

50

酸分子の配列をファミリーに分類し、ファミリーのメンバーが、核酸上の同じ開始点および終止点ならびに同じアダプターを有し、ファミリーについて、そのそれぞれのメンバーの配列から、コンセンサス配列を決定する、ステップと、(d) 参照配列におけるそれぞれの指定位置について、指定位置を含むコンセンサス配列を有するファミリーのサブセットを決定し、指定位置がバリエーションヌクレオチドによって占有されているコンセンサス配列を識別する、ステップと、(e) バリエーションヌクレオチドを有するサブセット内のコンセンサス配列が、コールをサポートするそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、(i) バリエーションヌクレオチドが、参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および(ii) バリエーションヌクレオチドが、(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または(2) 5'末端から、サブセット内のコンセンサス配列の指定位置におけるCからTへの変動の距離、または3'末端から、コンセンサス配列の指定位置におけるGからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップとを含む、方法に関する。

10

#### 【0037】

一部の実施形態では、ステップ(c)は、指定位置がバリエーションヌクレオチドによって占有されているサブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、変動を有するサブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる。

20

#### 【0038】

一態様では、本開示は、核酸の集団において偽陽性バリエーションヌクレオチドを識別するための方法であって、(a) 少なくとも1つが一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子である、配列がオーバーラップする核酸の集団を、5'-3'ポリメラーゼ活性および3'-5'エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、タンパク質が、一方または両方の末端において、3'オーバーハングを消化し、5'オーバーハングを相補的な核酸で充填して、二本鎖平滑末端化核酸を生成する、ステップと、(b) シーケンシングされた核酸を得るために、二本鎖平滑末端化核酸の配列を決定するステップと、(c) 参照配列におけるそれぞれの指定位置について、指定位置を含むシーケンシングされた核酸のサブセットを識別し、指定位置が参照ヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸、および指定位置がバリエーションヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸を識別する、ステップと、(d) 指定位置でCからTまたはGからAへの変動を有するシーケンシングされた核酸がコールをサポートするそれぞれの指定位置における偽陽性バリエーションヌクレオチドの存在をコールするステップであって、変動が、(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または(2) 指定位置が5'末端の規定される近接度内にあるサブセットの第1の比率内のシーケンシングされた核酸におけるCからTへの変換の過剰出現、もしくは指定位置が3'末端の規定される近接度内にあるサブセットの第2の比率内のシーケンシングされた核酸におけるGからAへの変換の過剰出現に基づいて、脱アミノ化エラーとして分類される、ステップとを含む、方法に関する。

30

40

#### 【0039】

一部の実施形態では、ステップ(c)は、指定位置がバリエーションヌクレオチドによって占有されているサブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、変動を有するサブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる。

#### 【0040】

一態様では、本開示は、参照配列内の指定位置における「C」から「T」または「G」から「A」へのバリエーションのマイナー対立遺伝子頻度を、指定位置にマッピングされるシ

50

シーケンシングされた核酸の集団において決定する方法であって、マイナー対立遺伝子頻度が、バリエーションを含む指定位置にマッピングされるシーケンシングされた核酸の数（「バリエーション数」）を、指定位置にマッピングされるシーケンシングされた核酸の総数と比較し、この方法は、脱アミノ化エラーの確率について、指定位置におけるTまたはAバリエーションのバリエーション数を調節するステップを含み、エラーの確率が、「T」の場合には分子の5'末端からおよび「A」の場合には分子の3'末端からのバリエーションの距離の関数である方法に関する。

#### 【0041】

一部の実施形態では、シーケンシングされたポリヌクレオチドの5'末端から選択された距離内に位置するCからTへのバリエーション、またはシーケンシングされた核酸の3'末端から選択された距離内に位置するGからAへのバリエーションは、バリエーション数にカウントされない。

10

#### 【0042】

一部の実施形態では、シーケンシングされたポリヌクレオチドの5'末端から選択された距離内に位置するCからTへのバリエーションの、シーケンシングされた核酸の5'末端から選択された距離外に位置するCからTへのバリエーションに対する比が、所定の比を上回る（例えば、50%を上回る）場合、またはシーケンシングされた核酸の3'末端から選択された距離内に位置するGからAへのバリエーションの、シーケンシングされた核酸の3'末端から選択された距離外に位置するGからAへのバリエーションに対する比が、所定の比を上回る（例えば、50%を上回る）場合、すべてのCからTへのバリエーションは、バリエーション数からデイスカウントされる。

20

#### 【0043】

一部の実施形態では、バリエーション数は、それぞれのCからTへのバリエーションまたはそれぞれのGからAへのバリエーションが、真のバリエーションである確率の合計として決定される。

#### 【0044】

一態様では、本開示は、前述の項目のいずれかに記載の方法によって、がんマーカーを有することが決定された対象に、がんマーカーにより特徴付けられるがんを処置するのに有効な治療介入を投与することを含む、方法に関する。

#### 【0045】

本開示は、さらに、先行する項目のいずれかに記載の方法を行うことによって、対象の無細胞核酸における1つまたは複数のバリエーションヌクレオチドの同一性に関するデータを受容するステップと、1つまたは複数のバリエーションヌクレオチドから、がんマーカーの存在を決定するステップと、がんマーカーにより特徴付けられるがんを処置するのに有効な治療介入を投与するステップとを含む、方法を提供する。

30

#### 【0046】

別の態様では、本開示は、システムに関する。

(1) 1つのそのようなシステムは、

(2) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

(3) 通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、1つまたは複数のコンピュータープロセッサによる実行の際に、

40

(a) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容するステップと、

(b) 参照配列におけるそれぞれの指定位置について、

(i) 指定位置を含む、シーケンシングリードのサブセットを識別し、

(ii) 指定位置がバリエーションヌクレオチドによって占有されているサブセット内のシーケンシングリードを識別する、ステップと、

(c) 変動を有するサブセット内のシーケンシングリードがコールをサポートするそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、

50

(i) バリエントが、参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および

(ii) バリエントヌクレオチドが、

(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) サブセット内のシーケンシングされた核酸の5'末端から、指定位置におけるCからTへの変動の距離、またはサブセット内のシーケンシングされた核酸の3'末端から、指定位置におけるGからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合にはコールされないことを除く、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む。

10

【0047】

一部の実施形態では、ステップ(c)は、指定位置がバリエントヌクレオチドによって占有されているサブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、変動を有するサブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエントヌクレオチドの存在がコールされる。

【0048】

本開示は、さらに、

(1) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

20

(2) 通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容するステップと、

(b) 決定した配列を、参照配列と比較するステップであって、決定した配列が、決定した配列の5'末端の20ヌクレオチドまたはそれ未満以内の少なくとも1つの指定位置における少なくとも1つのCからTへの変動、または決定した配列の3'末端の20ヌクレオチドまたはそれ未満以内における少なくとも1つのGからAへの変動を含む、ステップと、

(c) 核酸の配列を、決定した配列としてコールするステップであって、ただし、CからTへの変動が決定した配列の5'末端の20ヌクレオチドまたはそれ未満以内に存在するか、またはGからAへの変動が決定した配列の3'末端の20ヌクレオチドまたはそれ未満以内に存在する位置のうちの少なくとも1つにおいては、参照配列を占有しているヌクレオチドが、指定位置でコールされることを除く、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む、システムを提供する。

30

【0049】

本開示は、さらに、

(1) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

40

(2) 通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容するステップと、

(b) シーケンシングリードの配列をファミリーに分類するステップであって、ファミリーのメンバーが、核酸上の同じ開始点および終止点ならびに同じアダプターを有し、ファミリーについて、そのそれぞれのメンバーの配列から、コンセンサス配列を決定する、ステップと、

(c) 参照配列におけるそれぞれの指定位置について、指定位置を含むコンセンサス配列を有するファミリーのサブセットを決定し、指定位置がバリエントヌクレオチドによ

50

て占有されているコンセンサス配列を識別する、ステップと、

(d) それぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、バリエーションヌクレオチドを有するサブセット内のコンセンサス配列が、コールをサポートするが、指定位置におけるバリエーションヌクレオチドの存在は、

(i) バリエーションヌクレオチドが、参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および

(ii) バリエーションヌクレオチドが、

(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 5'末端から、サブセット内のコンセンサス配列の指定位置におけるCからTへの変動の距離、または3'末端から、コンセンサス配列の指定位置におけるGからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合

には、コールされないことを除く、ステップと

を含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターと

を含む、システムを提供する。

【0050】

一部の実施形態では、ステップ(c)は、指定位置がバリエーションヌクレオチドによって占有されているサブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、変動を有するサブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる。

【0051】

本開示は、さらに、

(1) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

(2) 通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容するステップと、

(b) 参照配列におけるそれぞれの指定位置について、指定位置を含むシーケンシングリードのサブセットを識別し、指定位置が参照ヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸、および指定位置がバリエーションヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸を識別する、ステップと、

(c) 指定位置にCからTまたはGからAへの変動指定位置でCからTまたはGからAへの変動を有するシーケンシングされた核酸がコールをサポートするそれぞれの指定位置における偽陽性バリエーションヌクレオチドの存在をコールするステップであって、変動が、

(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 指定位置が5'末端の規定される近接度内にあるサブセットの第1の比率内のシーケンシングされた核酸におけるCからTへの変換の過剰出現、もしくは指定位置が3'末端の規定される近接度内にあるサブセットの第2の比率内のシーケンシングされた核酸におけるGからAへの変換の過剰出現

に基づいて、脱アミノ化エラーとして分類される、ステップと

を含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターと

を含む、システムを提供する。

【0052】

一部の実施形態では、ステップ(c)は、指定位置がバリエーションヌクレオチドによって占有されているサブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、変動を有するサブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在が

10

20

30

40

50

コールされる。

【0053】

本開示は、さらに、

(1) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

(2) 通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容するステップと、

(b) 脱アミノ化エラーの確率に基づいてシーケンシングリードにおけるTまたはAバリエーションの数を調節するステップであって、エラーの確率が、「T」の場合には分子の5'末端からおよび「A」の場合には分子の3'末端からのバリエーションの距離の関数である、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む、システムを提供する。

10

【0054】

上述のシステムのいずれも、核酸シーケンサーをさらに含んでもよい。必要に応じて、核酸シーケンサーは、対象に由来する無細胞DNA分子から得られたシーケンシングライブラリーがシーケンシングし、ここで、シーケンシングライブラリーは、無細胞DNA分子およびアダプターを含み、アダプターは、バーコードを含む。必要に応じて、核酸シーケンサーは、シーケンシングライブラリーにシーケンシングバイセンシスを行って、シーケンシングリードを生成する。必要に応じて、核酸シーケンサーは、シーケンシングライブラリーにパイロシーケンシング、単分子シーケンシング、ナノポアシーケンシング、半導体シーケンシング、シーケンシングバイライゲーション、またはシーケンシングバイハイブリダイゼーションを行って、シーケンシングリードを生成する。必要に応じて、核酸シーケンサーは、シーケンシングライブラリーに由来するクローン単一分子アレイを使用して、シーケンシングリードを生成する。必要に応じて、核酸シーケンサーは、シーケンシングライブラリーをシーケンシングしてシーケンシングリードを生成するための、マイクロウェルのアレイを有するチップを含む。

20

【0055】

一部のシステムにおいて、コンピューター可読媒体は、メモリー、ハードドライブ、またはコンピューターサーバーを含む。一部のシステムにおいて、通信ネットワークは、遠隔通信ネットワーク、インターネット、エクストラネット、またはイントラネットを含む。一部のシステムにおいて、通信ネットワークは、分散コンピューティング、例えば、クラウドコンピューティングの可能な1つまたは複数のコンピューターサーバーを含む。一部のシステムにおいて、コンピューターは、核酸シーケンサーから遠隔設置されているコンピューターサーバー上に設置されている。一部のシステムにおいて、シーケンシングライブラリーは、試料を、1つまたは複数の試料と区別する試料バーコードをさらに含む。

30

【0056】

一部のシステムは、ネットワークを通じてコンピューターと通信する、電子ディスプレイであって、(a)～(c)を実施した際の結果を表示するためのユーザーインターフェース、例えば、グラフィカルユーザーインターフェース(GUI)またはウェブベースユーザーインターフェースを含む、電子ディスプレイをさらに含む。一部のシステムにおいて、電子ディスプレイは、パーソナルコンピューターにおいて存在する。一部のシステムにおいて、電子ディスプレイは、必要に応じてコンピューターから遠隔した位置にある、インターネット対応コンピューターにおいて存在する。

40

【0057】

一部の実施形態では、本明細書で開示されたシステムおよび方法の結果は、インプットとして使用され、紙形式でレポートが作成される。例えば、このレポートは、コールされたバリエーションおよび／または脱アミノ化のエラーと考えられるバリエーションのインジケーションを提供することができる。

50

## 【0058】

本明細書で開示された方法の様々なステップ、または本明細書で開示されたシステムによって実行されるステップは、同一もしくは異なる時間に、同一もしくは異なる地理的位置、例えば、国において、および／または同一もしくは異なる人々によって実行され得る。

## 【図面の簡単な説明】

## 【0059】

【図1】図1は、クレノウポリメラーゼによる末端修復の概要を示す。

## 【0060】

【図2】図2は、CからTへの脱アミノ化スキームを示す。

## 【0061】

【図3】図3は、分子の5'末端におけるCからTへの変換および分子の3'末端におけるGからAへの変換の優先度を示す。

## 【0062】

【図4】図4は、CからTおよびGからAへの変換のエラーの頻度および他の変換のものを比較するプロットを、分子の末端からの距離とともに示す。CからTおよびGからAへの変換のエラー頻度は、分子の末端の付近ではより高いが、一方で、他の変換のものは、分子の末端に対する位置から独立している。「CからTまたはGからA」と標識された点は、5'末端からの測定距離により層別化したCからTのエラー率および3'末端からの測定距離により層別化したGからAのエラー率の平均を示し、「他のエラー」と標識された点は、5'末端からの測定距離によって層別化したCからA + CからGのエラー率および3'末端からの測定距離によって層別化したGからT + GからCのエラー率の平均を示す。

## 【0063】

【図5】図5は、コンピューターシステムを示す。

## 【0064】

【図6】図6は、脱アミノ化エラーとして分類されるGからAへの置換を含む、5つのシーケンシングファミリーを示す。

## 【0065】

【図7】図7は、真実の変異として分類されるGからAへの置換を含む、5つのシーケンシングファミリーを示す。

## 【発明を実施するための形態】

## 【0066】

## 定義

対象は、動物、例えば、哺乳動物種（好ましくは、ヒト）もしくは鳥類（例えば、鳥）種、または他の生物、例えば、植物を指す。より具体的には、対象は、脊椎動物、例えば、哺乳動物、例えば、マウス、霊長類、サル、またはヒトであり得る。動物には、家畜動物、競技動物、およびペットが含まれる。対象は、健康な個体、症状もしくは徴候を有するか、または疾患もしくは疾患の傾向を有することが疑われる個体、または治療を必要とするかもしくは治療を必要とすることが疑われる個体であり得る。

## 【0067】

遺伝子バリエーションは、対象の核酸試料またはゲノムにおける、変更、バリエーション、または多型を指す。そのような変更、バリエーション、または多型は、参照ゲノムに対するものであり得、これは、種（例えば、ヒトについては、hg19もしくはhg38）、対象、または他の個体の参照ゲノムであり得る。変動としては、1つまたは複数の一塩基多型（SNV）、挿入、欠失、リピート、短い挿入、短い欠失、短いリピート、構造バリエーション、可変長タンDEMリピート、および／またはフランキング配列が挙げられ、コピー数バリエーション（CNV）、塩基転換、遺伝子融合、および他の再配列もまた、遺伝子変動の形態である。変動は、塩基の変化、挿入、欠失、リピート、コピー数多型、塩基転換、またはこれらの組合せであり得る。

## 【0068】

がんマーカーは、がんの存在またはそれを発症する危険性と関連付けられた、遺伝子バ

10

20

30

40

50

リアントである。がんマーカーは、対象が、がんを有するか、またはがんを発症する危険性が、年齢および性別が一致し、がんマーカーを有さない同じ種の対象よりも高いことという指標を提供することができる。がんマーカーは、がんの原因である場合もそうでない場合もある。

#### 【0069】

バーコードは、短い核酸（例えば、500未満、100未満、50未満、または10未満のヌクレオチド長）であり、核酸を、異なる種類または異なるプロセッシングを受けた、異なる試料（例えば、試料インデックスを提示する）、または同じ試料中の異なる核酸分子（例えば、バーコードを提示する）と区別するために、核酸分子を標識するのに使用される。タグは、一本鎖、二本鎖、または少なくとも部分的に二本鎖であり得る。タグは、同じ長さを有してもよく、または変動した長さを有してもよい。タグは、平滑末端であってもよく、またはオーバーハングを有してもよい。タグは、核酸の一方の末端または両方の末端に結合され得る。バーコードは、デコードされると、核酸の起源試料、形態、またはプロセッシングなどの情報を示し得る。タグを使用して、異なるバーコードおよび／または試料インデックスを有する核酸を含む複数の試料のプールおよび並行したプロセッシングを可能にすることができ、核酸は、後でバーコードを読み取ることによりデコンボリューションされる。バーコードはまた、分子識別子、試料識別子、インデックスタグ、および／またはタグとも称され得る。追加または代替として、バーコードは、同じ試料中の異なる分子を区別するために使用することができる。これは、試料中のそれぞれの異なる分子を固有にバーコーディングすること、またはそれぞれの分子を非固有にバーコーディングすることを含む。非固有バーコーディングの場合には、異なる分子が、少なくとも1つのタグと組み合わせて、参照ゲノム上にマッピングされるそれらの開始／終止位置に基づいて区別され得るように、限られた数のバーコードを使用して、それぞれの分子をバーコーディングしてもよい。そのため、典型的には、同じ開始／終止を有する任意の2つの分子が、同じバーコードも有する確率が低くなるように（例えば、10%未満、5%未満、1%未満、または0.1%未満）、十分な数の異なるバーコードが使用される。一部のバーコードは、複数の試料、1つの試料内の複数の分子形態、ならびに同じ開始点および終止点を有する1つの形態内の複数の分子を標識するための複数の分子識別子を含む。そのようなバーコードは、形式A1iで存在し得、ここで、文字は、試料の種類を示し、アラビア数字は、試料内の分子の形態を示し、ローマ数字は、形態内の分子を示す。

