(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2024-544597
(P2024-544597A)
(43)公表日　令和6年12月3日(2024.12.3)

(51)国際特許分類		FI			テーマコード（参考）
Ｃ１２Ｑ	*1/6886(2018.01)*	Ｃ１２Ｑ	1/6886	Ｚ	４Ｂ０２９
Ｃ１２Ｑ	*1/6806(2018.01)*	Ｃ１２Ｑ	1/6806	Ｚ ＺＮＡ	４Ｂ０６３
Ｃ１２Ｍ	*1/34 (2006.01)*	Ｃ１２Ｍ	1/34	Ｂ	
Ｃ１２Ｑ	*1/6869(2018.01)*	Ｃ１２Ｍ	1/34	Ａ	
Ｇ１６Ｂ	*40/00 (2019.01)*	Ｃ１２Ｑ	1/6869	Ｚ	

審査請求　未請求　予備審査請求　未請求　（全183頁）　最終頁に続く

(21)出願番号　特願2024-530567(P2024-530567)
(86)(22)出願日　令和4年11月23日(2022.11.23)
(85)翻訳文提出日　令和6年7月18日(2024.7.18)
(86)国際出願番号　PCT/US2022/080431
(87)国際公開番号　WO2023/097278
(87)国際公開日　令和5年6月1日(2023.6.1)
(31)優先権主張番号　63/282,509
(32)優先日　令和3年11月23日(2021.11.23)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　522105894
グレイル インコーポレイテッド
アメリカ合衆国 カリフォルニア ９４０２５， メンロー パーク， オブライエン ドライブ １５２５
(74)代理人　100147485
弁理士　杉村 憲司
(74)代理人　230118913
弁護士　杉村 光嗣
(74)代理人　100174001
弁理士　結城 仁美
(72)発明者　サミュエル エス グロス
アメリカ合衆国 カリフォルニア州 ９４０２５ メンロー パーク オブライエン ドライブ １５２５
最終頁に続く

(54)【発明の名称】　がん分類のための汚染片のサンプル汚染検出

(57)【要約】
がん分類のために生物学的サンプル中の汚染断片を検出する方法およびシステムを開示する。このシステムは、複数のＳＮＰ部位汚染マーカーおよびインデル部位汚染マーカーを同定する。複数のＳＮＰ部位汚染マーカーは、閾値距離内の少なくとも２つのＳＮＰ部位を含み、閾値頻度の範囲内の集団ハプロタイプ頻度を有し、グアニン－アデニン多型および／またはシトシン－チミン多型を除外し、ハーディーワインベルク平衡を確保し、または上記のパラメータの任意の組み合わせを含む。インデル部位汚染マーカーは、閾値長さ内にあり、高い複雑性を有し、集団ハプロタイプ頻度が閾値頻度の範囲内にあり、ハーディーワインベルク平衡を保証し、または上記のパラメータの任意の組み合わせを含む、インデル配列を含む。システムは、サンプルがホモ接合体である汚染マーカーを同定する。システムは、それぞれの汚染マーカー部位のホモ接合性ハプロタイプとは異なるハプロタイプを有する断片を同定することにより、サンプルの汚染レベルを推定する。

【特許請求の範囲】
【請求項１】
試験サンプル中のがんの存在を予測する方法であって、
前記試験サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む、前記試験サンプルを取得するステップと、
前記試験サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、
同定された汚染マーカーの１つにおいて、それぞれの汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記試験サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
いずれかの同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および
前記汚染レベルが閾値レベル以下であるか否かを判定するステップと、および
前記汚染レベルが前記閾値レベル以下であると判定したことに応答して、前記試験サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに対してがん分類を行い、がん予測を生成するステップ、
を備える、方法。
【請求項２】
請求項１に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項３】
請求項２に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項４】
請求項２～３のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項５】
請求項２～４のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項６】
請求項２～５のいずれか一項に記載の方法において、各複数のＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項７】
請求項２～６のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項８】
請求項２～７のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、方法。
【請求項９】
請求項１～８のいずれか一項に記載の方法において、前記複数の汚染マーカーは挿入－欠失（インデル）部位を含む、方法。
【請求項１０】
請求項９に記載の方法において、前記インデル部位は５ｂｐｓ～１０ｂｐｓである、方法。
【請求項１１】
請求項９～１０のいずれか一項に記載の方法において、前記インデル部位は４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する、方法。
【請求項１２】
請求項９～１１のいずれか一項に記載の方法において、各インデル部位の前記ハプロタ

イプはハーディーワインベルク平衡にある、方法。
【請求項１３】
請求項９～１２のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項１４】
請求項９～１３のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項１５】
請求項１～１４のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項１６】
請求項１～１５のいずれか一項に記載の方法において、前記汚染レベルの推定は、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記試験サンプルのシーケンス深度、前記試験サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項１７】
請求項１～１６のいずれか一項に記載の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、がんの分類を見送る、方法。
【請求項１８】
請求項１～１７のいずれか一項に記載の方法において、前記がん予測は、がんと非がんとの間の二値予測を含む、方法。
【請求項１９】
請求項１～１８のいずれか一項に記載の方法において、前記がん予測は、複数のがんタイプ間のマルチクラスがん予測を含む、方法。
【請求項２０】
請求項１～１９のいずれか一項に記載の方法において、前記がんの分類を実行することは、
前記試験サンプル中のｃｆＤＮＡ断片の配列リードに基づいて試験特徴ベクターを生成するステップと、および
前記試験特徴ベクターを分類モデルに入力し、前記試験サンプルの前記がん予測を生成するステップと、
を含む、方法。
【請求項２１】
請求項２０に記載の方法において、前記がんの分類を実行することは、さらに、
前記試験サンプルのｃｆＤＮＡ断片の初期セットをｐ値フィルタリングでフィルタリングして異常断片のセットを生成するステップであって、
前記フィルタリングは、前記初期セットから他の断片に関して閾値ｐ値を下回る断片を除去して前記異常断片のセットを生成するステップを備え、
ここで、前記試験特徴ベクターは、前記異常断片のセットの前記配列リードに基づく、方法。
【請求項２２】
請求項２０～２１のいずれか一項に記載の方法において、前記分類モデルは機械学習モデルである、方法。
【請求項２３】
コンピュータプロセッサによって実行されると、請求項１～２２のいずれか一項に記載の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項２４】
システムであって、

コンピュータプロセッサ、および
請求項２３に記載の前記非一過性のコンピュータ可読記憶媒体
を備える、システム。
【請求項２５】
試験サンプル中の疾患の存在を予測する方法であって、
前記試験サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む前記試験サンプルを取得するステップと、
前記試験サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、
同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記試験サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および
前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、および
前記汚染レベルが前記閾値レベル以下であると決定したことに応答して、前記試験サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに対して疾患分類を行い、疾患予測を生成するステップ
を備える、方法。
【請求項２６】
請求項２５に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項２７】
請求項２６に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項２８】
請求項２６～２７のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項２９】
請求項２６～２８のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項３０】
請求項２６～２９のいずれか一項に記載の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項３１】
請求項２６～３０のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項３２】
請求項２６～３１のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、方法。
【請求項３３】
請求項２５～３２のいずれか一項に記載の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、方法。
【請求項３４】
請求項３３に記載の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、方法。
【請求項３５】
請求項３３～３４のいずれか一項に記載の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する、方法。

【請求項３６】
請求項３３～３５のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項３７】
請求項３３～３６のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項３８】
請求項３３～３７のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項３９】
請求項２５～３８のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項４０】
請求項２５～３９のいずれか一項に記載の方法において、前記汚染レベルの推定は、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記試験サンプルのシーケンス深度、前記試験サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項４１】
請求項２５～４０のいずれか一項に記載の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、疾患分類を見送る、方法。
【請求項４２】
請求項２５～４１のいずれか一項に記載の方法において、前記疾患予測は、疾患と疾患なしとの間の二値予測を備える、方法。
【請求項４３】
請求項２５～４２のいずれか一項に記載の方法において、前記疾患予測は、複数の疾患間のマルチクラスがん予測を含む、方法。
【請求項４４】
請求項２５～４３のいずれか一項に記載の方法において、前記疾患分類を実行することを含む方法であって、
前記試験サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに基づいて試験特徴ベクターを生成するステップ、および
前記試験特徴ベクターを分類モデルに入力し、前記試験サンプルの前記疾患予測を生成するステップ
を備える、方法。
【請求項４５】
請求項４４に記載の方法において、前記疾患分類を実行することは、さらに、
異常断片の集合を生成するために、ｐ値フィルタリングを用いて前記試験サンプルのｃｆＤＮＡ断片の初期集合をフィルタリングするステップであって、前記フィルタリングは、前記異常断片の集合を生成するために、他の断片に関して閾値ｐ値を下回る断片を前記初期集合から除去するステップを備え、
ここで、前記試験特徴ベクターは、前記異常断片の集合の前記配列リードに基づく、
方法。
【請求項４６】
請求項４４～４５のいずれか一項に記載の方法において、前記分類モデルは機械学習モデルである、方法。
【請求項４７】
コンピュータプロセッサによって実行されると、請求項２５～４６のいずれか一項に記載の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体。

【請求項４８】
システムであって、
コンピュータプロセッサ、および
請求項４７に記載の前記非一過性のコンピュータ可読記憶媒体
を備える、システム。
【請求項４９】
試験サンプル中の汚染の有無を予測する方法であって、
前記試験サンプル中の複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片に由来する配列リードを取得するステップと、
前記配列リードに基づいて、前記試験サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、
同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記試験サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および
前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、および
前記汚染レベルが前記閾値レベル以上であると決定することに応答して、前記試験サンプルが汚染されていることを示す通知を生成するステップ
を備える、方法。
【請求項５０】
請求項４９に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項５１】
請求項４９に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項５２】
請求項５０～５１のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項５３】
請求項５０～５２のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項５４】
請求項５０～５３のいずれか一項に記載の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項５５】
請求項５０～５４のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項５６】
請求項５０～５５のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１からの前記複数のＳＮＰ部位を含む、方法。
【請求項５７】
請求項４９～５６のいずれか一項に記載の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、方法。
【請求項５８】
請求項５７に記載の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、方法。
【請求項５９】
請求項５７～５８のいずれか一項に記載の方法において、前記インデル部位は、４５％

～５５％の範囲内の前記集団ハプロタイプ頻度を有する、方法。
【請求項６０】
請求項５７～５９のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項６１】
請求項５７～６０のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項６２】
請求項５７～６１のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項６３】
請求項４９～６２のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項６４】
請求項４９～６３のいずれか一項に記載の方法において、前記汚染レベルの推定は、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記試験サンプルのシーケンス深度、前記試験サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項６５】
コンピュータプロセッサによって実行されると、請求項４９～６４のいずれか一項に記載の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項６６】
システムであって、
コンピュータプロセッサ、および
請求項６５に記載の前記非一過性のコンピュータ可読記憶媒体
を備える、システム。
【請求項６７】
がん分類モデルを訓練する方法であって、
第１の訓練サンプルを含む複数の訓練サンプルを取得し、各訓練サンプルは複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片を含むステップと、
各訓練サンプルについて、前記訓練サンプル中の前記ｃｆＤＮＡ断片に由来する配列リードを取得するステップと、
前記第１の訓練サンプルについて、
前記第１の訓練サンプルの前記配列リードに基づいて、前記第１の訓練サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから１つ以上の汚染マーカーを同定するステップと、
同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーの前記ホモ接合性ハプロタイプとは異なるハプロタイプを有する、前記第１の訓練サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
同定されたあらゆる汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、
前記汚染レベルが閾値レベル以下であるか否かを決定するステップ、および
前記第１の訓練サンプルの前記汚染レベルが前記閾値レベル以上であると決定することに応答して、前記複数の訓練サンプルから前記第１の訓練サンプルを除去するステップであって、ここで、前記第１の訓練サンプルを除く前記複数の訓練サンプルは、前記がん分類モデルを訓練して試験サンプルに対するがん予測を生成するために使用されるステップを備える、方法。
【請求項６８】

請求項６７に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項６９】
請求項６８に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項７０】
請求項６８～６９のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項７１】
請求項６８～７０のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項７２】
請求項６８～７１のいずれか一項に記載の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項７３】
請求項６８～７２のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項７４】
請求項６８～７３のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、方法。
【請求項７５】
請求項６７～７４のいずれか一項に記載の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、方法。
【請求項７６】
請求項７５に記載の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、方法。
【請求項７７】
請求項７５～７６のいずれか一項に記載の方法において、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項７８】
請求項７５～７７のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項７９】
請求項７５～７８のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項８０】
請求項７５～７９のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項８１】
請求項６７～８０のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項８２】
請求項６７～８１のいずれか一項に記載の方法において、汚染レベルを推定することは、さらに、前記第１の訓練サンプル中の同定された汚染ｃｆＤＮＡ断片の数、前記第１の訓練サンプルの配列決定深度、前記第１の訓練サンプル中のｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項８３】
請求項６７～８２のいずれか一項に記載の方法において、前記複数の訓練サンプルは、

非がんサンプルの第１のコホートおよびがんサンプルの第２のコホートを備え、前記がん分類モデルが、がんの存在の可能性を決定するために訓練される、方法。
【請求項８４】
請求項８３に記載の方法において、前記がんサンプルの第２のコホートは、第１のがんタイプを有する１つ以上のサンプルおよび、第２のがんタイプを有する１つ以上の追加サンプルを含み、ここで前記がん分類モデルは、前記第１のがんタイプの存在の第１の尤度および前記第２のがんタイプの存在の第２の尤度を決定するように訓練される、方法。
【請求項８５】
請求項６７～８４のいずれか一項に記載の方法において、前記がん分類モデルは、機械学習モデルである、方法。
【請求項８６】
請求項８５に記載の方法において、がん分類モデルは、決定木、ニューラルネットワーク、多層パーセプトロン、およびサポートベクターマシンのうちの少なくとも１つである、方法。
【請求項８７】
コンピュータプロセッサによって実行されると、請求項６７～８６のいずれか一項に記載の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項８８】
システムにおいて、
コンピュータプロセッサ、および
請求項８７に記載の前記非一過性のコンピュータ可読記憶媒体
を備える、システム。
【請求項８９】
コンピュータプログラム製品において、
訓練された前記がん分類モデルを記憶する前記非一過性のコンピュータ可読記憶媒体であって、前記コンピュータプログラム製品が、請求項６７～８６のいずれか一項に記載の方法により作成される、コンピュータプログラム製品。
【請求項９０】
治療キットにおいて、
個体からの遺伝物質を含む生物学的サンプルを保存するための１つ以上の収集容器、および
複数の汚染マーカーを標的とする複数のプローブであって、前記複数のプローブは、表２および表４のうちの少なくとも１つを含む、複数のプローブ。
【請求項９１】
請求項９０に記載の治療キットにおいて、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、治療キット。
【請求項９２】
請求項９１に記載の治療キットにおいて、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、治療キット。
【請求項９３】
請求項９１～９２のいずれか一項に記載の治療キットにおいて、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、治療キット。
【請求項９４】
請求項９１～９３のいずれか一項に記載の治療キットにおいて、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、治療キット。
【請求項９５】
請求項９１～９４のいずれか一項に記載の治療キットにおいて、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルグ平衡にある、治療キット。
【請求項９６】

請求項９１～９５のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、治療キット。
【請求項９７】
請求項９１～９６のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、治療キット。
【請求項９８】
請求項９１～９７のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、治療キット。
【請求項９９】
請求項９８に記載の治療キットにおいて、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、治療キット。
【請求項１００】
請求項９８～９９のいずれか一項に記載の治療キットにおいて、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、治療キット。
【請求項１０１】
請求項９８～１００のいずれか一項に記載の治療キットにおいて、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、治療キット。
【請求項１０２】
請求項９８～１０１のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、治療キット。
【請求項１０３】
請求項９８～１０２のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、表３の前記インデル部位を含む、治療キット。
【請求項１０４】
請求項９０～１０３のいずれか一項に記載の治療キットにおいて、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、治療キット。
【請求項１０５】
請求項９０～１０４のいずれか一項に記載の治療キットにおいて、さらに、
生物学的サンプル中の核酸断片を単離するための１つ以上の試薬を備える、治療キット。
【請求項１０６】
請求項９０～１０５のいずれか一項に記載の治療キットにおいて、さらに、
第１のコンピュータプログラム製品であって、
請求項２３に記載の、非一過性のコンピュータ可読記憶媒体、
請求項４７に記載の、非一過性のコンピュータ可読記憶媒体、
請求項６５に記載の、非一過性のコンピュータ可読記憶媒体、および
請求項８７に記載の、非一過性のコンピュータ可読記憶媒体
のうちの１つ以上を備える、治療キット。
【請求項１０７】
請求項９０～１０６のいずれか一項に記載の治療キットにおいて、さらに、
請求項８９に記載のコンピュータプログラム製品を備える、治療キット。
【発明の詳細な説明】
【背景技術】
【０００１】
[関連出願への相互参照]
本出願は、２０２１年１１月２３日に出願された米国仮出願第６３／２８２，５０９号の利益および優先権を主張し、この仮出願は、参照によりその全体が組み込まれる。

【０００２】
デオキシリボ核酸（ＤＮＡ）のメチル化は、遺伝子発現の制御において重要な役割を果たしている。異常なＤＮＡメチル化は、がんを含む多くの疾患プロセスに関与している。メチル化配列決定を用いるＤＮＡメチル化プロファイリング（例えば、全ゲノムバイサルファイト配列決定（ＷＧＢＳ））は、がんの検出、診断、および／またはがんのモニタリングのための貴重な診断ツールとしてますます認識されている。例えば、異なるメチル化領域の特異的パターンおよび／または対立遺伝子特異的メチル化パターンは、循環セルフリー（ｃｆ）ＤＮＡを用いた非侵襲的診断の分子マーカーとして有用である。しかし、本技術分野では、サンプルの個体に由来しないｃｆＤＮＡ断片の汚染検出方法の改善に対する必要性が残っている。

【０００３】
本開示は、上記の課題に対処することに向けられている。本明細書において提供する背景説明は、本開示の文脈を一般的に提示するためのものである。本明細書において特に示されない限り、本項に記載する材料は、本出願の特許請求の範囲に対する先行技術ではなく、本項に含めることによって先行技術であること、または先行技術の示唆であることを認めるものではない。

【先行技術文献】
【特許文献】
【０００４】
【特許文献１】米国特許出願第６２／６４２，４８０号
【特許文献２】米国特許出願第１６／７１９，９０２号
【特許文献３】米国特許出願第１６／３５２，６０２号
【特許文献４】米国特許出願第１５／９３１，０２２号
【特許文献５】米国仮特許出願第６２／９８５，２５８号
【特許文献６】米国特許出願１７／１１９，６０６号
【特許文献７】米国出願第１５／９３１，０２２号
【特許文献８】米国出願第１６／５７９，８０５号
【特許文献９】米国出願第１６／３５２，６０２号
【特許文献１０】米国出願第１６／７２３，７１６号

【発明の概要】
【発明が解決しようとする課題】
【０００５】
被験者における疾患状態（がんなど）の早期発見は、早期治療が可能となり、したがって生存の可能性が高まるため重要である。セルフリー（ｃｆ）ＤＮＡサンプル中のＤＮＡ断片の配列決定は、疾患の分類に使用できる特徴を同定するために使用することができる。例えば、がんの評価では、血液サンプルからのセルフリーＤＮＡに基づく特徴（体細胞変異、メチル化状態、または他の遺伝子異常の有無など）は、被験者ががんに罹患している可能性があるかどうかについての洞察、および被験者がどのようながんに罹患している可能性があるかについてのさらなる洞察を提供することができる。この目的のために、本明細書には、被験体が疾患に罹患している可能性を決定するためのセルフリーＤＮＡ（ｃｆＤＮＡ）配列決定データを解析するシステムおよび方法が含まれる。

【０００６】
本開示は、がん分類のための汚染断片のサンプル汚染検出のための改良されたシステムおよび方法を提供することによって、上記で特定された問題に対処する。本システムは、サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから１つ以上の汚染マーカーを同定する。本システムは、同定された汚染マーカーの１つにおいて、それぞれの汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有するサンプル中のあらゆるｃｆＤＮＡ断片を、汚染されたｃｆＤＮＡ断片として同定する。システムは、汚染されたｃｆＤＮＡ断片に基づいてサンプルの汚染レベルを推定する。本システムは、がん分類器の訓練に使用される訓練サンプルの両方に対してこの汚染検出を実施することができ

、また、がん分類器を展開する際のテストサンプルに対してもこの汚染検出を実施することができる。汚染マーカーには、複数の一塩基多型（ＳＮＰ）部位汚染マーカーおよび／またはインデル部位汚染マーカーが含まれる。いくつかの例において、複数のＳＮＰ部位混入マーカーは、閾値距離内にある少なくとも２つのＳＮＰ部位を含み、閾値頻度の範囲内にある集団ハプロタイプ頻度を有し、グアニン－アデニン多型および／またはシトシン－チミン多型を除外し、ハーディーワインベルク平衡を確保し、または上記のパラメータの任意の組み合わせを含む。いくつかの例では、インデル部位汚染マーカーは、閾値長さ内にあり、高い複雑性を有し、集団ハプロタイプ頻度が閾値頻度の範囲内にあり、ハーディーワインベルク平衡を保証する、または上記のパラメータの任意の組み合わせを有するインデル配列を含む。

【０００７】

第１の態様によれば、試験サンプル中のがんの存在を予測する方法が開示され、前記試験サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む前記試験サンプルを取得するステップと、前記試験サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、前記試験サンプルがホモ接合性ハプロタイプを有する、同定された１つ以上の汚染マーカーのそれぞれについて、同定された汚染マーカーの１つにおいて、それぞれの汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記試験サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、いずれかの同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記汚染レベルが閾値レベル以下であるか否かを判定するステップと、および前記汚染レベルが前記閾値レベル以下であると判定されたことに応答して、前記試験サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに対してがん分類を行い、がん予測を生成するステップを備える方法である。

【０００８】

第１の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。

【０００９】

第１の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。

【００１０】

第１の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００１１】

第１の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。

【００１２】

第１の態様の方法において、各複数のＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００１３】

第１の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。

【００１４】

第１の態様の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。

【００１５】

第１の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。

【００１６】

第１の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【００１７】
第１の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する。
【００１８】
第１の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【００１９】
第１の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。
【００２０】
第１の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【００２１】
第１の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【００２２】
第１の態様の方法において、前記汚染レベルを推定は、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記試験サンプルのシーケンス深度、前記試験サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる。
【００２３】
第１の態様の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、がんの分類を見送る。
【００２４】
第１の態様の方法において、前記がん予測は、がんと非がんとの間の二値予測を含む。
【００２５】
第１の態様の方法において、前記がん予測は、複数のがんタイプ間のマルチクラスがん予測を含む。
【００２６】
第１の態様の方法において、前記がんの分類を実行することは、さらに、前記試験サンプルのｃｆＤＮＡ断片の初期セットをｐ値フィルタリングでフィルタリングして異常断片のセットを生成するステップであって、前記フィルタリングは、前記初期セットから他の断片に関して閾値ｐ値を下回る断片を除去して前記異常断片のセットを生成するステップを備える。
【００２７】
第１の態様の方法において、前記分類モデルは、機械学習モデルである。
【００２８】
第２の態様によれば、コンピュータプロセッサによって実行されると、第１の態様の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体が開示される。
【００２９】
第３の態様によれば、コンピュータプロセッサ、および第２の態様の非一過性のコンピュータ可読記憶媒体を備えるシステムを開示する。
【００３０】
第４の態様によれば、試験サンプル中の疾患の存在を予測する方法が開示され、該方法は、前記試験サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む前記試験サンプルを得るステップと、前記試験サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、前記試験サンプルがホモ接合性ハプロタイプを有する１つ以上の汚染マーカーのそれぞれ

について、同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記試験サンプル中のあらゆる前記ｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、および前記汚染レベルが前記閾値レベル以下であると決定することに応答して、前記試験サンプル中のｃｆＤＮＡ断片の前記配列リードに対して疾患分類を行い、疾患予測を生成するステップと、を備える。

【００３１】

第４の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。

【００３２】

第４の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。

【００３３】

第４の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００３４】

第４の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。

【００３５】

第４の態様の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００３６】

第４の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。

【００３７】

第４の態様の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。

【００３８】

第４の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。

【００３９】

第４の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。

【００４０】

第４の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する。

【００４１】

第４の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００４２】

第４の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。

【００４３】

第４の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。

【００４４】

第４の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。

【００４５】
第４の態様の方法において、前記汚染レベルを推定することは、前記同定された汚染ｃｆＤＮＡ断片の数、前記試験サンプルのシーケンス深度、前記試験サンプル中のｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいてさらに行われる。
【００４６】
第４の態様の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、疾患分類を見送る。
【００４７】
第４の態様の方法において、前記疾患予測は、疾患と疾患なしとの間の二値予測を備える。
【００４８】
第４の態様の方法において、前記疾患予測は、複数の疾患間のマルチクラスがん予測からなることを特徴とする。
【００４９】
第４の態様の方法において、前記疾患分類を実行することは、前記試験サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに基づいて試験特徴ベクターを生成するステップ、および前記試験特徴ベクターを分類モデルに入力し、前記試験サンプルの前記疾患予測を生成するステップ、を備える。
【００５０】
第４の態様の方法において、前記疾患分類を実行することは、さらに、異常断片の集合を生成するためにｐ値フィルタリングを用いて前記試験サンプルのｃｆＤＮＡ断片の初期集合をフィルタリングするステップであって、前記フィルタリングは、前記異常断片の集合を生成するために、他の断片に関して閾値ｐ値を下回る断片を前記初期集合から除去するステップを備え、ここで、前記試験特徴ベクターは、前記異常断片の集合の前記配列リードに基づく。
【００５１】
第４の態様の方法において、前記分類モデルは、機械学習モデルである。
【００５２】
第５の態様によれば、コンピュータプロセッサによって実行されると、第４の態様の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体が開示される。
【００５３】
第６の態様によれば、コンピュータプロセッサ、および第５の態様の非一過性のコンピュータ可読記憶媒体を備えるシステムを開示する。
【００５４】
第７の態様によれば、試験サンプル中の汚染の存在を予測する方法が開示され、該方法は、前記試験サンプル中の複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片に由来する配列リードを取得するステップと、前記配列リードに基づいて、前記試験サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記試験サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、および前記汚染レベルが前記閾値レベル以下であると決定することに応答して、前記試験サンプルが汚染されていることを示す通知を生成するステップ、を備える。
【００５５】
第７の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。
【００５６】

第７の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。

【００５７】

第７の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００５８】

第７の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。

【００５９】

第７の態様の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００６０】

第７の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。

【００６１】

第７の態様の方法において、前記複数の汚染マーカーは、表１からの前記複数のＳＮＰ部位を含む。

【００６２】

第７の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。

【００６３】

第７の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。

【００６４】

第７の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する。

【００６５】

第７の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００６６】

第７の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。

【００６７】

第７の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。

【００６８】

第７の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。

【００６９】

第７の態様の方法において、前記汚染レベルの推定は、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記試験サンプルのシーケンス深度、前記試験サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる。

【００７０】

第８の態様によれば、コンピュータプロセッサによって実行されると、第７の態様の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体が開示される。

【００７１】

第９の態様によれば、コンピュータプロセッサおよび第８の態様の非一過性のコンピュータ可読記憶媒体を備えるシステムを開示する。

【００７２】
第１０の態様によれば、がん分類モデルを訓練する方法が開示され、該方法は、第１の訓練サンプルを含む複数の訓練サンプルを取得し、各訓練サンプルは複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片を含むステップと、各訓練サンプルについて、前記訓練サンプル中の前記ｃｆＤＮＡ断片に由来する配列リードを取得するステップと、前記第１の訓練サンプルについて、前記第１の訓練サンプルの前記配列リードに基づいて、前記第１の訓練サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから１つ以上の複数の汚染マーカーを同定するステップと、同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーの前記ホモ接合性ハプロタイプとは異なるハプロタイプを有する、前記第１の訓練サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、同定されたあらゆる汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、前記汚染レベルが閾値レベル以下であるか否かを決定するステップ、および前記第１の訓練サンプルの前記汚染レベルが前記閾値レベル以上であると決定することに応答して、前記複数の訓練サンプルから前記第１の訓練サンプルを除去するステップであって、ここで、前記第１の訓練サンプルを除く前記複数の訓練サンプルは、前記がん分類モデルを訓練して前記試験サンプルに対するがん予測を生成するために使用されるステップ、を備える。
【００７３】
第１０の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。
【００７４】
第１０の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。
【００７５】
第１０の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【００７６】
第１０の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。
【００７７】
第１０の態様の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【００７８】
第１０の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。
【００７９】
第１０の態様の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。
【００８０】
第１０の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。
【００８１】
第１０の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【００８２】
第１０の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【００８３】
第１０の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００８４】
第１０の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。
【００８５】
第１０の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【００８６】
第１０の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【００８７】
第１０の態様の方法において、前記複数の訓練サンプルは、非がんサンプルの第１のコホートおよびがんサンプルの第２のコホートを備え、前記がん分類モデルが、がんの存在の可能性を決定するように訓練される。
【００８８】
第１０の態様の方法において、前記がんサンプルの第２のコホートは、第１のがんタイプを有する１つ以上のサンプルおよび、第２のがんタイプを有する１つ以上の追加のサンプルを含み、ここで前記がん分類モデルは、前記第１のがんタイプの存在の第１の尤度および前記第２のがんタイプの存在の第２の尤度を決定するように訓練される。
【００８９】
第１０の態様の方法において、前記がん分類モデルは、機械学習モデルである。
【００９０】
第１０の態様の方法において、がん分類モデルは、決定木、ニューラルネットワーク、多層パーセプトロン、およびサポートベクターマシンのうちの少なくとも１つである。
【００９１】
第１１の態様によれば、コンピュータプロセッサによって実行されると、第１０の態様の方法をコンピュータプロセッサに実行させる命令を記憶した、非一過性のコンピュータ可読記憶媒体が開示される。
【００９２】
第１２の態様によれば、コンピュータプロセッサ、および第１１の態様の前記非一過性のコンピュータ可読記憶媒を備えるシステムを開示する。
【００９３】
第１３の態様によれば、訓練された前記がん分類モデルを記憶する前記非一過性のコンピュータ可読記憶媒体であって、前記コンピュータプログラム製品が、第１１の態様の方法によって作成される、非一過性のコンピュータ可読記憶媒体を備えるコンピュータプログラム製品を開示する。
【００９４】
第１４の態様によれば、個体からの遺伝物質を含む生物学的サンプルを保存するための１つ以上の収集容器、および複数の汚染マーカーを標的とする複数のプローブであって、前記複数のプローブは、表２および表４のうちの少なくとも１つを含む、複数のプローブを含む治療キットを開示する。
【００９５】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。
【００９６】
第１４の態様の治療キットにおいて、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。
【００９７】
第１４の態様の治療キットにおいて、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００９８】
第１４の態様の治療キットにおいて、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。
【００９９】
第１４の態様の治療キットにおいて、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【０１００】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。
【０１０１】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。
【０１０２】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。
【０１０３】
第１４の態様の治療キットにおいて、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【０１０４】
第１４の態様の治療キットにおいて、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【０１０５】
第１４の態様の治療キットにおいて、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【０１０６】
第１４の態様の治療キットにおいて、複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００のインデル部位を含む。
【０１０７】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【０１０８】
第１４の態様の治療キットにおいて、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【０１０９】
第１４の態様の治療キットにおいて、さらに、生物学的サンプル中の核酸断片を単離するための１つ以上の試薬を備える。
【０１１０】
第１４の態様の治療キットにおいて、さらに、第２の態様の非一過性のコンピュータ可読記憶媒体、第５の態様の非一過性のコンピュータ可読記憶媒体、第８の態様の非一過性のコンピュータ可読記憶媒体、および第１１の態様の非一過性のコンピュータ可読記憶媒体のうちの１つ以上を備える第１のコンピュータプログラム製品を備える。
【０１１１】
第１４の態様の治療キットは、さらに、第１３の態様のコンピュータプログラム製品を備える。
【図面の簡単な説明】
【０１１２】
【図１】１つ以上の実施形態による、サンプル中の汚染検出のプロセスを説明する例示的なフローチャートである。

【図２Ａ】１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための複数のＳＮＰ部位を同定するプロセスを説明する例示的なフローチャートである。
【図２Ｂ】１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための、インデル部位を同定するプロセスを説明する例示的なフローチャートである。
【図３Ａ】１つ以上の実施形態による、メチル化状態ベクターを得るためにセルフリー（ｃｆ）ＤＮＡの断片を配列決定するプロセスを説明する例示的なフローチャートである。
【図３Ｂ】１つ以上の実施形態による、メチル化状態ベクターを得るためにセルフリー（ｃｆ）ＤＮＡの断片を配列決定する図３Ａのプロセスの例示的な説明図である。
【図４Ａ】１つ以上の実施形態による、健常対照群のデータ構造を生成するプロセスを説明するフローチャートである。
【図４Ｂ】１つ以上の実施形態による、サンプルから異常にメチル化された断片を同定するプロセスを説明する例示的なフローチャートである。
【図５Ａ】１つ以上の実施形態による、がん分類器を訓練するプロセスを説明する例示的なフローチャートである。
【図５Ｂ】１つ以上の実施形態による、がん分類器の訓練に使用される特徴ベクターの生成例を示す図である。
【図６】図６Ａは、１つ以上の実施形態による、核酸サンプルを配列決定するための装置の例示的なフローチャートを示す。図６Ｂは、１つ以上の実施形態による、解析システムの例示的なブロック図である。
【図７】所与のサンプルについてホモ接合と判定された各汚染マーカーについて、基準または代替と判定された断片の総数に対する、汚染と判定された断片の数の割合の分布を示す図である。
【図８】例の結果の第１のグループによる、汚染マーカーの対立遺伝子頻度の散布図を示す。
【図９】例の結果の第１のグループによる、汚染マーカーの遺伝子型および接合性を示す２つのグラフを示す図である。
【図１０】第１のグループの例示的な結果に従って、ホモ接合体であり、所与のサンプルについて、そのサンプルの汚染を推定するために有用であると考えられるのに十分な断片が重なる汚染マーカーの割合を示すグラフである。
【図１１Ａ】第１の結果例に従った、異なるバッチのサンプルについての推定汚染レベルのグラフを示す。
【図１１Ｂ】第１の結果例に従った、サンプルの異なるバッチについての推定汚染レベルの追加グラフを示す。
【図１２】第２の結果例群による、表２および表４に列挙された汚染マーカーごとに得られたユニークなｃｆＤＮＡ断片の数の分布を示す図である。
【図１３】第２の結果例群による、品質チェックプロセスの結果を示す図である。
【図１４】第２の結果例群による、対立遺伝子頻度の散布図を示す。
【図１５】マーカーがヘテロ接合と呼ばれる頻度の散布図である。
【図１６】第２のグループの結果例による、ハーディーワインベルグ項の値の散布図を示す。
【図１７】第２のグループの結果例に従って、所与のサンプルについてホモ接合と呼ばれた各汚染マーカーについて、参照または代替と呼ばれた断片の総数に対する、汚染と分類された断片の数の割合の分布を示す図である。
【図１８】第２のグループの結果例に従って、仮想的な断片セットに対する汚染率を推定するための公式の適用を示す図である。
【図１９】第２のグループの結果例による、シミュレーションデータに汚染割合モデルを適用した結果を示す図である。
【図２０】第２の結果例群に従って、滴定レベルを変化させたｃｆＤＮＡサンプルの４つの滴定対に汚染分画モデルを適用した結果を示す図である。
【図２１】第２の結果例に従った、実験における８４個のｃｆＤＮＡサンプルについての

推定汚染割合の分布を示す図である。
【図２２】実験中の８４サンプルについて、複数のＳＮＰマーカーおよびインデルマーカーをそれ自体で別々に考慮した場合に得られた推定値が、第２のグループの例示的な結果に従って、高度に相関しており、有意な系統的偏りを示さないことを示す図である。
【図２３】１つ以上の実施形態による、複数のＳＮＰ汚染マーカーを含む表１を示す図である。
【図２４】１つ以上の実施形態による、複数のＳＮＰ汚染マーカーのプローブ配列リストを含む表２を示す。
【図２５】１つ以上の実施形態による、インデル汚染マーカーを含む表３を示す。
【図２６】１つ以上の実施形態による、インデル汚染マーカーのプローブ配列リストを含む表４を示す。
【０１１３】
図は、例示のみを目的とした様々な実施形態を描いている。当業者は、以下の考察から、本明細書に記載される原理から逸脱することなく、本明細書に例示される構造および方法の代替の実施形態が採用され得ることを容易に認識するであろう。
【０１１４】
[課題を解決するための手段]
［Ｉ．概要］
＜Ｉ．Ａ．メチル化の概要＞
本明細書に従って、個体からのｃｆＤＮＡ断片を、例えばメチル化されていないシトシンをウラシルに変換するなどの方法で処理し、塩基配列を決定し、配列リードを参照ゲノムと比較して、ＤＮＡ断片内の特定のＣｐＧ部位のメチル化状態を同定する。各ＣｐＧ部位は、メチル化されている場合とされていない場合がある。健常人と比較して異常なメチル化断片を同定することで、被験者のがんの状態を知ることができる。当技術分野でよく知られているように、ＤＮＡメチル化異常は（健常対照と比較して）さまざまな影響を引き起こし、それががんの一因となる可能性がある。異常にメチル化されたｃｆＤＮＡ断片の同定には様々な課題が生じる。第一に、ＤＮＡ断片が異常にメチル化されていると判定することは、対照群と比較して重要な意味を持つが、対照群の数が少なければ、対照群のサイズが小さいために統計的なばらつきが生じ、判定が信頼性を失う。さらに、対照群の中でもメチル化の状態は様々であり、被験者のＤＮＡ断片が異常にメチル化されていると判定する際に、これを考慮することは困難である。また、あるＣｐＧ部位のシトシンのメチル化は、その後に続くＣｐＧ部位のメチル化に影響を及ぼす可能性がある。この依存関係をカプセル化することは、それ自体が別の課題となりうる。
【０１１５】
メチル化は、デオキシリボ核酸（ＤＮＡ）において、シトシン塩基のピリミジン環上の水素原子がメチル基に変換され、５－メチルシトシンを形成するときに典型的に起こり得る。特に、メチル化は、本明細書で「ＣｐＧ部位（CPG site）」と呼ばれるシトシンとグアニンのジヌクレオチドで起こりうる。他の例では、メチル化は、ＣｐＧ部位の一部ではないシトシン、またはシトシンではない別のヌクレオチドで起こり得るが、これらはまれな現象である。本開示では、わかりやすくするために、メチル化はＣｐＧ部位に関連して議論される。異常なＤＮＡメチル化は、高メチル化または低メチル化として同定することができ、どちらもがんの状態を示す可能性がある。本開示を通して、ＤＮＡ断片が閾値以上のＣｐＧ部位を含み、それらのＣｐＧ部位の閾値以上の割合がメチル化または非メチル化である場合、ＤＮＡ断片について高メチル化および低メチル化を特徴付けることができる。
【０１１６】
本明細書に記載された原理は、非シトシンメチル化を含む非ＣｐＧコンテクストにおけるメチル化の検出にも同様に適用できる。そのような実施形態において、メチル化を検出するために使用される湿式実験室アッセイは、本明細書に記載されるものとは異なる場合がある。さらに、本明細書で議論されるメチル化状態ベクターは、一般的にメチル化が起

こった部位または起こっていない部位（それらの部位が特にＣｐＧ部位でなくても）である要素を含むことができる。この置換により、本明細書に記載される残りのプロセスは同じであり得、その結果、本明細書に記載される本発明の概念は、それらの他の形態のメチル化に適用され得る。

【０１１７】

＜Ｉ．Ｂ．定義＞

「細胞遊離核酸（cell free nucleic acid）」または「ｃｆＮＡ」という用語は、個体の体内（例えば血液）を循環する核酸断片であって、１つ以上の健常細胞から、および／または１つ以上の不健全細胞（例えばがん細胞）から起源する核酸断片を指す。「細胞遊離ＤＮＡ（cell free DNA）」または「ｃｆＤＮＡ」という用語は、個体の体内（例えば血液）を循環するデオキシリボ核酸断片を指す。さらに、ｃｆＮＡまたは個体の体内のｃｆＤＮＡは、ヒト以外の他の供給源に由来することもある。

【０１１８】

「ゲノム核酸（genomic nucleic acid）」、「ゲノムＤＮＡ（genomic DNA）」、または「ｇＤＮＡ」という用語は、１つ以上の細胞から得られる核酸分子またはデオキシリボ核酸分子を指す。様々な実施形態において、ｇＤＮＡは、健康な細胞（例えば、非腫瘍細胞）または腫瘍細胞（例えば、生検サンプル）から抽出され得る。いくつかの実施形態において、ｇＤＮＡは、白血球などの血球系由来の細胞から抽出され得る。

【０１１９】

「循環腫瘍ＤＮＡ（circulating tumor DNA）」または「ｃｔＤＮＡ」という用語は、腫瘍細胞または他のタイプのがん細胞に由来し、瀕死細胞のアポトーシスまたはネクローシスなどの生物学的プロセスの結果として、または生存腫瘍細胞によって能動的に放出される、個体の体液（例えば、血液、汗、尿、または唾液）中に放出され得る核酸断片を指す。

【０１２０】

「ＤＮＡ断片（DNA fragment）」、「断片（fragment）」、または「ＤＮＡ分子（DNA molecule）」という用語は、一般に、任意のデオキシリボ核酸断片、すなわち、ｃｆＤＮＡ、ｇＤＮＡ、ｃｔＤＮＡなどを指す場合がある。

【０１２１】

「異常な断片（anomalous fragment）」、「異常なメチル化断片（anomalously methylated fragment）」、または「異常なメチル化パターンを有する断片（fragment with an anomalous methylation pattern）」という用語は、ＣｐＧ部位の異常なメチル化を有する断片を指す。断片の異常なメチル化は、対照群における断片のメチル化パターンを観察することの意外性を同定するために、確率モデルを用いて決定することができる。

【０１２２】

「極端なメチル化を伴う異常な断片（unusual fragment with extreme methylation）」または「ＵＦＸＭ」という用語は、低メチル化断片または高メチル化断片を指す。低メチル化断片および高メチル化断片とは、それぞれメチル化または非メチル化の割合が閾値（例えば９０％）を超えるＣｐＧ部位を少なくともいくつか（例えば５つ）有する断片を指す。

【０１２３】

「アノマリスコア（anomaly score）」という用語は、そのＣｐＧ部位に重なるサンプルからの異常な断片（または、いくつかの実施形態ではＵＦＸＭ）の数に基づく、ＣｐＧ部位のスコアを指す。異常スコアは、分類のためのサンプルのフィーチャライズの文脈で使用される。

【０１２４】

本明細書で使用される場合、用語「約（about）」または「おおよそ（approximately）」は、当業者によって決定される特定の値について許容可能な誤差範囲内であることを意味し得、これは、値が測定または決定される方法、例えば測定システムの限界に部

分的に依存し得る。例えば、「約（about）」は、当業者の慣行に従って、１標準偏差以内または１標準偏差以上を意味することができる。「約（About）」は、所与の値の±２０％、±１０％、±５％、または±１％の範囲を意味することができる。用語「約（about）」または「おおよそ（approximately）」は、値の１桁以内、５倍以内、または２倍以内を意味し得る。本願明細書および特許請求の範囲に特定の値が記載されている場合、特に断りのない限り、特定の値の許容誤差範囲内を意味する「約（about）」という用語が想定されるべきである。用語「約（about）」は、当業者が一般的に理解する意味を有することができる。用語「約（about）」は、±１０％を指すことができる。「約（about）」という用語は、±５％を指すことができる。

【０１２５】

本明細書で使用される場合、「生物学的サンプル（biological sample）」、「患者サンプル（patient sample）」、または「サンプル（sample）」という用語は、被験体から採取され、被験体に関連する生物学的状態を反映することができ、セルフリーＤＮＡを含む任意のサンプルを指す。生物学的サンプルの例としては、被験体の血液、全血、血漿、血清、尿、脳脊髄液、糞便、唾液、汗、涙、胸水、心嚢液、または腹膜液が挙げられるが、これらに限定されない。生物学的サンプルには、生死を問わず、被験体由来のあらゆる組織または物質が含まれ得る。生物学的サンプルは、セルフリーサンプルであり得る。生物学的サンプルは、核酸（例えば、ＤＮＡまたはＲＮＡ）またはその断片を含むことができる。核酸（nucleic acid）」という用語は、デオキシリボ核酸（ＤＮＡ）、リボ核酸（ＲＮＡ）、またはそれらのハイブリッドもしくは断片を指すことができる。サンプル中の核酸は、セルフリー核酸であり得る。サンプルは、液体サンプルまたは固体サンプル（例えば、細胞サンプルまたは組織サンプル）であり得る。生物学的サンプルは、体液、例えば、血液、血漿、血清、尿、膣液、（例えば、精巣の）水腫からの液、膣洗浄液、胸水、腹水、脳脊髄液、唾液、汗、涙、喀痰、気管支肺胞洗浄液、乳頭からの排出液、身体の異なる部位（例えば、甲状腺、乳房）からの吸引液などであることができる。生物学的サンプルは便サンプルでもよい。様々な実施形態において、セルフリーＤＮＡについて濃縮された生物学的サンプル（例えば、遠心分離プロトコルを介して得られた血漿サンプル）中のＤＮＡの大部分は、セルフリーであり得る（例えば、ＤＮＡの５０％、６０％、７０％、８０％、９０％、９５％、または９９％より多くがセルフリーであり得る）。生物学的サンプルは、組織または細胞の構造を物理的に破壊するために処理され（例えば、遠心分離および／または細胞溶解）、その結果、細胞内成分を溶液中に放出することができる。この溶液は、さらに酵素、緩衝液、塩、洗浄剤などを含むことができ、この酵素、緩衝液、塩、洗浄剤などは、解析のためにサンプルを調製するために使用することができる。

【０１２６】

本明細書で使用される場合、「対照（control）」、「対照サンプル（control sample）」、「参照（reference）」、「参照サンプル（reference sample）」、「正常（normal）」、および「正常サンプル（normal sample）」という用語は、特定の状態を有しないか、そうでなければ健康である被験体からのサンプルを表す。一例において、本明細書に開示されるような方法は、腫瘍を有する被験体に対して実施され得、ここで、参照サンプルは、被験体の健康な組織から採取されたサンプルである。参照サンプルは、対象から取得することも、データベースから取得することもできる。参照は、例えば、被験体からのサンプルの配列決定から得られた核酸断片配列をマッピングするために使用される参照ゲノムであり得る。参照ゲノムは、生物学的サンプルおよび体質サンプルからの核酸断片配列が整列および比較され得る、ハプロイドまたは二倍体ゲノムを指し得る。体質サンプルの例としては、被験者から得られた白血球のＤＮＡを挙げることができる。ハプロイドゲノムの場合、各遺伝子座には１つのヌクレオチドしか存在しない。２倍体ゲノムの場合、ヘテロ接合性遺伝子座を同定することができる。各ヘテロ接合性遺伝子座は２つの対立遺伝子を持つことができ、どちらの対立遺伝子もその遺伝子座へのアライメントに一致させることができる。

【０１２７】
本明細書において、「がん（cancer）」または「腫瘍（tumor）」という用語は、組織の異常な腫瘤を指し、その腫瘤の成長は正常組織の成長を上回り、正常組織の成長と協調していない。
【０１２８】
本明細書で使用される場合、「健康な（healthy）」という語句は、良好な健康状態を有する被験体を指す。健康な対象は、悪性疾患または非悪性疾患がないことを示すことができる。「健康な個体（healthy individual）」は、アッセイされる状態とは無関係の、通常「健康（healthy）」とはみなされない他の疾患または状態を有することができる。
【０１２９】
本明細書において、「メチル化（methylation）」という用語は、シトシン塩基のピリミジン環上の水素原子がメチル基に変換され、５－メチルシトシンを形成するデオキシリボ核酸（ＤＮＡ）の修飾を指す。特に、メチル化は、本明細書で「ＣｐＧ部位（CpG sites）」と呼ばれるシトシンとグアニンのジヌクレオチドで起こる傾向がある。ＣｐＧ部位以外のシトシンやシトシン以外のヌクレオチドでメチル化が起こることもあるが、これはまれである。ｃｆＤＮＡのメチル化異常は、高メチル化または低メチル化として同定することができ、どちらもがんの状態を示す可能性がある。ＤＮＡメチル化異常は（健常対照と比較して）さまざまな影響を引き起こす可能性があり、がんの一因となる可能性がある。本明細書に記載された原理は、ＣｐＧコンテクストにおけるメチル化の検出にも、非シトシンメチル化を含む非ＣｐＧコンテクストにおけるメチル化の検出にも同様に適用可能である。さらに、メチル化状態ベクターは、一般にメチル化が起こった部位または起こっていない部位のベクターである要素を含むことができる（それらの部位が特にＣｐＧ部位でなくても）。
【０１３０】
本明細書で互換的に使用されるように、用語「メチル化断片（methylation fragment）」または「核酸メチル化断片（nucleic acid methylation fragment）」は、核酸（例えば、核酸分子および／または核酸断片）のメチル化配列決定によって決定される、複数のＣｐＧ部位における各ＣｐＧ部位のメチル化状態の配列を指す。メチル化断片において、核酸断片中の各ＣｐＧ部位の位置およびメチル化状態は、参照ゲノムに対する配列リード（例えば、核酸の配列決定から得られる）のアライメントに基づいて決定される。核酸メチル化断片は、複数のＣｐＧ部位の各ＣｐＧ部位のメチル化状態（例えば、メチル化状態ベクター）からなり、参照ゲノムにおける核酸断片の位置（例えば、ＣｐＧインデックス、または別の類似の指標を用いて核酸断片中の最初のＣｐＧ部位の位置によって指定される）および核酸断片中のＣｐＧ部位の数を特定する。核酸分子のメチル化配列決定に基づく、参照ゲノムに対する配列リードのアライメントは、ＣｐＧインデックスを用いて実施することができる。本明細書で使用する場合、「ＣｐＧインデックス（CpG index）」という用語は、ヒト参照ゲノムなどの参照ゲノムにおける複数のＣｐＧ部位（例えば、ＣｐＧ　１、ＣｐＧ　２、ＣｐＧ　３など）の各ＣｐＧ部位のリストを意味し、これは電子フォーマットであり得る。ＣｐＧインデックスはさらに、ＣｐＧインデックスにおけるそれぞれのＣｐＧ部位について、対応する参照ゲノムにおける、対応するゲノム上の位置を含んでいる。このようにして、それぞれの核酸メチル化断片中のそれぞれのＣｐＧ部位は、それぞれの参照ゲノム中の特定の位置にインデックス付けされ、ＣｐＧインデックスを用いて決定することができる。
【０１３１】
本明細書で使用される場合、「真陽性（true positive）」（ＴＰ）という用語は、状態を有する被験体を指す。「真陽性（True positive）」は、腫瘍、がん、前がん状態（例えば、前がん病変）、限局性もしくは転移性がん、または非悪性疾患を有する被験体を指すことができる。「真陽性（True positive）」は、状態を有する被験体を指し得、本開示のアッセイまたは方法によって状態を有することが同定される。本明細書で使

用される場合、「真陰性（True negative）」（ＴＮ）という用語は、病態を有さないか、または検出可能な病態を有さない被験体を指す。真陰性は、腫瘍、がん、前がん状態（例えば、前がん病変）、限局性または転移性がん、非悪性疾患、またはその他の健康な被験体などの疾患または検出可能な疾患を有さない被験体を指すことができる。真陰性は、病態を有さないか、検出可能な病態を有さないか、または本開示のアッセイもしくは方法によって病態を有さないと同定される被験体を指すことができる。

【０１３２】

本明細書で使用される場合、「参照ゲノム（reference genome）」という用語は、被験体から同定された配列を参照するために使用され得る、任意の生物またはウイルスの、部分的であるか完全であるかにかかわらず、任意の特定の既知の、配列決定された、または特徴付けられたゲノムを指す。ヒト被験体および他の多くの生物に使用される例示的な参照ゲノムは、アメリカ国立生物工学情報センター（「ＮＣＢＩ」）またはカリフォルニア大学サンタクルーズ校（「ＵＣＳＣ」）がホストするオンラインゲノムブラウザで提供されている。「ゲノム（genome）」とは、核酸配列で表される生物またはウイルスの完全な遺伝情報を指す。本明細書で使用される場合、参照配列または参照ゲノムは、多くの場合、個体または複数の個体からアセンブルされた、または部分的にアセンブルされたゲノム配列である。いくつかの実施形態において、参照ゲノムは、１つ以上のヒト個体からのアセンブルされた、または部分的にアセンブルされたゲノム配列である。参照ゲノムは、ある種の遺伝子セットの代表例とみなすことができる。いくつかの実施形態では、参照ゲノムは染色体に割り当てられた配列からなる。例示的なヒト参照ゲノムとしては、ＮＣＢＩビルド３４（ＵＣＳＣ相当：ｈｇ１６）、ＮＣＢＩビルド３５（ＵＣＳＣ相当：ｈｇ１７）、ＮＣＢＩビルド３６．１（ＵＣＳＣ相当：ｈｇ１８）、ＧＲＣｈ３７（ＵＣＳＣ相当：ｈｇ１９）、およびＧＲＣｈ３８（ＵＣＳＣ相当：ｈｇ３８）が挙げられるが、これらに限定されない。