#### 【0070】

アダプターは、通常、試料核酸分子のいずれか一端または両端への連結のために少なくとも部分的に二重鎖の短い核酸（例えば、500、100または50ヌクレオチド未満の長さ）である。アダプターは、両端においてアダプターに隣接する核酸分子の増幅を可能とするプライマー結合部位、および／または次世代シーケンシング（NGS）に対するプライマー結合部位を含むシーケンシングプライマー結合部位を含み得る。アダプターは、フローセル支持体に付着したオリゴヌクレオチドなどの捕捉用プローブに対する結合部位も含み得る。アダプターは、上述のバーコードも含み得る。好ましくは、バーコードは、核酸分子のアンプリコンおよびシーケンシングリードに含まれるように、プライマーおよびシーケンシングプライマー結合部位に対して配置される。核酸分子の各末端に、同一または異なるアダプターを連結することができる。同一のアダプターは、バーコードが異なることを除いて、各末端に連結されることがある。好ましいアダプターは、一端が、1つまたは複数の相補的ヌクレオチドに関しても平滑末端またはテイルである核酸分子に接合するために、本明細書に記載されているように平滑末端またはテイルである、Y型アダプターである。別の好ましいアダプターは、解析される核酸に接合するために平滑またはテイル末端を同様に有する、釣鐘型アダプターである。

#### 【0071】

本明細書で使用される場合、用語「シーケンシング」は、生体分子、例えば、DNAまたはRNAなどの核酸分子の配列を決定するために使用されるいくつかの技術のうちのいずれかを指す。例示的なシーケンシング方法として、これらに限定されないが、ターゲッ

トシーケンシング、単分子リアルタイムシーケンシング、エクソンシーケンシング、電子顕微鏡に基づくシーケンシング、パネルシーケンシング、トランジスタ媒介型シーケンシング、直接シーケンシング、ランダムショットガンシーケンシング、サンガーのジデオキシ末端シーケンシング、全ゲノムシーケンシング、ハイブリダイゼーションによるシーケンシング、パイロシーケンシング、キャピラリー電気泳動、ゲル電気泳動、デュプレックスシーケンシング、サイクルシーケンシング、一塩基伸長シーケンシング、固相シーケンシング、ハイスループットシーケンシング、超並列署名シーケンシング、エマルジョンPCR、低変性温度における共増幅PCR（COLD-PCR）、マルチプレックスPCR、可逆的色素ターミネーターによるシーケンシング、ペアドエンドシーケンシング、短期シーケンシング、エクソヌクレアーゼシーケンシング、ライゲーションによるシーケンシング、短リードシーケンシング、単分子シーケンシング、シーケンシングバイシンセシス、リアルタイムシーケンシング、リバースターミネーターシーケンシング、ナノポアシーケンシング、454シーケンシング、Solexaゲノムアナライザーシーケンシング、SOLID（商標）シーケンシング、MS-PETシーケンシング、およびこれらの組合せが挙げられる。一部の実施形態では、シーケンシングは、例えば、IlluminaまたはApplied Biosystemsから市販されている遺伝子アナライザーなどの遺伝子アナライザーにより実施することができる。

【0072】

表現「次世代シーケンシング」またはNGSは、従来のサンガーおよびキャピラリー電気泳動に基づく手法と比較してスループットが増加したシーケンシング技術を指し、例えば、一度に数十万もの比較的小さな配列リードを作成する能力を有する。次世代シーケンシング技法のいくつかの例として、これらに限定されないが、シーケンシングバイシンセシス、シーケンシングバイライゲーション、およびシーケンシングバイハイブリダイゼーションが挙げられる。

【0073】

表現「シーケンシングラン」は、少なくとも1つの生体分子（例えば、DNAまたはRNAなどの核酸分子）に関する一部の情報を決定するために実施されるシーケンシング実験の任意のステップまたは部分を指す。

【0074】

DNA（デオキシリボ核酸）は、4種のヌクレオチド；アデニン（A）、チミン（T）、シトシン（C）、およびグアニン（G）を含むヌクレオチド鎖である。RNA（リボ核酸）は、4種のヌクレオチド；A、ウラシル（U）、G、およびCを含むヌクレオチド鎖である。特定のヌクレオチド対は、相補的様式で互いに特異的に結合する（相補的塩基対合と称される）。DNAでは、アデニン（A）はチミン（T）と対合し、シトシン（C）はグアニン（G）と対合する。RNAでは、アデニン（A）はウラシル（U）と対合し、シトシン（C）はグアニン（G）と対合する。第1の核酸鎖が、第1の鎖におけるヌクレオチドと相補的であるヌクレオチドから構成される第2の核酸鎖に結合する場合、2つの鎖は結合して二本鎖を形成する。本明細書で使用される場合、「核酸シーケンシングデータ」、「核酸シーケンシング情報」、「核酸配列」、「ヌクレオチド配列」、「ゲノム配列」、「遺伝子配列」、または「断片配列」、または「核酸シーケンシングリード」は、DNAまたはRNAなどの核酸分子（例えば、全ゲノム、全トランスクリプトーム、エキソーム、オリゴヌクレオチド、ポリヌクレオチド、または断片）におけるヌクレオチド塩基（例えば、アデニン、グアニン、シトシン、およびチミンまたはウラシル）の順序を表示する任意の情報またはデータを示す。本発明の教示によって、これらに限定されないが、キャピラリー電気泳動、マイクロアレイ、ライゲーションに基づくシステム、ポリメラーゼに基づくシステム、ハイブリダイゼーションに基づくシステム、直接的または間接的なヌクレオチド同定システム、パイロシーケンシング、イオンまたはpHベース検出システム、および電子署名ベースのシステムを含む技法、プラットフォームまたは技術のすべての利用可能な変形を使用して得られた配列情報が考慮されることが理解されるべきである。

## 【0075】

「ポリヌクレオチド」、「核酸」、「核酸分子」、または「オリゴヌクレオチド」は、ヌクレオシド間の連結によって接合されたヌクレオシド（デオキシリボヌクレオシド、リボヌクレオシド、またはその類似体を含む）の線状ポリマーを指す。典型的には、ポリヌクレオチドは、少なくとも3つのヌクレオシドを含む。オリゴヌクレオチドのサイズは、少数のモノマー単位、例えば、3～4から数百のモノマー単位の範囲であることが多い。ポリヌクレオチドが、「ATGCCCTG」などの文字配列で表される場合は常に、ヌクレオチドは、別段に注記されていなければ、左から右への5'から3'の順であり、「A」はデオキシアデノシンを示し、「C」はデオキシシチジンを示し、「G」はデオキシグアノシンを示し、「T」はチミジンを示すことが理解されよう。当技術分野の標準であるように、塩基自体、塩基を含むヌクレオシド、またはヌクレオチドを指すために、文字A、C、G、およびTを使用することができる。

10

## 【0076】

参照配列は、実験的に決定された配列と比較するために使用される既知の配列である。例えば、既知の配列は、ゲノム全体、染色体、またはその任意のセグメントであり得る。参照は、典型的には、少なくとも20、50、100、200、250、300、350、400、450、500、1000、またはそれより多いヌクレオチドを含む。参照配列は、ゲノムもしくは染色体の単一の連続配列と整列させることができるか、またはゲノムもしくは染色体の異なる領域と整列させられている非連続的セグメントを含むことができる。参照ヒトゲノムとして、例えば、hg19およびhg38が挙げられる。

20

## 【0077】

参照配列における「指定位置」という用語は、参照配列におけるゲノム座標を指す。

## 【0078】

第1の核酸配列またはその相補体と第2の核酸配列またはその相補体が、ヒト染色体の配列などの連続参照配列の非相同セグメントを除いて、重複して整列される場合、第1の単鎖核酸配列は、第2の単鎖核酸配列と重複する。全体的にまたは部分的に二本鎖の核酸は、その鎖のいずれかが他の核酸の鎖と重複する場合、別の全体的にまたは部分的に二本鎖の核酸と重複する。

## 【0079】

「C」から「T」へのバリエーションまたは変換は、参照配列では塩基「C」によって占有されている座標位置における、シーケンシングされたポリヌクレオチドでの塩基「T」の存在を指す。「G」から「A」へのバリエーションまたは変換は、参照配列では塩基「G」によって占有されている座標位置における、シーケンシングされたポリヌクレオチドでの塩基「A」の存在を指す。

30

## 【0080】

核酸分子は、概念として、5'末端、内部部分、および3'末端に分割することができる。末端は、末端から所定数のヌクレオチドに基づいて表記することができる。例えば、5'末端は、例えば、5'末端までの20個の末端ヌクレオチドで表される。3'末端は、例えば、3'末端までの20個の末端ヌクレオチドで表される。あるいは、核酸分子は、記載されるような末端部分と、残部とに分割してもよい。

40

## 【0081】

「マイナー対立遺伝子頻度」という用語は、所与の核酸集団、例えば、試料において、マイナー対立遺伝子（例えば、もっとも一般的な対立遺伝子ではないもの）が生じる頻度を指す。低いマイナー対立遺伝子頻度の遺伝子バリエーションは、試料における比較的低い存在頻度を有し得る。

## 【0082】

「マイナー対立遺伝子比率」（MAF）は、所与の試料における対立遺伝子変異を有するDNA分子の比率を指す。体細胞バリエーションのMAFは、0.5未満、0.1未満、0.05未満、または0.01未満であり得る。例えば、体細胞バリエーションのMAFは、0.05未満である。

50

## 【0083】

用語「プロセシングすること」、「算出すること」、および「比較すること」は、交換可能に使用される。この用語は、差、例えば、数または配列の差を決定することを指すことができる。例えば、遺伝子発現、コピー数変動（CNV）、インデル、および／または単一ヌクレオチド変異（SNV）値または配列をプロセシングすることができる。

## 【0084】

アダプターは、ライゲーション、ハイブリダイゼーション、および／または増幅を含む、任意のアプローチによって、核酸分子またはポリヌクレオチド配列に連結することができる、人工的に合成された配列である。アダプターは、通常、試料核酸分子のいずれかの末端または両方の末端への連結のために、少なくとも部分的に二本鎖になっている、短い核酸（例えば、500ヌクレオチド長未満、100ヌクレオチド長未満、または50ヌクレオチド長未満）である。アダプターは、両方の末端においてアダプターが隣接した核酸分子の増幅を可能にするプライマー結合部位、および／または次世代シーケンシング（NGS）のためのプライマー結合部位を含むシーケンシングプライマー結合部位を含み得る。アダプターはまた、捕捉用プローブ、例えば、フローセル支持体に結合したオリゴヌクレオチドの結合部位も含み得る。アダプターはまた、上述のようなバーコードも含み得る。タグがアンプリコンおよび核酸分子のシーケンシングリードに含まれるように、タグは、好ましくは、プライマーおよびシーケンシングプライマー結合部位に対して位置付けられる。同じかまたは異なるアダプターが、核酸分子のそれぞれの末端に連結され得る。同じアダプターが、それぞれの末端に連結されることがあるが、ただし、タグは異なる。好ましいアダプターは、Y字型アダプターであり、その場合、一方の末端が、平滑末端であるかまたは本明細書に記載されるように核酸分子への結合のための尾部があり、この核酸分子もまた、平滑末端であるかまたは1つもしくは複数の相補的ヌクレオチドを有する尾部を有する。別の好ましいアダプターは、釣鐘型アダプターであり、同様に、平滑末端または分析しようとする核酸への結合のための尾部を有する末端を有する。

## I. 概要

## 【0085】

核酸をシーケンシングすることにより、疾患の存在、易罹患性、または予後と関連する変動を識別することができる。しかしながら、そのような情報に関する値は、シーケンシングのための核酸を調製することを含むシーケンシングプロセスまたは他の因子、例えば、輸送および／もしくは最初の研究室でのプロセシング中に核酸試料の品質に影響を及ぼす環境条件によって導入されるエラーによって、被害を受ける可能性がある。品質に影響を及ぼす環境条件としては、温度およびプロセシング前の保管期間の長さが挙げられる。本開示は、ある特定の態様において、試料中の核酸における一本鎖オーバーハングを平滑化することが、核酸鎖の5'末端においてシトシン（C）がチミン（T）に変化し、その結果相補的核酸鎖の3'末端においてグアニン（G）からアデニン（A）への相補的塩基の変化が生じる、脱アミノ化に誘導されるシーケンシングエラーが導入される重大な傾向を有するという観察を根拠としている。本開示は、そのような脱アミノ化に誘導されるシーケンシングエラーを識別し、それらを、がんまたは他の疾患と関連し得る実際の配列変動と区別する方法を提供する。

## II. 脱アミノ化に誘導されるエラーを識別し、補正する方法

## 【0086】

核酸は、塩基「C」が塩基「T」に変換される脱アミノ化に供され得る。この場合、二本鎖分子において、一方の鎖は、「T」を有し、相補鎖は、「G」を有することになる。そのようなエラーは、異なる鎖の配列を追跡すれば、シーケンシングにより検出することができる。

## 【0087】

本方法は、少なくとも一本鎖オーバーハングを有する部分的に二本鎖である任意の核酸、またはそのような核酸を含む集団に、行うことができる。典型的に、本方法は、集団の少なくとも一部が、一方または両方の末端に一本鎖オーバーハングを有する部分的に二本

鎖である核酸の集団に行われる。本方法は、例えば、少なくとも2個、10,000個、1,000,000個、1,000,000,000個、10,000,000,000個、またはそれよりも多くの異なるそのような核酸を含む集団に、行うことができる。通常、集団における、一本鎖オーバーハングを有するものを含む少なくとも一部の核酸は、オーバーラップする配列のものである。そのような集団は、天然に存在し得るかもしくは試料の調製中の断片化の結果として存在し得るか、または酵素により、例えば、部分的な制限消化によって、生成され得る。

#### 【0088】

核酸集団の好ましい形態は、無細胞核酸、例えば、血液および他の体液中に存在するものである。そのような核酸は、典型的に、一方または両方の末端に一本鎖オーバーハングを有する二本鎖DNA、ならびに一本鎖DNAおよびRNAを含む、不均質な形態にある。二本鎖平滑末端化DNAもまた、存在し得る。

#### 【0089】

核酸集団は、一方または両方の末端に一本鎖オーバーハングを有する二本鎖核酸の酵素による平滑末端化によって、シーケンシングのために調製することができる。集団は、ヌクレオチド（例えば、A、C、G、およびTまたはU）の存在下において、5'-3' DNAポリメラーゼ活性および3'-5'エキソヌクレアーゼ活性を有するタンパク質で処置され得る。例示的なタンパク質は、DNAポリメラーゼ、例えば、クレノウ大型断片およびT4 DNAポリメラーゼである。5'オーバーハングにおいて、タンパク質は、相補鎖における凹んだ3'末端を、5'末端と同じになるまで延ばし、平滑末端が得られる。3'オーバーハングにおいて、タンパク質は、3'末端から、反対側の鎖の5'末端まで、およびときには5'末端を越えて、消化する。消化が、反対側の鎖の5'末端を越えて進行した場合、ギャップが、5'オーバーハングのように、ポリメラーゼ活性によって充填され得る。二本鎖核酸の平滑末端化により、アダプターの結合および後続の増幅が容易となる。

#### 【0090】

しかしながら、平滑末端化プロセスにより、図1〜3に示されるように、脱アミノ化エラーも導入され得る。図1は、クレノウ酵素が、5'オーバーハングを充填し、3'オーバーハングを消化する、スキームを示す。図2および3は、C-T脱アミノ化に誘導されるエラーが、ワトソン鎖の5'末端に導入され、相補的なG-Aエラーが相補的なクリック鎖の3'末端に導入される、スキームを示す。脱アミノ化に誘導されるCからTへの変換は、丸で囲んだTによって示される。丸で囲んだAは、相補鎖における対応する変化を表す。5'ワトソン鎖における脱アミノ化に誘導されるエラーは、ワトソン鎖の5'オーバーハングに基づく3'末端の伸長、例えば、ワトソン鎖におけるCからTへの変換およびクリック鎖におけるGからAへの変換に起因して、相補的ヌクレオチドとしてクリック鎖の3'末端に再現される。二本鎖領域における脱アミノ化に誘導されるエラーは、充填または消化プロセスによって再現されず、2つの鎖は、その位置またはヌクレオチドにおいて、非相補的ヌクレオチドを有する。ワトソン鎖の3'末端における脱アミノ化に誘導されるエラーは、消化により除去される。クリック鎖の5'末端付近の脱アミノ化により誘導されるエラーは、ワトソン鎖の3'末端が、脱アミノ化に誘導されるエラーに相補的なヌクレオチドの充填を必要とするように、消化して戻される場合、保持され得る。一部の実施形態では、鎖の5'末端におけるCからTへの変動および鎖の3'末端におけるGからAへの変動のみが、核酸分子の両方の鎖に出現する。