【０１３３】

本明細書で使用される場合、「配列リード（sequence reads）」または「リード（reads）」という用語は、本明細書に記載されるまたは当該技術分野で公知の任意の配列決定プロセスによって生成されるヌクレオチド配列を指す。リードは、核酸断片の一端から生成することができ（「シングルエンドリード（single-end reads）」）、核酸の両端から生成することもある（例えば、ペアエンドリード（paired-end reads）、ダブルエンドリード（double-end reads））。一部の実施形態では、配列リード（例えば、シングルエンドリードまたはペアエンドリード）は、標的核酸断片の片鎖または両鎖から生成することができる。配列リードの長さは、特定の配列決定技術に関連することが多い。例えば、ハイスループットメソッドは、数十から数百塩基対（ｂｐ）の大きさで変化し得る配列リードを提供する。いくつかの実施形態において、配列リードは、約１５ｂｐ～９００ｂｐの長さの平均、中央値、または平均の長さである（例えば約２０ｂｐ、約２５ｂｐ、約３０ｂｐ、約３５ｂｐ、約４０ｂｐ、約４５ｂｐ、約５０ｂｐ、約５５ｂｐ、約６０ｂｐ、約６５ｂｐ、約７０ｂｐ、約７５ｂｐ、約８０ｂｐ、約８５ｂｐ、約９０ｂｐ、約９５ｂｐ、約１００ｂｐ、約１１０ｂｐ、約１２０ｂｐ、約１３０、約１４０ｂｐ、約１５０ｂｐ、約２００ｂｐ、約４５０ｂｐ、約３００ｂｐ、約３５０ｂｐ、約４００ｂｐ、約４５０ｂｐ、または約５００ｂｐである）。いくつかの実施形態では、配列リードは、約１０００ｂｐ、２０００ｂｐ、５０００ｂｐ、１０，０００ｂｐ、または５０，０００ｂｐ以上の平均、中央値、または平均長である。例えば、ナノポアシークエンシングは、数十から数百から数千塩基対のサイズで変化し得る配列リードを提供することができる。イルミナのパラレル配列では、それほど変化しない配列リードを提供することができ、例えば、配列リードのほとんどは２００ｂｐより小さくなり得る。配列リード（sequence read）（または配列リード（sequencing read））は、核酸分子（例えば、ヌクレオチドの文字列）に対応する配列情報を指すことができる。例えば、配列リードは、核酸断片の一部からのヌクレオチドの文字列（例えば、約２０～約１５０）に対応し得、核酸断片の一端または両端のヌクレオチドの文字列に対応し得、または核酸断片全体

のヌクレオチドに対応し得る。配列読み取りは、様々な方法、例えば、配列決定技術、またはプローブ、例えば、ハイブリダイゼーションアレイまたはキャプチャープローブ、または増幅技術、例えば、ポリメラーゼ連鎖反応（ＰＣＲ）、または単一プライマーを用いる線形増幅、または等温増幅を用いて得ることができる。

【０１３４】

本明細書で使用される「配列決定（sequencing）」などの用語は、一般に、核酸またはタンパク質などの生物学的高分子の順序を決定するために使用され得るあらゆる生化学的プロセスを指す。例えば、配列決定データは、ＤＮＡ断片などの核酸分子中のヌクレオチド塩基の全てまたは一部を含み得る。

【０１３５】

本明細書で使用する場合、「シーケンス深度（sequencing depth）」という用語は、「カバレッジ（coverage）」という用語と互換的に使用され、遺伝子座にアライメントされたユニークな核酸標的分子に対応するコンセンサス配列リードによって遺伝子座がカバーされる回数を指し、例えば、シーケンス深度は、遺伝子座をカバーするユニークな核酸標的分子の数に等しい。遺伝子座は、ヌクレオチドのように小さくても、染色体アームのように大きくても、ゲノム全体のように大きくてもよい。シーケンス深度は、「Ｙｘ」、例えば、５０ｘ、１００ｘなどで表すことができ、ここで「Ｙ」は、核酸標的に対応する配列で遺伝子座がカバーされる回数、例えば、特定の遺伝子座をカバーする独立した配列情報が得られる回数を指す。一部の実施形態では、シーケンス深度は、配列決定されたゲノムの数に対応する。シーケンス深度はまた、複数の遺伝子座、または全ゲノムに適用され得、この場合、Ｙは、それぞれ、遺伝子座、ハプロイドゲノム、または全ゲノムが配列決定された平均または平均回数を指し得る。平均深度が引用されている場合、データセットに含まれる異なる遺伝子座の実際の深度は、値の範囲に及ぶ可能性がある。超深度シーケンスとは、遺伝子座におけるシーケンス深度が少なくとも１００倍であることを指す。

【０１３６】

本明細書で使用される場合、「感度（sensitivity）」または「真陽性率（true positive rate）」（ＴＰＲ）という用語は、真陽性の数を真陽性と偽陰性の数の合計で割ったものを指す。感度は、ある状態を真に有する集団の割合を正しく同定するアッセイまたは方法の能力を特徴付けることができる。例えば、感度は、集団内のがんを有する被験者の数を正しく同定する方法の能力を特徴付けることができる。別の例では、感度は、がんを示す１つ以上のマーカーを正しく同定する方法の能力を特徴付けることができる。

【０１３７】

本明細書で使用される場合、「特異度（specificity）」または「真陰性率（true negative rate）」（ＴＮＲ）という用語は、真陰性の数を真陰性と偽陽性の数の合計で割ったものを指す。特異性は、ある状態を真に有していない集団の割合を正しく同定するアッセイまたは方法の能力を特徴付けることができる。例えば、特異性は、ある集団の中でがんでない被験者の数を正しく同定する方法の能力を特徴付けることができる。別の例では、特異性は、がんを示す１つ以上のマーカーを正しく同定する方法の能力を特徴付ける。

【０１３８】

本明細書で使用する場合、「被験体（subject）」という用語は、ヒト（例えば、男性ヒト、女性ヒト、胎児、妊娠中の女性、子供など）、非ヒト動物、植物、細菌、真菌または原生生物を含むがこれらに限定されない、任意の生物または非生物を指す。ヒトまたは非ヒト動物であれば、哺乳動物、爬虫類、鳥類、両生類、魚類、偶蹄類、反芻動物、ウシ科（例えば、ウシ）、ウマ科（例えば、ウマ）、ヤギおよびヒツジ（例えば、ヒツジ、ヤギ）、ブタ（例えば、ブタ）、ラクダ科（例えば、ラクダ、ラマ、アルパカ）、サル、類人猿（例えば、ゴリラ、チンパンジー）、クマ科（例えば、クマ）、家禽、イヌ、ネコ、マウス、ラット、魚、イルカ、クジラ、およびサメが被験体になることができる。いくつかの実施形態において、被験体は、あらゆる段階の男性または女性（例えば、男性、女性

または子供）である。サンプルが採取される、または本明細書に記載される方法もしくは組成物のいずれかによって処置される対象は、任意の年齢であり得、成人、幼児または子供であり得る。

【０１３９】

本明細書で使用する場合、「組織（tissue）」という用語は、機能的な単位として一緒になっている細胞群に対応することができる。一つの組織には複数の種類の細胞が存在し得る。異なる種類の組織は、異なる種類の細胞（例えば、肝細胞、肺胞細胞または血球）から構成されることがあるが、異なる生物（母親対胎児）の組織、または健康な細胞対腫瘍細胞に対応することもある。「組織（tissue）」という用語は、一般に、ヒトの体内で見出されるあらゆる細胞群（例えば、心臓組織、肺組織、腎臓組織、鼻咽頭組織、口腔咽頭組織）を指すことができる。いくつかの態様において、「組織（tissue）」または「組織型（tissue type）」という用語は、セルフリー核酸が由来する組織を指すために使用することができる。一例では、ウイルス核酸断片は血液組織に由来し得る。別の例では、ウイルス核酸断片は腫瘍組織に由来し得る。

【０１４０】

本明細書で使用する場合、「ゲノムの（genomic）」という用語は、生物のゲノムの特性を指す。ゲノム特性の例としては、ゲノムの全部または一部の一次核酸配列（例えば、ヌクレオチド多型、インデル、配列再配列、変異頻度などの有無）、ゲノム内の１つ以上の特定のヌクレオチド配列のコピー数（例えば、コピー数、対立遺伝子頻度分率、単一染色体または全ゲノムの倍数性など）、ゲノムの全部または一部のエピジェネティック状態（例えば、メチル化などの共有結合核酸修飾、ヒストン修飾、ヌクレオソームの位置決めなど）、生物のゲノムの発現プロファイル（例えば、遺伝子発現レベル、アイソタイプ発現レベル、遺伝子発現比など）に関するものが挙げられるが、これらに限定されない。

【０１４１】

本明細書で使用する用語は、特定の場合を説明するためのものであり、限定を意図するものではない。本明細書で使用される場合、単数形「a」、「an」および「the」は、文脈上明らかにそうでないことが示されない限り、複数形も含むことが意図される。さらに、「含む（including）」、「含む（includes）」、「有する（having）」、「有する（has）」、「有する（with）」、またはそれらの変形が、詳細な説明および／または特許請求の範囲のいずれかで使用される限り、そのような用語は、用語「含む（comprising）」と同様の方法で包括的であることが意図される。

【０１４２】

＜Ｉ．Ｃ．がん分類の概要＞

一般に、がん分類では、生物学的サンプルが配列決定され、生物学的サンプル中の遺伝物質の配列リードからがんを予測するために解析される。ワークフローは、例えば、医療従事者、配列決定装置、解析システムなどを含む１つ以上の実在物による作用を含むことができる。ワークフローの目的には、個人のがんの検出および／またはモニタリングが含まれる。ヘルスケアの観点からは、ワークフローは他の既存のがん診断ツールを補完する役割を果たすことができる。ワークフローは、がんと診断された個人の治療計画をより良く知らせるために、がんの早期発見および／または日常的ながんモニタリングを提供する役割を果たすことができる。一般的なワークフローは、代替的に、より一般的に疾患分類に適用することができる。

【０１４３】

医療従事者がサンプル採取を行う。がん分類を受けようとする個人が医療機関を訪れる。医療従事者は、がん分類を実施するためのサンプルを採取する。生物学的サンプルの例としては、被験者の組織生検、血液、全血、血漿、血清、尿、脳脊髄液、糞便、唾液、汗、涙、胸水、心嚢液、または腹膜液が挙げられるが、これらに限定されない。サンプルには、その個人に属する遺伝物質が含まれ、がん分類のために抽出され、塩基配列が決定される場合がある。サンプルが採取されると、サンプルは配列決定装置に提供される。サンプルとともに、医療従事者は、個人に関連する他の情報、例えば、生物学的性別、年齢、

民族性、喫煙状態、過去の診断などを収集することができる。１つ以上の実施形態において、医療従事者は、治療キットを利用することができる。治療キットは、１つ以上のサンプル採取容器を含むことができる。治療キットは、試料の処理および解析に使用するための試薬、プローブ、コンピュータプログラム製品、指示書などをさらに含むことができる。

【０１４４】

シークエンシングデバイスは、サンプルに対してサンプルシークエンシングを実行する。研究室の臨床医は、シーケンシングの準備として、サンプルに対して１つ以上の処理ステップを実行することができる。臨床医は、試薬、プローブなどを含む処理キットを利用することもできる。準備完了後、臨床医はサンプルを配列装置にロードする。シーケンシングに使用される装置の例を、図６Ａおよび図６Ｂに関連してさらに説明する。配列決定装置は一般に、断片に対応する核酸塩基の配列を決定するために配列決定される核酸の断片を抽出および単離する。配列決定はまた、核物質の増幅を含み得る。様々な配列決定プロセスには、サンガー配列決定、断片解析、および次世代配列決定が含まれる。配列決定は、全ゲノム配列決定または標的パネルを用いた標的配列決定の場合がある。ＤＮＡメチル化の文脈では、バイサルファイト配列決定（例えば、図３Ａおよび図３Ｂにさらに記載）は、ＣｐＧ部位における非メチル化シトシンのバイサルファイト変換を通じてメチル化の状態を決定することができる。サンプル配列決定により、サンプル中の複数の核酸断片の配列が得られる。１つ以上の実施形態において、配列は、メチル化状態ベクターを含み得、各メチル化状態ベクターは、断片上のＣｐＧ部位についてのメチル化状態を記載する。

【０１４５】

次いで、解析システムが配列リードを処理してがん予測を生成する。

【０１４６】

解析システムは、解析前処理を実行することができる。解析前処理には、配列リードの重複除去、カバレッジに関する測定基準の決定、サンプルが汚染されているかどうかの決定、汚染断片の除去、配列決定エラーの呼び出しなどが含まれるが、これらに限定されない。汚染されていると判定されたサンプルは、さらなる解析から差し控えられる場合がある。言い換えれば、解析システムは、汚染されたサンプルに対してさらなる解析（疾患またはがんの分類など）を行うことを差し控えることができる。一部の実施形態では、汚染サンプルは物理的に廃棄される場合がある。

【０１４７】

解析システムは１つ以上の解析を実行する。解析は、統計解析または１つ以上の訓練されたモデルの適用であり、サンプルが由来する個体の少なくともがんの状態を予測する。ＣｐＧ部位のメチル化、一塩基多型（ＳＮＰ）、挿入または欠失（インデル）、他のタイプの遺伝子変異など、様々な遺伝的特徴が評価および考慮され得る。メチル化の文脈では、解析は、異常メチル化同定（例えば、図４Ａおよび図４Ｂにさらに記載）、特徴抽出（例えば、図５Ａおよび図５Ｂにさらに記載）、およびがん予測を決定するためのがん分類器の適用（例えば、図５Ａおよび図５Ｂにさらに記載）を含み得る。がん分類器は、がん予測を決定するために抽出された特徴を入力する。がん予測はラベルまたは値となる。ラベルは特定のがん状態を示すことがあり、例えば、二値ラベルはがんの有無を示すことができ、多クラスラベルはスクリーニングされる複数のがん種から１つ以上のがん種を示すことができる。値は、特定のがん状態の可能性、例えばがんの可能性、および／または特定のがん種の可能性を示すことができる。予測値はさらに、がんシグナルの定量化を示す場合があり、これには１つ以上の特定の起源組織のシグナルの定量化が含まれる場合がある。

【０１４８】

解析システムは医療従事者に予測を返す。医療従事者は、がん予測に基づいて治療計画を確立または調整することができる。治療の最適化についてはセクションＶＩ．Ｃ．治療でさらに説明する。

【０１４９】
［ＩＩ．サンプル処理］
＜ＩＩ．Ａ．汚染検出＞
図１は、１つ以上の実施形態による、サンプル中の汚染検出プロセス１００を説明する例示的なフローチャートである。一般に、サンプルは、健康であるか、がんであることが知られているか、がんであることが疑われているか、または事前情報が知られていない個人からのものであってよい。サンプルは、血液、血漿、血清、尿、糞便、および唾液サンプルからなる群から選択され得る。あるいは、サンプルは、全血、血液分画（例えば、白血球（ＷＢＣ））、組織生検、胸水、心嚢液、脳脊髄液、および腹膜液からなる群から選択することもできる。サンプルを用いたがん分類または他の分類／解析に先立ち、汚染検出プロセス１００は試料の汚染レベルを決定する。例えば、プロセス１００は、規則（例えば、閾値汚染レベルまたは間隔）と比較されたときに、サンプルが汚染されているかどうかを決定する汚染推定値および／または信頼区間を出力することができる。汚染検出は、個体のホモ接合性対立遺伝子とは異なる対立遺伝子を有する断片を同定するための汚染マーカーとして、一組の遺伝子配列を利用する。本明細書において、プロセス１００は、解析システム（その例は、図６Ａおよび図６Ｂならびに対応する説明に提供される）によって実行されるものとして記載されるが、ステップの一部または全部は、他の比較可能に記載された配列決定装置および／またはコンピュータプロセッサによって実行されてもよい。

【０１５０】
解析システムは、汚染マーカープローブを含む標的パネルで、サンプル中のｃｆＤＮＡ断片を配列決定する。汚染マーカーは、ヒトゲノム中の遺伝子配列であり、挿入－欠失（インデル）部位および複数の一塩基多型（ＳＮＰ）部位を含み得る。いくつかの実施形態において、汚染マーカーは、表１に列挙された複数のＳＮＰ部位および表３に列挙されたインデル部位の任意の組み合わせからなる。例えば、汚染マーカーは、表１に列挙された複数のＳＮＰ部位のうち少なくとも１００、２００、３００、４００、５００、６００、７００、８００、９００、または１，０００部位および、表３に列挙されたインデル部位のうち少なくとも１００、２００、３００、４００、５００、６００、７００、８００、９００、または１，０００部位を含む。汚染マーカープローブは、ＤＮＡの一本鎖上の汚染マーカー配列またはＤＮＡの両鎖上の汚染マーカー配列を標的とするように設計することができる。メチル化バイサルファイト配列決定の文脈では、プローブは、高メチル化断片、低メチル化断片、またはその両方を標的とするように設計され得る。プローブはまた、汚染マーカー配列の１つ、いくつか、または全てのハプロタイプを標的とするように設計され得る。例えば、２つのＳＮＰ部位からなる複数のＳＮＰ部位（最大４つの潜在的ハプロタイプ）の場合、プローブは、１つ、いくつか、または全てのハプロタイプを標的とし得る。汚染マーカーの潜在的なハプロタイプの全てを標的とするプローブを設計することにより、いずれか１つの特定のハプロタイプへの参照の偏りを回避することができる。いくつかの実施形態において、汚染マーカー用に設計されたプローブは、表２に列挙された複数のＳＮＰ部位用のプローブおよび表４に列挙されたインデル部位用のプローブの任意の組み合わせからなる。汚染マーカーの選択およびその後のプローブ設計については、図２Ａおよび図２Ｂで後述する。配列決定の結果、解析システムはサンプル中のｃｆＤＮＡ断片の配列リードを有する。

【０１５１】
解析システムは、試料がホモ接合性ハプロタイプを有する複数の汚染マーカーから１２０個の１つ以上の汚染マーカーを同定する。各汚染マーカーにおけるサンプルのハプロタイプを決定するために、解析システムは、その汚染マーカーにおけるｃｆＤＮＡ断片の全配列リードのハプロタイプを評価する。配列リードのハプロタイプは、汚染マーカー部位に対して設計された全てのプローブにリードを再整列させ、最も良好な整列を有するプローブに対応する対立遺伝子を特定することにより、より正確に決定することができる。解析システムが十分な数のリード、例えば少なくとも１０、２０、３０、４０、５０を観察

し、１つのハプロタイプを有する配列リードの割合が非常に高い、例えば９０％、９１％、９２％、９３％、９４％、９５％、９６％、９７％、９８％、または９９％を超える場合、解析システムは、サンプルがそのハプロタイプについてホモ接合体である（すなわち、サンプル中の汚染マーカーのコピーが両方とも同じハプロタイプである）と判定する。高いパーセンテージを超えない汚染マーカーは、ホモ接合体ではないと判定され、むしろヘテロ接合体（すなわち、サンプル中の汚染マーカーの２つのコピーは、異なるハプロタイプである）と判定される場合がある。

【０１５２】

解析システムは、同定された汚染マーカーの一つにおいて、それぞれの汚染マーカーのホモ接合体ハプロタイプとは異なるハプロタイプを有するあらゆるｃｆＤＮＡ断片を汚染断片として同定する。同定された汚染マーカーごとに、解析システムは、同定された汚染マーカーにおいてサンプルのホモ接合性ハプロタイプとは異なるハプロタイプを有する場合、ｃｆＤＮＡ断片を汚染（「汚染断片（複数可）（contaminated fragment(s)）」とも呼ばれる）として標識またはその他の方法で同定する（ステップ１３０）。利用される複数の汚染マーカーのうち、任意のサンプルは、最初の複数のサブセットについてホモ接合のハプロタイプを有する可能性が高い。

【０１５３】

解析システムは、特定された汚染断片に基づいて、汚染レベルを推定する（ステップ１４０）。例えば、同定された汚染断片がない場合、汚染レベルは０または０％となる。解析システムは、サンプルのシーケンス深度、サンプル中のｃｆＤＮＡ断片の総数、実施された汚染マーカーの数、同定された汚染断片の総数、またはそれらの組み合わせに基づいて、汚染レベルをさらに推定することができる。いくつかの例では、解析システムは、ある信頼区間内の汚染を推定するために、異なるハプロタイプを有するｃｆＤＮＡ断片をカウントする。

【０１５４】

解析システムは、汚染レベルを閾値と比較することにより、サンプルが汚染されているか否かを判断する（ステップ１５０）。いくつかの例では、解析システムは推定汚染レベルを閾値汚染レベル、限界値、または間隔と比較する。閾値汚染レベルまたは間隔は、用途に応じて調整または変化させることができる。例えば、機密性の高いアプリケーションでは、非常に低い汚染閾値が必要となる場合がある。単なる例として、閾値汚染レベルは０．１～１．０％または０．０１～１．０％の間の値または間隔とすることができる。いくつかの例では、閾値汚染限界は０．１％である。いくつかの例では、閾値汚染限界は０．０１％である。一部の実施例では、サンプルの推定汚染レベルが閾値汚染限界値を超える場合、解析システムは、がん分類結果の生成を拒否し（例えば、サンプルをがんまたは非がんと呼ぶことを拒否する）、分類器を訓練するための訓練セットにサンプルを含めることを拒否し（例えば、がん／非がんまたはがんの様々な起源組織を呼ぶための二値またはマルチクラス分類器）、および／または汚染サンプルに基づく疾患または非疾患分類結果を防止することができる。追加的および／または代替的に、解析システムは比較を使用して、サンプルが汚染されているか否かを判定する。例えば、サンプルが閾値未満である場合、解析システムはサンプルが汚染されていないと判定することができる。逆に、試料が閾値以上である場合、解析システムは試料が汚染されていると判定することができる。

【０１５５】

図２Ａは、１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための複数のＳＮＰ部位を同定するプロセス２００を説明する例示的なフローチャートである。プロセス２００は、本明細書では、解析システムによって実行されるものとして記載されるが、ステップの一部または全部は、他の比較可能に記載された配列決定デバイスおよび／またはコンピュータプロセッサによって実行されてもよい。

【０１５６】

解析システムは、閾値距離内にある複数のＳＮＰ部位を特定する（ステップ２０５）。複数のＳＮＰ部位は、少なくとも２つのＳＮＰ部位を含む遺伝子配列である。１つ以上の

実施形態において、複数のＳＮＰ部位は、２、３、４、または５つのＳＮＰ部位を含む。解析システムは、閾値距離、例えば、５塩基対（ｂｐ）、１０ｂｐ、１５ｂｐ、２０ｂｐ、または２５ｂｐを設定する。ＳＮＰ部位が近ければ近いほど、単一断片上に存在する確率が高くなるが、閾値距離が小さいほど、実行可能な複数のＳＮＰ部位の数も制限される。解析システムは、上記の考慮事項および／またはターゲットアッセイパネルに含めることができる汚染マーカープローブの予算を考慮して、閾値距離を調整することができる。

【０１５７】

解析システムは、閾値範囲内の集団ハプロタイプ頻度を有するハプロタイプを有する複数のＳＮＰ部位を含む（ステップ２１０）。２－ＳＮＰ部位では、それぞれ２つのバリアント対立遺伝子を有する２つのＳＮＰが存在する。つまり、両方のＳＮＰ部位が置換を持たない［０、０］の第１のハプロタイプ、両方のＳＮＰ部位が置換を持つ［１、１］の第２のハプロタイプ、第２のＳＮＰ部位だけが置換を持つ［０、１］の第３のハプロタイプ、および第１のＳＮＰ部位だけが置換を持つ［１、０］の第４のハプロタイプである。解析システムは、ステップ２１０に含まれる複数のＳＮＰ部位から、集団ハプロタイプ頻度が約５０％の閾値範囲内にある２つのハプロタイプ（２－ＳＮＰ部位における４つの潜在的ハプロタイプのうちの）を有するものを特定する。集団ハプロタイプ頻度は、ゲノムデータベースから得るか、集団を代表するサンプルセットを用いて決定することができる。閾値範囲は、５０％の±１％、±２％、±３％、±４％、±５％、±６％、±７％、±８％、±９％、または±１０％とすることができる。例えば、解析システムは、ハプロタイプ［０、０］および［１、１］が閾値範囲内の実質的な集団ハプロタイプ頻度を有する一方、［１、０］および［０、１］が小さな集団ハプロタイプ頻度を有する２－ＳＮＰ部位を含む。別の例では、解析システムは、ハプロタイプ［１、０］および［０、１］が閾値範囲内の実質的な集団ハプロタイプ頻度を有する一方、［０、０］および［１、１］が小さな集団ハプロタイプ頻度を有する２－ＳＮＰ部位を含む。

【０１５８】

解析システムは、グアニン－アデニン多型およびシトシン－チミン多型を有する複数のＳＮＰ部位を除外する（ステップ２１５）。解析システムは、例えばメチル化配列決定のためのバイサルファイト配列決定での問題を回避するために、このような複数のＳＮＰ部位を除外する。他のタイプの配列決定に依存する実施形態では、解析システムはステップ２１５をスキップすることができる。

【０１５９】

解析システムは、各複数ＳＮＰ部位におけるハプロタイプ、がハーディーワインベルク平衡にあることを保証する（ステップ２２０）。複数のＳＮＰ部位の各ＳＮＰ部位について、解析システムはＳＮＰ部位の各対立遺伝子の対立遺伝子頻度を計算する。対立遺伝子頻度は、遺伝子データベースから得られたデータ、または集団を代表するサンプルセットを用いて計算することができる。次に、解析システムは、各ＳＮＰ部位の対立遺伝子頻度がハーディーワインベルグ平衡にあるかどうかをチェックする。