#### 【0091】

したがって、末端修復された二本鎖分子の場合、もとの分子のワトソン鎖における5'オーバーハングに位置する「C」から「T」への変換は、Tエラーで表され、すべての増幅された分子において、相補鎖におけるAとして伝播される。もとの分子の二本鎖部分に位置する「C」から「T」への変換は、一方の鎖においてGで表され、相補鎖においてはAとして表される。増幅の際に、エラーは、一方の鎖、例えば、ワトソン鎖において「T」として、および相補鎖、例えば、クリック鎖の同じ位置において「A」および「G」の混合物として伝播される可能性が高い。もとの分子のワトソン鎖において3'オーバーハング

に位置する「C」から「T」への変換は、消化され、オーバーハングから排除されて、平滑末端化二本鎖分子が形成される。ワトソン鎖に3'オーバーハングを有する分子のクリック鎖の5'末端付近に位置する「C」から「T」への変換は、5'オーバーハングが消化して戻されており、充填されると、ワトソン／クリック鎖において、T／Aとして出現し得る。これは、すべての増幅された分子において、T／Aとして伝播される可能性が高い。したがって、シーケンシングすると、もとのワトソン鎖からのリードは、Tを含むが、もとのクリック鎖からのリードは、Gを含むため、もとの分子の二本鎖部分における「C」から「T」への変換は、エラーとして検出され得る。対照的に、もとの分子のワトソン鎖における5'オーバーハングに位置する「C」から「T」への変換は、ワトソン／クリック鎖において、それぞれ、相補的なT／Aをもたらす。したがって、5'および3'両方のオーバーハングにおけるヌクレオチドの変換は、典型的に、自明のエラーも、二本鎖によるサポート、例えば、A／T（ワトソン／クリック）またはC／G（ワトソン／クリック）も、提供しない。

10

#### 【0092】

核酸集団は、一本鎖核酸から二本鎖への変換、および／またはRNAからDNAへの変換といった、さらなるプロセッシングに供することができる。これらの形態の核酸もまた、アダプターに連結させ、増幅させることができる。

#### 【0093】

前の増幅の有無にかかわらず、上述のように平滑末端化に供される核酸、および必要に応じて試料中の他の核酸を、シーケンシングして、シーケンシングされた核酸が得られる。シーケンシングされた核酸は、核酸を冗長的にシーケンシングした後に（例えば、単一分子の増幅もしくは再リーディングによって）産生された配列リードを含む、核酸の配列、または配列が決定された核酸のいずれかを指し得る。シーケンシングは、試料中の個々の核酸分子の配列データが、直接的または間接的のいずれかで、試料中の個々の核酸分子の増幅産物のコンセンサス配列から得られるように、行われる。

20

#### 【0094】

一部の方法では、平滑末端化を行った後の試料中の一本鎖オーバーハングを有する二本鎖核酸は、両方の末端において、バーコードまたはタグを含むアダプターに連結され（ライゲーションもしくはプライマー伸長によって結合される）、シーケンシングによって、核酸配列、ならびにアダプター内のバーコードが決定される。平滑末端化されたDNA分子は、平滑末端が、少なくとも部分的に二本鎖のアダプター（例えば、Y字型または釣鐘型のアダプター）の平滑末端とライゲーションされ得る。あるいは、試料核酸およびアダプターの平滑末端は、ライゲーションを容易にするために相補的なヌクレオチドが尾部にあってもよい。例えば、アダプターは、尾部、例えば、鎖のうちの一方に結合または連結された少なくとも1つのヌクレオチドを有してもよく、少なくとも1つのヌクレオチドは、目的の核酸分子に導入されるオーバーハングに対して相補的である。アダプターの尾部は、ヌクレオチドA、T、C、またはGのうちのいずれか1つまたは複数であり得る。

30

#### 【0095】

試料は、同じ核酸の任意の2つのインスタンスが、一方の末端または両方の末端に連結されたアダプターから、同じバーコードの組合せを受容する確率が低くなるように（例えば、1%未満または0.1%未満）十分な数のアダプターと接触され得る。アダプターをこの様式で使用するにより、参照核酸上の同じ開始点および終止点を有し、同じバーコードの組合せに連結された配列を、同じもとの分子から生成されたリードのファミリーにグループ分けすることが可能となる。そのようなファミリーは、増幅前の試料中の核酸の増幅産物の配列を提示する。ファミリーメンバーの配列をコンパイルして、平滑末端化およびアダプター結合によって改変された、もとの試料中の核酸分子のコンセンサスヌクレオチドまたは完全なコンセンサス配列を導出することができる。換言すると、試料中の核酸の特定の位置を占有しているヌクレオチドは、ファミリーメンバー配列においてその対応する位置を占有しているヌクレオチドのコンセンサスであると決定される。コンセンサスヌクレオチドは、2つの方法を挙げると、投票または信頼性スコアなどの方法によっ

40

50

て、決定することができる。ファミリーには、二本鎖核酸の一方の鎖または両方の鎖の配列が含まれ得る。ファミリーのメンバーが、二本鎖核酸由来の両方の鎖の配列を含む場合、一方の鎖の配列を、すべての配列をコンパイルする目的でそれらの相補物に変換して、コンセンサヌクレオチドまたは配列を導出する。一部のファミリーは、単一メンバー配列のみを含み得る。この場合には、この配列は、増幅前の試料中の核酸の配列として解釈され得る。あるいは、単一メンバー配列のみを有するファミリーは、後続の分析から排除してもよい。

#### 【0096】

シーケンシングされた核酸におけるヌクレオチド変動は、シーケンシングされた核酸を参照配列と比較することによって、決定することができる。参照配列は、公知の配列、例えば、ある対象由来の公知の全体的または部分的ゲノム配列、ヒト対象の全ゲノム配列であることが多い。参照配列は、hg19であってもよい。シーケンシングされた核酸は、試料中の核酸について直接的に決定された配列、または上述のように、そのような核酸の増幅産物の配列のコンセンサスを提示し得る。比較は、参照配列における1つまたは複数の指定位置で行われ得る。それぞれの配列を最大にアライメントしたときに、参照配列の指定位置に対応する位置を含む、シーケンシングされた核酸のサブセットを、識別することができる。そのようなサブセット内で、あるとすればどのシーケンシングされた核酸が、指定位置にヌクレオチド変動を含むか、ならびに必要に応じて、あるとすればどれが参照ヌクレオチド（すなわち、参照配列におけるものと同じもの）を含むか、および／または参照ヌクレオチドを含むサブセット内の配列の数を、決定することができる。ヌクレオチド変動を含むシーケンシングされた核酸によってサポートされる場合に、バリエーションがコールされ得る。例えば、ヌクレオチドバリエーションを含むサブセット内のシーケンシングされた核酸の数が、閾値を超えた場合、バリエーションヌクレオチドが、指定位置においてコールされ得る。閾値は、単純な数字、例えば、少なくとも1、2、3、4、5、6、7、8、9、もしくは10個の、ヌクレオチドバリエーションを含むサブセット内のシーケンシングされた核酸であってもよく、または閾値は、比、例えば、他の可能性の中でもとりわけ、少なくとも0.5、1、2、3、4、5、10、15、もしくは20%の、ヌクレオチドバリエーションを含むサブセット内のシーケンシングされた核酸であってもよい。比較は、参照配列における任意の目的とされる指定位置について、繰り返すことができる。しばしば、比較は、参照配列上の少なくとも20個、100個、200個、または300個の連続した位置、例えば、20～500個または50～300個の連続した位置を占有する指定位置について、行われ得る。他の変動をコールするのに使用した同じ信頼性を有するサブセット内のシーケンシングされた核酸によってサポートされるCからTまたはGからAへの変動は、それにもかかわらず、脱アミノ化に誘導されるシーケンシングエラーを含み得る。

#### 【0097】

脱アミノ化により誘導されるシーケンシングエラーは、コールされるバリエーションヌクレオチドからそれらを排除するための措置をとらない限り、意図せずともコールされるバリエーションヌクレオチドに含まれてしまう。脱アミノ化に誘導されるエラーは、2つの基本的な基準のうちのいずれかまたは両方によって認識することができる。第1に、脱アミノ化エラーは、コンテキスト依存性である。シトシンからチミンへの脱アミノ化は、シトシンに、チミンおよびグアニンが隣接している場合に（すなわち、TCGとして）、他のヌクレオチドが隣接している場合よりも多く生じる。同様に、（相補鎖における）グアニンからアデニンへの変動は、グアニンに、CGAとしてCおよびAが隣接している場合に、他のヌクレオチドが隣接している場合よりも高い頻度で生じる。したがって、脱アミノ化に誘導されるエラーは、CからTまたはGからAへの変動が、それぞれ、TCGからTTGまたはCGAからCAAのコンテキストで生じる場合に、コールされ得る。一部の方法では、脱アミノ化エラーのうちの約90%は、これらのコンテキストで生じる。

#### 【0098】

第2に、脱アミノ化に誘導されるエラーは、指定位置と、シーケンシングされた核酸の

末端との間の距離、すなわち換言すると、これらの位置間のヌクレオチドの数に依存する。例えば、配列の内部部分で生じる脱アミノ化に誘導されるエラーは、一方の鎖からのリードでは「T」として、相補鎖からのリードでは「G」として検出可能である可能性が高い。しかしながら、シーケンシングされている核酸の端部（末端）の近傍で生じる脱アミノ化に誘導されるエラーは、そのようなエラーが、平滑末端修復のプロセスによって導入され、2つの完全に相補的な鎖が生じ得るため、明白ではない可能性がある。より具体的には、シトシンからチミンへの脱アミノ化を含む配列リードは、シーケンシングされた核酸の5'末端の近傍において、より高頻度で生じ得、グアニンからアデニンへの脱アミノ化は、3'末端の近傍において、より高頻度で生じ得る。したがって、指定位置において脱アミノ化により生じるCからTへの変動とシーケンシングされた核酸の5'末端との間の平均距離は、指定位置における参照ヌクレオチドとシーケンシングされた核酸の5'末端との間の平均距離よりも短い。同様に、指定位置において脱アミノ化により生じるGからAへの変動とシーケンシングされた核酸の3'末端との間の平均距離は、指定位置における参照ヌクレオチドとシーケンシングされた核酸の3'末端との間の平均距離よりも短い。平均距離の差が大きいほど、脱アミノ化エラーの確率が高い。対照的に、指定位置におけるGからAまたはCからTへの変動が、シーケンシングエラーではなく実際の変動を表す場合、これらの変動とシーケンシングされた核酸の末端との平均距離と、指定位置における参照ヌクレオチドのものとの比較において、ランダムな因子に起因して生じ得る体系的な差はないはずである。

#### 【0099】

ある特定の態様では、参照配列内の指定位置における「C」から「T」または「G」から「A」へのバリエーションのマイナー対立遺伝子頻度を、指定位置にマッピングされるシーケンシングされたポリヌクレオチドの集団において決定する方法であって、マイナー対立遺伝子頻度が、バリエーションを含む指定位置にマッピングされるシーケンシングされたポリヌクレオチドの数（「バリエーション数」）を、指定位置にマッピングされるシーケンシングされたポリヌクレオチドの総数と比較するものであり、この方法が、脱アミノ化エラーの確率について、ゲノム座標におけるTまたはAバリエーションのバリエーション数を調節するステップを含み、エラーの確率が、「T」の場合には分子の5'末端から、および「A」の場合には分子の3'末端からのバリエーションの距離の関数である、方法が、本明細書に提供される。

#### 【0100】

図4に示されるように、脱アミノ化エラーにより分子において「T」バリエーションが生じる可能性は、バリエーションがある位置の分子の5'末端からの距離の関数である。より具体的には、バリエーションが分子の5'末端に近いほど、バリエーションがCからTへの塩基転換である可能性が高い。これは、充填される5'オーバーハングがある場所にエラーが伝播され、5'末端におけるオーバーハングが短い方が、長いオーバーハングよりもその可能性が高いためである。同様に、分子の3'末端におけるGからAへのバリエーションは、同様の理由から、位置が分子の3'末端に近いほど、可能性が高い。

#### 【0101】

任意の試料において、分子の末端からの距離の関数として、CからTへのバリエーションの統計学的決定を行うことができる。この関数は、分子の一部の内部距離において、漸近線に達する。漸近量は、全般的な脱アミノ化率を表す。この比率は、試料ごとに変動し得る。

#### 【0102】

脱アミノ化に誘導されるエラーが生じる可能性が高いシーケンシングされた核酸の末端への関連近接度は、シーケンシングされている核酸集団における一本鎖オーバーハングの長さにほぼ対応しているが、相補鎖の末端を越える消化およびそれに続く充填に起因して、3'オーバーハングの場合には、わずかに長い場合がある。近接度は、例えば、シーケンシングされた核酸鎖の3'または5'末端から、30ヌクレオチドもしくはそれ未満、25ヌクレオチドもしくはそれ未満、20ヌクレオチドもしくはそれ未満、15ヌクレオチドもしくはそれ未満、10ヌクレオチドもしくはそれ未満、または5ヌクレオチドもしくは

それ未満として規定され得る（「末端近接度」）。近接度は、3'末端または5'末端で同じかまたは異なって規定されてもよい。

#### 【0103】

前述のように、それぞれの配列を最大にアライメントしたときに、参照配列の指定位置に対応する位置を含む、シーケンシングされた核酸のサブセットが、識別される。このサブセット内のシーケンシングされた核酸のうちの一部は、5'末端の規定される近接度以内に生じる指定位置を有する。これらのシーケンシングされた核酸は、サブセットの第1の比率と称され得る。サブセット内のシーケンシングされた核酸のうちの一部は、3'末端の規定される近接度内に生じる指定位置を有する。これらのシーケンシングされた核酸は、サブセットの第2の比率と称され得る。「C」から「T」への変換は、次いで、第1の比率を構成するシーケンシングされた核酸におけるその出現によって認識され得、「G」から「A」への変換は、第2の比率を構成するシーケンシングされた核酸におけるその出現によって認識され得る。出現は、単純に、関連する比率内の指定位置におけるCからTまたはGからAへの変動を含む、存在するシーケンシングされた核酸の数として規定され得る。例えば、CからTへの脱アミノ化エラーは、第1の比率のある特定の数の、例えば、少なくとも1、2、3、4、5、または6つのシーケンシングされた核酸が、指定位置にCからTへの変動を含む場合に、コールされ得る。同様に、GからAへの脱アミノ化エラーは、第2の比率のある特定の数の、例えば、少なくとも1、2、3、4、5、または6つのシーケンシングされた核酸が、指定位置にGからAへの変動を含む場合に、コールされ得る。

#### 【0104】

出現は、それぞれ第1の比率または第2の比率の外側の割合と比較した、指定位置でCからTまたはGからAへの変動を含む第1または第2の比率内の核酸の割合によって規定することもできる。例えば、脱アミノ化エラーは、関連する比率の指定位置におけるCからTまたはGからAへの変動の出現が、関連する比率内のシーケンシングされた核酸のうちの少なくとも25、30、40、50、60、または70%である場合に、コールされ得る。過剰出現は、指定位置でCからTまたはGからAへの変動を有する関連比率内のシーケンシングされた核酸の、同じサブセット内であるが比率の外側にあるCからTまたはGからAへの変動を有するシーケンシングされた核酸の対応する割合と比較した、相対的な割合によって規定することもできる。関連する比率内のCからTまたはGからAへの変動を有するシーケンシングされた核酸の出現が、比率の外部よりも高いことは、変動が脱アミノ化エラーであるという指標である。例えば、サブセットの第1の比率内のシーケンシングされた核酸のうちの50%、およびサブセット内ではあるが比率の外部の核酸のうちの1%のみ（指定位置が5'末端の規定される近接度内にない場合）が、指定位置にCからTへの転位を含む場合、CからTへの転位は、おそらくは、脱アミノ化に誘導されるエラーである。

#### 【0105】

マイナー対立遺伝子比率を決定することは、指定位置にマッピングされる特定のバリエーションを含む分子の、指定位置にマッピングされる全分子に対する比を計算することを含み得る。そのため、例えば、100個の分子が、ゲノム座標にマッピングされ、そのうちの13個が、バリエーションを含む場合、マイナー対立遺伝子頻度は、13%として計算することができる。しかしながら、ある特定のバリエーションが、脱アミノ化エラーの結果であると考えられる場合、これらは、カウントからディスカウントされ得る。そのため、例えば、13個のバリエーションのうちの7個が、エラーとして指定される場合、比率は、 $6/93$ 、または6.4%として計算され得る。ある特定の事例では、例えば、分子の5'末端に位置する座標におけるバリエーションの比が、座標におけるすべてのバリエーションのうちの50%よりも多くを占める場合、指定位置におけるすべてのバリエーションが、ディスカウントされ得る。