【数１】

$$p^2 + 2pq + q^2 = 1$$

ここで、ｐはＳＮＰ部位の第１の対立遺伝子の対立遺伝子頻度であり、ｑはＳＮＰ部位の第２の対立遺伝子の対立遺伝子頻度である。複数のＳＮＰ部位の各ＳＮＰ部位がハーディーワインベルグ平衡にあるかどうかをチェックすることにより、進化または変異している可能性のあるＳＮＰ部位、すなわち平衡にないＳＮＰ部位から生じるアーチファクトを確実に防止することができる。解析システムは、ハーディーワインベルグ平衡の計算にある程度の許容誤差を組み込むことができる。

【０１６０】

解析システムは、ステップ２０５、２１０、２１５、および２２０のうちの１つ以上を省略することができる。前述のように、バイサルファイト配列決定を行わない実施形態では、ステップ２１５を省略することができる。他の実施形態では、ステップ２１０および／またはステップ２２０が省略され得る。

【０１６１】

この時点で、解析システムは、ステップ２０５、２１０、２１５、および２２０の１つ、いくつか、または全ての後に、複数のＳＮＰ部位を汚染マーカーとして使用可能であると同定した。解析システムは、配列決定パネルに実装され得る汚染マーカーの予算に基づいて、実行可能な複数のＳＮＰ部位の数をさらに切り詰めることができる。解析システムは、ゲノム全体にわたる多重ＳＮＰ部位汚染マーカーの分布を最適化することができる。解析システムはまた、ステップ２０５、２１０、２１５、および２２０における様々なパラメータを調整して、汚染マーカーとして選択される複数のＳＮＰ部位の数を最適化することができる。例えば、解析システムは、ステップ２０５において閾値距離を増加させ、ステップ２１０、２１５、および２２０において考慮され得る複数のＳＮＰ部位の数を増加させることができる。別の例として、解析システムは、ステップ２１０において閾値範囲を減少させて、ステップ２１５および２２０において考慮され得る複数のＳＮＰ部位の数を減少させることができる。表１は、例示的な実施態様に従って、汚染マーカーとして使用するために選択された複数のＳＮＰ部位のリストを含む。１つ以上の実施態様において、複数のＳＮＰ部位は、基準のリストに従ってランク付けされ得る。基準には、（１）ＳＮＰ位置を取り囲む配列の複雑さ（ｋ－ｍｅｒエントロピー）、（２）設計されたプローブとゲノム内の他の領域との類似性、（３）理想値０．５からの集団ハプロタイプ頻度の偏差、（４）実際の配列決定されたサンプルで観察される、その部位での読み取り重複率、などが含まれ得る。

【０１６２】

解析システムは、複数のＳＮＰ部位汚染マーカーの各ハプロタイプを標的とする汚染マーカープローブを設計する（ステップ２２５）。各複数のＳＮＰ部位汚染マーカーについて、ステップ２１０でどの２つのハプロタイプが考慮されるかに応じて、解析システムは、２つのハプロタイプの各々を標的とするプローブを設計する。解析システムは、各ハプロタイプの両方のＤＮＡ鎖を標的とするプローブを設計することもできる。汚染マーカーの各ハプロタイプを標的とするプローブを設計することにより、配列決定における参照または代替バイアスを回避することができる。別の実施形態では、解析システムは、各複数ＳＮＰ部位汚染マーカーの参照配列を標的とする単一のプローブを設計する。表２は、例示的な実施態様に従って、表１の複数のＳＮＰ部位汚染マーカーに対して設計されたプローブのリストを含む。

【０１６３】

図２Ｂは、１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための、インデル部位を同定するステップ２３０を説明する例示的なフローチャートである。ステップ２３０は、本明細書では、解析システムによって実行されるものとして記載されるが、ステップの一部または全部は、他の比較可能に記載された配列決定デバイスおよび／またはコンピュータプロセッサによって実行されてもよい。

【０１６４】

解析システムは、長さの範囲内でインデル部位を同定する（ステップ２３５）。インデル部位とは、個体のゲノムから挿入または欠失された遺伝子配列のことである。長さの範囲は、例えば、５～１０ｂｐ、５～１５ｂｐ、５～２０ｂｐ、５～２５ｂｐ、５～５０ｂｐ、５～１００ｂｐ、１０～１５ｂｐ、１０～２０ｂｐ、１０～２５ｂｐ、１０～５０ｂｐ、１０～１００ｂｐ、１５～２０ｂｐ、１５～２５ｂｐ、１５～５０ｂｐ、または１５～１００ｂｐであり得る。

【０１６５】

解析システムには、複雑性の高いインデル部位が含まれる（ステップ２４０）。複雑度の低いインデル部位には、ホモポリマーおよび単純なタンデムリピートが含まれる。ホモ

ポリマーは、１つの反復ヌクレオチドの文字列であり、例えば、ＡＣＧＴＴＴＴＴＴＴＴＴＴＴＴＴＴＴＡＣＧは、１５個のチミンのホモポリマーを含む。ホモポリマーについては、低複雑性とみなされるための閾値反復数が存在する可能性があり、例えば、５回以上の反復は低複雑性とみなされる。単純タンデムリピートは、繰り返されるヌクレオチドタンデムの文字列であり、例えば、ＡＣＧＴＣＡＴＣＡＴＣＡＴＣＡＴＣＡＴＣＡＴＣＡＴＡＣＧＴは、ヌクレオチドタンデムＣＡＴの７つの繰り返される実例を含む。単純なタンデム反復の場合、低複雑性とみなされる反復の閾値数が存在する可能性があり、例えば、低複雑性とみなされる反復は３回以上である。高複雑性配列には、ホモポリマーや単純タンデム反復を含まない、より長い配列が含まれる。例えば、高複雑性配列コンテクストは、特に長いインデルまたは特に特異的な長い挿入、例えば、ＡＣＧＴＡＣＣＧＧＧＴＴＴＴＣＡ（ここで、「ＡＣＣＧＧＧＴＴＴＴ」は挿入配列である）を含み得る。高複雑度インデルの配列は、例えばポリメラーゼ連鎖反応（ＰＣＲ）または配列決定などのサンプル処理によってもたらされるエラーとは対照的に、汚染された断片に対するスクリーニングを確実にする。

【０１６６】

解析システムは、閾値範囲内の集団対立遺伝子頻度を有するインデル部位を含む（ステップ２４５）。集団対立遺伝子頻度は、ゲノムデータベースから得るか、集団を代表するサンプルセットを用いて決定することができる。閾値範囲は、５０％の±１％、±２％、±３％、±４％、±５％、±６％、±７％、±８％、±９％、または±１０％とすることができる。挿入配列の場合、第１の対立遺伝子は挿入配列を含まないが、第２の対立遺伝子は挿入配列を含む。欠失配列の場合、第１の対立遺伝子は欠失配列を含むが、第２の対立遺伝子は欠失配列を含まない。

【０１６７】

解析システムは、各インデル部位の対立遺伝子がハーディーワインベルク平衡にあることを保証する（ステップ２５０）。各インデル部位について、解析システムはインデル部位の各対立遺伝子の対立遺伝子頻度を計算する。対立遺伝子頻度は、遺伝子データベースから得られたデータ、または母集団を代表するサンプルセットを用いて計算することができる。次に、解析システムは、各インデル部位の対立遺伝子頻度がハーディーワインベルク平衡にあるかどうかをチェックする。

【数２】

$$p^2 + 2pq + q^2 = 1$$

ここで、ｐはインデル部位の第１の対立遺伝子の対立遺伝子頻度であり、ｑはインデル部位の第２の対立遺伝子の対立遺伝子頻度である。各インデル部位がハーディーワインベルグ平衡にあるかどうかをチェックすることにより、進化または変異している可能性のあるインデル部位、すなわち平衡にないインデル部位から生じるアーチファクトを確実に防ぐことができる。解析システムはハーディーワインベルグ平衡の計算にある程度の許容誤差を組み込むことができる。

【０１６８】

解析システムは、ステップ２３５、２４０、２４５、および２５０の１つ以上を省略することができる。

【０１６９】

この時点で、解析システムは、ステップ２３５、２４０、２４５、および２５０の１つ、いくつか、または全ての後に、汚染マーカーとして使用可能なインデル部位を同定している。解析システムは、配列決定パネルに実装され得る汚染マーカーの予算に基づいて、実行可能なインデル部位の数をさらに削減することができる。解析システムは、ゲノム全体にわたるインデル部位汚染マーカーの分布を最適化することができる。解析システムは

また、ステップ２３５、２４０、２４５、および２５０における様々なパラメータを調整して、汚染マーカーとして選択されるインデル部位の数を最適化することができる。例えば、解析システムはステップ２３５で長さの範囲を広げ、ステップ２４０、２４５、２５０で考慮される可能性のあるインデル部位の数を増やすことができる。表３は、例示的な実施態様に従って、汚染マーカーとして使用するために選択されたインデル部位のリストを含む。１つ以上の実施態様において、インデル部位は１つ以上の基準に従ってランク付けすることができる。基準には、（１）設計されたプローブとゲノムの他の領域との類似性、（２）集団ハプロタイプ頻度の理想値０．５からの偏差、（３）実際の配列決定されたサンプルで観察される、その部位での読み取り重複率、他の基準、またはそれらの組み合わせが含まれる。インデル部位は、ランキングに従って、またインデル部位の予算に従って選択することができる。

【０１７０】

解析システムは、インデル部位汚染マーカーの各対立遺伝子を標的とする２５５個の汚染マーカープローブを設計する。解析システムは、２つの対立遺伝子のそれぞれを標的とするプローブを設計する。解析システムはまた、各対立遺伝子の両方のＤＮＡ鎖を標的とするプローブを設計してもよい。汚染マーカーの各対立遺伝子を標的とするプローブを設計することにより、配列決定における参照または代替バイアスを回避することができる。別の実施形態では、解析システムは、各インデル部位汚染マーカーの参照配列を標的とする単一のプローブを設計する。表４は、例示的な実施態様に従って、表３のインデル部位汚染マーカー用に設計されたプローブのリストを含む。

【０１７１】

＜ＩＩ．Ｂ．ＤＮＡ断片のメチル化状態ベクターの生成＞

図３Ａは、１つ以上の実施形態による、メチル化状態ベクターを得るためにｃｆＤＮＡの断片を配列決定するプロセス３００を説明する例示的なフローチャートである。ＤＮＡメチル化を解析するために、解析システムはまず、複数のｃｆＤＮＡ分子からなる個人からのサンプルを取得する（ステップ３１０）。さらなる実施形態において、プロセス３００は、他のタイプのＤＮＡ分子の配列決定にも適用され得る。

【０１７２】

サンプルから、解析システムは各ｃｆＤＮＡ分子を単離することができる。ｃｆＤＮＡ分子は、非メチル化シトシンをウラシルに変換するために処理することができる。一実施形態では、この方法は、メチル化されたシトシンを変換することなく、メチル化されていないシトシンをウラシルに変換するＤＮＡの重亜硫酸塩処理を使用する。例えば、ＥＺ ＤＮＡ Ｍｅｔｈｙｌａｔｉｏｎ（商標）－Ｇｏｌｄ、ＥＺ ＤＮＡ Ｍｅｔｈｙｌａｔｉｏｎ（商標）－ＤｉｒｅｃｔまたはＥＺ ＤＮＡ Ｍｅｔｈｙｌａｔｉｏｎ（商標）－Ｌｉｇｈｔｎｉｎｇキット（Ｚｙｍｏ Ｒｅｓｅａｒｃｈ Ｃｏｒｐから入手可能、カリフォルニア州アーバイン）などの市販のキットを、重亜硫酸塩変換に使用する。別の実施形態では、非メチル化シトシンのウラシルへの変換は、酵素反応を用いて達成される。例えば、ＡＰＯＢＥＣ－Ｓｅｑ（ＮＥＢｉｏｌａｂｓ、マサチューセッツ州イプスウィッチ）のような、非メチル化シトシンのウラシルへの変換のための市販のキットを使用することができる。

【０１７３】

変換されたｃｆＤＮＡ分子から、配列決定ライブラリーを調製することができる（ステップ３３０）。ライブラリー調製中、アダプターライゲーションにより、核酸分子（例えば、ＤＮＡ分子）に一意的分子識別子（ＵＭＩ）を付加することができる。ＵＭＩは、アダプターライゲーション中にＤＮＡ断片（例えば、物理的剪断、酵素消化、および／または化学的断片化によって断片化されたＤＮＡ分子）の末端に付加される短い核酸配列（例えば、４～１０塩基対）であり得る。ＵＭＩは、特定のＤＮＡ断片に由来する配列リードを同定するために使用できるユニークなタグとして機能する縮退塩基対であり得る。アダプターライゲーション後のＰＣＲ増幅において、ＵＭＩは結合したＤＮＡ断片と共に複製される。これにより、下流の解析において、同じオリジナル断片に由来する配列リードを

同定することができる。

【０１７４】

任意に、配列決定ライブラリーは、複数のハイブリダイゼーションプローブを用いて、がんの状態について有益なｃｆＤＮＡ分子またはゲノム領域について濃縮され得る。ハイブリダイゼーションプローブは、特に特定されたｃｆＤＮＡ分子または標的領域にハイブリダイズすることができる短いオリゴヌクレオチドであり、その後の配列決定および解析のために、それらの断片または領域を濃縮する。ハイブリダイゼーションプローブは、研究者の関心のある特定のＣｐＧ部位のセットを対象とした高深度解析を行うために使用することができる。ハイブリダイゼーションプローブは、１倍、２倍、３倍、４倍、５倍、６倍、７倍、８倍、９倍、１０倍、または１０倍以上のカバレッジであり、１つ以上の標的配列にわたってタイリングすることができる。例えば、２倍のカバレッジでタイリングされたハイブリダイゼーションプローブは、標的配列の各部分が２つの独立したプローブにハイブリダイズするようなオーバーラッププローブからなる。ハイブリダイゼーションプローブは、１倍未満のカバレッジで、１つ以上の標的配列にわたってタイリングすることができる。

【０１７５】

一実施形態では、ハイブリダイゼーションプローブは、非メチル化シトシンをウラシルに変換するために（例えば、重亜硫酸塩を使用して）処理されたＤＮＡ分子を濃縮するように設計される。濃縮中、ハイブリダイゼーションプローブ（本明細書では「プローブ（probes）」とも呼ぶ）は、がん（または疾患）の有無、がんの状態、またはがんの分類（例えば、がんクラスまたは起源組織）について情報提供する核酸断片を標的としてプルダウンするために使用することができる。プローブは、ＤＮＡの標的鎖（相補的）にアニール（またはハイブリダイズ）するように設計することができる。標的鎖は、「ポジティブ（positive）」鎖（例えば、ｍＲＮＡに転写され、その後タンパク質に翻訳される鎖）または相補的な「ネガティブ（negative）」鎖である。プローブの長さは１０塩基対、１００塩基対、１０００塩基対のいずれでもよい。プローブは、メチル化部位パネルに基づいて設計することができる。プローブは、特定の変異や、特定のがんや他のタイプの疾患に対応すると疑われるゲノム（例えば、ヒトや他の生物）の標的領域を解析するために、標的遺伝子パネルに基づいて設計することができる。さらに、プローブは標的領域の重複部分をカバーすることもできる。

【０１７６】

一旦調製されると、配列決定ライブラリーまたはその一部は、複数の配列リードを得るために配列決定され得る。配列リードは、コンピュータソフトウェアによる処理および解釈のために、コンピュータ読み取り可能なデジタル形式であってもよい。配列リードは、アライメント位置情報を決定するために参照ゲノムにアライメントされ得る。アライメント位置情報は、所定の配列リードの開始ヌクレオチド塩基および終了ヌクレオチド塩基に対応する参照ゲノム中の領域の開始位置および終了位置を示す場合がある。アライメント位置情報は、開始位置および終了位置から決定され得る配列リード長も含み得る。参照ゲノム中の領域は、遺伝子または遺伝子のセグメントに関連する場合がある。配列リードは、R_1およびR_2と表記されるリードペアで構成される。例えば、第１リードR_1は核酸断片の第１末端から配列決定され、第２リードR_2は核酸断片の第２末端から配列決定される。したがって、第１のリードR_1および第２のリードR_2のヌクレオチド塩基対は、参照ゲノムのヌクレオチド塩基と一貫して（例えば、反対の向きで）整列される場合がある。リードペアR_1およびR_2から得られるアライメント位置情報には、第１のリード（例えば、R_1）の末端に対応する参照ゲノム内の開始位置および、第２のリード（例えば、R_2）の末端に対応する参照ゲノム内の終了位置が含まれる場合がある。言い換えれば、参照ゲノムにおける開始位置と終了位置は、核酸断片が対応する参照ゲノム内の可能性の高い位置を表すことができる。ＳＡＭ（配列アライメントマップ）形式またはＢＡＭ（二値）形式の出力ファイルが生成され、メチル化状態の決定などのさらなる解析のために出力される。

【０１７７】
解析システムは、配列リードから、参照ゲノムとのアライメントに基づいて、各ＣｐＧ部位の位置およびメチル化状態を決定する（ステップ３５０）。解析システムは、参照ゲノムにおける断片の位置（例えば、各断片の最初のＣｐＧ部位の位置で指定されたもの、または他の類似の指標）、断片内のＣｐＧ部位の数、および断片内の各ＣｐＧ部位のメチル化（例えば、Ｍと表記）、非メチル化（例えば、Ｕと表記）、または不確定（例えば、Ｉと表記）のいずれかのメチル化状態を指定する、各断片のメチル化状態ベクターを生成する（ステップ３６０）。観察される状態には、メチル化された状態およびメチル化されていない状態があるが、観察されない状態は不確定である。不確定なメチル化状態は、塩基配列の誤りおよび／またはＤＮＡ断片の相補鎖のメチル化状態の不一致に由来する可能性がある。メチル化状態ベクターは、後で使用および処理するために、一時的または永続的なコンピュータメモリに保存することができる。さらに、解析システムは、単一のサンプルから重複するリードまたは重複するメチル化状態ベクターを除去することができる。解析システムは、１つ以上のＣｐＧ部位を有する特定の断片が、閾値数またはパーセンテージを超える不確定メチル化状態を有することを決定してもよく、そのような断片を除外してもよく、またはそのような断片を選択的に含んでもよいが、そのような不確定メチル化状態を考慮したモデルを構築してもよく、そのようなモデルの１つについては、図４と関連して後述する。

【０１７８】
図３Ｂは、１つ以上の実施形態による、メチル化状態ベクターを得るためにｃｆＤＮＡ分子を配列決定する図３Ａのプロセス３００の例示的な図である。一例として、解析システムは、この例では３つのＣｐＧ部位を含むｃｆＤＮＡ分子を受け取る（ステップ３１２）。示されるように、ｃｆＤＮＡ分子の第１および第３のＣｐＧ部位は（ステップ３１２）、メチル化されている（ステップ３１４）。処理の間に（ステップ３２０）、ｃｆＤＮＡ分子は変換されて（ステップ３１２）、変換ｃｆＤＮＡ分子を生成する（ステップ３２２）。処理の間に（ステップ３２０）、メチル化されていなかった第２のＣｐＧ部位は、そのシトシンがウラシルに変換される。しかし、第１および第３のＣｐＧ部位は変換されなかった。

【０１７９】
変換後、配列決定ライブラリーが調製され（ステップ３３０）、配列決定が行われ（ステップ３４０）、配列リードが生成される（ステップ３４２）。解析システムは、配列リード（ステップ３４２）を参照ゲノム（ステップ３４４）にアライメントする（ステップ３５０）。参照ゲノム（ステップ３４４）は、断片ｃｆＤＮＡがヒトゲノムのどの位置に由来するかについてのコンテクストを提供する。この単純化された例では、解析システムは、３つのＣｐＧ部位がＣｐＧ部位２３、２４および２５（説明の便宜のために使用される任意の参照識別子）に相関するように、配列リード（ステップ３４２）をアライメントする（ステップ３５０）。このようにして、解析システムは、ｃｆＤＮＡ分子（ステップ３１２）上の全てのＣｐＧ部位のメチル化状態、およびＣｐＧ部位がマッピングされるヒトゲノム内の位置の両方の情報を生成することができる。示されるように、メチル化されている配列リード（ステップ３４２）上のＣｐＧ部位は、シトシンとして読み取られる。この例では、シトシンは配列リード（ステップ３４２）の１番目および３番目のＣｐＧ部位にのみ現れているため、元のｃｆＤＮＡ分子の１番目および３番目のＣｐＧ部位がメチル化されていると推測できる。一方、第２のＣｐＧ部位はチミンとして読み取れる（Ｕは配列決定過程でＴに変換される）ので、第２のＣｐＧ部位は元のｃｆＤＮＡ分子ではメチル化されていないと推測できる。これら２つの情報であるメチル化の状態および位置を用いて、解析システムは断片ｃｆＤＮＡ（ステップ３１２）のメチル化状態ベクター（ステップ３５２）を生成する（ステップ３６０）。この例では、得られたメチル化状態ベクター（ステップ３５２）は＜M_{23}、U_{24}、M_{25}＞であり、Ｍはメチル化ＣｐＧ部位に対応し、Ｕは非メチル化ＣｐＧ部位に対応し、添え字番号は参照ゲノムにおける各ＣｐＧ部位の位置に対応する。

【０１８０】
生物学的サンプル中の核酸から配列リードを得るために、1つ以上の代替的な配列決定方法を使用することができる。１つ以上の配列決定方法は、核酸（例えば、セルフリー核酸）から測定された配列リードの数を得るために使用され得る任意の形態の配列決定を含み得、これには、Ｒｏｃｈｅ　４５４プラットフォーム、Ａｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓ　ＳＯＬＩＤプラットフォーム、Ｈｅｌｉｃｏｓ　Ｔｒｕｅ　Ｓｉｎｇｌｅ　Ｍｏｌｅｃｕｌｅ　ＤＮＡ配列決定技術、Ａｆｆｙｍｅｔｒｉｘ社のシークエンシング・バイ・ハイブリダイゼーション・プラットフォーム、Ｐａｃｉｆｉｃ　Ｂｉｏｓｃｉｅｎｃｅｓの１分子リアルタイム（ＳＭＲＴ）技術、４５４　Ｌｉｆｅ　Ｓｃｉｅｎｃｅｓ、Ｉｌｌｕｍｉｎａ／Ｓｏｌｅｘａ、Ｈｅｌｉｃｏｓ　Ｂｉｏｓｃｉｅｎｃｅｓのシーケンシング・バイ・シンセシス・プラットフォーム、Ａｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓのシーケンシング・バイ・ライゲーション・プラットフォームなどがある。Ｌｉｆｅ　ｔｅｃｈｎｏｌｏｇｉｅｓのＩＯＮ　ＴＯＲＲＥＮＴテクノロジーおよびＮａｎｏｐｏｒｅシーケンスも、生物学的サンプル中の核酸（例えば、セルフリー核酸）から配列リードを得るために使用できる。シーケンシング・バイ・シンセシスおよび可逆ターミネーターベースの配列（例えば、ＩｌｌｕｍｉｎａのＧｅｎｏｍｅ　Ａｎａｌｙｚｅｒ、Ｇｅｎｏｍｅ　Ａｎａｌｙｚｅｒ　ＩＩ、ＨＩＳＥＱ　２０００、ＨＩＳＥＱ　４５００（Ｉｌｌｕｍｉｎａ、カリフォルニア州サンディエゴ））を使用して、遺伝子型データセットを形成するために、訓練対象の生物学的サンプルから得られたセルフリー核酸から配列リードを得ることができる。何百万ものセルフリー核酸（例えば、ＤＮＡ）断片を並行して配列決定することができる。この種の配列決定技術の一例では、光学的に透明なスライドを含むフローセルが使用され、その表面には８つの個別レーンがあり、オリゴヌクレオチドアンカー（例えば、アダプタープライマー）が結合している。セルフリー核酸サンプルは、検出を容易にするシグナルまたはタグを含むことができる。生物学的サンプルから得られたセルフリー核酸からの配列リードの取得は、例えば、フローサイトメトリー、定量的ポリメラーゼ連鎖反応（ｑＰＣＲ）、ゲル電気泳動、遺伝子チップ分析、マイクロアレイ、質量分析、サイトフルオリメトリー分析、蛍光顕微鏡、共焦点レーザー走査顕微鏡、レーザー走査型サイトメトリー、アフィニティークロマトグラフィー、手動バッチモード分離、電場懸濁、配列決定、およびこれらの組み合わせなどの様々な技術を用いて、シグナルまたはタグの定量情報を得ることを含むことができる。

【０１８１】
１つ以上の配列決定方法は、全ゲノム配列決定アッセイを含み得る。全ゲノムシーケンスアッセイは、コピー数変異またはコピー数異常のような大きな変異を決定するために使用され得る全ゲノムまたは全ゲノムの実質的部分についての配列リードを生成する物理的アッセイを含み得る。このような物理的アッセイは、全ゲノム配列決定技術または全エクソーム配列決定技術を用いることができる。全ゲノムシークエンシングアッセイは、被験体のゲノム全体にわたって、少なくとも１倍、２倍、３倍、４倍、５倍、６倍、７倍、８倍、９倍、１０倍、少なくとも２０倍、少なくとも３０倍、または少なくとも４０倍の平均シークエンシング深度を有し得る。いくつかの実施形態では、シーケンス深度は約３０，０００倍である。１つ以上の配列決定方法は、標的化パネル配列決定アッセイを含み得る。標的化パネル配列決定アッセイは、遺伝子の標的化パネルについて、少なくとも５０，０００倍、少なくとも５５，０００倍、少なくとも６０，０００倍、または少なくとも７０，０００倍の平均シーケンス深度を有し得る。遺伝子の標的化パネルは、４５０～５００の遺伝子を含み得る。遺伝子の標的化パネルは、５００±５遺伝子の範囲、５００±１０遺伝子の範囲、または５００±２５遺伝子の範囲を含み得る。

【０１８２】
１つ以上の配列決定方法は、ペアエンドシーケンスを含み得る。１つ以上の配列決定方法は、複数の配列リードを生成し得る。複数の配列リードは、１０～７００、５０～４００、または１００～３００の範囲の平均長を有し得る。１つ以上の配列決定方法は、メチル化配列決定アッセイを含み得る。メチル化配列決定は、ｉ）全ゲノムメチル化配列決定

、または、ｉｉ）複数の核酸プローブを使用する標的ＤＮＡメチル化配列決定であり得る。例えば、メチル化配列決定は、全ゲノムバイサルファイト配列決定（例えば、ＷＧＢＳ）である。メチル化配列決定は、メチロームの最も情報量の多い領域を標的とする複数の核酸プローブ、独自のメチル化データベース、および先行するプロトタイプの全ゲノムおよび標的配列決定アッセイを用いる標的ＤＮＡメチル化配列決定であり得る。

【０１８３】

メチル化配列決定は、それぞれの核酸メチル化断片中の１つ以上の５－メチルシトシン（５ｍＣ）および／または５－ヒドロキシメチルシトシン（５ｈｍＣ）を検出し得る。メチル化配列決定は、それぞれの核酸メチル化断片中の、１つ以上の非メチル化シトシンまたは１つ以上のメチル化シトシンの、対応する１つ以上のウラシルへの変換を含み得る。１つ以上のウラシルは、メチル化配列決定中に、１つ以上の対応するチミンとして検出され得る。１つ以上の非メチル化シトシンまたは１つ以上のメチル化シトシンの変換は、化学的変換、酵素的変換、またはそれらの組み合わせからなり得る。

【０１８４】

例えば、バイサルファイト変換は、メチル化されたシトシン（例えば、５－メチルシトシンまたは５－ｍＣ）をそのまま残して、シトシンをウラシルに変換することを含む。ＤＮＡによっては、約９５％のシトシンがメチル化されていない場合があり、得られるＤＮＡ断片には、チミンで表される多くのウラシルが含まれる場合がある。配列決定の前に核酸を処理するために酵素的変換プロセスを使用することができ、これは様々な方法で行うことができる。重亜硫酸塩を含まない変換の一例は、非修飾シトシンに影響を与えることなく５－メチルシトシンおよび５－ヒドロキシメチルシトシンを非破壊的かつ直接検出するための、重亜硫酸塩を含まない塩基分解能配列決定法、ＴＥＴ－アシステッドピリジンボラン配列決定（ＴＡＰＳ）からなる。それぞれの核酸メチル化断片中の対応する複数のＣｐＧ部位におけるＣｐＧ部位のメチル化状態は、メチル化配列決定によってＣｐＧ部位がメチル化されていると判定された場合にはメチル化され、メチル化配列決定によってＣｐＧ部位がメチル化されていないと判定された場合にはメチル化されていないことがある。

【０１８５】

メチル化配列決定アッセイ（例えば、ＷＧＢＳおよび／または標的メチル化配列決定）は、約１，０００倍、２，０００倍、３，０００倍、５，０００倍、１０，０００倍、１５，０００倍、２０，０００倍、または３０，０００倍までを含むが、これらに限定されない平均シーケンス深度を有し得る。メチル化配列決定は、３０，０００倍を超えるシーケンス深度、例えば、少なくとも４０，０００倍または５０，０００倍を有することができる。全ゲノムバイサルファイト配列決定法は、２０倍～５０倍の平均シーケンス深度を有し得、標的化メチル化配列決定法は、１００倍～１０００倍の平均有効深度を有し得、ここで、有効深度は、標的化メチル化配列決定によって得られる同じ数の配列リードを得るための等価全ゲノムバイサルファイト配列決定カバレッジであり得る。

【０１８６】

メチル化配列決定（例えば、ＷＧＢＳおよび／または標的化メチル化配列決定）に関するさらなる詳細については、例えば、２０１８年３月１３日に出願された「メチル化断片異常検出」と題する米国特許出願第６２／６４２，４８０号、および２０１９年１２月１８日に出願された「メチル化情報を用いて細胞源分画を推定するシステムおよび方法」と題する米国特許出願第１６／７１９，９０２号を参照されたく、これらの各々は参照により本明細書に組み込まれる。断片メチル化パターンを得るために、本明細書に開示されたもの、および／またはその任意の改変、置換、または組み合わせを含む、メチル化配列決定のための他の方法を使用することができる。メチル化配列決定は、例えば、２０１９年３月１３日に出願された「異常断片の検出および分類」と題する米国特許出願第１６／３５２，６０２号に記載されるように、または２０２０年５月１３日に出願された「モデルベースの特徴化および分類」と題する米国特許出願第１５／９３１，０２２号に開示される技術のいずれかに従って、１つ以上のメチル化状態ベクターを同定するために使用され

得、これらの各々は、参照により本明細書に組み込まれる。

【０１８７】

核酸のメチル化配列決定および得られた１つ以上のメチル化状態ベクターを使用して、複数の核酸メチル化断片を得ることができる。それぞれの対応する複数の核酸メチル化断片（例えば、それぞれの遺伝子型データセットについて）は、１００個以上の核酸メチル化断片から構成され得る。各対応する複数の核酸メチル化断片にわたる核酸メチル化断片の平均数は、１０００個以上の核酸メチル化断片、５０００個以上の核酸メチル化断片、１０，０００個以上の核酸メチル化断片、２０，０００個以上の核酸メチル化断片、または３０，０００個以上の核酸メチル化断片からなり得る。各対応する複数の核酸メチル化断片にわたる核酸メチル化断片の平均数は、１０，０００個～５０，０００個の核酸メチル化断片であり得る。対応する複数の核酸メチル化断片は、１０００個以上、１万個以上、１０万個以上、１００万個以上、１０００万個以上、１億個以上、５億個以上、１０億個以上、２０億個以上、３０億個以上、４０億個以上、５０億個以上、６０億個以上、７０億個以上、８０億個以上、９０億個以上、または１００億個以上の核酸メチル化断片からなり得る。対応する複数の核酸メチル化断片の平均長さは、１４０～４８０ヌクレオチドであり得る。

【０１８８】

核酸を配列決定する方法およびメチル化配列決定データに関するさらなる詳細は、２０２０年３月４日に出願された「オートエンコーダを用いたがん状態判定システムおよび方法」と題する米国仮特許出願第６２／９８５，２５８号に開示されており、その全体が参照により本明細書に組み込まれる。

【０１８９】

＜ＩＩ．Ｃ．異常断片の特定＞

解析システムは、サンプルのメチル化状態ベクターを使用して、サンプルの異常断片を決定することができる。解析システムは、サンプル中の各断片について、その断片に対応するメチル化状態ベクターを用いて、その断片が異常断片であるかどうかを判定することができる。いくつかの実施形態では、解析システムは、各メチル化状態ベクターについて、そのメチル化状態ベクターまたは健常対照群においてさらに確率の低い他のメチル化状態ベクターを観察する確率を記述するｐ値スコアを計算する。ｐ値スコアを計算するプロセスについては、セクションＩＩ．Ｃ．ｉ．のＰ値フィルタリングでさらに後述する。解析システムは、閾値ｐ値スコアを下回るメチル化状態ベクターを有する断片を異常断片として決定することができる。一部の実施形態では、解析システムはさらに、メチル化または非メチル化の割合がある閾値を超える少なくともいくつかのＣｐＧ部位を有する断片を、それぞれ高メチル化および低メチル化断片として標識する。高メチル化断片または低メチル化断片は、極端なメチル化を伴う異常な断片（ＵＦＸＭ）とも呼ばれることがある。他の実施形態では、解析システムは、異常断片を決定するための様々な他の確率モデルを実装することができる。他の確率モデルの例には、混合モデル、深層確率モデルなどが含まれる。いくつかの実施形態では、解析システムは、異常な断片を識別するために、以下に説明するプロセスの任意の組み合わせを使用することができる。同定された異常断片を用いて、解析システムは、他の処理、例えば、がん分類器の訓練および展開における使用のために、サンプルのメチル化状態ベクターのセットをフィルタリングすることができる。

【０１９０】

＜ＩＩ．Ｃ．Ｉ．Ｐ値フィルタリング＞

いくつかの実施形態では、解析システムは、各メチル化状態ベクターについて、健常対照群の断片からのメチル化状態ベクターと比較したｐ値スコアを計算する。ｐ値スコアは、そのメチル化状態ベクターに一致するメチル化状態を観察する確率、または健常対照群においてさらに確率の低い他のメチル化状態ベクターを観察する確率を表すことができる。ＤＮＡ断片が異常にメチル化されていると判定するために、解析システムは、正常にメチル化されている断片が大多数を占める健常対照群を使用することができる。異常な断片

を判定するためにこの確率的解析を行う場合、健常対照群を構成する対照被験者群と比較して、判定に重みを持たせることができる。健常対照群における頑健性を確保するために、解析システムは、ＤＮＡ断片を含むサンプルを供給するために、ある閾値数の健常個体を選択することができる。以下の図４Ａは、解析システムがｐ値スコアを計算し得る健常対照群のデータ構造を生成する方法を説明する。図４Ｂは、生成されたデータ構造を用いてｐ値スコアを計算する方法を説明する。

【０１９１】

図４Ａは、実施形態による、健常対照群のデータ構造を生成するプロセス４００を説明するフローチャートである。健常対照群のデータ構造を作成するために、解析システムは、複数の健常個体から複数のＤＮＡ断片（例えば、ｃｆＤＮＡ）を受け取ることができる。メチル化状態ベクターは、例えば、プロセス３００を介して、各断片について同定することができる。

【０１９２】

各断片のメチル化状態ベクターを用いて、解析システムはメチル化状態ベクターをＣｐＧ部位の文字列に細分化することができる。いくつかの実施形態では、解析システムは、得られた文字列が全て所定の長さ未満になるように、メチル化状態ベクターを細分化する（ステップ４０５）。例えば、長さ１１のメチル化状態ベクターを長さ３以下の文字列に細分化すると、長さ３の文字列が９個、長さ２の文字列が１０個、長さ１の文字列が１１個となる。別の例では、長さ７のメチル化状態ベクターを長さ４以下の文字列に細分化すると、長さ４の文字列が４個、長さ３の文字列が５個、長さ２の文字列が６個、長さ１の文字列が７個となる。メチル化状態ベクターが、指定された文字列の長さより短いか、同じ長さである場合、メチル化状態ベクターは、そのベクターの全てのＣｐＧ部位を含む単一の文字列に変換されてもよい。

【０１９３】

解析システムは、ベクター内の可能性のあるＣｐＧ部位およびメチル化状態の可能性ごとに、指定されたＣｐＧ部位を文字列の最初のＣｐＧ部位として有し、そのメチル化状態の可能性を有する対照群に存在する文字列の数をカウントすることによって、文字列を集計する（ステップ４１０）。例えば、あるＣｐＧ部位において、文字列の長さが３であるとすると、２＾３、または８通りの文字列構成があり得る。その与えられたＣｐＧ部位で、８つの可能性のある文字列構成のそれぞれについて、解析システムは、各メチル化状態のベクターの可能性が対照群にいくつ出現するかを集計する（ステップ４１０）。この例を続けると、これは、＜M_x、M_{x+1}、M_{x+2}＞、＜M_x、M_{x+1}、U_{x+2}＞、．．．、＜U_x、U_{x+1}、U_{x+2}＞を参照ゲノム中の各開始ＣｐＧ部位ｘについて計算する量を集計することを含み得る。解析システムは、各開始ＣｐＧ部位と文字列の可能性について集計されたカウントを格納するデータ構造を作成する（ステップ４１５）。

【０１９４】

文字列の長さに上限を設定することには、いくつかの利点がある。第一に、文字列の最大長によっては、解析システムによって作成されるデータ構造のサイズが劇的に増大する可能性がある。例えば、文字列の最大長が４である場合、長さ４の文字列を集計するために、全てのＣｐＧ部位が少なくとも２＾４の数を持つことを意味する。文字列の最大長を５に増やすと、全てのＣｐＧ部位が集計するためにさらに２＾４または１６の数を持つことを意味し、以前の文字列長に比べて集計する数（および必要なコンピュータ・メモリ）が２倍になる。文字列サイズを小さくすることにより、データ構造の作成と性能（例えば、後述するように後でアクセスするために使用する）を、計算とストレージの点で合理的に保つことができる。第二に、最大文字列長を制限する統計的な考慮点として、文字列カウントを使用する下流モデルのオーバーフィッティングを避けることができる。ＣｐＧ部位の長い文字列が生物学的に結果に強い影響を与えない場合（例えば、がんの存在を予測する異常性の予測）、ＣｐＧ部位の大きな文字列に基づいて確率を計算することは、利用可能でない可能性のあるかなりの量のデータを使用するため問題となる可能性があり、したがって、モデルが適切に実行するには疎らになりすぎる可能性がある。例えば、事前１

００個のＣｐＧ部位を条件として異常／がんの確率を計算する場合、長さ１００のデータ構造中の文字列のカウントを使用することができ、理想的には事前１００個のメチル化状態に正確に一致するものがある。長さ１００の文字列が疎らにしかカウントされていない場合、試験サンプルに含まれる長さ１００の文字列が異常か否かを判定するのに十分なデータがない可能性がある。

【０１９５】

図４Ｂは、１つ以上の実施形態による、サンプルから異常にメチル化された断片を同定するためのステップ４２０を説明するフローチャートである。ステップ４２０において、解析システムは、被験体のｃｆＤＮＡ断片からメチル化状態ベクターを生成する（プロセス３００）。解析システムは、各メチル化状態ベクターを以下のように処理することができる。

【０１９６】

所与のメチル化状態ベクターについて、解析システムは、メチル化状態ベクター内の同じ開始ＣｐＧ部位および同じ長さ（すなわち、ＣｐＧ部位のセット）を有するメチル化状態ベクターの全ての可能性を列挙する（ステップ４３０）。各メチル化状態は一般にメチル化または非メチル化のいずれかであるため、各ＣｐＧ部位には事実上２つの可能性があり、したがってメチル化状態ベクターの異なる可能性の数は２のべき乗に依存し、長さｎのメチル化状態ベクターは2^nの可能性のあるメチル化状態ベクターと関連することになる。メチル化状態ベクターが１つ以上のＣｐＧ部位の不確定状態を含む場合、解析システムは、観察された状態を有するＣｐＧ部位のみを考慮して、メチル化状態ベクターの可能性を列挙することができる（ステップ４３０）。

【０１９７】

解析システムは、識別された開始ＣｐＧ部位およびメチル化状態ベクター長について、健常対照群データ構造にアクセスすることにより、メチル化状態ベクターの各可能性を観察する確率を計算する（ステップ４４０）。いくつかの実施形態において、所与の可能性を観察する確率を計算することは、共同確率計算をモデル化するためにマルコフ連鎖確率を使用する。マルコフモデルは、少なくとも部分的に、対応する複数のＣｐＧ部位を有する健康な非がんコホートデータセット中のそれらの核酸メチル化断片にわたる、それぞれの断片（例えば、核酸メチル化断片）の対応する複数のＣｐＧ部位における各ＣｐＧ部位のメチル化状態の評価に基づいて、訓練され得る。例えば、マルコフモデル（例えば、隠れマルコフモデルまたはＨＭＭ）を用いて、複数の核酸メチル化断片中の核酸メチル化断片について、配列中の各状態について、配列中の次の状態を観察する可能性を決定する確率のセットが与えられた場合に、メチル化状態の配列（例えば、「Ｍ」または「Ｕ」からなる）が観察され得る確率を決定する。確率のセットは、ＨＭＭを訓練することによって得ることができる。このような訓練は、観察されたメチル化状態配列（例えば、メチル化パターン）の初期訓練データセットが与えられた場合に、統計的パラメータ（例えば、第１の状態が第２の状態に遷移し得る確率（遷移確率）および／または所定のメチル化状態がそれぞれのＣｐＧ部位について観察され得る確率（放出確率））を計算することを含み得る。ＨＭＭは、教師あり訓練（例えば、基礎となる配列ならびに観察された状態が既知であるサンプルを使用する）および／または教師なし訓練（例えば、ビタビ学習、最尤推定、期待値最大化訓練、および／またはバウム－ウェルチ訓練）を使用して訓練することができる。他の実施形態において、マルコフ連鎖確率以外の計算方法が、メチル化状態ベクターの各可能性を観察する確率を決定するために使用される。例えば、そのような計算方法は、学習された表現を含み得る。ｐ値閾値は、０．０１～０．１０、または０．０３～０．０６であり得る。ｐ値閾値は０．０５とすることができる。ｐ値の閾値は０．０１未満、０．００１未満、または０．０００１未満とすることができる。

【０１９８】

解析システムは、各可能性について計算された確率を使用して、メチル化状態ベクターのｐ値スコアを計算する（ステップ４５０）。いくつかの実施形態において、これは、問題のメチル化状態ベクターに一致する可能性に対応する計算された確率を特定することを

含む。具体的には、これは、メチル化状態ベクターと同じＣｐＧ部位のセット、または同様に同じ開始ＣｐＧ部位および長さを有する可能性であり得る。解析システムは、特定された確率以下の確率を有するあらゆる可能性の計算された確率を合計して、ｐ値スコアを生成することができる。

【０１９９】

このｐ値は、断片のメチル化状態ベクター、または健常対照群でさらに確率の低い他のメチル化状態ベクターを観察する確率を表すことができる。低いｐ値スコアは、一般に、健常対照群に対して、健常個体ではまれであり、断片が異常にメチル化されたラベルを付けられるメチル化状態ベクターに対応し得る。高いｐ値スコアは、一般に、健康な個体において相対的な意味で存在すると予想されるメチル化状態ベクターに関連し得る。例えば、健常対照群ががんでない群である場合、低いｐ値は、がんでない群と比較して断片が異常にメチル化されていることを示し、したがって、被験体におけるがんの存在を示す可能性がある。

【０２００】

上記のように、解析システムは、複数のメチル化状態ベクターのそれぞれについてｐ値スコアを計算することができ、それぞれが検査サンプル中のｃｆＤＮＡ断片を表す。どの断片が異常にメチル化されているかを同定するために、解析システムは、ｐ値スコアに基づいてメチル化状態ベクターのセットをフィルタリングすることができる（ステップ４６０）。いくつかの実施形態では、ｐ値スコアを閾値と比較し、閾値以下の断片のみを残すことによってフィルタリングが実行される。この閾値ｐ値スコアは、０．１、０．０１、０．００１、０．０００１などのオーダーであり得る。

【０２０１】

プロセス４００の結果例によると、解析システムは、訓練においてがんでない参加者について、異常なメチル化パターンを有する断片の中央値（範囲）２，８００（１，５００～１２，０００）を得ることができ、訓練においてがんである参加者について、異常なメチル化パターンを有する断片の中央値（範囲）３，０００（１，２００～４２０，０００）を得ることができる。これらの異常なメチル化パターンを持つ断片のフィルターセットは、セクションＩＩＩで後述するように、下流の解析に使用することができる。

【０２０２】

いくつかの実施形態では、解析システムは、メチル化状態ベクターの可能性を決定し、ｐ値を計算するためにスライディングウィンドウを使用する（ステップ４５５）。可能性を列挙し、メチル化状態ベクター全体についてｐ値を計算するのではなく、解析システムは、可能性を列挙し、連続するＣｐＧ部位のウィンドウのみについてｐ値を計算することができ、ウィンドウは、少なくともいくつかの断片よりも（ＣｐＧ部位の）長さが短い（そうでなければ、ウィンドウは目的を果たさない）。ウィンドウの長さは、静的、使用者による決定、動的、またはその他の方法で選択することができる。

【０２０３】

ウィンドウより大きいメチル化状態ベクターについてｐ値を計算する場合、ウィンドウは、ベクター内の最初のＣｐＧ部位から始まるウィンドウ内のベクターからのＣｐＧ部位の連続セットを識別することができる。解析システムは、最初のＣｐＧ部位を含むウィンドウのｐ値スコアを計算することができる。解析システムは次に、ウィンドウをベクター内の第２のＣｐＧ部位まで「スライド（ｓｌｉｄｅ）」させ、第２のウィンドウについて別のｐ値スコアを計算することができる。このように、ウィンドウサイズｌとメチル化ベクター長ｍに対して、各メチル化状態ベクターはｍ－ｌ＋１のｐ値スコアを生成することができる。ベクターの各部分についてｐ値計算を完了した後、全てのスライディングウィンドウから最も低いｐ値スコアを、メチル化状態ベクターの全体的なｐ値スコアとすることができる。他の実施形態では、解析システムは、メチル化状態ベクターのｐ値スコアを集約して、総合ｐ値スコアを生成する。

【０２０４】

スライディングウィンドウを使用することで、メチル化状態ベクターの列挙された可能

性および、それに対応する確率計算の数を減らすことができる。現実的な例を挙げると、５４個以上のＣｐＧ部位を持つ断片がある。単一のｐスコアを生成するために２＾５４（～１．８×１０＾１６）の可能性について確率を計算する代わりに、分析システムは、その断片のメチル化状態ベクターの５０のウィンドウのそれぞれについて５０のｐ値計算をもたらすサイズ５のウィンドウ（例えば）を使用することができる。５０の計算のそれぞれは、２＾５（３２）のメチル化状態ベクターの可能性を列挙することができ、合計で５０×２＾５（１．６×１０＾３）の確率計算になる。その結果、実行すべき計算が大幅に削減され、異常断片の正確な同定には意味がなくなる。

【０２０５】

不確定状態を有する実施形態では、解析システムは、断片のメチル化状態ベクター中の不確定状態を有するＣｐＧ部位を合計したｐ値スコアを計算することができる。解析システムは、不確定状態を除くメチル化状態ベクターの全てのメチル化状態とコンセンサスを有する全ての可能性を同定することができる。解析システムは、特定された可能性の確率の合計として、メチル化状態ベクターに確率を割り当てることができる。一例として、解析システムは、ＣｐＧ部位１および３に対するメチル化状態が観察され、ＣｐＧ部位１および３における断片のメチル化状態とコンセンサスが得られているので、＜M_1、M_2、U_3＞および＜M_1、U_2、U_3＞のメチル化状態ベクターの可能性の確率の合計として、＜M_1、I_2、U_3＞のメチル化状態ベクターの確率を計算することができる。不確定状態を有するＣｐＧ部位を合計するこの方法は、２＾ｉまでの可能性の確率の計算を使用することができ、ここでｉは、メチル化状態ベクター中の不確定状態の数を示す。追加の実施形態において、動的プログラミングアルゴリズムが、１つ以上の不確定状態を有するメチル化状態ベクターの確率を計算するために実施され得る。

【０２０６】

いくつかの実施形態では、確率および／またはｐ値スコアを計算する計算負荷は、少なくとも一部の計算をキャッシュすることによってさらに低減され得る。例えば、解析システムは、メチル化状態ベクター（またはそのウィンドウ）の可能性に関する確率の計算を、一時的または永続的なメモリにキャッシュすることができる。他の断片が同じＣｐＧ部位を有する場合、可能性確率をキャッシュすることにより、基礎となる可能性確率を再計算する必要なく、ｐスコア値を効率的に計算することができる。同様に、解析システムは、ベクター（またはそのウィンドウ）からのＣｐＧ部位のセットに関連するメチル化状態ベクターの可能性の各々についてｐ値スコアを計算することができる。解析システムは、同じＣｐＧ部位を含む他の断片のｐ値スコアを決定する際に使用するために、ｐ値スコアをキャッシュしてもよい。一般に、同じＣｐＧ部位を有するメチル化状態ベクターの可能性のｐ値スコアは、同じＣｐＧ部位のセットからの可能性の異なるもののｐ値スコアを決定するために使用され得る。

【０２０７】

１つ以上の核酸メチル化断片は、領域モデルまたはがん分類器を訓練する前にフィルタリングすることができる。核酸メチル化断片をフィルタリングすることは、対応する複数の核酸メチル化断片から、１つ以上の選択基準（例えば、１つの選択基準を下回るまたは上回る）を満たさないそれぞれの核酸メチル化断片を除去することを含み得る。１つ以上の選択基準は、ｐ値閾値を含み得る。それぞれの核酸メチル化断片の出力ｐ値は、それぞれの核酸メチル化断片の対応するメチル化パターンと、それぞれの核酸メチル化断片の対応する複数のＣｐＧ部位を有する健康な非がんコホートデータセット中のそれらの核酸メチル化断片のメチル化パターンの対応する分布との比較に基づいて、少なくとも部分的に決定され得る。

【０２０８】

複数の核酸メチル化断片をフィルタリングすることは、ｐ値閾値を満足しないそれぞれの核酸メチル化断片を除去することを含み得る。フィルターは、第１の複数の核酸メチル化断片にわたって観察されたメチル化パターンを用いて、それぞれの核酸メチル化断片のそれぞれのメチル化パターンに適用することができる。それぞれの核酸メチル化断片（例

えば、断片１、．．．、断片Ｎ）のそれぞれのメチル化パターンは、メチル化部位の識別子と対応するメチル化パターンで識別され、１と０の配列で表され、対応する１つ以上のメチル化部位（例えば、ＣｐＧ部位）を含み、ここで、各「１」は、１つ以上のＣｐＧ部位におけるメチル化ＣｐＧ部位を表し、各「０」は、１つ以上のＣｐＧ部位における非メチル化ＣｐＧ部位を表す。第１の複数の核酸メチル化断片にわたって観察されたメチル化パターンは、第１の複数の核酸メチル化断片によって集合的に表されるＣｐＧ部位状態（例えば、ＣｐＧ部位Ａ、ＣｐＧ部位Ｂ、．．．、ＣｐＧ部位ＺＺＺ）についてのメチル化状態分布を構築するために使用することができる。核酸メチル化断片の処理に関するさらなる詳細は、２０２０年３月４日に出願された「オートエンコーダを用いたがん状態判定システムおよび方法」と題する米国仮特許出願第６２／９８５，２５８号に開示されており、その全体が参照により本明細書に組み込まれる。

【０２０９】

それぞれの核酸メチル化断片が、異常メチル化スコア閾値未満の異常メチル化スコアを有する場合、それぞれの核酸メチル化断片は、１つ以上の選択基準における選択基準を満足しないことがある。この状況において、異常メチル化スコアは、混合モデルによって決定され得る。例えば、混合モデルは、同じ長さで同じ対応するゲノム位置の可能なメチル化状態ベクターの数に基づいて、それぞれの核酸メチル化断片のメチル化状態ベクター（例えば、メチル化パターン）の尤度を決定することによって、核酸メチル化断片の異常なメチル化パターンを検出することができる。これは、参照ゲノムの各ゲノム位置において、指定された長さのベクターについて複数の可能なメチル化状態を生成することによって実行することができる。複数の可能性のあるメチル化状態を用いて、可能性のあるメチル化状態の総数と、それに続くゲノム位置における各予測メチル化状態の確率を決定することができる。次に、参照ゲノム内のゲノム位置に対応するサンプル核酸メチル化断片の尤度を、サンプル核酸メチル化断片を予測された（例えば、可能性のある）メチル化状態にマッチングさせ、予測されたメチル化状態の計算された確率を検索することによって決定することができる。その後、サンプル核酸メチル化断片の確率に基づいて異常メチル化スコアを計算することができる。