#### 【0106】

脱アミノ化に誘導されるエラーは、コンテキストもしくは出現のいずれかまたは両方に

基づいて、分類することができる。例えば、CからTまたはGからAの転位が、上記に示される脱アミノ化エラーを示唆するコンテキストで生じる場合、転位を脱アミノ化エラーとして分類するのに必要なサブセットの関連比率における過剰出現の程度は、分類が過剰出現単独に基づいて行われる場合に必要とされるであろうものと比較して、低減され得る。  
【0107】

見かけのバリエーションが脱アミノ化エラーとしてコールされるかどうかは、いくつかの因子に基づき得る。バリエーション分子の絶対数が、ある特定の閾値（または比もしくは割合）を上回る場合、ある遺伝子座におけるバリエーションの存在が、該当し得る。また、バリエーションの存在は、対立遺伝子比率（バリエーションを有する遺伝子座にマッピングされる分子のパーセント）が、例えば、対照試料において予測される比率によって決定される、閾値を上回る場合、外して報告され得る。外して報告される場合、バリエーションの存在およびバリエーションのマイナー対立遺伝子比率の両方が、外して報告され得る。外して報告する際、脱アミノ化エラーは、いくつかの異なる様式のうちのいずれかで処理され得る。一実施形態では、所定の末端近接度内に位置する任意の「T」バリエーションは、単純に、エラーに帰属し、ディスカウントされ得る。この場合には、所定の末端近接度外の「T」バリエーションのみが、実際のバリエーションとしてカウントされ、報告要件の対象となる。別の方法では、所定の末端近接度内に位置する「T」バリエーションの、所定の末端近接度外に位置するものに対する比率が、決定される。その量が、ある特定の閾値量を上回る場合、例えば、20%を上回る、30%を上回る、40%を上回る、50%を上回る場合、エラー率は、その位置におけるバリエーションを報告しないのに十分に高いと考えられる。量が閾値レベルよりも低い場合、バリエーションは、通常の報告要件の対象となる。別の方法では、マイナー対立遺伝子比率が、予測される全般エラー率を上回る場合、バリエーションは、エラーの存在に関係なく外して報告され、エラーは補正されてもよく、またはされなくてもよい。別の実施形態では、選択された位置（すべてが近傍ゾーン内にあってもよく、または近傍ゾーン外のバリエーションを含んでもよい）において、「T」バリエーションは、バリエーションの確率がエラーであるとしてスコアされ、すべての位置におけるスコアを加算して、マイナー対立遺伝子比率に組み込まれる数が得られる。そのため、例えば、最初の（末端の）5'ヌクレオチドにおけるバリエーションが真のバリエーションである可能性は、50%であり得る。10番目の5'ヌクレオチドにおけるバリエーションの可能性は、75%であり得る。20番目を越える5'ヌクレオチドにおけるバリエーションの可能性は、95%であり得る。試料が、1位に1つ、10位に1つ、および50位に1つのバリエーションを示す場合、これらは、 $0.5 + 0.75 + 0.95 = 2.2$ としてスコア付けすることができ、マイナー対立遺伝子比率に対して2.2のバリエーションでカウントされる。そのような確率は、例えば、少なくとも10個、少なくとも50個、少なくとも100個、または少なくとも500個の対照試料を試験することによって、経験的に決定することができる。

【0108】

シーケンシングが、もとの試料中の単一のシーケンシングされた核酸の増幅したコピーを提示するファミリーを識別することによって行われる場合、典型的に、もとの試料中の核酸の両方の鎖を提示するファミリーを含む、ファミリー内のそれぞれのファミリーメンバーは、脱アミノ化エラーを含む。異なる鎖が、異なるヌクレオチドを有する場合、エラーは、自明である。

【0109】

バリエーションヌクレオチドが特定の試料において脱アミノ化エラーとして分類される、参照配列における指定位置の数は、変動し得る。例えば、そのような指定位置の数は、他の可能性の中でもとりわけ、少なくとも1、2、3、4、5、6、7、8、9、または10であり得る。

### III. コンピューター実施

【0110】

本方法は、湿式化学のステップ以外の本明細書または添付の特許請求の範囲に記載されたステップのいずれかまたはすべてが、好適なプログラムされたコンピューターで実施さ

10

20

30

40

50

れ得るように、コンピューターで実施することができる。コンピューターは、メインフレーム、パーソナルコンピューター、タブレット、スマートフォン、クラウド、オンラインデータストレージ、リモートデータストレージなどであり得る。コンピューターは、1つまたは複数の位置で操作することができる。

#### 【0111】

核酸集団を分析するためのコンピュータープログラムは、本明細書または添付の特許請求の範囲に記載される、湿式化学以外のステップのいずれかを行うためのコード、例えば、二本鎖平滑末端化核酸の配列を決定して、シーケンシングされた核酸を得るためのコード；指定位置を含む、シーケンシングされた核酸のサブセットを識別し、指定位置が、参照配列におけるそれぞれの指定位置において、バリエーションヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸の数を識別するためのコード；ならびに変動を有するサブセット内のシーケンシングされた核酸の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするためのコードであって、指定位置におけるバリエーションヌクレオチドの存在は、(i) バリエーションが、参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および(ii) バリエーションヌクレオチドが、(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または(2) サブセット内のシーケンシングされた核酸の5'末端から、指定位置におけるCからTへの変動の距離、またはサブセット内のシーケンシングされた核酸の3'末端から、指定位置におけるGからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、コードを含み得る。コンピュータープログラムはまた、データベースまたはシーケンシング装置から配列データを受容し、計算されたデータ、例えば、バリエーションヌクレオチドまたは脱アミノ化に誘導されるシーケンシングエラーを、ディスプレイまたはプリンターに出力するためのコードも含み得る。

#### 【0112】

本方法は、核酸集団を分析するためのシステム（例えば、データ処理システム）において実施され得る。システムはまた、本明細書または添付の特許請求の範囲に記載されるステップ、例えば、以下の：二本鎖平滑末端化核酸の配列を決定して、シーケンシングされた核酸を得るステップ；指定位置を含む、シーケンシングされた核酸のサブセットを識別し、指定位置が、参照配列におけるそれぞれの指定位置において、バリエーションヌクレオチドによって占有されているサブセット内のシーケンシングされた核酸の数を識別する、ステップ；ならびに変動を有するサブセット内のシーケンシングされた核酸の数が、閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、指定位置におけるバリエーションヌクレオチドの存在は、(i) バリエーションが、参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および(ii) バリエーションヌクレオチドが、(1) 指定位置の周囲のヌクレオチドコンテキスト、および／または(2) サブセット内のシーケンシングされた核酸の5'末端から、指定位置におけるCからTへの変動の距離、またはサブセット内のシーケンシングされた核酸の3'末端から、指定位置におけるGからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップのうちの1つまたは複数を行うための、互いに連結されたプロセッサ、システムバス、主メモリー、および必要に応じて補助メモリーも含み得る。システムはまた、他の付属機器の中でもとりわけ、結果、例えば、バリエーションヌクレオチドおよび脱アミノ化に誘導されるエラーを出力するためのディスプレイまたはプリンター、ユーザー入力を提供するため、例えば、閾値または規定される近接性を設定するためのキーボードおよび／またはポインターも含み得る。システムはまた、未加工のシーケンシングデータを提供するために、メモリーに連結されたシーケンシング装置を含み得る。

#### 【0113】

本方法の様々なステップは、コンピューター可読媒体（例えば、ハードドライブ、補助メモリー、外部メモリー、サーバー；データベース、ポータブルメモリーデバイス（例えば、CD-R、DVD、ZIPディスク、フラッシュメモリーカード）などに記憶された

、情報および／またはプログラムを利用し、結果を生成することができる。例えば、コンピューター可読媒体に記憶され得る、本方法に使用される情報および本方法によって生成される結果としては、参照配列、ヌクレオチドバリエーションもしくは脱アミノ化に誘導されるエラーのコールのための閾値または規定される近接度、未加工のシーケンシングデータ、シーケンシングされた核酸、バリエーションヌクレオチドおよびそれらの疾患との関連付け、ならびに脱アミノ化に誘導されるエラーが挙げられる。

#### 【0114】

本開示はまた、核酸集団を分析するための製品であって、実行されると、本方法のステップを実施する、1つまたは複数のプログラムを含む、マシン可読媒体を含む、製品も含む。

#### 【0115】

本開示は、ハードウェアおよび／またはソフトウェアで実施することができる。例えば、本開示の異なる態様は、クライアントサイドロジックまたはサーバーサイドロジックのいずれかで実施することができる。本開示またはその構成成分は、適切に設定されたコンピューティングデバイス中にロードされた場合に、そのデバイスに、本開示に従って実施させるロジック命令および／またはデータを含有する固定媒体プログラムコンポーネントにおいて具現化することができる。ロジック命令を含有する固定媒体は、ビューワーのコンピューターに物理的にローディングするために固定媒体のビューワーに送達され得るか、またはロジック命令を含有する固定媒体は、ビューワーが、プログラムコンポーネントをダウンロードするために通信媒体を介してアクセスするリモートサーバーに存在してもよい。

#### 【0116】

本開示は、本開示の方法を実施するためにプログラミングされたコンピューターコントロールシステムを提供する。図5は、本開示の方法を実施するためにプログラミングされたか、またはそうでなければ本開示の方法を実施するよう構成されたコンピューターシステム901を示す。コンピューターシステム901は、セントラルプロセッシングユニット（CPU、本明細書ではまた「プロセッサ」および「コンピュータープロセッサ」）905（シングルコアもしくはマルチコアプロセッサ、または並行プロセッシングのための複数のプロセッサであってもよい）を含む。コンピューターシステム901は、メモリーまたはメモリー位置910（例えば、ランダムアクセスメモリー、リードオンリーメモリー、フラッシュメモリー）、電子保存ユニット915（例えば、ハードディスク）、1つまたは複数の他のシステムと通信するための通信インターフェース920（例えばネットワークアダプタ）、ならびにキャッシュメモリー、他のメモリー、データストレージ、および／または電子ディスプレイアダプターなどの周辺デバイス925も含む。メモリー910、保存ユニット915、インターフェース920および周辺デバイス925は、

#### 【0117】

マザーボードなどの通信バス（実線）を通じてCPU905と通信する。保存ユニット915は、データを保存するためのデータ保存ユニット（またはデータ保管庫）であり得る。コンピューターシステム901は、通信インターフェース920の補助により、コンピューターネットワーク（「ネットワーク」）930に作動可能に接続され得る。ネットワーク930は、インターネット、インターネットおよび／もしくはエクストラネット、またはインターネットと通信するイントラネットおよび／もしくはエクストラネットであり得る。一部の事例では、ネットワーク930は、電気通信および／またはデータネットワークである。ネットワーク930は、ローカルエリアネットワークであり得る。ネットワーク930は、クラウドコンピューティングなどの分散コンピューティングを可能にし得る1つまたは複数のコンピューターサーバーを含み得る。ネットワーク930は、一部の事例では、コンピューターシステム901の補助により、コンピューターシステム901に接続されたデバイスがクライアントまたはサーバーとして機能することを可能にし得るピアツーピアネットワークを実施することができる。

#### 【0118】

C P U 9 0 5 は、プログラムまたはソフトウェアにおいて具現化され得るマシン可読命令のシーケンスを実行することができる。命令は、メモリー 9 1 0 などのメモリー位置に保存され得る。命令は、C P U 9 0 5 を対象とすることができ、これは、次に C P U 9 0 5 が本開示の方法を実施するようにプログラムするか、またはそうでなければ本開示の方法を実施するように構成することができる。C P U 9 0 5 により実施される動作の例としては、取り出し、デコード、実行、および書き戻しを挙げることができる。

#### 【0119】

C P U 9 0 5 は、回路、例えば、集積回路の一部であり得る。システム 9 0 1 の 1 つまたは複数の他のコンポーネントは、回路に含まれ得る。一部の事例では、回路は、アプリケーション特異的集積回路 (A S I C) である。

#### 【0120】

保存ユニット 9 1 5 は、ドライバ、ライブラリー、および保存済みプログラムなどのファイルを保存することができる。保存ユニット 9 1 5 は、ユーザーデータ、例えば、ユーザープリファレンスおよびユーザープログラムを保存することができる。コンピューターシステム 9 0 1 は、一部の事例では、イントラネットまたはインターネットを通じてコンピューターシステム 9 0 1 と通信するリモートサーバー上に位置するなど、コンピューターシステム 9 0 1 の外部の 1 つまたは複数の追加のデータ保存ユニットを含み得る。

#### 【0121】

コンピューターシステム 9 0 1 は、ネットワーク 9 3 0 を通じて、1 つまたは複数のリモートコンピューターシステムと通信することができる。例えば、コンピューターシステム 9 0 1 は、ユーザーのリモートコンピューターシステムと通信することができる。リモートコンピューターシステムの例としては、パーソナルコンピューター（例えば、ポータブル P C）、スレートまたはタブレット P C（例えば、A p p l e（登録商標）の i P a d（登録商標）、S a m s u n g（登録商標）の G a l a x y T a b）、電話、スマートフォン（例えば、A p p l e（登録商標）の i P h o n e（登録商標）、アンドロイド（登録商標）使用可能デバイス、B l a c k b e r r y（登録商標））、または携帯情報端末が挙げられる。ユーザーは、ネットワーク 9 3 0 を介してコンピューターシステム 9 0 1 にアクセスすることができる。

#### 【0122】

本明細書に記載されている方法は、コンピューターシステム 9 0 1 の電子保存場位置、例えば、メモリー 9 1 0 または電子保存ユニット 9 1 5 上に保存されたマシン（例えば、コンピュータープロセッサ）実行可能コードにより実施され得る。マシン実行可能またはマシン可読コードはソフトウェアの形態で提供され得る。使用中、コードはプロセッサ 9 0 5 によって実行され得る。一部の事例では、コードは、保存ユニット 9 1 5 から検索され、プロセッサ 9 0 5 によって容易にアクセスするためにメモリー 9 1 0 上に保存され得る。一部の状況では、電子保存ユニット 9 1 5 は除外されてもよく、マシン実行可能命令がメモリー 9 1 0 上に保存される。

#### 【0123】

コードは予めコンパイルされ、コードを実行するように適応されたプロセッサを有するマシンで使用するために構成され得るか、またはランタイム中にコンパイルされ得る。コードは、コードが予めコンパイルされたかまたはコンパイルされてすぐの様式で実行することを可能にするように選択され得るプログラム言語で供給され得る。

#### 【0124】

コンピューターシステム 9 0 1 などの本明細書において提供されるシステムおよび方法の様子は、プログラミングにおいて具現化され得る。技術の様々な様子は、典型的には、一種のマシン可読媒体において保有または具現化されるマシン（またはプロセッサ）実行可能コードおよび／または関連データの形態で、「製品」または「製造物品」として考えられ得る。マシン実行可能コードは、電子保存ユニット、例えば、メモリー（例えば、リードオンリーメモリー、ランダムアクセスメモリー、フラッシュメモリー）、またはハードディスク上に保存され得る。

10

20

30

40

50

## 【0125】

「保存」式媒体は、コンピューター、プロセッサなどの有形メモリーのいずれかもしくはすべて、またはそれらの関連モジュール（様々な半導体メモリー、テープドライブ、ディスクドライブなど）を含み得、これは常にソフトウェアプログラミングの非一時的な保存を提供することができる。ソフトウェアのすべてまたは一部は、時折、インターネットまたは様々な他の電気通信ネットワークを通じて通信され得る。例えば、このような通信は、1つのコンピューターまたはプロセッサから別のコンピューターまたはプロセッサに、例えば、管理サーバーまたはホストコンピューターからアプリケーションサーバーのコンピュータープラットフォームへのソフトウェアの読み込みを可能にし得る。よって、ソフトウェア要素を有し得る別の種類の媒体は、有線および光学地上通信ネットワークを通して、ならびに様々な無線リンク上で、ローカルデバイス間の物理インターフェースなどにわたって使用される光学波、電波および電磁波を含む。有線またはワイヤレスリンク、光学リンクなどのこのような波を運ぶ物理的要素もソフトウェアを保有する媒体と考えることができる。本明細書で使用される場合、非一時的な有形

10

## 【0126】

「保存」媒体に限定されない限り、コンピューターまたはマシン「可読媒体」などの用語は、実行のための命令をプロセッサに提供する際に関与する任意の媒体を指す。

## 【0127】

よって、コンピューター実行可能コードなどのマシン可読媒体は、有形保存媒体、搬送波媒体または物理伝送媒体を含むが、これらに限定されない多くの形態を取ることができる。不揮発性保存媒体は、例えば、図面に示されるデータベースなどを実施するために使用され得る任意のコンピューターなどの保存デバイスのいずれかなどの光学または磁気ディスクを含む。揮発性保存媒体は、このようなコンピュータープラットフォームのメインメモリーなどの動的メモリーを含む。有形伝送媒体は、同軸ケーブル；コンピューターシステム内にバスを含むワイヤを含む、銅線および光学ファイバを含む。搬送波伝送媒体は、電気もしくは電磁シグナル、または高周波（RF）赤外（IR）データ通信中に生成されるものなどの音もしくは光波の形態を取ることができる。したがって、コンピューター可読媒体の一般的な形態は、例えば、フロッピー（登録商標）ディスク、フレキシブルディスク、ハードディスク、磁気テープ、任意の他の磁気媒体、CD-ROM、DVDもしくはDVD-ROM、任意の他の光学媒体、パンチカード、紙テープ、穴のパターンを有する任意の他の物理的保存媒体、RAM、ROM、PROMおよびEPROM、FLASH（登録商標）-EPROM、任意の他のメモリーチップもしくはカートリッジ、搬送波輸送データもしくは命令、このような搬送波などを輸送するケーブルもしくはリンク、またはコンピューターがプログラミングコードおよび／もしくはデータを読むことができる任意の他の媒体を含む。コンピューター可読媒体のこれらの形態の多くは、実行のためにプロセッサに対する1つまたは複数の命令の1つまたは複数のシーケンスの保有に関与し得る。

20

30

## 【0128】

コンピューターシステム901は、例えば、レポートを提供するための、ユーザーインターフェース（UI）940を含む、電子ディスプレイ935を含み得るか、またはそれと通信し得る。UIの例としては、グラフィカルユーザーインターフェース（GUI）およびウェブベースユーザーインターフェースが挙げられるが、これらに限定されない。

40

## 【0129】

本開示の方法およびシステムは、1つまたは複数のアルゴリズムを用いて実施することができる。アルゴリズムは、中央処理装置905により実行すると、ソフトウェアによって実施され得る。

## IV. 方法の一般的な特徴

## 1. 試料

## 【0130】

試料は、対象から単離された任意の生物学的試料であり得る。試料としては、体組織、

50

例えば、公知または疑いのある固形腫瘍、全血、血小板、血清、血漿、糞便、赤血球、白血球 (white blood cell) もしくは白血球 (leucocyte)、内皮細胞、組織生検、脳脊髄液、関節液、リンパ液、腹水、間質液もしくは細胞外液、細胞間の空間の流体が挙げられ、歯肉溝滲出液、骨髓、胸水、脳脊髄液、唾液、粘液、喀痰、精液、汗、尿が挙げられる。試料は、好ましくは、体液、特に、血液およびその比率、ならびに尿である。そのような試料には、腫瘍から流出した核酸が含まれる。核酸としては、DNAおよびRNAを挙げることができ、二本鎖形態および／または一本鎖形態であり得る。試料は、対象からもともと単離した形態であってもよく、または成分、例えば、細胞を除去もしくは追加するため、1つの成分を別の成分と比べて濃縮するため、もしくは1つの形態の核酸を別のものに、例えば、RNAからDNAに、もしくは一本鎖核酸から二本鎖に変換するために、さらなるプロセッシングに供されていてもよい。したがって、例えば、分析のための体液は、無細胞核酸、例えば、無細胞DNA (cfDNA) を含む、血漿または血清である。

#### 【0131】

血漿の体積は、シーケンシングされる領域の所望されるリード深度に依存し得る。例示的な体積は、0.4~40ml、5~20ml、10~20mlである。例えば、体積は、0.5ml、1ml、5ml、10ml、20ml、30ml、または40mlであり得る。サンプリングされる血漿の体積は、5~20mlであり得る。

#### 【0132】

試料は、ゲノム等価物を含む様々な量の核酸を含み得る。例えば、約30ngのDNAの試料は、約10,000 ( $10^4$ ) 個のハプロイドヒトゲノム等価物を含み得、cfDNAの場合には、約2000億 ( $2 \times 10^4$ ) 個の個々のポリヌクレオチド分子を含み得る。同様に、約100ngのDNAの試料は、約30,000個のハプロイドヒトゲノム等価物を含み得、cfDNAの場合には、約6000億個の個々の分子を含み得る。

#### 【0133】

試料は、異なる源、例えば、細胞および無細胞に由来する核酸を含み得る。試料は、変異を有する核酸を含み得る。例えば、試料は、生殖細胞系列変異および／または体細胞変異を有するDNAを含み得る。試料は、がん関連変異 (例えば、がん関連体細胞変異) を有するDNAを含み得る。

#### 【0134】

増幅前の試料中の無細胞核酸の例示的な量は、約1fg~約1μg、例えば、1pg~200ng、1ng~100ng、10ng~1000ngの範囲である。例えば、量は、最大約600ng、最大約500ng、最大約400ng、最大約300ng、最大約200ng、最大約100ng、最大約50ng、または最大約20ngの無細胞核酸分子であり得る。量は、少なくとも1fg、少なくとも10fg、少なくとも100fg、少なくとも1pg、少なくとも10pg、少なくとも100pg、少なくとも1ng、少なくとも10ng、少なくとも100ng、少なくとも150ng、または少なくとも200ngの無細胞核酸分子であり得る。量は、最大1フェムトグラム (fg)、10fg、100fg、1ピコグラム (pg)、10pg、100pg、1ng、10ng、100ng、150ng、または200ngの無細胞核酸分子であり得る。本方法は、1フェムトグラム (fg) ~200ngを得ることを含み得る。

#### 【0135】

無細胞核酸試料は、無細胞核酸を含む試料を指す。無細胞核酸は、細胞内に含まれていないかしくは細胞に結合していない核酸、または換言すると、インタクトな細胞を除去した試料中に残存している核酸である。無細胞核酸は、対象に由来する体液 (例えば、血液、尿、CSFなど) から得られたすべての非封入核酸を指し得る。無細胞核酸としては、DNA (cfDNA)、RNA (cfRNA)、およびそれらのハイブリッドが挙げられ、ゲノムDNA、ミトコンドリアDNA、循環DNA、siRNA、miRNA、循環RNA (cRNA)、tRNA、rRNA、核小体RNA (snRNA)、Piwi相互作用RNA (piRNA)、長い非コーディングRNA (長いncRNA)、またはこ

10

20

30

40

50

これらのうちのいずれかの断片が含まれる。無細胞核酸は、二本鎖であっても、一本鎖であっても、それらのハイブリッドであってもよい。無細胞核酸は、分泌または細胞死プロセス、例えば、細胞壊死およびアポトーシスを通じて、体液中に放出され得る。一部の無細胞核酸、例えば、循環腫瘍DNA (ctDNA) は、がん細胞から、体液中に放出される。その他のものは、健常細胞から放出される。ctDNAは、非カプセル化腫瘍由来断片化DNAであり得る。無細胞胎児DNA (cffDNA) は、母体血流中に自由に循環している胎児DNAである。

#### 【0136】

無細胞核酸またはそれに関連するタンパク質は、1つまたは複数のエピジェネティックな改変を有し得る。例えば、無細胞核酸は、アセチル化、5-メチル化、ユビキチン化、リン酸化、SUMO化、リボシル化、および／またはシトルリン化され得る。

10

#### 【0137】

無細胞核酸は、約100～500ヌクレオチドの例示的なサイズ分布を有し、110～約230ヌクレオチドの分子が、分子の約90%に相当し、ヒトにおける最頻値は約168ヌクレオチドであり、第2のマイナーピークは、240～440ヌクレオチドの範囲にある。無細胞核酸は、約160～約180ヌクレオチド、または約320～約360ヌクレオチド、または約440～約480ヌクレオチドであり得る。

#### 【0138】

無細胞核酸は、溶液に見出される無細胞核酸を、インタクトな細胞および体液の他の不溶性成分から分離する、分割ステップによって、体液から単離することができる。分割には、遠心分離または濾過などの技法が含まれ得る。あるいは、体液中の細胞を、溶解し、無細胞核酸および細胞核酸を、一緒にプロセッシングしてもよい。一般に、緩衝液の添加および洗浄ステップの後に、無細胞核酸は、アルコールで沈降させることができる。混入物質または塩を除去するために、シリカベスカラムなどのさらなる洗浄ステップを使用してもよい。手順のある特定の態様、例えば、収率を最適化するために、例えば、非特異的バルク担体核酸を、反応全体にわたり添加してもよい。

20

#### 【0139】

そのようなプロセッシングの後、試料は、二本鎖DNA、一本鎖DNA、および一本鎖RNAを含む、様々な形態の核酸を含み得る。必要に応じて、一本鎖DNAおよびRNAは、後続のプロセッシングおよび分析ステップに含められるように、二本鎖形態に変換してもよい。

30

### 2. 増幅

#### 【0140】

アダプターが隣接する試料核酸を、PCR、および典型的に、増幅させようとするDNA分子に隣接するアダプターにおけるプライマー結合部位にプライマーが結合することによりプライミングされる他の増幅方法によって増幅させることができる。増幅方法は、サーマルサイクリングの結果としての伸長、変性、およびアニーリングのサイクルを含み得るか、または転写に媒介される増幅にあるような等温のものであってもよい。他の増幅方法としては、リガーゼ連鎖反応、鎖置換増幅、核酸配列に基づく増幅、および自己持続性配列に基づく複製が挙げられる。

40

#### 【0141】

従来的な核酸増幅方法を使用して、核酸分子にバーコードを導入するために、1回または複数回の増幅を適用することができる。増幅は、1つまたは複数の反応混合物中で行うことができる。分子タグおよび試料インデックス／タグは、同時、または任意の連続的順序で導入することができる。分子タグおよび試料インデックス／タグは、配列捕捉の前および／または後に導入することができる。一部の場合には、分子タグのみが、プローブ捕捉の前に導入され、一方で試料インデックス／タグは、配列捕捉の後に導入される。一部の場合には、分子タグおよび試料インデックス／タグの両方が、プローブ捕捉の前に導入される。一部の場合には、試料インデックス／タグは、配列捕捉の後に導入される。通常、配列捕捉は、標的配列、例えば、ゲノム領域のコーディング配列に相補的な一本鎖核酸

50

分子を導入することを含み、そのような領域の変異は、がん型と関連している。典型的に、増幅により、200 nt～700 nt、250 nt～350 nt、または320 nt～550 ntの範囲のサイズで、分子タグおよび試料インデックス／タグを有する非固有または固有にタグ付けされた複数の核酸アンプリコンが生成される。一部の実施形態では、アンプリコンは、約300 ntのサイズを有する。一部の実施形態では、アンプリコンは、約500 ntのサイズを有する。

### 3. バーコード

#### 【0142】

バーコードは、他の方法の中でもとりわけ、化学合成、ライゲーション、オーバーラップ伸長PCRによって、アダプターに組み込まれ得るか、またはそれ以外では結合され得る。一般に、反応における固有または非固有のバーコードの割当ては、米国特許出願第20010053519号、同第20110160078号、ならびに米国特許第6,582,908号および同第7,537,898号およびUS9,598,731に記載される方法およびシステムに従う。

#### 【0143】

タグは、ランダムまたは非ランダムで、試料核酸に連結され得る。一部の場合には、それらは、予測される識別子比（例えば、バーコードの組合せ）で、マイクロウェルに導入される。バーコード集団は、固有であってもよく、例えば、すべてのバーコードが、同じヌクレオチド配列を有する。バーコード集団は、非固有であってもよく、例えば、バーコードのうちの一部分が、同じヌクレオチド配列を有し、バーコードのうちの一部分が、異なるヌクレオチド配列を有する。例えば、ゲノム試料当たり1個を上回る、2個を上回る、3個を上回る、4個を上回る、5個を上回る、6個を上回る、7個を上回る、8個を上回る、9個を上回る、10個を上回る、20個を上回る、50個を上回る、100個を上回る、500個を上回る、1000個を上回る、5000個を上回る、10000個を上回る、50,000個を上回る、100,000個を上回る、500,000個を上回る、1,000,000個を上回る、または1,000,000,000個を上回る識別子がロードされるように、識別子がロードされ得る。一部の場合には、ゲノム試料当たり2個を下回る、3個を下回る、4個を下回る、5個を下回る、6個を下回る、7個を下回る、8個を下回る、9個を下回る、10個を下回る、20個を下回る、50個を下回る、100個を下回る、500個を下回る、1000個を下回る、5000個を下回る、10000個を下回る、50,000個を下回る、100,000個を下回る、500,000個を下回る、1,000,000個を下回る、50,000,000個を下回る、100,000,000個を下回る、500,000,000個を下回る、1,000,000,000個を下回る、または1,000,000,000,000個を下回る識別子がロードされるように、識別子がロードされ得る。一部の場合には、試料ゲノム当たりのロードされる識別子の平均数は、ゲノム試料当たり約1個、2個、3個、4個、5個、6個、7個、8個、9個、10個、20個、50個、100個、500個、1000個、5000個、10000個、50,000個、100,000個、500,000個、1,000,000個、50,000,000個、100,000,000個、500,000,000個、または1,000,000,000,000個の識別子を下回るか、または上回る。

#### 【0144】

好ましい形式は、標的分子の両方の末端にライゲーションされる20～50個の異なるタグを使用し、20～50×20～50個のタグ、例えば、400～2500個のタグが作製される。そのような数のタグは、同じ開始点および終止点を有する異なる分子が、高い確率（例えば、少なくとも94%、99.5%、99.99%、99.999%）で異なる組合せのタグを受容するのに十分である。

#### 【0145】

一部の場合には、識別子は、所定のまたはランダムなまたはセミランダムな配列のオリゴヌクレオチドであり得る。他の場合には、バーコードが複数の中で必ずしも互いに固有とならないような、複数のバーコードが使用されてもよい。この例では、バーコードは、

バーコードおよびそれが結合され得る配列の組合せが、個別にトラッキングすることができ、固有の配列を作製するように、（例えば、ライゲーションまたはPCR増幅によって）個々の分子に結合され得る。本明細書に記載されるように、配列リードの最初の（開始）点および終わりの（終止）部分の配列データと組み合わせて非固有にタグ付けされたバーコードの検出により、特定の分子に固有の同一性を割り当てることが可能となり得る。個々の配列リードの長さまたは塩基対の数もまた、そのような分子に固有の同一性を割り当てるために使用することができる。本明細書に記載されるように、固有の同一性が割り当てられた核酸の一本鎖に由来する断片は、それによって、それに続く親鎖に由来する断片および／または相補鎖の識別を可能にし得る。

#### 4. シーケンシング

##### 【0146】

前の増幅の有無にかかわらず、アダプターが隣接している試料核酸は、シーケンシングに供することができる。シーケンシング方法としては、例えば、サンガーシーケンシング、高スループットシーケンシング、パイロシーケンシング、シーケンシングバイシンセシス、単分子シーケンシング、ナノポアシーケンシング、半導体シーケンシング、シーケンシングバイライゲーション、シーケンシングバイハイブリダイゼーション、RNA-Seq (Illumina)、Digital Gene Expression (Helicos)、次世代シーケンシング、合成による単分子シーケンシング (SMSS) (Helicos)、大規模並列シーケンシング、クローン単一分子アレイ (Solexa)、ショットガンシーケンシング、Ion Torrent、Oxford Nanopore、Roche Genia、Maxim-Gilbertシーケンシング、プライマーウォーキング、PacBio、SOLiD、Ion Torrent、またはNanoporeプラットフォームを使用したシーケンシングが挙げられる。シーケンシング反応は、様々な試料プロセッシングユニットで行うことができ、これは、複数レーン、複数チャネル、複数ウェル、または複数の試料セットを実質的に同時にプロセッシングする他の手段であり得る。試料プロセッシングユニットにはまた、複数の実行を同時にプロセッシングすることを可能にする複数の試料チャンバが含まれ得る。

##### 【0147】

シーケンシング反応は、他の疾患のがんのマーカーを含むことが既知の1つまたは複数の断片型に行うことができる。シーケンシング反応はまた、試料中に存在する任意の核酸断片に行うこともできる。シーケンシング反応は、少なくとも5%、10%、15%、20%、25%、30%、40%、50%、60%、70%、80%、90%、95%、99%、99.9%、または100%のゲノムの配列カバレッジを提供し得る。他の場合には、ゲノムの配列カバレッジは、5%未満、10%未満、15%未満、20%未満、25%未満、30%未満、40%未満、50%未満、60%未満、70%未満、80%未満、90%未満、95%未満、99%未満、99.9%未満、または100%未満であり得る。

##### 【0148】

多重シーケンシングを使用して、同時のシーケンシング反応を行ってもよい。一部の場合には、無細胞ポリヌクレオチドは、少なくとも1000回、2000回、3000回、4000回、5000回、6000回、7000回、8000回、9000回、10000回、50000回、100,000回のシーケンシング反応でシーケンシングされ得る。他の事例では、無細胞ポリヌクレオチドは、1000回未満、2000回未満、3000回未満、4000回未満、5000回未満、6000回未満、7000回未満、8000回未満、9000回未満、10000回未満、50000回未満、100,000回未満のシーケンシング反応でシーケンシングされ得る。シーケンシング反応は、逐次的に行われてもよく、または同時に行われてもよい。後続のデータ分析は、シーケンシング反応のすべてまたは一部に行われ得る。一部の場合には、データ分析は、少なくとも1000回、2000回、3000回、4000回、5000回、6000回、7000回、8000回、9000回、10000回、50000回、100,000回のシーケンシング反応に行われ得る。他の場合には、データ分析は、1000回未満、2000回未満、3

10

20

30

40

50

000回未満、4000回未満、5000回未満、6000回未満、7000回未満、8000回未満、9000回未満、10000回未満、50000回未満、100,000回未満のシーケンシング反応に行われ得る。例示的なリード深度は、遺伝子座（塩基）1つ当たり1000～50000リードである。

## 5. 分析

### 【0149】

本方法は、対象における状態、特に、がんの存在を診断するため、状態を特徴付ける（例えば、がんのステージを決定するかもしれないがんの不均質性を決定する）ため、状態の処置に対する応答をモニタリングするため、状態を発症する危険性または状態の後続の経過の予後診断を達成するために、使用することができる。