【０２１０】

それぞれの核酸メチル化断片は、それぞれの核酸メチル化断片の残基数が閾値未満である場合、１つ以上の選択基準における選択基準を満たすことができない。残基の閾値数は、１０～５０、５０～１００、１００～１５０、または１５０以上であり得る。残基数の閾値は、２０～９０の固定値とすることができる。それぞれの核酸メチル化断片が閾値未満のＣｐＧ部位を有する場合、それぞれの核酸メチル化断片は、１つ以上の選択基準における選択基準を満たさないことがある。ＣｐＧ部位の閾値数は、４、５、６、７、８、９、または１０であり得る。それぞれの核酸メチル化断片は、それぞれの核酸メチル化断片のゲノム開始位置およびゲノム終了位置が、それぞれの核酸メチル化断片がヒトゲノム参照配列において閾値未満のヌクレオチド数を表すことを示す場合、１つ以上の選択基準における選択基準を満たすことができない。

【０２１１】

フィルタリングは、対応する複数の核酸メチル化断片中の核酸メチル化断片のうち、対応する複数の核酸メチル化断片中の別の核酸メチル化断片と同じ対応するメチル化パターンを有し、同じ対応するゲノム開始位置およびゲノム終了位置を有する核酸メチル化断片を除去することができる。このフィルタリングステップは、場合によってはＰＣＲ重複を含む、完全な重複である冗長な断片を除去することができる。このフィルタリングにより、対応する複数の核酸メチル化断片中の別の核酸メチル化断片と、対応するゲノム開始位置およびゲノム終了位置が同じで、異なるメチル化状態の数が閾値未満である核酸メチル化断片を除去することができる。核酸メチル化断片の保持に使用される異なるメチル化状態の閾値数は、１、２、３、４、５、または５以上であり得る。例えば、第２の核酸メチル化断片と同じ対応するゲノムの開始および終了位置を有するが、それぞれのＣｐＧ部位（例えば、参照ゲノムにアライメントされる）において少なくとも１、少なくとも２、少

なくとも３、少なくとも４、または少なくとも５の異なるメチル化状態を有する第１の核酸メチル化断片が保持される。別の例として、同じメチル化状態ベクターを有するが（例えば、メチル化パターン）、第２の核酸メチル化断片と対応するゲノムの開始および終了位置が異なる第１の核酸メチル化断片も保持される。

【０２１２】

フィルタリングは、複数の核酸メチル化断片中のアッセイアーチファクトを除去することができる。アッセイアーチファクトの除去は、配列決定されたハイブリダイゼーションプローブから得られた配列リードおよび／またはバイサルファイト変換中に変換を受けなかった配列から得られた配列リードを除去することを含み得る。フィルタリングは、汚染物質を除去し得る（例えば、配列決定、核酸単離、および／またはサンプル調製に起因する）。

【０２１３】

フィルタリングは、複数の訓練被験者にわたるがん状態に対するそれぞれのメチル化断片の相互情報フィルタリングに基づいて、複数のメチル化断片からメチル化断片のサブセットを除去することができる。例えば、相互情報は、同時にサンプリングされた関心のある２つの状態間の相互依存性の尺度を提供することができる。相互情報は、１つ以上のデータセットからＣｐＧ部位の独立したセット（例えば、核酸メチル化断片の全部または一部内）を選択し、２つのサンプルグループ間（例えば、遺伝子型データセットのサブセットおよび／またはグループ、生物学的サンプル、および／または被験者）でＣｐＧ部位のセットに対するメチル化状態の確率を比較することによって決定することができる。相互情報スコアは、スライディングウィンドウのそれぞれのフレームにおけるそれぞれの領域における第１の条件対第２の条件のメチル化パターンの確率を示すことができ、したがって、それぞれの領域の識別力を示す。相互情報スコアは、選択されたＣｐＧ部位のセットおよび／または選択されたゲノム領域を横切って進行するにつれて、スライディングウィンドウの各フレームの各領域について同様に計算され得る。相互情報フィルタリングに関するさらなる詳細は、２０２０年１２月１１日に出願された「パッチ畳み込みニューラルネットワークを用いたがん分類」と題する米国特許出願１７／１１９，６０６号に開示されており、その全体が参照により本明細書に組み込まれる。

【０２１４】

＜ＩＩ．Ｃ．ＩＩ．過剰メチル化断片および過剰メチル化断片＞

いくつかの実施形態では、解析システムは、ＣｐＧ部位の数が閾値以上であり、ＣｐＧ部位のメチル化率が閾値以上であるか、またはＣｐＧ部位の非メチル化率が閾値以上である断片として異常断片を決定する。断片（またはＣｐＧ部位）の長さの閾値の例としては、３以上、４以上、５以上、６以上、７以上、８以上、９以上、１０以上などがある。メチル化または非メチル化のパーセンテージの閾値の例としては、８０％以上、８５％以上、９０％以上、９５％以上、または５０％～１００％の範囲内の他の割合が挙げられる。

【０２１５】

＜ＩＩ．Ｄ．例示的解析システム＞

図６Ａは、１つ以上の実施形態による核酸サンプルを配列決定するための装置の例示的なフローチャートである。この例示的なフローチャートは、シーケンサー６２０および解析システム６００などのデバイスを含む。シーケンサー６２０および解析システム６００は、図３Ａのプロセス３００、図４Ａのプロセス４００、図４Ｂのステップ４２０、および本明細書に記載される他のステップにおける１つ以上のステップを実行するために、連動して動作し得る。

【０２１６】

様々な実施形態において、シーケンサー６２０は、濃縮核酸サンプル６１０を受け取る。図６Ａに示すように、シーケンサー６２０は、特定のタスク（例えば、シーケンシングの開始またはシーケンシングの終了）とのユーザー対話を可能にするグラフィカルユーザーインターフェース６２５、ならびに濃縮断片サンプルを含むシーケンシングカートリッジをロードするため、および／またはシーケンシングアッセイを実行するために必要なバ

ッファーをロードするための１つ以上のロードステーション６３０を含むことができる。したがって、シーケンサー６２０のユーザーがシーケンサー６２０のローディングステーション６３０に必要な試薬およびシーケンシングカートリッジを提供すると、ユーザーはシーケンサー６２０のグラフィカルユーザーインターフェース６２５と対話することによってシーケンシングを開始することができる。一旦開始されると、シーケンサー６２０はシーケンシングを実行し、核酸サンプル６１０からの濃縮断片の配列リードを出力する。

【０２１７】

いくつかの実施形態では、シーケンサー６２０は、解析システム６００と通信可能に結合される。解析システム６００は、１つ以上のＣｐＧ部位におけるメチル化状態の評価、バリアントコールまたは品質管理などの様々な用途のために配列リードを処理するために使用されるいくつかのコンピューティングデバイスを含む。シーケンサー６２０は、配列リードをＢＡＭファイル形式で解析システム６００に提供することができる。分析システム６００は、無線、有線、または無線および有線の組み合わせの通信技術を通じてシーケンサー６２０に通信可能に結合され得る。一般に、解析システム６００は、プロセッサおよび、プロセッサによって実行されると、プロセッサに配列リードを処理させるか、または本明細書に開示される方法もしくはプロセスのいずれかの１つ以上のステップを実行させるコンピュータ命令を記憶する非一過性のコンピュータ可読記憶媒体を備えるように構成される。

【０２１８】

いくつかの実施形態では、配列リードは、例えば、図３Ａのプロセス３００のうちのステップ３４０を介して、アライメント位置情報を決定するために、当該分野で公知の方法を使用して参照ゲノムにアライメントされ得る。アライメント位置は、一般に、所定の配列リードの開始ヌクレオチド塩基および終了ヌクレオチド塩基に対応する参照ゲノム内の領域の開始位置および終了位置を記述し得る。メチル化配列決定に対応して、アライメント位置情報は、参照ゲノムへのアライメントに従って配列リードに含まれる最初のＣｐＧ部位および最後のＣｐＧ部位を示すように一般化されてもよい。アライメント位置情報はさらに、所定の配列リード内の全てのＣｐＧ部位のメチル化状態および位置を示す場合がある。参照ゲノム内の領域は、遺伝子または遺伝子のセグメントと関連付けられることがあり、そのようなものとして、解析システム６００は、配列リードに、その配列リードにアライメントする１つ以上の遺伝子をラベル付けすることがある。一実施形態では、断片の長さ（またはサイズ）は、開始位置および終了位置から決定される。

【０２１９】

様々な実施形態において、例えば、ペアエンド配列決定プロセスが使用される場合、配列リードは、Ｒ＿１およびＲ＿２と示されるリードペアで構成される。例えば、第１のリードＲ＿１は、二本鎖ＤＮＡ（ｄｓＤＮＡ）分子の第１の末端から配列決定される場合があり、一方、第２のリードＲ＿２は、二本鎖ＤＮＡ（ｄｓＤＮＡ）の第２の末端から配列決定される場合がある。したがって、第１のリードＲ＿１および第２のリードＲ＿２のヌクレオチド塩基対は、参照ゲノムのヌクレオチド塩基と一貫して（例えば、反対方向に）整列される可能性がある。リードペアＲ＿１およびＲ＿２から得られるアライメント位置情報には、第１のリード（例えば、Ｒ＿１）の末端に対応する参照ゲノム内の開始位置と、第２のリード（例えば、Ｒ＿２）の末端に対応する参照ゲノム内の終了位置とが含まれる場合がある。言い換えれば、参照ゲノムにおける開始位置と終了位置は、核酸断片が対応する参照ゲノム内の可能性の高い位置を表すことができる。ＳＡＭ（配列アライメントマップ）形式またはＢＡＭ（二値）形式の出力ファイルが生成され、さらなる解析のために出力される。

【０２２０】

次に図６Ｂを参照すると、図６Ｂは、一実施形態によるＤＮＡサンプルを処理するための解析システム６００のブロック図である。解析システムは、ＤＮＡサンプルの解析に使用するための１つ以上のコンピューティングデバイスを実装する。解析システム６００は、配列プロセッサ６４０、配列データベース６４５、モデルデータベース６５５、モデル

６５０、パラメータデータベース６６５、およびスコアエンジン６６０を含む。いくつかの実施形態において、解析システム６００は、図３Ａのプロセス３００および図４Ａのプロセス４００の一部または全部を実行する。

【０２２１】

配列プロセッサ６４０は、サンプルからの断片についてメチル化状態ベクターを生成する。断片上の各ＣｐＧ部位において、配列プロセッサ６４０は、図３Ａのプロセス３００を介して、参照ゲノムにおける断片の位置、断片中のＣｐＧ部位の数、および断片中の各ＣｐＧ部位のメチル化、非メチル化、または不確定のいずれかのメチル化状態を指定する、各断片のメチル化状態ベクターを生成する。配列プロセッサ６４０は、配列データベース６４５に断片のメチル化状態ベクターを格納することができる。配列データベース６４５のデータは、サンプルからのメチル化状態ベクターが互いに関連付けられるように組織化されてもよい。

【０２２２】

さらに、複数の異なるモデル６５０が、モデルデータベース６５５に保存されるか、またはテストサンプルと共に使用するために検索されることがある。一実施例では、モデルは、異常断片から導出された特徴ベクターを使用して、テストサンプルのがん予測を決定するための訓練されたがん分類器である。がん分類器の訓練と使用については、「セクションＩＩＩ．がんを判定するためのがん分類器」に関連してさらに説明する。解析システム６００は、１つ以上のモデル６５０を訓練し、様々な訓練されたパラメータをパラメータデータベース６６５に格納することができる。解析システム６００は、モデル６５０を機能と共にモデルデータベース６５５に格納する。

【０２２３】

推論の間、スコアエンジン６６０は、出力を返すために１つ以上のモデル６５０を使用する。スコアエンジン６６０は、パラメータデータベース６６５からの訓練されたパラメータと共に、モデルデータベース６５５内のモデル６５０にアクセスする。各モデルに従って、スコアエンジンはモデルのための適切な入力を受信し、受信した入力、パラメータ、および入力と出力とを関連付ける各モデルの関数に基づいて出力を計算する。一部の使用例では、スコアエンジン６６０はさらに、モデルから計算された出力に対する信頼度に関連するメトリクスを計算する。他の使用例では、スコアエンジン６６０はモデルで使用するための他の中間値を計算する。

【０２２４】

［ＩＩＩ．がんを判定するためのがん分類器］

＜ＩＩＩ．Ａ．概要＞

がん分類器は、テストサンプルの特徴ベクターを受け取り、テストサンプルががん、より具体的には特定のがんタイプを有するテスト対象からのものであるか否かを決定するように訓練することができる。がん分類器は、複数の分類パラメータおよび、入力としての特徴ベクターと、分類パラメータを用いて入力特徴ベクター上で動作する関数によって決定される出力としてのがん予測との間の関係を表す関数とを含むことができる。一部の実施形態では、がん分類器に入力する特徴ベクターは、試験サンプルから決定された異常断片のセットに基づく。異常断片は、図４Ｂのステップ４２０を介して決定されてもよく、より具体的には、ステップ４２０のステップ４７０を介して決定される高メチル化断片および低メチル化断片、または何らかの他のステップに従って決定される異常断片であってもよい。がん分類器を配備する前に、解析システムはがん分類器を訓練することができる。

【０２２５】

＜ＩＩＩ．Ｂ．がん分類器の訓練＞

図５Ａは、実施形態による、がん分類器を訓練するプロセス５００を説明するフローチャートである。解析システムは、異常断片のセットおよびがんタイプのラベルを各々が有する複数の訓練サンプルを取得する（ステップ５１０）。複数の訓練サンプルは、「非がん（non-cancer）」という一般的なラベルを有する健常人からのサンプル、「がん（c

ancer）」という一般的なラベルを有する被験者からのサンプル、または特定のラベル（例えば、「乳がん（breast cancer）」、「肺がん（lung cancer）」など）を有する被験者からのサンプルの任意の組み合わせを含むことができる。あるがん種の被験者からの訓練サンプルは、そのがん種のコホートまたはがん種コホートと呼ばれることがある。解析システムは、がん分類器の訓練に使用される訓練サンプルが汚染されていないことを保証することができる。訓練サンプルが汚染されているかどうかを決定するために、解析システムは図１のプロセス１００を実行することができる。

【０２２６】

解析システムは、各訓練サンプルについて、訓練サンプルの異常断片セットに基づく特徴ベクターを決定する（ステップ５２０）。解析システムは、ＣｐＧ部位の初期セット内の各ＣｐＧ部位について異常スコアを計算することができる。ＣｐＧ部位の初期セットは、ヒトゲノム中の全てのＣｐＧ部位またはその一部であってもよく、10^4、10^5、10^6、10^7、10^8などのオーダーであってもよい。一実施形態では、解析システムは、異常断片セット中にＣｐＧ部位を包含する異常断片が存在するか否かに基づいて、二値スコアリングで特徴ベクターの異常スコアを定義する。別の実施形態では、解析システムは、ＣｐＧ部位と重なる異常断片のカウントに基づいて異常スコアを定義する。一例では、解析システムは、異常断片が存在しない場合の第１のスコア、少数の異常断片が存在する場合の第２のスコア、および少数以上の異常断片が存在する場合の第３のスコアを割り当てる３値スコアリングを使用することができる。例えば、解析システムは、ＣｐＧ部位と重なるサンプル中の異常断片を５個カウントし、５個のカウントに基づいて異常スコアを算出する。

【０２２７】

一旦全ての異常スコアが訓練サンプルについて決定されると、分析システムは、各要素について、初期セット内のＣｐＧ部位の１つに関連する異常スコアの１つを含む要素のベクターとして、特徴ベクターを決定することができる。解析システムは、サンプルのカバレッジに基づいて、特徴ベクターの異常スコアを正規化することができる。ここで、カバレッジは、分類器に使用されるＣｐＧ部位の初期セットによってカバーされる全てのＣｐＧ部位にわたる、または所与の訓練サンプルについての異常断片セットに基づく、シーケンス深度の中央値または平均値を指すことができる。

【０２２８】

例として、訓練特徴ベクターの行列（ステップ５２２）を示す図５Ｂを参照する。この例では、解析システムは、がん分類器のための特徴ベクターを生成する際に考慮するＣｐＧ部位［Ｋ］（ステップ５２６）を同定した。解析システムは、訓練サンプル［Ｎ］（ステップ５２４）を選択する。解析システムは、訓練サンプル［ｎ１］の特徴ベクターにおいて使用される第１の任意のＣｐＧ部位［ｋ１］の第１の異常スコア（ステップ５２８）を決定する。解析システムは、異常断片のセット内の各異常断片をチェックする。解析システムが、第１のＣｐＧ部位を含む少なくとも１つの異常断片を識別した場合、解析システムは、図５Ｂに示されるように、第１のＣｐＧ部位に対する第１の異常スコア（ステップ５２８）を１として決定する。第２の任意のＣｐＧ部位［ｋ２］を考慮すると、解析システムは同様に、第２のＣｐＧ部位［ｋ２］を含む少なくとも１つの異常断片のセットをチェックする。解析システムが、第２のＣｐＧ部位を含むそのような異常断片を発見しない場合、解析システムは、図５Ｂに示されるように、第２のＣｐＧ部位［ｋ２］に対する第２の異常スコア（ステップ５２９）を０に決定する。解析システムがＣｐＧ部位の初期セットに対する全ての異常スコアを決定すると、解析システムは、第１のＣｐＧ部位［ｋ１］に対する１の第１の異常スコア（ステップ５２８）、第２のＣｐＧ部位［ｋ２］に対する０の第２の異常スコア（ステップ５２９）、およびそれ以降の異常スコアとを含む異常スコアを含む第１のトレーニングサンプル［ｎ１］に対する特徴ベクターを決定し、これにより特徴ベクター［１、０、．．．］を形成する。

【０２２９】

サンプルの特徴づけに対するその他のアプローチは、以下に記載されている。「モデル

ベースのフィーチャライゼーションおよび分類」と題する米国出願第１５／９３１，０２２号、「ターゲット配列決定のための混合モデル」と題する米国出願第１６／５７９，８０５号、「異常断片の検出および分類」と題する米国出願第１６／３５２，６０２号、および「セルフリーＤＮＡサンプルのメチル化断片に基づく起源判定」と題する米国出願第１６／７２３，７１６号は、全て参照によりその全体が組み込まれる。

【０２３０】

解析システムは、がん分類器において使用するために考慮されるＣｐＧ部位をさらに制限することができる。解析システムは、ＣｐＧ部位の初期セット内の各ＣｐＧ部位について、訓練サンプルの特徴ベクターに基づいて情報利得を計算する（ステップ５３０）。ステップ５２０から、各訓練サンプルは、ＣｐＧ部位の初期セットにおける全てのＣｐＧ部位の異常スコアを含む特徴ベクターを有する。しかし、ＣｐＧ部位の初期セット中の一部のＣｐＧ部位は、がん種を区別する上で他のＣｐＧ部位ほど有益でない可能性があり、あるいは他のＣｐＧ部位と重複している可能性がある。

【０２３１】

一実施形態では、解析システムは、がんの種類ごとに、また初期セット内のＣｐＧ部位ごとに情報利得を計算し、そのＣｐＧ部位を分類器に含めるかどうかを決定する（ステップ５３０）。情報利得は、所与のがん種を有する訓練サンプルについて、他の全てのサンプルと比較して計算される。例えば、２つの確率変数「異常断片（anomalous fragment）」（「ＡＦ」）および「がんタイプ（cancer type）」（「ＣＴ」）が使用される。一実施形態では、ＡＦは、上記の異常スコア／特徴ベクターについて決定されたように、所与のサンプルにおいて所与のＣｐＧ部位に重なる異常断片が存在するか否かを示す二値変数である。ＣＴは、がんが特定のタイプであるかどうかを示すランダム変数である。解析システムは、ＡＦが与えられたＣＴに関する相互情報を計算する。すなわち、特定のＣｐＧ部位に重なる異常断片があるかどうかがわかれば、がんのタイプについて何ビットの情報が得られるかを計算する。実際には、第一のがん種について、解析システムは他の各がん種に対する一対の相互情報利得を計算し、他の全てのがん種にわたる相互情報利得を合計する。

【０２３２】

所与のがん型について、解析システムはこの情報を使用して、ＣｐＧ部位がどの程度がんに特異的であるかに基づいてＣｐＧ部位をランク付けすることができる。この手順は、検討中の全てのがん種について繰り返すことができる。ある特定の領域が、あるがんの訓練用サンプルでは一般的に異常メチル化されているが、他のがん型の訓練用サンプルまたは健康な訓練用サンプルではメチル化されていない場合、それらの異常断片と重なるＣｐＧ部位は、あるがん型に対して高い情報利得を持つ可能性がある。各がん種のランク付けされたＣｐＧ部位は、がん分類器で使用するために、そのランクに基づいて選択されたＣｐＧ部位セットに貪欲に追加（選択）することができる（ステップ５４０）。

【０２３３】

追加の実施形態では、解析システムは、がん分類器で使用される有益なＣｐＧ部位を選択するための他の選択基準を考慮することができる。一つの選択基準は、選択されたＣｐＧ部位が他の選択されたＣｐＧ部位から閾値以上の分離を有することである。例えば、選択されたＣｐＧ部位は、他の選択されたＣｐＧ部位から閾値以上の塩基対数（例えば、１００塩基対）離れていることであり、閾値以内の分離であるＣｐＧ部位は、がん分類器における検討のために両方とも選択されないようにする。

【０２３４】

一実施形態では、初期セットから選択されたＣｐＧ部位のセットに従って、解析システムは、必要に応じて、訓練サンプルの特徴ベクターを修正することができる（ステップ５５０）。例えば、解析システムは、選択されたＣｐＧ部位のセットに含まれないＣｐＧ部位に対応する異常スコアを除去するために、特徴ベクターを切り捨てることができる。

【０２３５】

解析システムは、訓練サンプルの特徴ベクターを用いて、多数の方法のいずれかでがん

分類器を訓練することができる。特徴ベクターは、ステップ５２０からのＣｐＧ部位の初期セット、またはステップ５５０からのＣｐＧ部位の選択セットに対応することができる。一実施形態では、解析システムは、訓練サンプルの特徴ベクターに基づいてがんと非がんを区別するために、二値がん分類器を訓練する（ステップ５６０）。このように、解析システムは、健常人からの非がんサンプルおよび被験者からのがんサンプルの両方を含む訓練サンプルを使用する。各訓練サンプルは、「がん（cancer）」または「非がん（non-cancer）」の２つのラベルのいずれかを有することができる。この実施形態では、分類器はがんの存在または非存在の可能性を示すがん予測を出力する。

【０２３６】

別の実施形態では、解析システムは、多くのがん種（原発組織（ＴＯＯ）ラベルとも呼ばれる）を区別するために、マルチクラスがん分類器を訓練する（ステップ５７０）。がんタイプは、１つ以上のがんを含むことができ、非がんタイプを含むこともできる（任意の追加の他の疾患または遺伝的障害なども含むことができる）。そうするために、解析システムは、がん型コホートを使用することができ、また、非がん型コホートを含んでも含まなくてもよい。この多がんの実施形態では、がん分類器は、分類されるがん型の各々についての予測値からなるがん予測（より具体的には、ＴＯＯ予測）を決定するように訓練される。予測値は、所与の訓練サンプル（および推論中のテストサンプル）が各がん種を有する尤度に対応し得る。ある実施態様では、予測値は０～１００でスコア化され、予測値の累積は１００に等しい。例えば、がん分類器は、乳がん、肺がん、非がんの予測値を含むがん予測を返す。例えば、分類器は、試験サンプルが乳がんの可能性が６５％、肺がんの可能性が２５％、非がんの可能性が１０％であるというがん予測を返すことができる。解析システムは、さらに予測値を評価して、１つ以上のＴＯＯラベル、例えば、最も高い予測値を有する第１のＴＯＯラベル、２番目に高い予測値を有する第２のＴＯＯラベルなどを示すＴＯＯ予測とも呼ばれ得る、サンプル中の１つ以上のがんの存在予測を生成することができる。上記の例を継続し、割合が与えられた場合、この例では、システムは、乳がんが最も高い可能性を有することから、サンプルが乳がんに罹患していると判定することができる。

【０２３７】

いずれの実施形態においても、解析システムは、特徴ベクターを有する訓練サンプルのセットをがん分類器に入力し、分類器の関数が訓練特徴ベクターを対応するラベルに正確に関連付けるように分類パラメータを調整することによって、がん分類器を訓練する。解析システムは、がん分類器の反復バッチ訓練のために、訓練サンプルを１つ以上の訓練サンプルのセットにグループ化することができる。訓練特徴ベクターを含む訓練サンプルの全てのセットを入力し、分類パラメータを調整した後、がん分類器を十分に訓練して、ある程度の誤差の範囲内でその特徴ベクターに従ってテストサンプルをラベル付けすることができる。解析システムは、多数の方法のいずれか１つに従ってがん分類器を訓練することができる。一例として、二値がん分類器は、対数損失関数を用いて訓練されるＬ２－正則化ロジスティック回帰分類器であってもよい。別の例として、多がん分類器は多項ロジスティック回帰であってもよい。実際には、いずれのタイプのがん分類器も、他の技術を用いて訓練することができる。これらの技法は、カーネル法、ランダムフォレスト分類器、混合モデル、オートエンコーダモデル、多層ニューラルネットワークなどの機械学習アルゴリズムなどを使用する可能性があるなど、多数ある。

【０２３８】

分類器には、ロジスティック回帰アルゴリズム、ニューラルネットワークアルゴリズム、サポートベクターマシンアルゴリズム、ナイーブベイズアルゴリズム、最近傍アルゴリズム、ブースティング木アルゴリズム、ランダムフォレストアルゴリズム、決定木アルゴリズム、多項ロジスティック回帰アルゴリズム、線形モデル、または線形回帰アルゴリズムを含めることができる。

【０２３９】

＜ＩＩＩ．Ｃ．がん分類器の展開＞

がん分類装置の使用中、解析システムは、がんの種類が未知の被験者から試験サンプルを得ることができる。解析システムは、プロセス３００、４００、および４２０の任意の組み合わせでＤＮＡ分子から構成される試験サンプルを処理して、異常断片のセットを得ることができる。解析システムは、プロセス５００で議論された同様の原理に従って、がん分類器で使用するための試験特徴ベクターを決定することができる。解析システムは、がん分類器によって使用される複数のＣｐＧ部位中の各ＣｐＧ部位について異常スコアを計算することができる。例えば、がん分類器は、１，０００個の選択されたＣｐＧ部位の異常スコアを含む特徴ベクターを入力として受け取る。したがって、解析システムは、異常断片のセットに基づいて、１，０００個の選択されたＣｐＧ部位の異常スコアを含むテスト特徴ベクターを決定することができる。解析システムは、訓練サンプルと同じ方法で異常スコアを計算することができる。一部の実施形態では、解析システムは、異常断片のセット中にＣｐＧ部位を包含する高メチル化断片または低メチル化断片が存在するか否かに基づいて、異常スコアを二値スコアとして定義する。