### 【0150】

様々ながんを、本方法を使用して検出することができる。がん細胞は、ほとんどの細胞と同様に、古い細胞が死亡し新しい細胞に置き換わる、代謝回転率によって特徴付けることができる。一般に、死亡した細胞は、所与の対象において脈管構造と接触すると、DNAまたはDNAの断片を血流中に放出し得る。これは、疾患の様々なステージにあるがん細胞についても当てはまる。がん細胞はまた、疾患のステージに応じて、様々な遺伝子の異常、例えば、コピー数多型ならびにまれな変異によって、特徴付けることができる。この現象は、本明細書に記載される方法およびシステムを使用して、個体におけるがんの存在または不在を検出するために使用することができる。

### 【0151】

検出することができるがんの型および数としては、血液のがん、脳のがん、肺がん、皮膚がん、鼻のがん、喉のがん、肝臓がん、骨がん、リンパ腫、脾臓がん、皮膚がん、大腸がん、直腸がん、甲状腺がん、膀胱がん、腎臓がん、口腔がん、胃がん、固形状態の腫瘍、不均質腫瘍、均質腫瘍などを挙げることができる。

### 【0152】

がんは、変異、まれな変異、挿入欠失、コピー数多型、塩基転換、転位、逆転、欠失、異数性、部分異数性、倍数性、染色体不安定性、染色体構造変更、遺伝子融合、染色体融合、遺伝子短縮、遺伝子増幅、遺伝子重複、染色体病変、DNA病変、核酸化学改変における異常な変化、エピジェネティックパターンにおける異常な変化を含む、遺伝子変動により検出することができる。

### 【0153】

遺伝子データはまた、がんの特定の形態を特徴付けるためにも使用することができる。がんは、組成およびステージ決定の両方において、不均質であることが多い。遺伝子プロファイルデータにより、特定のサブタイプの診断または処置に重要であり得るその特定のサブタイプのがんの特徴付けが可能となり得る。この情報はまた、対象または従事者に特定のがん型の予後に関する手がかりを提供し、対象または従事者のいずれかが、処置選択肢を疾患の進行に合わせて適合させることを可能にし得る。一部のがんは、進行すると、より侵襲的かつ遺伝子的に不安定になる。他のがんは、良性、不活性、または休眠状態のままであり得る。本開示のシステムおよび方法は、疾患の進行を決定するのに有用であり得る。

### 【0154】

本分析はまた、特定の処置選択肢の有効性を決定するのににも有用である。奏功処置選択肢は、処置が成功している場合、より多くのがんが死滅し、DNAが流出し得るため、対象の血液中に検出されるコピー数多型またはまれな変異の量が増加し得る。他の例では、これは生じない場合がある。別の例では、おそらくはある特定の処置選択肢は、時間とともに、がんの遺伝子プロファイルと相関し得る。この相関性は、治療法を選択するのに有用であり得る。加えて、がんが、処置後に寛解状態にあることが観察された場合、本方法を使用して、残存疾患または疾患の再発をモニタリングすることができる。

### 【0155】

本方法はまた、がん以外の状態における遺伝子変動を検出するためにも使用することが

10

20

30

40

50

できる。免疫細胞、例えば、B細胞は、ある特定の疾患の存在時には、急速なクローン増殖を受け得る。クローン増殖は、コピー数多型検出を使用してモニタリングすることができる。ある特定の免疫状態をモニタリングすることができる。この例では、コピー数多型分析を、経時的に行って、特定の疾患がどのように進行し得るかのプロファイルを得ることができる。コピー数多型またはさらにはまれな変異の検出でさえも、病原体の集団が、感染の過程においてどのように変化しているかを決定するために使用することができる。これは、慢性感染症、例えば、HIV/AIDSまたは肝炎感染症では、特に重要であり得、それによって、ウイルスは、生命サイクル状態を変化させる、および／または感染の過程においてより悪性な形態へと変異し得る。本方法は、免疫細胞が移植組織を破壊しようとする場合の宿主身体の拒絶活性を決定またはプロファイリングして、移植組織の状態をモニタリングするため、ならびに処置過程の変更または拒絶の予防に使用することができる。

10

#### 【0156】

さらに、本開示の方法は、対象における異常な状態の不均質性を特徴付けるために使用することができる。この方法は、対象における細胞外ポリヌクレオチドの遺伝子プロファイルを生成することを含み、ここで、遺伝子プロファイルは、コピー数多型およびまれな変異の分析から得られる複数のデータを含む。一部の場合には、がんを含むがこれに限定されない、疾患は、不均質であり得る。疾患細胞は、同一ではない場合がある。がんの例では、一部の腫瘍は、異なる型の腫瘍細胞を含むことが既知であり、一部の細胞は、がんの異なるステージにある。他の例では、不均質性は、複数病巣の疾患を構成し得る。さらに、がんの例では、おそらくは、1つまたは複数の病巣が、原発部位から拡がった転移の結果である、複数の腫瘍病巣が存在し得る。

20

#### 【0157】

本方法は、不均質性疾患における異なる細胞に由来する遺伝子情報のまとめであるデータのプロファイル、フィンガープリント、またはセットを生成するために使用することができる。このデータのセットは、コピー数多型およびまれな変異分析を、単独または組合せで含み得る。

#### 【0158】

本方法は、がんまたは胎児起源の他の疾患を診断、予後診断、モニタリング、または観察するために使用することができる。すなわち、これらの手法は、DNAおよび他のポリヌクレオチドが母体分子とともに共循環していてもよい、出生前の対象におけるがんまたは他の疾患を診断、予後診断、モニタリング、または観察するために、妊娠している対象において用いることができる。

30

### 6. 処置

#### 【0159】

試料中のバリエーションヌクレオチドの数および種類は、試料を提供した対象の、処置、すなわち、治療介入に対する適合性に関する指標を提供し得る。例えば、多数のバリエーションヌクレオチドの存在は、そのような変異の存在が、免疫療法の標的を形成しているネオエピトープと関連しているため、免疫療法の肯定的な指標である。免疫療法には、他の処置の中でもとりわけ、PD-1、PD-2、PD-L1、PD-L2、CTLA-40、OX40、B7.1、B7He、LAG3、CD137、KIR、CCR5、CD27、またはCD40のいずれかに対する抗体の使用が含まれ得る。免疫療法のための他の例示的な作用物質としては、炎症促進性サイトカイン、例えば、IL-1 $\beta$ 、IL-6、およびTNF- $\alpha$ が挙げられる。他の例示的な作用物質としては、腫瘍に対する活性化されたT細胞、例えば、T細胞からの腫瘍抗原を標的とするキメラ抗原の発現によるものである。免疫療法は、変異の存在によって野生型対応物とは区別される、腫瘍抗原を攻撃するように、免疫系を刺激する。

40

#### 【0160】

他のバリエーションヌクレオチドは、既存の薬物の標的を提供するか、またはそのような薬物に対する耐性を示す。脱アミノ化に誘導されるシーケンシングエラーに起因する偽陽性

50

を排除することにより、バリエーションヌクレオチドの数および種類を決定することができる正確さが増加する。したがって、本方法によって分析される対象は、続いて、発見されたヌクレオチドバリエーションに応じて、異なる処置レジメンに供され得る。したがって、例えば、決定されたバリエーションヌクレオチドの数が、閾値にあるかまたは閾値を上回る対象では、決定されたバリエーションヌクレオチドの数が閾値を下回る対象よりも、免疫療法を受容し得る割合が多い。

#### 【0161】

シーケンシングされた核酸におけるヌクレオチドの変異は、シーケンシングされた核酸を参照配列と比較することによって決定することができる。参照配列は、既知の配列、例えば、目的物由来の既知の全ゲノム配列または部分的ゲノム配列、ヒト対象の全ゲノム配列であることが多い。参照配列は、hg19であってもよい。シーケンシングされた核酸は、上述のように、試料中の核酸に対して直接的に決定される配列、またはこのような核酸の増幅産物の配列のコンセンサスを示すことができる。比較は、参照配列の1つまたは複数の指定された位置で実施することができる。参照配列の指定された位置に対応する位置を含むシーケンシングされた核酸のサブセットは、各配列が最大限に配列された場合に特定することができる。このようなサブセット内で、存在する場合、どのシーケンシングされた核酸が指定された位置にヌクレオチド変異を含むか、および必要に応じて、存在する場合、どれが参照ヌクレオチドを含むか（すなわち、参照配列におけるものと同じか）を決定することができる。ヌクレオチド変異を含むサブセットにおけるシーケンシングされた核酸の数が閾値を超える場合、バリエーションヌクレオチドが指定された位置でコールされ得る。閾値は、単純な数、例えば、ヌクレオチドバリエーションを含むサブセット内の少なくとも1、2、3、4、5、6、7、8、9、または10個のシーケンシングされた核酸であってもよく、または比、例えば、少なくとも0.5であってもよく、サブセット内のシーケンシングされた核酸のうちの1、2、3、4、5、10、15、または20個は、他の可能性の中で、ヌクレオチドバリエーションを含む。比較は、参照配列における目的の任意の指定された位置に対して繰り返すことができる。比較は、参照配列における少なくとも20、100、200、または300の連続する位置、例えば、20～500、または50～300の連続する位置を占める指定された位置に対して実施できることもある。

#### 【0162】

上記または下記で引用されるすべての特許出願、ウェブサイト、他の刊行物、受託番号などは、各個別の項目が参照によってそのように組み込まれることが具体的かつ個別に示されているのと同程度に、すべての目的で、参照によりその全体が組み込まれる。様々な配列バージョンが、異なる時点の受託番号に関連する場合、この出願の有効出願日の受託番号に関連するバージョンを意味する。有効出願日は、該当する場合、実際の出願日または受託番号に言及する優先出願の出願日の早いほうを意味する。同様に、刊行物、ウェブサイトなどの異なるバージョンが異なる時点で公開されている場合、別段示されていない場合は、本出願の有効出願日に最も近く公開されたバージョンを意味する。本開示の任意の構成、ステップ、エレメント、実施形態、または態様は、別段具体的に示されていない場合は、いずれかの他のものと組み合わせて使用することができる。本開示は、明確化および理解を目的として、例証および実施例によって幾分詳細に記載されているが、特定の変更および修正が添付の特許請求の範囲の範囲内で実践され得ることは明らかであろう。

#### 【実施例】

#### 【0163】

##### （実施例1）

図6は、無細胞DNAのシーケンシングリードファミリーを示す。シーケンシングリードは、ヒト染色体2のALK遺伝子（CD246）の様々なセグメントにマッピングされる。ALK遺伝子の関連領域の参照配列は、図の下部に示されている（配列中のギャップは、図面の簡潔さのために示されていない追加のヌクレオチドを表す）。この図は、上から下に、それぞれ、2、3、6、3、および6つのリードを有する、5つのシーケンシングリードファミリーを示す。一方の向きからのリードを、黒色で示し、他方の向きからの

リードを、白色で示す。ファミリーのそれぞれは、ファミリーのそれぞれのリードにおいて、GからAへのミスマッチを示す。別々に見ると、これらのシーケンシングリードファミリーは、GからAへの変異をコールするのに十分な根拠を提供する。しかしながら、この図は、GからAへの変異の位置が、配列リードの3'末端に対して、以下：

(1) ファミリー1：第1の鎖：2つのリード、第2の鎖：リードなし、GからAへの変異は、3'末端から70塩基に位置している

(2) ファミリー2：第1の鎖：リードなし、第2の鎖：3つのリード、GからAへの変異は、3'末端から2塩基に位置している

(3) ファミリー3：第1の鎖：2つのリード、第2の鎖：4つのリード、GからAへの変異は、3'末端から6塩基に位置している

(4) ファミリー4：第1の鎖：1つのリード、第2の鎖：2つのリード、GからAへの変異は、3'末端から1塩基に位置している

(5) ファミリー5：第1の鎖：5つのリード、第2の鎖：1つのリード、GからAへの変異は、3'末端から3塩基に位置している

とみなされる場合に变化する。

【0164】

GからAへの置換が見られた5つのファミリーのうちの4つ(80%)において、置換は、シーケンシングリードの3'末端から20塩基以内に生じている。80%という数字は、シーケンシングリードの3'末端から20塩基以内に置換を有するGからAへの置換を有するファミリーの50%を上回る。したがって、置換は、真実の変異(これは、がんの診断または予後診断に関係を有し得る)以外は、機能的有意性のない脱アミノ化に誘導されるエラーとして分類される。

(実施例2)

【0165】

図7は、図6と同様の形式で提示されており、それぞれ、8、4、2、5、および4つのメンバーを有する5つのファミリーからのシーケンシングリードを示す。ここでも、5つのファミリーのそれぞれは、そのリードのそれぞれに、見かけのGからAへの置換を有する。しかしながら、この場合には、シーケンシングリードの3'末端への置換の相対的な位置は、以下に示されるように異なっている。

(1) ファミリー1：第1の鎖：8つのリード、第2の鎖：リードなし、GからAへの変異は、3'末端から62塩基に位置している

(2) ファミリー2：第1の鎖：2つのリード、第2の鎖：2つのリード、GからAへの変異は、3'末端から2塩基に位置している

(3) ファミリー3：第1の鎖：2つのリード、第2の鎖：リードなし、GからAへの変異は、3'末端から72塩基に位置している

(4) ファミリー4：第1の鎖：1つのリード、第2の鎖：4つのリード、GからAへの変異は、3'末端から63塩基に位置している

(5) ファミリー5：第1の鎖：リードなし、第2の鎖：4つのリード、GからAへの変異は、3'末端から79塩基に位置している

【0166】

この事例では、5つのファミリーのうちの1つ(20%)のみが、そのシーケンシングリードの3'末端から20塩基以内に見かけのGからAへの置換を有する。20%は、変異を脱アミノ化エラーとして分類するための50%のカットオフを上回っていない。したがって、見かけのGからAへの置換は、真実の変異として分類される。

本発明は、例えば、以下の項目を提供する。

(項目1)

核酸の集団におけるバリエーションヌクレオチドを識別するための方法であって、

(a) 一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子を含む核酸の集団を、5'-3'ポリメラーゼ活性および3'-5'エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、前記タンパク質が、一方または両方の末端において、

10

20

30

40

50

3'オーバーハングを消化し、5'オーバーハングを相補的な核酸で充填して、二本鎖平滑末端化核酸を生成する、ステップと、

(b) シーケンシングされた核酸を得るために、前記二本鎖平滑末端化核酸の配列を決定するステップと、

(c) 参照配列におけるそれぞれの指定位置について、

(i) 前記指定位置を含む、シーケンシングされた核酸のサブセットを識別し、

(ii) 前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸を識別する、ステップと、

(d) ステップ(c)の(ii)における前記シーケンシングされた核酸が、前記コールをサポートするそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、

(i) 前記バリエーションが、前記参照ヌクレオチドと比較して、CからTまたはGからAへの変動である場合、および

(ii) 前記バリエーションヌクレオチドが、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 前記サブセット内のシーケンシングされた核酸の5'末端から、前記指定位置における前記CからTへの変動の距離、または前記サブセット内のシーケンシングされた核酸の3'末端から、前記指定位置における前記GからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップを含む、方法。

(項目2)

ステップ(c)の(ii)が、前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内の核酸の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、前記変動を有する前記サブセット内のシーケンシングされた核酸の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる、項目1に記載の方法。

(項目3)

前記バリエーションヌクレオチドが、前記サブセット内のシーケンシングされた核酸において5'末端の規定される近接度内にある前記指定位置における前記CからTへの変動の出現、または前記サブセット内のシーケンシングされた核酸において3'末端の規定される近接度内にある前記指定位置における前記GからAへの変動の出現に基づいて、脱アミノ化エラーとして分類される、項目1に記載の方法。

(項目4)

ステップ(c)の(ii)が、前記指定位置が参照ヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸の数を識別することをさらに含む、先行する項目のいずれかに記載の方法。

(項目5)

ステップ(b)が、前記二本鎖平滑末端化核酸の両方の鎖の配列を決定することを含む、先行する項目のいずれかに記載の方法。

(項目6)

ステップ(c)が、少なくとも1つの指定位置について行われ、前記変動を有する前記サブセット内の前記シーケンシングされた核酸が、前記二本鎖平滑末端化核酸のシーケンシングされた核酸の両方の鎖の配列を含む、項目5に記載の方法。

(項目7)

ステップ(b)が、鎖の両方の末端から配列を決定することを含む、先行する項目のいずれかに記載の方法。

(項目8)

前記二本鎖平滑末端化核酸を、バーコードを含むアダプターに連結させること、前記アダプターに結合するプライマー分子からプライミングされる前記核酸を増幅させることをさらに含む、

10

20

30

40

50

ステップ（b）が、増幅した核酸分子の配列を決定すること、および前記増幅した核酸分子の配列をファミリーにグループ分けすることであって、ファミリーのメンバーが、前記核酸上の同じ開始点および終止点ならびに同じバーコードを有し、前記ファミリーについて、そのそれぞれのメンバーの配列から、複数の位置のそれぞれにおけるコンセンサスヌクレオチドを決定することを含む、先行する項目のいずれかに記載の方法。

(項目 9)

前記核酸の集団が、対象の無細胞核酸を含む試料に由来する、先行する項目のいずれかに記載の方法。

(項目 10)