【０２４０】

分析システムは次に、試験特徴ベクターをがん分類器に入力することができる。その後、がん分類器の機能は、プロセス５００で訓練された分類パラメータおよび試験特徴ベクターに基づいてがん予測を生成することができる。第一の態様において、がん予測は二値であり、「がん（cancer）」または「非がん（non-cancer）」からなる群から選択され得、第二の態様において、がん予測は多くのがん種および「非がん（non-cancer）」からなる群から選択される。付加的な実施形態において、がん予測は、多数のがんタイプの各々についての予測値を有する。さらに、解析システムは、検査サンプルががんの種類の一つである可能性が最も高いと判定することができる。検査サンプルのがん予測が、乳がんの可能性６５％、肺がんの可能性２５％、非がんの可能性１０％である上記の例に従うと、解析システムは検査サンプルが乳がんに罹患している可能性が最も高いと判定することができる。別の例では、がん予測が、がんでない可能性６０％、がんの可能性４０％という二値である場合、解析システムは、検査サンプルががんでない可能性が最も高いと判定する。追加の実施形態において、最も高い尤度を有するがん予測は、被験体をそのがん種を有すると呼ぶために、やはり閾値（例えば、４０％、５０％、６０％、７０％）と比較され得る。最も高い尤度を有するがん予測がその閾値を超えない場合、解析システムは決定的でない結果を返すことがある。

【０２４１】

追加の実施形態では、解析システムは、プロセス５００のステップ５６０で訓練されたがん分類器を、プロセス５００のステップ５７０またはプロセス５００で訓練された別のがん分類器と連鎖させる。解析システムは、プロセス５００のステップ５６０において二値分類器として訓練されたがん分類器に試験特徴ベクターを入力することができる。解析システムはがん予測の出力を受け取ることができる。がん予測は、検査対象ががんに罹患している可能性が高いか、またはがんに罹患していない可能性が高いかについての二値であってもよい。他の実施態様では、がん予測は、がんの可能性および非がんである可能性を記述する予測値を含む。例えば、がん予測は８５％のがん予測値および１５％の非がん予測値を有する。解析システムは、被験者ががんである可能性が高いと判定することができる。解析システムが、被験者ががんである可能性が高いと判定すると、解析システムは、異なるがんタイプを区別するように訓練された多クラスがん分類器に試験特徴ベクターを入力することができる。多クラスがん分類器は、試験特徴ベクターを受け取り、複数のがんタイプのがんタイプのがん予測を返すことができる。例えば、多クラスがん分類器は、被験体が卵巣がんである可能性が最も高いことを指定するがん予測を提供する。別の実施態様では、多クラスがん分類器は、複数のがん種の各がん種の予測値を提供する。例えば、がん予測値は、４０％の乳がんタイプ予測値、１５％の大腸がんタイプ予測値、および４５％の肝臓がん予測値を含むことができる。

【０２４２】

二値がん分類の一般化された実施形態によれば、解析システムは、試験サンプルの配列

決定データ（例えば、メチル化配列決定データ、ＳＮＰ配列決定データ、他のＤＮＡ配列決定データ、ＲＮＡ配列決定データなど）に基づいて、試験サンプルのがんスコアを決定することができる。解析システムは、サンプルががんである可能性が高いかどうかを予測するための二値閾値カットオフに対して、試験サンプルのがんスコアを比較することができる。二値閾値カットオフは、１つ以上のＴＯＯサブタイプクラスに基づくＴＯＯ閾値化を用いて調整することができる。解析システムはさらに、１つ以上の可能性の高いがん種を示すがん予測を決定するために、多クラスがん分類器において使用するための試験サンプルの特徴ベクターを生成することができる。

【０２４３】

この分類器は、被験体、例えば疾患状態が未知の被験体の疾患状態を決定するために使用され得る。本方法は、被験体から得られた生物学的サンプル中の対応する複数の核酸断片の複数のゲノム特性中の各ゲノム特性についての値を含む、試験ゲノムデータ構築物（例えば、単一時点試験データ）を電子的形態で取得することを含み得る。次いで、本方法は、試験ゲノムデータ構成体を試験分類器に適用し、それにより被験体における疾患状態の状態を決定することを含み得る。被験体は、以前に疾患状態と診断されたことがない場合がある。

【０２４４】

分類器は、少なくとも（ｉ）第１の時点で被験体から取得された第１の生物学的サンプルから生成された第１の試験ゲノムデータ構成体、および（ｉｉ）第２の時点で被験体から取得された第２の生物学的試料から生成された第２の試験ゲノムデータ構成体を使用する時間的分類器とすることができる。

【０２４５】

訓練された分類器は、被験体、例えば、疾患状態が未知の被験体の疾患状態を決定するために使用することができる。この場合、本方法は、被験体についての試験時系列データセットを電子形式で取得することを含むことができ、試験時系列データセットには、複数の時点におけるそれぞれの時点について、それぞれの時点において被験体から得られた対応する生物学的サンプル中の対応する複数の核酸断片の複数の遺伝子型特性についての値を含む対応する試験遺伝子型データ構築物および、複数の時点における連続する時点のそれぞれの対について、連続する時点のそれぞれの対の間の時間の長さの表示を含む。次いで、本方法は、試験遺伝子型データ構成体を試験分類器に適用し、それにより被験体における疾患状態の状態を決定することを含むことができる。被験体は、以前に疾患状態と診断されたことがない場合がある。

【０２４６】

＜ＩＶ．応用＞

いくつかの実施形態において、本発明の方法、解析システムおよび／または分類器は、がんの存在を検出するため、がんの進行または再発をモニターするため、治療反応または有効性をモニターするため、最小残存病変（ＭＲＤ）の存在を決定またはモニターするため、またはそれらの任意の組み合わせに使用することができる。例えば、本明細書に記載されるように、分類器を使用して、試験特徴ベクターががんを有する被験体からのものである可能性を記述する確率スコア（例えば、０～１００）を生成することができる。一部の実施形態では、確率スコアは、被験体ががんであるか否かを判定するために閾値確率と比較される。他の実施形態では、尤度スコアまたは確率スコアは、疾患の進行を監視するため、または治療の有効性（例えば、治療効果）を監視するために、複数の異なる時点（例えば、治療前または治療後）で評価することができる。さらに他の実施形態では、尤度スコアまたは確率スコアは、臨床的決定（例えば、がんの診断、治療の選択、治療効果の評価など）を行うため、または臨床的決定に影響を及ぼすために使用することができる。例えば、一実施形態では、確率スコアが閾値を超えた場合、医師は適切な治療を処方することができる。

【０２４７】

＜ＩＶ．Ａ．がんの早期発見＞

いくつかの実施形態では、本発明の方法および／または分類器は、がんを有する疑いのある対象におけるがんの有無を検出するために使用される。例えば、分類器（例えば、セクションＩＩＩで上述し、セクションＶで検討する）を使用して、試験特徴ベクターががんを有する対象からのものである可能性を記述するがん予測を決定することができる。

【０２４８】

一実施形態では、がん予測は、試験サンプルががんに罹患しているかどうかの尤度（例えば、０～１００でスコア化される）である（すなわち、二値分類）。したがって、解析システムは、被験体ががんであるかどうかを決定するための閾値を決定することができる。例えば、６０以上または等しいがん予測は、被験体ががんに罹患していることを示すことができる。さらに他の実施形態では、６５以上または等しい、７０以上または等しい、７５以上または等しい、８０以上または等しい、８５以上または等しい、９０以上または等しい、または９５以上または等しいがん予測は、被験体ががんに罹患していることを示す。他の実施形態では、がん予測は疾患の重症度を示すことができる。例えば、８０のがん予測は、８０未満のがん予測（例えば、７０の確率スコア）と比較して、より重篤な形態、またはより後期のがんを示すことがある。同様に、経時的ながん予測値の増加（例えば、２つ以上の時点で採取された同一被験者からの複数のサンプルからの試験特徴ベクターを分類することによって決定される）は、疾患の進行を示すことがあり、または経時的ながん予測値の減少は、治療の成功を示すことがある。

【０２４９】

別の実施形態では、がん予測は多数の予測値から構成され、分類され（すなわち、多クラス分類）、複数のがんタイプの各々が予測値を有する（例えば、０～１００でスコア化される）。予測値は、所与の訓練サンプル（および推論中、訓練サンプル）ががんタイプの各々を有する尤度に対応し得る。解析システムは、最も高い予測値を有するがん種を特定し、被験体がそのがん種を有する可能性が高いことを示すことができる。他の実施形態では、解析システムはさらに、最高予測値を閾値（例えば、５０、５５、６０、６５、７０、７５、８０、８５など）と比較して、被験体がそのがん種を有する可能性が高いと判定する。他の実施形態では、予測値は疾患の重症度も示すことができる。例えば、８０を超える予測値は、６０の予測値と比較して、がんのより重篤な形態、またはより後期の段階を示すことがある。同様に、経時的な予測値の増加（例えば、２つ以上の時点で採取された同一被験者からの複数のサンプルからの試験特徴ベクターを分類することによって決定される）は、疾患の進行を示すことができ、または経時的な予測値の減少は、治療の成功を示すことができる。

【０２５０】

本発明の態様によれば、本発明の方法およびシステムは、複数のがん徴候を検出または分類するように訓練することができる。例えば、本発明の方法、システムおよび分類器は、１個以上、２個以上、３個以上、５個以上、１０個以上、１５個以上、または２０個以上の異なるタイプのがんの存在を検出するために使用することができる。

【０２５１】

本発明の方法、システムおよび分類器を使用して検出できるがんの例には、がん腫、リンパ腫、芽細胞腫、肉腫、および白血病またはリンパ系悪性腫瘍が含まれる。このようながんのより特定の例としては、扁平上皮がん（例えば、上皮性扁平上皮がん）、皮膚がん、悪性黒色腫、小細胞肺がんを含む肺がん、非小細胞肺がん（「ＮＳＣＬＣ」）、肺腺がんおよび肺扁平上皮がんを含む肺がん、腹膜のがん、胃腸がんを含む胃がんまたは胃腸がん、膵臓がん（例えば、膵管腺がん）、子宮頸がん、卵巣がん（例えば、高悪性度漿液性卵巣がん）、肝臓がん（例えば、肝細胞がん（ＨＣＣ））、肝細胞がん、肝がん、膀胱がん（例えば、尿路上皮性膀胱がん）、精巣（胚細胞腫瘍）がん、乳がん（例えば、ＨＥＲ２陽性乳がん、ＨＥＲ２陰性乳がん、トリプルネガティブ乳がん）、脳腫瘍（例えば、星細胞腫、神経膠腫（例えば、膠芽腫））、結腸がん、直腸がん、大腸がん、子宮内膜がんまたは子宮体がん、唾液腺がん、腎臓がんまたは腎臓がん（例えば、腎細胞がん、腎芽細胞腫またはウィルムス腫瘍）、前立腺がん、外陰がん、甲状腺がん、肛門がん、陰茎がん

、頭頸部がん、食道がん、および上咽頭がん（ＮＰＣ）が含まれるが、これらに限定されない。がんの追加的な例としては、網膜芽細胞腫、卵胞膜細胞腫、男化腫瘍、非ホジキンリンパ腫（ＮＨＬ）、多発性骨髄腫および急性血液悪性腫瘍を含むがこれらに限定されない血液悪性腫瘍、子宮内膜症、線維肉腫、絨毛がん、喉頭がん、カポジ肉腫、神経鞘腫、乏突起膠腫、神経芽腫、横紋筋肉腫、骨原性肉腫、平滑筋肉腫、尿路がんが含まれるが、これらに限定されない。

【０２５２】

いくつかの実施形態において、がんは、肛門がん、膀胱がん、乳がん、子宮頸がん、大腸がん、食道がん、胃がん、頭頸部がん、肝胆道がん、白血病、肺がん、リンパ腫、悪性黒色腫、多発性骨髄腫、卵巣がん、膵臓がん、前立腺がん、腎臓がん、甲状腺がん、子宮がん、またはそれらの任意の組み合わせのうちの１つ以上である。

【０２５３】

いくつかの実施形態において、１つ以上のがんは、肛門がん、大腸がん、食道がん、頭頸部がん、肝胆道がん、肺がん、卵巣がん、膵臓がん、ならびにリンパ腫および多発性骨髄腫などの「高シグナル（high-signal）」がん（５年がん特異的死亡率が５０％を超えるがんと定義される）であり得る。高シグナルがんはより侵攻性の傾向があり、一般的に患者から採取した試験さンプル中のセルフリー核酸濃度は平均以上である。

【０２５４】

＜ＩＶ．Ｂ．がんおよび治療のモニタリング＞

いくつかの実施形態において、がん予測は、疾患の進行をモニターするため、または治療の有効性（例えば、治療効果）をモニターするために、複数の異なる時点（例えば、または治療の前後）で評価することができる。例えば、本発明は、第１の時点でがん患者から第１のサンプル（例えば、第１の血漿ｃｆＤＮＡサンプル）を取得し、そこから（本明細書に記載されるように）第１のがん予測を決定し、第２の時点でがん患者から第２の試験サンプル（例えば、第２の血漿ｃｆＤＮＡサンプル）を取得し、そこから（本明細書に記載されるように）第２のがん予測を決定することを含む方法を含む。

【０２５５】

特定の実施形態では、第１の時点はがん治療前（例えば、切除手術または治療的介入前）であり、第２の時点はがん治療後（例えば、切除手術または治療的介入後）であり、分類器は治療の有効性を監視するために利用される。例えば、２回目のがん予測値が１回目のがん予測値に比べて減少した場合、治療は成功したとみなされる。しかし、第２のがん予測値が第１のがん予測値に比べて増加した場合は、治療は成功しなかったとみなされる。他の実施形態では、第１および第２の時点の両方が、がん治療の前（例えば、切除手術または治療介入の前）である。さらに他の実施形態では、第１の時点および第２の時点の両方が、がん治療後（例えば、切除手術後または治療介入後）である。さらに他の実施形態において、ｃｆＤＮＡサンプルは、第１および第２の時点でがん患者から取得され、分析され得る。例えば、がんの進行をモニターするため、がんが寛解状態にあるかどうかを決定するため（例えば、治療後に）、残存疾患または疾患の再発をモニターまたは検出するため、または治療（例えば、治療上の）効果をモニターするためである。

【０２５６】

当業者であれば、患者におけるがんの状態をモニターするために、任意の所望の時点のセットにわたってがん患者から試験サンプルを取得し、本発明の方法に従って分析することができることを容易に理解するであろう。いくつかの実施形態において、第１および第２の時点は、約１５分～約３０年の時間量、例えば、約１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、または約２４時間、例えば約１、２、３、４、５、１０、１５、２０、２５、または約５０日、または例えば約１、２、３、４、５、６、７、８、９、１０、１１、または約１２ヶ月、または例えば約１、１．５、２、２．５、３、３．５、４、４．５、５、５．５、６、６．５、７、７．５、８、８．５、９、９．５、１０、１０．５、１１、１１．５、１２、１２．５、１３、１３．５、１４、１４．５、１５、１５．５、１６、

１６．５、１７、１７．５、１８、１８．５、１９、１９．５、２０、２０．５、２１、２１．５、２２、２２．５、２３、２３．５、２４、２４．５、２５、２５．５、２６、２６．５、２７、２７．５、２８、２８．５、２９、２９．５または約３０年である。他の実施形態では、少なくとも５ヶ月に１回、少なくとも６ヶ月に１回、少なくとも１年に１回、少なくとも２年に１回、少なくとも３年に１回、少なくとも４年に１回、または少なくとも５年に１回、患者から試験サンプルを得ることができる。

【０２５７】

＜ＩＶ．Ｃ．治療＞

さらに別の実施形態では、がん予測は、臨床的決定（例えば、がんの診断、治療選択、治療効果の評価など）を行うため、または臨床的決定に影響を及ぼすために使用することができる。例えば、一実施形態では、がん予測（例えば、がんについて、または特定のがん種について）が閾値を超えた場合、医師は適切な治療（例えば、切除手術、放射線療法、化学療法、および／または免疫療法）を処方することができる。

【０２５８】

分類器（本明細書に記載）は、サンプル特徴ベクターががんを有する対象からのものであるというがん予測を決定するために使用することができる。一実施形態では、がん予測値が閾値を超えると、適切な治療（例えば、切除手術または治療）が処方される。例えば、一実施形態では、がん予測値が６０以上または等しい場合、１つ以上の適切な治療法が処方される。別の実施形態では、がん予測値が６５以上または等しい、７０以上または等しい、７５以上または等しい、８０以上または等しい、８５以上または等しい、９０以上または等しい、または９５以上または等しい場合、１つ以上の適切な治療法が処方される。他の実施形態では、がん予測は疾患の重症度を示すことができる。次いで、疾患の重症度に適合する適切な治療が処方され得る。

【０２５９】

いくつかの実施形態において、治療は、化学療法剤、標的化がん治療剤、分化療法剤、ホルモン療法剤、および免疫療法剤からなる群から選択される１つ以上のがん治療剤である。例えば、治療は、アルキル化剤、代謝拮抗剤、アントラサイクリン、抗腫瘍抗生物質、細胞骨格破壊剤（タキサン）、トポイソメラーゼ阻害剤、分裂阻害剤、コルチコステロイド、キナーゼ阻害剤、ヌクレオチドアナログ、白金系薬剤、およびそれらの任意の組み合わせからなる群から選択される１つ以上の化学療法剤であり得る。いくつかの実施形態において、治療は、シグナル伝達阻害剤（例えば、チロシンキナーゼおよび成長因子受容体阻害剤）、ヒストン脱アセチル化酵素（ＨＤＡＣ）阻害剤、レチノイン受容体作動薬、プロテオソーム阻害剤、血管新生阻害剤、およびモノクローナル抗体結合体からなる群から選択される１つ以上の標的化がん治療剤である。いくつかの実施形態では、治療は、トレチノイン、アリトレチノイン、ベキサロテンなどのレチノイドを含む１つ以上の分化療法剤である。いくつかの実施形態では、治療は、抗エストロゲン剤、アロマターゼ阻害剤、黄体ホルモン剤、エストロゲン剤、抗アンドロゲン剤、およびＧｎＲＨアゴニストまたはアナログからなる群から選択される１つ以上のホルモン療法剤である。一実施形態において、治療は、リツキシマブ（ＲＩＴＵＸＡＮ）およびアレムツズマブ（ＣＡＭＰＡＴＨ）のようなモノクローナル抗体療法、ＢＣＧ、インターロイキン－２（ＩＬ－２）およびインターフェロン－アルファのような非特異的免疫療法およびアジュバント、免疫調節薬、例えばサリドマイドおよびレナリドミド（ＲＥＶＬＩＭＩＤ）からなる群から選択される１つ以上の免疫療法剤である。腫瘍の種類、がんの病期、以前のがん治療または治療薬への曝露およびがんの他の特徴などの特徴に基づいて、適切ながん治療薬を選択することは、熟練した医師やがん専門医の能力の範囲内である。

【０２６０】

＜Ｖ．キットの実施＞

また、本明細書には、がん分類器に関する方法を含む上述の方法を実施するためのキットも開示する。キットは、遺伝物質を含む個体からサンプルを収集するための１つ以上の収集容器を含み得る。サンプルは、血液、血漿、血清、尿、糞便、唾液、他のタイプの体

液、またはそれらの任意の組み合わせを含むことができる。このようなキットには、サンプルから核酸を単離するための試薬を含めることができる。試薬は、緩衝液および検出剤を含む、核酸を配列決定するための試薬をさらに含み得る。１つ以上の実施形態において、キットは、特定のゲノム領域、特定の変異、特定の遺伝子変異、またはそれらの組み合わせを標的とするためのプローブを含む１つ以上の配列決定パネルを含み得る。１つ以上の実施形態において、キットは、例えば、表２、表４、またはそれらのいくつかの組み合わせからの、汚染を標的とするプローブを含む少なくとも１つのパネルを含む。他の実施形態において、キットを介して収集されたサンプルは、サンプル中の核酸を配列決定するために配列決定パネルを使用し得る配列決定研究所に提供される。

【０２６１】

キットはさらに、キットに含まれる試薬の使用説明書を含むことができる。例えば、キットはサンプルを採取し、試験サンプルから核酸を抽出するための説明書を含むことができる。説明書の例としては、試薬を添加する順序、被験試料から核酸を単離するために使用する遠心速度、核酸の増幅方法、核酸の配列決定方法、またはそれらの組み合わせを挙げることができる。説明書はさらに、記載された方法のいずれかのステップを実行する目的で、分析システム２００としてのコンピューティングデバイスを操作する方法を明らかにすることができる。

【０２６２】

上記の構成要素に加えて、キットは、本開示全体を通して記載される様々な方法を実行するためのコンピュータソフトウェアを記憶したコンピュータ読み取り可能な記憶媒体を含むことができる。これらの説明書が存在し得る１つの形態は、適切な媒体または基材に印刷された情報、例えば、情報が印刷された紙片または複数の紙片、キットの包装中、パッケージ挿入物中に印刷された情報としてである。さらに別の手段としては、コンピュータ可読媒体、例えばディスケット、ＣＤ、ハードドライブ、ネットワークデータ記憶装置であり、その上にコンピュータコードの形で命令が保存されている。さらに別の手段としては、インターネットを経由して、削除されたサイトの情報にアクセスできるウェブサイトのアドレスがある。

【０２６３】

［ＶＩ．結果例］

＜ＶＩ．Ａ．サンプルの収集および処理＞

研究デザインおよびサンプルであるＣＣＧＡ（ＮＣＴ０２８８９９７８）は、長期追跡調査を伴う、他施設共同研究、症例対照研究、観察研究である。３４２施設から約１５，０００人の参加者から非同定生物検体が収集された。サンプルはトレーニングセット（１，７８５例）とテストセット（１，０１５例）に分けられ、サンプルは各コホートにおける部位間のがん種および非がん種の分布が事前に規定された通りになるように選択され、がんおよび非がんのサンプルは性別で年齢を頻度的に一致させた。

【０２６４】

全ゲノムバイサルファイト配列決定は、血漿からｃｆＤＮＡを単離し、ｃｆＤＮＡの解析に全ゲノムバイサルファイト配列決定（ＷＧＢＳ；深さ３０倍）を採用した。ｃｆＤＮＡを、修正ＱＩＡａｍｐ循環核酸キット（Ｑｉａｇｅｎ、メリーランド州ジャーマンタウン）を用いて、患者１人につき２本の血漿チューブ（合計量１０ｍｌまで）から抽出した。最大７５ｎｇの血漿ｃｆＤＮＡを、ＥＺ－９６　ＤＮＡメチル化キット（Ｚｙｍｏ　Ｒｅｓｅａｒｃｈ社、Ｄ５００３）を用いて重亜硫酸塩に変換した。変換されたｃｆＤＮＡは、Ａｃｃｅｌ－ＮＧＳ　Ｍｅｔｈｙｌ－Ｓｅｑ　ＤＮＡライブラリー調製キット（Ｓｗｉｆｔ　ＢｉｏＳｃｉｅｎｃｅｓ、ミシガン州アナーバー）を用いてデュアルインデックスシーケンスライブラリーを調製するために使用され、構築されたライブラリーはＫＡＰＡ　イルミナプラットフォーム用ライブラリー定量キット（Ｋａｐａ　Ｂｉｏｓｙｓｔｅｍｓ、マサチューセッツ州ウィルミントン）を用いて定量された。１０％のＰｈｉＸ　ｖ３ライブラリー（Ｉｌｌｕｍｉｎａ、ＦＣ－１１０－３００１）とともに４つのライブラリーをプールし、Ｉｌｌｕｍｉｎａ　ＮｏｖａＳｅｑ　７０００　Ｓ２フローセル上でク

ラスター化した後、１５０ｂｐペアエンドシーケンス（３０倍）を行った。
【０２６５】
各サンプルについて、ＷＧＢＳ断片セットを、異常なメチル化パターンを有する断片の小さなサブセットに縮小した。さらに、高メチル化または低メチル化ｃｆＤＮＡ断片が選択された。異常なメチル化パターンを有し、高メチル化または低メチル化、すなわちＵＦＸＭを有するｃｆＤＮＡ断片が選択された。がんでない個体で高頻度に生じる断片や、不安定なメチル化を有する断片は、がんの状態を分類するための高度に識別的な特徴を生み出す可能性は低い。そこで、ＣＣＧＡ研究から得られた、がんのない非喫煙者１０８人（年齢：５８±１４歳、女性７９人［７３％］）からなる独立した参照セット（すなわち、参照ゲノム）を用いて、統計モデルと典型的な断片のデータ構造を作成した。これらのサンプルは、セクションＩＩ．Ｃ．で前述したように、断片内のＣｐＧメチル化状態の所与の配列の尤度を推定するマルコフ連鎖モデル（次数３）の訓練に使用された。このモデルは、正常な断片範囲（ｐ値＞０．００１）内で較正されることが実証され、マルコフモデルからのｐ値が＞０．００１である断片を、異常性が不十分であるとして拒絶するために使用された。
【０２６６】
上述のように、さらなるデータ削減ステップでは、少なくとも５つのＣｐＧがカバーされ、平均メチル化が０．９以上（高メチル化）または０．１未満（低メチル化）の断片のみを選択した。この手順により、訓練においてがんを発症していない参加者のＵＦＸＭ断片の中央値（範囲）は２，８００（１，５００～１２，０００）、訓練においてがんを発症した参加者のＵＦＸＭ断片の中央値（範囲）は３，０００（１，２００～４２０，０００）となった。このデータ削減手順は参照セットデータのみを使用するため、この段階は各サンプルに１回のみ適用する必要があった。
【０２６７】
＜ＶＩ．Ｂ．汚染検出結果＞
結果例の最初のグループとして、表２および表４の汚染マーカープローブの性能を調べるために、コードベースの開発にさらに力を入れる前に、意図的な汚染の有無にかかわらず、異なる内部研究からの血漿ｃｆＤＮＡサンプルおよびｇＤＮＡサンプルの複数のバッチを予備分析に使用した。
【０２６８】
図７は、所与のサンプルについてホモ接合と呼ばれた各汚染マーカーについて、基準または代替と呼ばれた断片の総数に対する、汚染と分類された断片の数の割合の分布を、第一の結果例群に従って示したものである。分析システムは、表２および表４の汚染マーカープローブを利用して、サンプル中の汚染断片を同定した。汚染断片の割合は、汚染マーカープローブと重なる各サンプル中の断片の総数に対する同定された汚染断片の比率として計算した。グラフ７００に示すように、各サンプルの汚染断片の割合をプロットし、Ｙ軸は汚染断片の割合を示した。０．００１を超えるサンプルもいくつかあったが、０．００１以下のサンプルが圧倒的に多かった。グラフ７００の箱ひげ図では、下位四分位値、中央値、上位四分位値の全てが０．０００５を大きく下回っている。グラフ７１０は、Ｘ軸に汚染片の割合を、Ｙ軸にサンプル数をプロットしたもので、同じ結果を示している。第２グループの結果例に対応する図については、図１７を参照されたい。
【０２６９】
図８は、第一の結果例に従った、汚染マーカーの対立遺伝子頻度の散布図である。解析システムは、ＣＣＧＡ研究とそのフォローアップ研究からのサンプルを用いて、汚染マーカーの対立遺伝子頻度を計算した。解析システムはまた、ゲノムデータベースＧｅｎｏｍｅｓ　ＰｒｏｊｅｃｔおよびｇｎｏｍＡＤから集団対立遺伝子分率を得た（グラフ１００）。各汚染マーカーは、ｘ軸が集団対立遺伝子分率を示し、ｙ軸がサンプルから計算した対立遺伝子分率を示してプロットしている。第２グループの結果例に対応する図については、図１４を参照されたい。
【０２７０】