前記無細胞核酸が、がんを有するかまたはがんを有することに一致する徴候または症状を有する対象の体液に由来する、項目 9 に記載の方法。

10

(項目 11)

前記体液が、血液、血漿、唾液、尿、および脳脊髄液からなる群から選択される、項目 10 に記載の方法。

(項目 12)

前記指定位置における前記 C から T への変動は、その出現が、前記指定位置が前記 5' 末端の規定される近接度内にある前記サブセットの第 1 の比率において少なくとも 50% である場合に、脱アミノ化エラーとして分類されるか、または前記指定位置における前記 G から A への変動は、その出現が、前記指定位置が前記 3' 末端の規定される近接度内にある前記サブセットの第 2 の比率において少なくとも 50% である場合に、脱アミノ化エラーとして分類される、先行する項目のいずれかに記載の方法。

20

(項目 13)

前記指定位置における前記 C から T への変動は、前記変動が、前記指定位置が前記 5' 末端の規定される近接度内にある前記サブセットの第 1 の比率において、前記サブセット内の他のシーケンシングされた核酸におけるよりも少なくとも 2 倍の出現を有することに基づいて、脱アミノ化エラーとして分類されるか、または前記指定位置における前記 G から A への変動は、前記変動が、前記指定位置が前記 3' 末端の規定される近接度内にある前記サブセットの第 2 の比率において、前記サブセット内の他のシーケンシングされた核酸におけるよりも少なくとも 2 倍の出現を有することに基づいて、脱アミノ化エラーとして分類される、項目 12 に記載の方法。

30

(項目 14)

前記閾値は、前記変動が、前記サブセット内のシーケンシングされた核酸の少なくとも 1% で存在することである、先行する項目のいずれかに記載の方法。

(項目 15)

前記 C から T または G から A への変動が、少なくとも、前記周囲のコンテキストが T C G から T T G または C G A から C A A であることに基づいて、脱アミノ化エラーとして分類される、先行する項目のいずれかに記載の方法。

(項目 16)

前記 5' 末端へ前記規定される近接度が、前記 5' 末端まで 20 ヌクレオチド以内であるか、またはそれよりも少ない数のヌクレオチド以内であるとして規定され、前記 3' 末端への前記規定される近接度が、前記 3' 末端まで 20 ヌクレオチド以内であるか、またはそれよりも少ない数のヌクレオチド以内であるとして規定される、項目 3 に記載の方法。

40

(項目 17)

前記 5' 末端への前記規定される近接度が、前記 5' 末端まで 20 ヌクレオチド以内であるとして規定され、前記 3' 末端への前記規定される近接度が、前記 3' 末端まで 20 ヌクレオチド以内であるとして規定される、項目 16 に記載の方法。

(項目 18)

前記タンパク質が、クレノウである、先行する項目のいずれかに記載の方法。

(項目 19)

ステップ（c）および（d）が、これらのステップを実行するコンピューターにより作

50

動されるシステムなどにおいて行われる、先行する項目のいずれかに記載の方法。

(項目 2 0)

前記参照配列が、ヒトゲノムの配列である、先行する項目のいずれかに記載の方法。

(項目 2 1)

前記参照配列が、ヒト染色体の配列である、項目 2 0 に記載の方法。

(項目 2 2)

前記参照配列が、前記ヒトゲノムの非連続領域を含む、項目 2 0 に記載の方法。

(項目 2 3)

前記コールされるバリエーションヌクレオチドのうちの少なくとも 1 つが、がんに関連していることが既知である、先行する項目のいずれかに記載の方法。

(項目 2 4)

がんを有するかまたはがんを有することが疑われる対象の集団の試料に由来する核酸集団に行われ、前記集団内の対象が、その後、前記個々の対象においてどのバリエーションヌクレオチドがコールされたかに応じて、異なる処置を受容する、先行する項目のいずれかに記載の方法。

(項目 2 5)

脱アミノ化エラーとして分類されるバリエーションヌクレオチドが、前記コールされたバリエーションヌクレオチドのうちの少なくとも 1 % である、先行する項目のいずれかに記載の方法。

(項目 2 6)

脱アミノ化エラーとして分類されるバリエーションヌクレオチドが、前記コールされたバリエーションヌクレオチドのうちの少なくとも 1 0 % である、先行する項目のいずれかに記載の方法。

(項目 2 7)

バリエーションの存在は、少なくとも 5 個のバリエーションヌクレオチドが脱アミノ化エラーとして分類される場合、コールされない、先行する項目のいずれかに記載の方法。

(項目 2 8)

前記核酸の集団が、固形組織に由来する、先行する項目のいずれかに記載の方法。

(項目 2 9)

前記体液が、血漿である、項目 1 1 に記載の方法。

(項目 3 0)

前記 5 ' 末端に連結されるバーコードを含む前記アダプターが、前記 3 ' 末端に連結されるバーコードを含む前記アダプターとは異なる、項目 8 に記載の方法。

(項目 3 1)

前記脱アミノ化エラーの頻度が、少なくとも 1 % である、先行する項目のいずれかに記載の方法。

(項目 3 2)

前記脱アミノ化エラーの頻度が、少なくとも 1 0 % である、先行する項目のいずれかに記載の方法。

(項目 3 3)

前記バリエーションヌクレオチドは、前記サブセット内のシーケンシングされた核酸の 5 ' 末端から、前記指定位置における前記 C から T への変動の平均距離が、前記指定位置における前記参照ヌクレオチドの平均距離よりも小さいこと、または前記サブセット内のシーケンシングされた核酸の 3 ' 末端から、前記指定位置における前記 G から A への変動が、前記指定位置における前記参照ヌクレオチドの平均距離よりも小さいことに基づいて、脱アミノ化エラーとして分類される、先行する項目のいずれかに記載の方法。

(項目 3 4)

前記バリエーションヌクレオチドが、単一バリエーション ( S N V ) である、先行する項目のいずれかに記載の方法。

(項目 3 5)

10

20

30

40

50

核酸においてバリエーションヌクレオチドを識別する方法であって、

(a) 一本鎖オーバーハングを有する二本鎖核酸を、5' - 3' ポリメラーゼ活性および 3' - 5' エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、それによって、二本鎖平滑末端化核酸を生成するステップと、

(b) 前記二本鎖平滑末端化核酸の配列を決定するステップと、

(c) 前記決定した配列を参照配列と比較するステップであって、前記決定した配列が、前記決定した配列の 5' 末端の 20 ヌクレオチドまたはそれ未満以内の少なくとも 1 つの指定位置における少なくとも 1 つの C から T への変動、または前記決定した配列の 3' 末端の 20 ヌクレオチドまたはそれ未満以内における少なくとも 1 つの G から A への変動を含む、

ステップと、  
(d) 前記核酸の配列を、前記決定した配列としてコールするステップであって、ただし、C から T への変動が前記決定した配列の 5' 末端の 20 ヌクレオチドまたはそれ未満以内に存在するか、または G から A への変動が前記決定した配列の 3' 末端の 20 ヌクレオチドまたはそれ未満以内に存在する前記位置のうちの少なくとも 1 つにおいて、前記参照配列を占有しているヌクレオチドが、前記指定位置でコールされることを除く、ステップとを含む、方法。

(項目 36)

前記 C から T または G から A への変動が、TCG から TTG または CGA から CAA の周囲コンテキストにおいて生じる、項目 35 に記載の方法。

(項目 37)

核酸の集団においてバリエーションヌクレオチドを識別する方法であって、

(a) 少なくとも 1 つが一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子である、オーバーラップする配列の核酸の集団を、5' - 3' ポリメラーゼ活性および 3' - 5' エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、前記タンパク質が、3' オーバーハングを消化し、5' オーバーハングを充填して、二本鎖平滑末端化核酸を生成する、ステップと、

(b) 前記二本鎖平滑末端化核酸を、バーコードを含むアダプターに連結させ、前記アダプターに結合するプライマー分子からプライミングされる前記核酸を増幅させるステップと、

(c) 増幅した核酸分子の配列を決定し、前記増幅した核酸分子の配列をファミリーに分類し、ファミリーのメンバーが、前記核酸上の同じ開始点および終止点ならびに同じバーコードを有し、前記ファミリーについて、そのそれぞれのメンバーの配列から、コンセンサス配列を決定するステップと、

(d) 参照配列におけるそれぞれの指定位置について、前記指定位置を含むコンセンサス配列を有するファミリーのサブセットを決定するステップであって、前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内のコンセンサス配列を識別する、ステップと、

(e) 前記バリエーションヌクレオチドを有する前記サブセット内の前記コンセンサス配列が前記コールをサポートする指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエーションヌクレオチドの存在は、

(i) 前記バリエーションヌクレオチドが、前記参照ヌクレオチドと比較して、C から T または G から A への変動である場合、および

(i i) 前記バリエーションヌクレオチドが、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および / または

(2) 前記 5' 末端から、前記サブセット内のコンセンサス配列の前記指定位置における前記 C から T への変動の距離、または前記 3' 末端から、コンセンサス配列の前記指定位置における前記 G から A への変動の距離に基づいて、脱アミノ化エラーとして分類される場合

には、コールされないことを除く、ステップと

を含む、方法。

10

20

30

40

50

(項目 3 8)

ステップ (c) の (i i) が、前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内の核酸の数を識別し、ステップ (d) の (i) および (i i) で指定される場合を除き、前記変動を有する前記サブセット内のシーケンシングされた核酸の数が閾値に合致するとき、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる、項目 3 7 に記載の方法。

(項目 3 9)

核酸の集団において偽陽性バリエーションヌクレオチドを識別するための方法であって、

(a) 少なくとも 1 つが一方または両方の末端に一本鎖オーバーハングを有する二本鎖分子である、配列がオーバーラップする核酸の集団を、5' - 3' ポリメラーゼ活性および 3' - 5' エキソヌクレアーゼ活性を有するタンパク質と接触させるステップであって、前記タンパク質が、一方または両方の末端において、3' オーバーハングを消化し、5' オーバーハングを相補的な核酸で充填して、二本鎖平滑末端化核酸を生成する、ステップと、

(b) シーケンシングされた核酸を得るために、前記二本鎖平滑末端化核酸の配列を決定するステップと、

(c) 参照配列におけるそれぞれの指定位置について、前記指定位置を含むシーケンシングされた核酸のサブセットを識別し、そして前記指定位置が参照ヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸、および前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸の数を識別する、ステップと、

(d) 前記指定位置に C から T または G から A への変動指定位置で C から T または G から A への変動を有する前記シーケンシングされた核酸が、前記コールに合致してサポートするそれぞれの指定位置における偽陽性バリエーションヌクレオチドの存在をコールするステップであって、そして前記変動が、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および / または

(2) 前記指定位置が前記 5' 末端の規定される近接度内にある前記サブセットの第 1 の比率内のシーケンシングされた核酸における前記 C から T への変換の過剰出現、もしくは前記指定位置が前記 3' 末端の規定される近接度内にある、前記サブセットの第 2 の比率内のシーケンシングされた核酸における前記 G から A への変換の過剰出現に基づいて、脱アミノ化エラーとして分類される、ステップと

を含む、方法。

(項目 4 0)

ステップ (c) の (i i) が、前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内の核酸の数を識別し、ステップ (d) の (i) および (i i) で指定される場合を除き、前記変動を有する前記サブセット内のシーケンシングされた核酸の数が閾値に合致するとき、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる、項目 3 9 に記載の方法。

(項目 4 1)

参照配列内の指定位置における「C」から「T」または「G」から「A」へのバリエーションのマイナー対立遺伝子頻度を、前記指定位置にマッピングされるシーケンシングされた核酸の集団において決定する方法であって、マイナー対立遺伝子頻度が、前記バリエーションを含む前記指定位置にマッピングされるシーケンシングされた核酸の数（「バリエーション数」）を、前記指定位置にマッピングされるシーケンシングされた核酸の総数と比較し、前記方法が、脱アミノ化エラーの確率について、前記指定位置における T または A バリエーションの数を調節するステップを含み、エラーの確率が、「T」の場合には分子の 5' 末端からおよび「A」の場合には分子の 3' 末端からの前記バリエーションの距離の関数である、方法。

(項目 4 2)

シーケンシングされたポリヌクレオチドの 5' 末端から選択された距離内に位置する「T」バリエーション、またはシーケンシングされた核酸の 3' 末端から選択された距離内に位置する「A」バリエーションが、前記バリエーション数にカウントされない、項目 4 1 に記載の方法。

(項目 4 3)

シーケンシングされたポリヌクレオチドの 5' 末端から選択された距離内に位置する「T」バリエントの、シーケンシングされた核酸の 5' 末端から前記選択された距離外に位置する「T」バリエントに対する比が、所定の比を上回る（例えば、50%を上回る）場合、またはシーケンシングされた核酸の 3' 末端から選択された距離内に位置する「A」バリエントの、シーケンシングされた核酸の 3' 末端から前記選択された距離外に位置する「A」バリエントに対する比が、所定の比を上回る（例えば、50%を上回る）場合、すべての「T」バリエントは、前記バリエント数からディスカウントされる、項目 4 1 に記載の方法。

(項目 4 4)

前記バリエント数が、それぞれの「T」バリエントまたはそれぞれの「A」バリエントが、真のバリエントである確率の合計として決定される、項目 4 1 に記載の方法。

(項目 4 5)

先行する項目のいずれかに記載の方法によって、がんマーカーを有することが決定された対象に、前記がんマーカーにより特徴付けられるがんを処置するのに有効な治療介入を投与することを含む、方法。

(項目 4 6)

項目 1 から 4 5 のいずれか 1 項に記載の方法を行うことによって、対象の無細胞核酸における 1 つまたは複数のバリエントヌクレオチドの同一性に関するデータを受容するステップと、

前記 1 つまたは複数のバリエントヌクレオチドから、がんマーカーの存在を決定するステップと、

前記がんマーカーにより特徴付けられるがんを処置するのに有効な治療介入を投与するステップとを含む、方法。

(項目 4 7)

核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

前記通信インターフェースと通信するコンピューターであって、1 つまたは複数のコンピュータープロセッサと、前記 1 つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 前記核酸シーケンサーによって生成された前記シーケンシングリードを、前記通信ネットワークを通じて受容するステップと、

(b) 参照配列におけるそれぞれの指定位置について、

(i) 前記指定位置を含む、シーケンシングリードのサブセットを識別し、

(i i) 前記指定位置がバリエントヌクレオチドによって占有されている、前記サブセット内のシーケンシングリードを識別する、ステップと、

(c) 前記変動を有する前記サブセット内の前記シーケンシングリードが、前記コールをサポートするそれぞれの指定位置におけるバリエントヌクレオチドの存在をコールするステップであって、ただし、指定位置におけるバリエントヌクレオチドの存在は、

(i) 前記バリエントが、前記参照ヌクレオチドと比較して、C から T または G から A への変動である場合、および

(i i) 前記バリエントヌクレオチドが、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 前記サブセット内のシーケンシングされた核酸の 5' 末端から、前記指定位置における前記 C から T への変動の距離、または前記サブセット内のシーケンシングされた核酸の 3' 末端から、前記指定位置における前記 G から A への変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップと

を含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターと

を含む、システム。

10

20

30

40

50

(項目 4 8)

ステップ (c) の (i i) が、前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内の核酸の数を識別し、ステップ (d) の (i) および (i i) で指定される場合を除き、前記変動を有する前記サブセット内のシーケンシングされた核酸の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる、項目 4 7 に記載のシステム。

(項目 4 9)

核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

前記通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、前記1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 前記核酸シーケンサーによって生成された前記シーケンシングリードを、前記通信ネットワークを通じて受容するステップと、

(b) 前記決定した配列を、参照配列と比較するステップであって、前記決定した配列が、前記決定した配列の 5' 末端の 20 ヌクレオチドまたはそれ未満以内の少なくとも 1 つの指定位置における少なくとも 1 つの C から T への変動、または前記決定した配列の 3' 末端の 20 ヌクレオチドまたはそれ未満以内における少なくとも 1 つの G から A への変動を含む、ステップと、

(c) 前記核酸の配列を前記決定した配列としてコールするステップであって、ただし、C から T への変動が前記決定した配列の 5' 末端の 20 ヌクレオチドまたはそれ未満以内に存在するか、または G から A への変動が前記決定した配列の 3' 末端の 20 ヌクレオチドまたはそれ未満以内に存在する前記位置のうちの少なくとも 1 つにおいては、前記参照配列を占有しているヌクレオチドが、前記指定位置でコールされることを除く、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む、システム。

(項目 5 0)

核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

前記通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、前記1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 前記核酸シーケンサーによって生成された前記シーケンシングリードを、前記通信ネットワークを通じて受容するステップと、

(b) 前記シーケンシングリードの配列をファミリーにグループ分けするステップであって、ファミリーのメンバーが、前記核酸上の同じ開始点および終止点ならびに同じバーコードを有し、前記ファミリーについて、そのそれぞれのメンバーの配列から、コンセンサス配列を決定する、ステップと、

(c) 参照配列におけるそれぞれの指定位置について、前記指定位置を含むコンセンサス配列を有するファミリーのサブセットを決定し、前記指定位置がバリエーションヌクレオチドによって占有されている、前記コンセンサス配列を識別するステップと、

(d) 前記バリエーションヌクレオチドを有する前記サブセット内の前記コンセンサス配列がそれぞれの指定位置におけるバリエーションヌクレオチドの存在をコールするステップであって、ただし、前記コールをサポートするが、指定位置におけるバリエーションヌクレオチドの存在は、

(i) 前記バリエーションヌクレオチドが、前記参照ヌクレオチドと比較して、C から T または G から A への変動である場合、および

(i i) 前記バリエーションヌクレオチドが、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 5' 末端から、前記サブセット内のコンセンサス配列内の前記指定位置における

10

20

30

40

50

前記CからTへの変動の距離、または3'末端から、コンセンサス配列の前記指定位置における前記GからAへの変動の距離に基づいて、脱アミノ化エラーとして分類される場合には、コールされないことを除く、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む、システム。

(項目51)

ステップ(c)が、前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内のコンセンサス配列の数を識別し、ステップ(d)の(i)および(ii)で指定される場合を除き、前記変動を有する前記サブセット内のコンセンサス配列の数が閾値に合致すると、それぞれの指定位置におけるバリエーションヌクレオチドの存在がコールされる、項目50に記載のシステム。

10

(項目52)

核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、前記通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、前記1つまたは複数のコンピュータープロセッサによる実行の際に、

(a) 前記核酸シーケンサーによって生成された前記シーケンシングリードを、前記通信ネットワークを通じて受容するステップと、

(b) 参照配列におけるそれぞれの指定位置について、前記指定位置を含むシーケンシングリードのサブセットを識別し、前記指定位置が参照ヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸、および前記指定位置がバリエーションヌクレオチドによって占有されている前記サブセット内のシーケンシングされた核酸の数を識別する、ステップと、

20

(c) 前記指定位置でCからTまたはGからAへの変動を有する前記シーケンシングされた核酸が前記コールをサポートするそれぞれの指定位置における偽陽性バリエーションヌクレオチドの存在をコールするステップであって、前記変動が、

(1) 前記指定位置の周囲のヌクレオチドコンテキスト、および／または

(2) 前記指定位置が5'末端の規定される近接度内にある、前記サブセットの第1の比率内のシーケンシングされた核酸における前記CからTへの変換の過剰出現、もしくは前記指定位置が3'末端の規定される近接度内にある前記サブセットの第2の比率内のシーケンシングされた核酸における前記GからAへの変換の過剰出現に基づいて、脱アミノ化エラーとして分類される、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む、システム。

30

(項目53)

核酸シーケンサーによって生成されたシーケンシングリードを、通信ネットワークを通じて受容する通信インターフェースと、

前記通信インターフェースと通信するコンピューターであって、1つまたは複数のコンピュータープロセッサと、前記1つまたは複数のコンピュータープロセッサによる実行の際に、

40

(a) 前記核酸シーケンサーによって生成された前記シーケンシングリードを、前記通信ネットワークを通じて受容するステップと、

(b) 脱アミノ化エラーの確率に基づいて前記シーケンシングリードにおけるTまたはAバリエーションの数を調節するステップであって、エラーの確率が、「T」の場合には分子の5'末端からおよび「A」の場合には前記分子の3'末端からの前記バリエーションの距離の関数である、ステップとを含む方法を実施するマシン実行可能コードを含むコンピューター可読媒体とを含む、コンピューターとを含む、システム。

(項目54)

前記核酸シーケンサーをさらに含む、項目47から53のいずれか1項に記載のシステム。

50

(項目 5 5)

前記核酸シーケンサーが、対象に由来する無細胞DNA分子から生成されたシーケンシングライブラリーをシーケンシングし、前記シーケンシングライブラリーが、前記無細胞DNA分子およびバーコードを含むアダプターを含む、項目47から53のいずれか1項に記載のシステム。

(項目 5 6)

前記核酸シーケンサーが、前記シーケンシングライブラリーにシーケンシングバイシセシスを行って、前記シーケンシングリードを生成する、項目47から53のいずれか1項に記載のシステム。

(項目 5 7)

前記核酸シーケンサーが、前記シーケンシングライブラリーにパイロシーケンシング、単分子シーケンシング、ナノポアシーケンシング、半導体シーケンシング、シーケンシングバイライゲーション、またはシーケンシングバイハイブリダイゼーションを行って、前記シーケンシングリードを生成する、項目47から53のいずれか1項に記載のシステム。

(項目 5 8)

前記核酸シーケンサーが、前記シーケンシングライブラリーに由来するクローン単分子アレイを使用して、前記シーケンシングリードを生成する、項目47から53のいずれか1項に記載のシステム。

(項目 5 9)

前記核酸シーケンサーが、前記シーケンシングライブラリーをシーケンシングして前記シーケンシングリードを生成するための、マイクロウェルのアレイを有するチップを含む、項目47から53のいずれか1項に記載のシステム。

(項目 6 0)

前記コンピューター可読媒体が、メモリー、ハードドライブ、またはコンピューターサーバーを含む、項目47から53のいずれか1項に記載のシステム。

(項目 6 1)

前記通信ネットワークが、遠隔通信ネットワーク、インターネット、エクストラネット、またはイントラネットを含む、項目47から53のいずれか1項に記載のシステム。

(項目 6 2)

前記通信ネットワークが、分散コンピューティングの可能な1つまたは複数のコンピューターサーバーを含む、項目47から53のいずれか1項に記載のシステム。

(項目 6 3)

分散コンピューティングが、クラウドコンピューティングである、項目64に記載のシステム。

(項目 6 4)

前記コンピューターが、前記核酸シーケンサーから遠隔設置されているコンピューターサーバー上に設置されている、項目47から53のいずれか1項に記載のシステム。

(項目 6 5)

前記シーケンシングライブラリーが、試料を、1つまたは複数の試料と区別する試料バーコードをさらに含む、項目64に記載のシステム。

(項目 6 6)

ネットワークを通じて前記コンピューターと通信する、電子ディスプレイであって、(a)～(c)を実施した際の結果を表示するためのユーザーインターフェースを含む、電子ディスプレイをさらに含む、項目47から53のいずれか1項に記載のシステム。

(項目 6 7)

前記ユーザーインターフェースが、グラフィカルユーザーインターフェース(GUI)またはウェブベースユーザーインターフェースである、項目66に記載のシステム。

(項目 6 8)

前記電子ディスプレイが、パーソナルコンピューターにおいて存在する、項目66に記載のシステム。

10

20

30

40

50

(項目 6 9 )

前記電子ディスプレイが、インターネット対応コンピューターにおいて存在する、項目 6 6 に記載のシステム。

(項目 7 0 )

前記インターネット対応コンピューターが、前記コンピューターから遠隔した位置に設置されている、項目 6 9 に記載のシステム。

【図面】

【図 1】

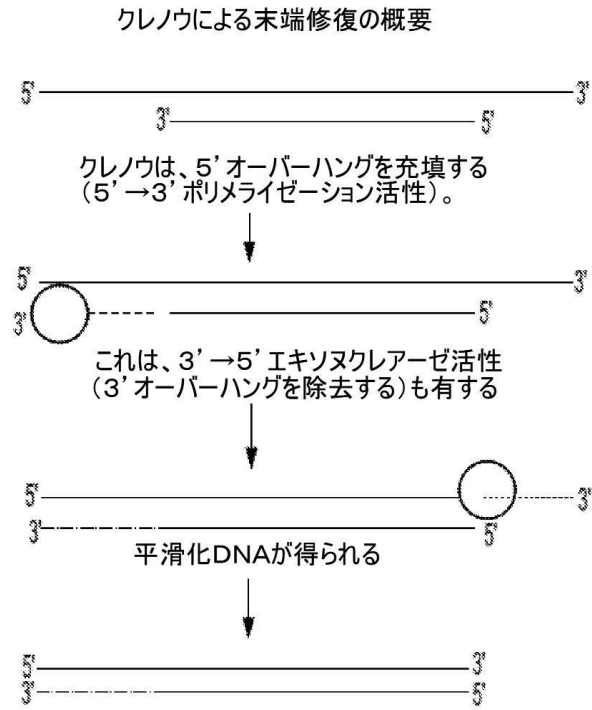
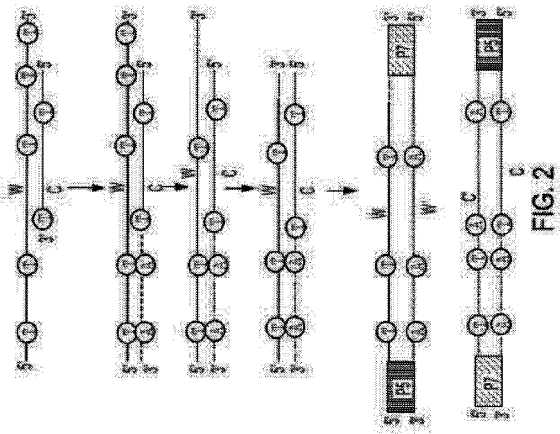


FIG. 1

【図 2】



10

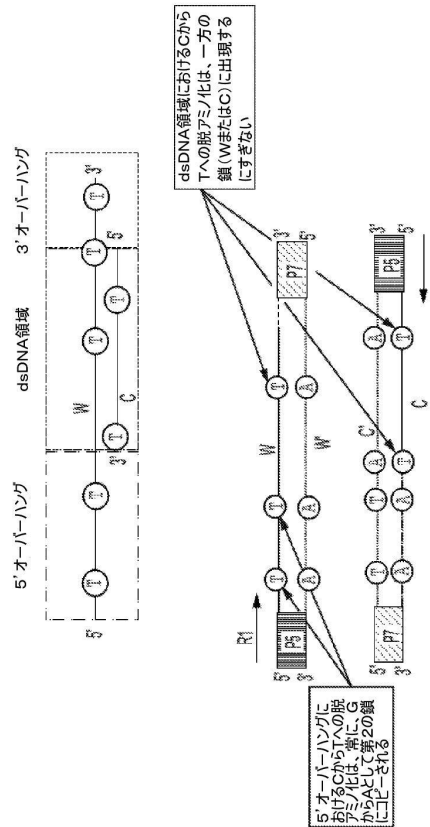
20

30

40

50

【図 3】



【図 4】

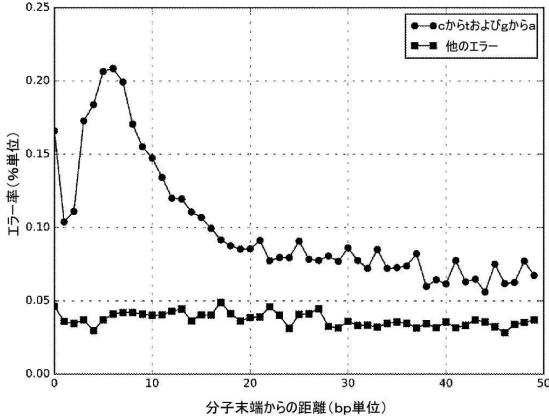


Fig. 4

FIG. 3

【図 5】

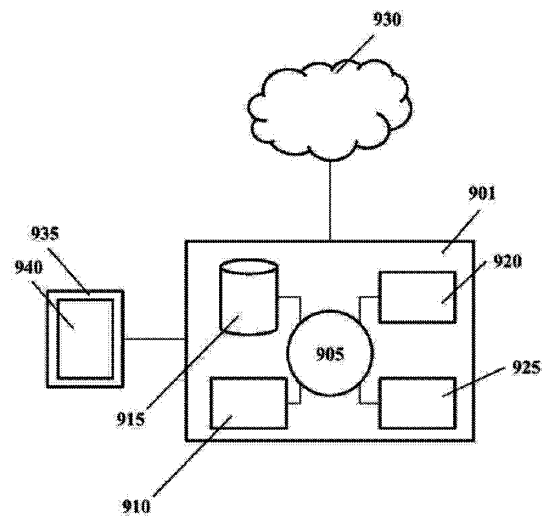


Fig. 5

【図 6】

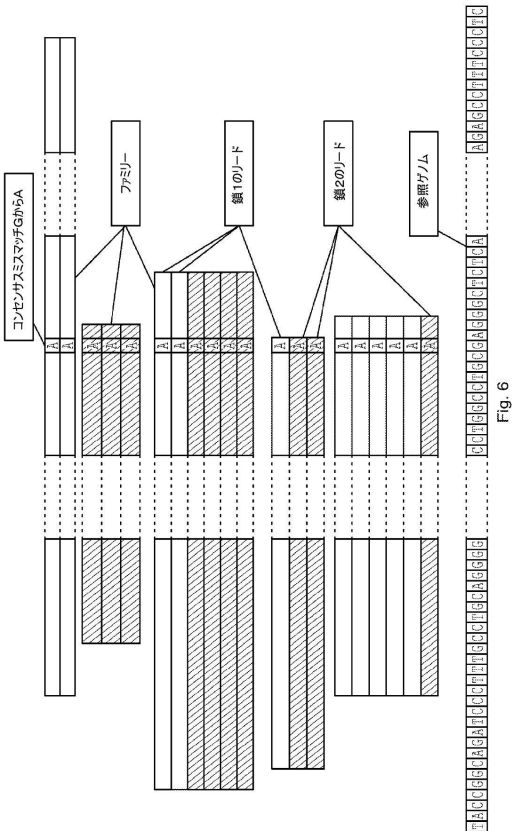


Fig. 6

10

20

30

40

50

【図 7】

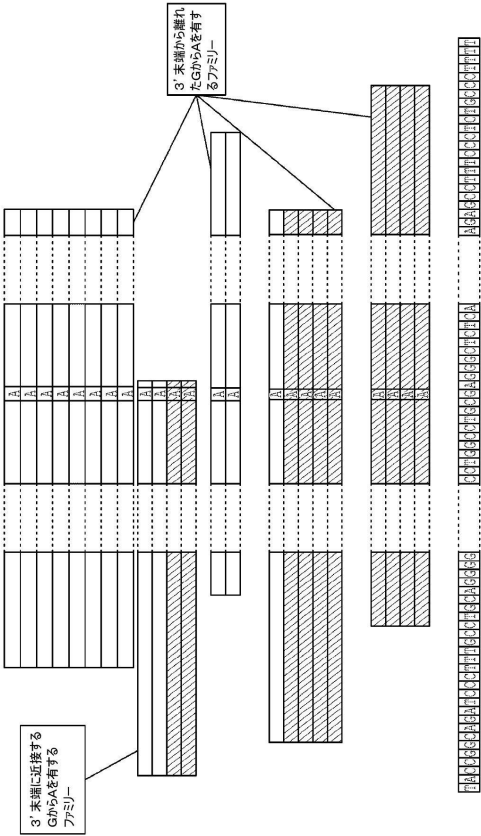


Fig. 7

10

20

30

40

50

## フロントページの続き

## (51)国際特許分類

F I

C 1 2 Q 1/6853(2018.01)  
C 1 2 N 15/09 (2006.01)C 1 2 Q 1/6853 Z  
C 1 2 N 15/09 Z

弁護士 山本 健策

(72)発明者 シコラ, マーシン

アメリカ合衆国 カリフォルニア 9 4 0 6 3, レッドウッド シティ, ペノブスコット ドライブ  
5 0 5, ガーダント ヘルス, インコーポレイテッド 気付

(72)発明者 ケネディ, アンドリュー

アメリカ合衆国 カリフォルニア 9 4 0 6 3, レッドウッド シティ, ペノブスコット ドライブ  
5 0 5, ガーダント ヘルス, インコーポレイテッド 気付

(72)発明者 ジャイモビッチ, アリエル

アメリカ合衆国 カリフォルニア 9 4 0 6 3, レッドウッド シティ, ペノブスコット ドライブ  
5 0 5, ガーダント ヘルス, インコーポレイテッド 気付

(72)発明者 チュドヴァ, ダーリヤ

アメリカ合衆国 カリフォルニア 9 4 0 6 3, レッドウッド シティ, ペノブスコット ドライブ  
5 0 5, ガーダント ヘルス, インコーポレイテッド 気付

(72)発明者 フェアクロー, スティーブン

アメリカ合衆国 カリフォルニア 9 4 0 6 3, レッドウッド シティ, ペノブスコット ドライブ  
5 0 5, ガーダント ヘルス, インコーポレイテッド 気付

審査官 松田 芳子

## (56)参考文献

国際公開第2016/149261 (WO, A1)

米国特許出願公開第2015/0066385 (US, A1)

特開2017-033046 (JP, A)

国際公開第2016/128758 (WO, A1)

国際公開第2017/136603 (WO, A1)

Chen L. et.al., DNA damage is a major cause of sequencing errors, directly confounding var  
iant identification, Biorxiv, 2016年08月23日, doi: <https://doi.org/10.1101/070334>Chen, G. et.al., Cytosine deamination is a major cause of baseline noise in next generation s  
equencing, Mol. diagn. ther., 2014年, vol.18, no.5, pp.587-593Briggs A. et.al., Preparation of next-generation sequencing libraries from damaged DNA,  
Methods in molecular biology, vol.840, 2012年, pp.143-154Starrett, G.J. et.al., The DNA cytosine deaminase APOBEC3H haplotype I likely contributes  
to breast and lung cancer mutagenesis, Nature communications, 2016年, vol.7, pp.1-13,  
doi: 10.1038/ncomms12918

(58)調査した分野 (Int.Cl., DB名)

C 1 2 Q 1/68

JSTPlus/JMEDPlus/JST7580 (JDreamIII)

Caplus/MEDLINE/EMBASE/BIOSIS (STN)