図９は、第１のグループの例示的な結果に従って、ＳＮＰ汚染マーカー（グラフ９００）およびインデル汚染マーカー（グラフ９１０）の遺伝子型および接合性を示す２つのグラフを示す。複数のＳＮＰ汚染マーカーについてのグラフ９００において、左側の３つの箱ひげ図は、サンプルの第１のセット（基準ホモ接合体、代替ホモ接合体、およびヘテロ接合体）の接合性の比例内訳を表し、右側の３つの箱ひげ図は、サンプルの第２のセット（基準ホモ接合体、代替ホモ接合体、およびヘテロ接合体）の接合性の比例内訳を表す。同様に、インデル汚染マーカーのグラフ９１０も、２セットのサンプル間で分割されたインデル汚染メーカーの接合性の比例内訳を表す。図１５および図１６を参照されたい。

【０２７１】

図１０は、第１の結果例群に従って、所定のサンプルについて、ホモ接合性であり、そのサンプルの汚染を推定するのに有用であると考えられるのに十分な断片が重複している汚染マーカーの割合を示すグラフである。分析システムは、各サンプルの汚染マーカーの接合性を決定した。次に、分析システムは、サンプルにおいてホモ接合であることが判明した汚染マーカーの割合を計算した。サンプル中のホモ接合体であることが判明した汚染マーカーの数が多いほど、汚染検出能力が高くなる。さらに、これらのホモ接合部位は、サンプルの汚染を推定するのに有用と考えられるだけのサンプル断片が重なるなど、様々な品質チェック基準を経て、所定のサンプルに有用な部位の最終的な割合が得られる。その割合は、多重ＳＮＰ部位汚染マーカー、インデル部位汚染マーカー、多重ＳＮＰ部位汚染マーカーおよびインデル部位汚染マーカーの両方について、特にプロットした。最初の箱ひげ図１０００では、割合の中央値は約０．５２５であった。複数のＳＮＰ部位汚染マーカーに関する２番目の箱ひげ図１０１０では、中央値の割合は約０．５７５であった。３番目の箱ひげ図１０２０では、インデル部位汚染マーカーであった。中央値は約０．４７５であった。図１５および図１３を参照されたい。

【０２７２】

図１１Ａおよび１１Ｂは、第１の結果例群に従った、サンプルの異なるバッチについての推定汚染レベル検出のグラフを示す。グラフ１１００は、サンプルの「脱リスク（Derisking）」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１１０は、サンプルの「ｄｏｐｐｌｅｒ＿ｐｒｅｌｉｍ＿ｔｅｓｔ」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１２０は、サンプルの「Ｈｙｂ１＿ＳＯＰ＿１２ｐｌｅｘ」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１３０は、サンプルの「ＭＲＤ」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１４０は、「ｃｆＤＮＡ滴定（cfDNA titration）」バッチサンプルからの推定サンプル汚染レベルの分布を示す。グラフ１１５０は、サンプルの「ｇＤＮＡ滴定（gDNA titration）」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１６０は、サンプルの「ダウンサンプリングｇＤＮＡ滴定（downsampled gDNA titration）（ｖ１．０化学）」バッチからの推定サンプル汚染レベルの分布を示す。また、グラフ１１７０は、「ダウンサンプルｇＤＮＡ滴定（downsampled gDNA titration）（ｖ１．５化学）」バッチ試料からの推定サンプル汚染レベルの分布を示す。第２グループの結果例に対応する図については、図２０および図２１を参照されたい。

【０２７３】

第２の結果例群として、表２および表４の汚染マーカープローブの性能を調べるために、意図的な汚染のない８４人の一個人からの血漿ｃｆＤＮＡサンプルのコホートを、改訂され成熟したコードベースを用いた最終分析に用いた。

【０２７４】

図１２は、実験中の８４サンプルについて集計した、表２および表４に列挙した汚染マーカーごとに得られた特異なｃｆＤＮＡ断片の数の分布を示す。ｘ軸は、汚染マーカーを標的とするように設計されたプローブと重複する特異なｃｆＤＮＡ断片分子の数を示し、全８４サンプルにわたって集計し、ｙ軸は、ｘ軸によって示されるビンに該当する汚染マーカーの割合を示す。

【０２７５】

表２および表４に記載されているような断片マーカーは、分析システムのステップ２２０および２５０なしで選択した。ハーディーワインベルグ平衡にないマーカーは、参照ハプロタイプおよび代替ハプロタイプの集団データベース頻度に基づいてフィルタリングした。この場合、ハーディーワインベルグ項（ｐ２＋２ｐｑ＋ｑ２）の値は［０．９、１．５］の範囲であった（１より高い値への偏りは、集団の非ランダム交配条件において予想される）。表２および表４の１０００マーカーのうち、この条件では４マーカーが不合格であった。

【０２７６】

各汚染マーカーは、そのマーカーのプローブと重複するｃｆＤＮＡ断片について、個々のサンプルに関してさらなる品質チェック（ＱＣ）を受けた。特定のサンプルについて品質チェック条件に合格しなかったマーカーは、汚染割合の推定を含む、特定のマーカーについてのゲノムバリアントコールに依存するあらゆる分析について、そのサンプルについて後に廃棄された。

【０２７７】

実験の８４サンプルからの汚染マーカーおよび１サンプルの各ペアについて、複数のＳＮＰマーカーおよびインデルマーカーのハプロタイプ、および対立遺伝子をそれぞれ参照または代替として確実に呼び出すために使用することができる、ゲノム変異体の位置範囲に重複する特異なｃｆＤＮＡ断片分子の最小数、この場合２０が、一定の割合（例えば、複数のＳＮＰマーカーでは０．７、インデルマーカーでは０．５）以上のｃｆＤＮＡ断片について、参照または代替マーカーとして使用することができる。後者の条件は、マーカーおよびサンプルのペアに期待される理論的な二項確率分布（ヘテロ接合と呼ばれるマーカーの平均は０．５、またはホモ接合と呼ばれるサンプル中の全マーカーの総汚染率に依存する平均）に従って、参照ハプロタイプと呼ばれるｃｆＤＮＡ断片の数と代替ハプロタイプと呼ばれるｃｆＤＮＡ断片の数の比が、非常にあり得ないものであってはならないこと（例えば、ｐ値＜１０－５）をチェックするために拡張する。

【０２７８】

図１３は、この品質チェックプロセスの結果を示している。前述の基準に基づいて、８４サンプル全てについて８５１マーカーが合格し、８３サンプルについて６８マーカーが合格し、８２サンプルについて２０マーカーが合格した。不合格マーカーは、８４サンプル全てで７マーカーであった（ハーディーワインベルグ平衡の合格条件を満たさなかった４個を含む）。残りのマーカーの不合格率は、この両極端の間でばらつきがあった。

【０２７９】

図１４～１６は、８４サンプル中少なくとも８２サンプルについて品質チェック基準に合格した汚染マーカーについて、表２および表４の汚染マーカーの基となるゲノム変異体の特徴を示す図である。図１４は、これらの汚染マーカーを選択するための対立遺伝子頻度範囲である［０．３、０．７］の範囲にある対立遺伝子頻度の散布図を示す。図１５は、マーカーがヘテロ接合と呼ばれる頻度の散布図を示す。図１６は、ハーディーワインベルグ項の値の散布図を示す。これらのグラフのｘ軸は、集団データベース１０００　Ｇｅｎｏｍｅｓ（複数のＳＮＰマーカー用）またはｇｎｏｍＡＤ（インデルマーカー用）で利用可能なデータから得られた値または計算された値を表し、ｙ軸は、我々の実験における８４個のｃｆＤＮＡサンプルから計算された値を表す。全てのポイントは、ｙ＝ｘ線（グラフでは破線で表現）の周りに多少のノイズを伴って横たわることが予想された。

【０２８０】

ｃｆＤＮＡ断片のセットにおける汚染率を推定するために、まず、汚染源に由来するｃｆＤＮＡ断片を同定した。マーカーと呼ばれるホモ接合型バリアントと反対のバリアントとして呼ばれるｃｆＤＮＡ断片は、汚染断片として分類した。断片が汚染マーカー部位と重ならない、またはヘテロ接合体として遺伝子型決定された汚染マーカー部位と重なる断片のサブセットは、汚染断片の分類および汚染率を推定する目的で無視した。汚染率を推定する能力は、少なくとも１つの汚染断片を見つけることを条件とした。汚染断片は分数化できないため、検出可能な汚染断片の下限は、対象となる断片の総数に反比例した。

【０２８１】
図１７は、全８４検体にわたる各マーカーについて集計した、所定の検体についてホモ接合体と呼ばれた各汚染マーカーについて、参照または代替と呼ばれた断片の総数に対する、汚染である分類された断片の数の割合の分布を示す。グラフ１７１０は、ジッター散布図および、分布の極値および四分位数を示すこれらの値の箱ひげ図を示す。これらの値の算術平均は2×10^{-4}である。検出下限より低い値は０に抑圧され、その結果、０に小さなバンプが生じる。グラフ１７２０は、同じデータの平滑化密度線（実線）、および二項パラメータ$p = 2 \times 10^{-4}$、サンプルサイズは特定の汚染マーカーに対するｃｆＤＮＡ断片の集合数として、データに二項分布フィットを用いて得られたシミュレーション値（破線）を示す。全体として、０と最頻値付近で若干のずれはあるものの、よく適合している。

【０２８２】
汚染断片の概念を発展させ、あるｃｆＤＮＡサンプルの汚染率を、外部ソースからのｃｆＤＮＡ汚染の根本的プロセスをモデル化することによって推定した。サンプルに対してホモ接合体として呼ばれるマーカーだけを考慮すると、汚染率がｃｆで、サンプルに対して呼ばれるホモ接合体とは反対のバリアントの集団対立遺伝子頻度がａｆｉである場合、マーカーに対する汚染断片が観察される確率はｃｆｘａｆｉである。このマーカーでの断片の総数がｎｉであれば、このマーカーにおける汚染断片の期待数は、ｃｆｘａｆｉｘｎｉである。確率論の期待値の直線性の性質を用いると、全てのマーカーについてこれらの期待数を合計して、全てのマーカーにわたる汚染断片の期待総数の式、$nc = \Sigma_i (cf \times af_i \times n_i)$を得ることができる。全ての既知の変数を片側に持ってくるように再配列すると、汚染割合の推定値は、$cf = nc / (\Sigma_i (af_i \times n_i))$となる。

【０２８３】
図１８は、仮想的な断片セットに対する汚染率を推定するための式の適用を示す図である。この例では、４つの複数のＳＮＰおよび２つのインデル汚染マーカーがあり、そのうち３つの複数のＳＮＰおよび１つのインデルマーカーだけがホモ接合と呼ばれる。３つの汚染断片の同定、各マーカーの断片の総数、反対ハプロタイプの各マーカーでの対立遺伝子頻度を考慮すると、サンプル全体の汚染断片の推定値が得られる。

【０２８４】
図１９は、シミュレーションデータに汚染割合モデルを適用した結果を示している。各ポイントについて、表２および表４の各汚染マーカーについて、それらの集団頻度に基づいて、バックグラウンドゲノムおよび汚染源ゲノムの両方についてバリアントコール（ホモ接合参照、ホモ接合代替、またはヘテロ接合）をサンプリングすることによってデータを生成した。次に、参照および代替ハプロタイプ断片のカウントを、実験から得られた８４サンプルのうち、このマーカーについて観察されたｃｆＤＮＡ断片の平均数をサイズとする二項分布からサンプリングし、確率を２つのバリアントコールおよびシミュレーションパラメータとして入力された汚染割合の関数とした。推定値のばらつきを捉えるため、各シミュレーションは１００回繰り返した。グラフのｘ軸はシミュレーションのパラメータとして使用された汚染率を表し、ｙ軸はモデルによって推定された汚染率の値を表す。破線は$y = x$線を表す。各箱ひげ図は、指定された汚染率パラメータで実行された１００のシミュレーションの推定汚染率の範囲を示す。この範囲は、中央値が期待値に非常に近く、汚染度が高くなるにつれて四分位範囲が小さくなることを示す。これらの結果は、汚染率を推定するために本明細書で説明したモデルを検証するものである。

【０２８５】
図２０は、それぞれ滴定レベルが異なるｃｆＤＮＡサンプルの４つの滴定対に、汚染率モデルを適用した結果を示す。滴定対におけるドナーサンプルは汚染として扱い、その滴定レベルは汚染率として扱う。汚染率の推定値が確実に得られていることを確認するため、各組の各滴定レベルも、インプット材料の入手可能性に応じて複数回繰り返した。Ｘ軸は滴定レベル、Ｙ軸はそのレベルで滴定した各組の複製に対する推定汚染率を示す。実線は全観測値に対する線形モデルフィットを示し、破線は$y = x$線を示す。このグラフから

、線形モデルフィットはｙ＝ｘ線とほぼ同じ傾きを有するが、２本の線の切片は異なっており、全てのサンプルに約２．５ｘ１０$^{-4}$という一定の意図しないバックグラウンド汚染があることがわかる。これらのサンプル内の汚染断片を詳しく調べると、この一定の残留汚染は実在し、バリアントコーリングなどの技術的ノイズによるものではないことが明らかになった（ここでは図示していない）。これらの結果は、汚染率を推定するための本明細書に記載のモデルをさらに検証するものである。

【０２８６】

図２１は、実験における８４個のｃｆＤＮＡサンプルの推定汚染率の分布を示す。このグラフは、ジッター散布図ならびに推定汚染率値の箱ひげ図を示す。中央値は１．８ｘ１０$^{-4}$、平均値は３．２ｘ１０$^{-4}$である。この平均値は、同じサンプルセットで得られた汚染断片の割合の平均値（２ｘ１０$^{-4}$）とは異なり、高いことに注意すべきであり、これは、汚染断片モデルが、マーカーが両方のゲノムで同じハプロタイプを持つ場合に同定できなかった汚染断片も考慮するためであると予想される。

【０２８７】

図２２は、複数のＳＮＰマーカーおよびインデルマーカーを別々に考慮した場合、実験における８４サンプルについて得られた推定値が高い相関を示し、有意な系統的偏りを示さないことを示す。このことは、両タイプの汚染マーカーの性能が同等であることを証明している。

【０２８８】

＜ＶＩＩ．追加考察＞

以上の実施形態の詳細な説明は、本開示の特定の実施形態を示す添付の図面を参照する。異なる構造および動作を有する他の実施形態は、本開示の範囲から逸脱しない。「本発明（the invention）」等の用語は、本明細書に記載される出願人の発明の多くの代替的な態様または実施形態の特定の具体例を参照して使用され、その使用またはその不存在は、出願人の発明の範囲または特許請求の範囲を限定することを意図するものではない。

【０２８９】

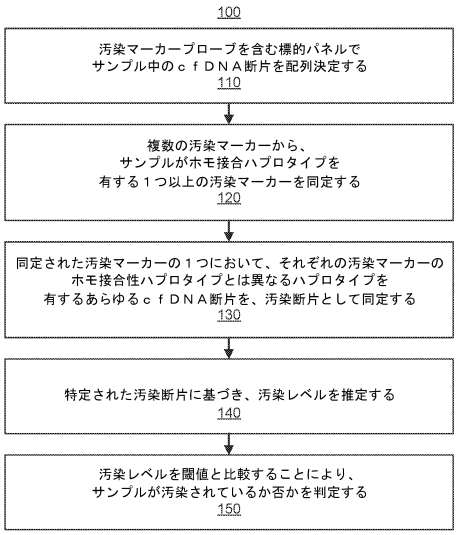
本発明の実施形態は、本明細書における操作を実行するための装置に関することもできる。この装置は、必要な目的のために特別に構成されてもよく、および／または、コンピュータに格納されたコンピュータプログラムによって選択的に起動または再構成される汎用のコンピューティングデバイスで構成されてもよい。このようなコンピュータプログラムは、非一過性の、有形のコンピュータ可読記憶媒体、または、コンピュータシステムバスに結合され得る、電子命令を記憶するのに適した任意のタイプの媒体に記憶され得る。さらに、本明細書で言及されるコンピューティングシステムは、単一のプロセッサを含んでもよいし、コンピューティング能力を高めるために複数のプロセッサ設計を採用する構造であってもよい。

【０２９０】

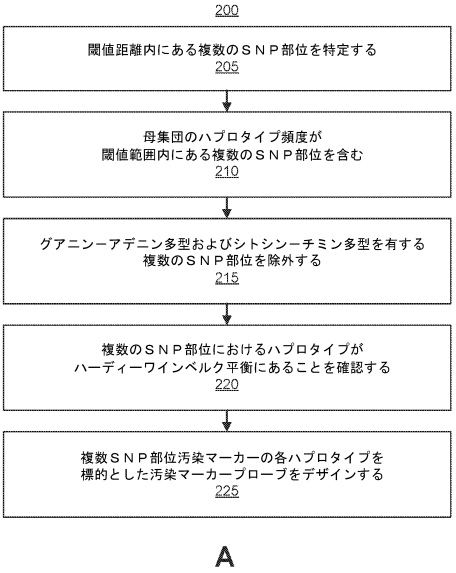
分析システムによって実行されるものとして本明細書で説明されるステップ、操作、またはプロセスのいずれかは、単独で、または他のコンピューティングデバイスと組み合わせて、装置の１つ以上のハードウェアまたはソフトウェアモジュールを用いて実行または実装され得る。一実施形態では、ソフトウェアモジュールは、プロセスのいずれかまたは全てを実行するためにコンピュータプロセッサによって実行され得る、コンピュータプログラムコードを含むコンピュータ可読媒体を含むコンピュータプログラム製品、操作、または説明されたステップ、で実装される。

【図面】

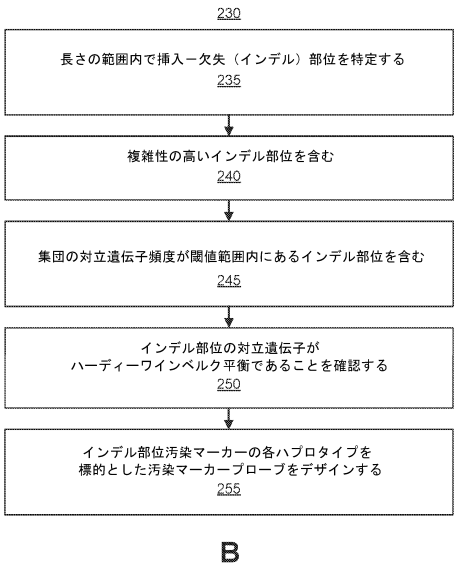
【図１】



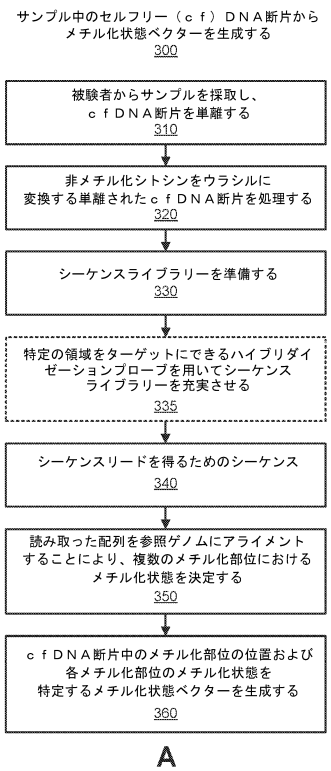
【図２Ａ】



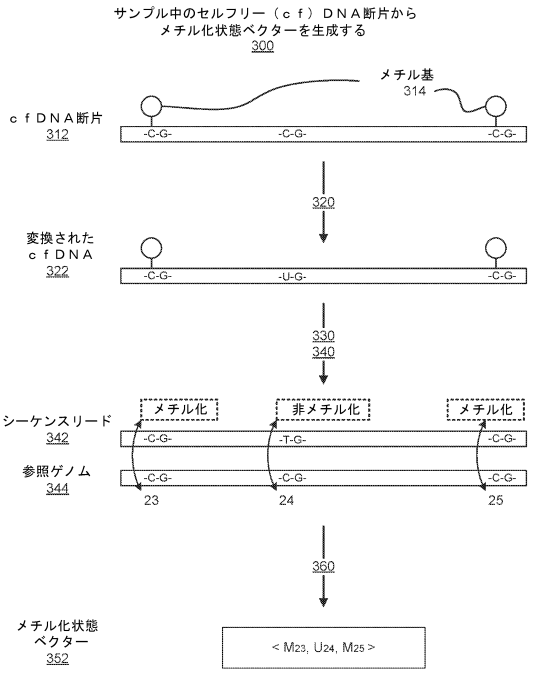
【図２Ｂ】



【図３Ａ】

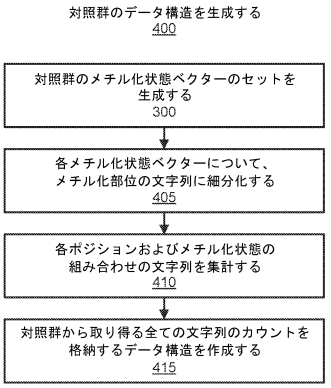


【図３Ｂ】



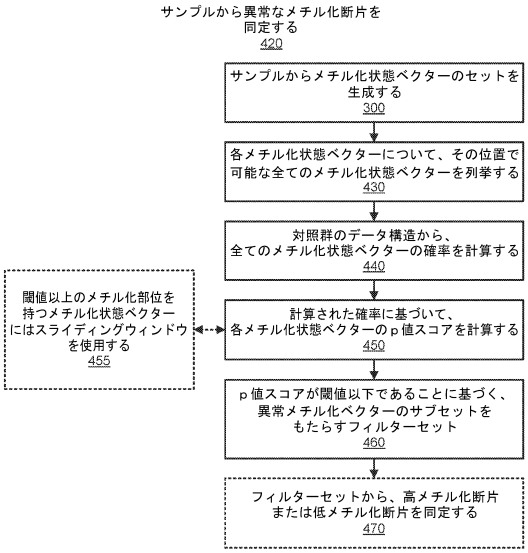
B

【図４Ａ】



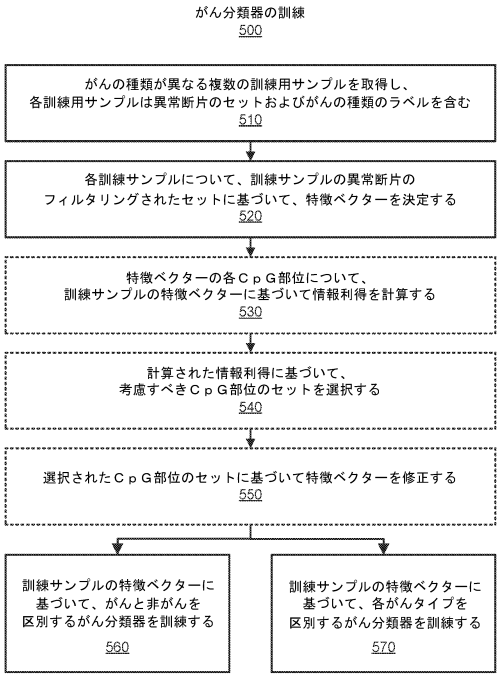
A

【図４Ｂ】



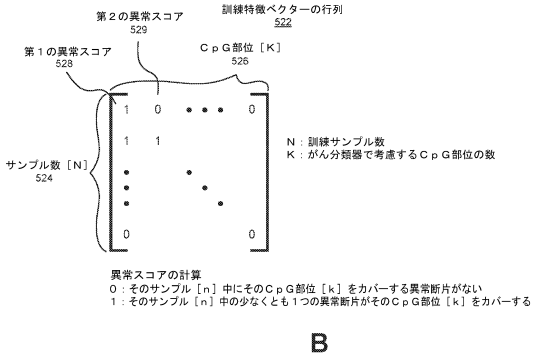
B

【図５Ａ】



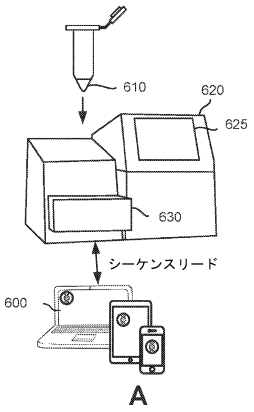
A

【図５Ｂ】

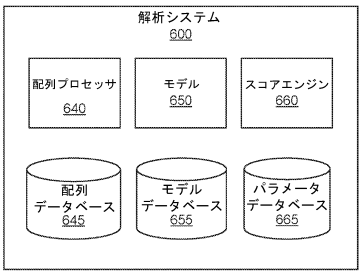


Ｂ

【図６】

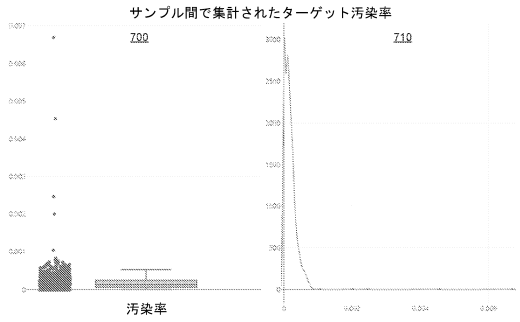


Ａ

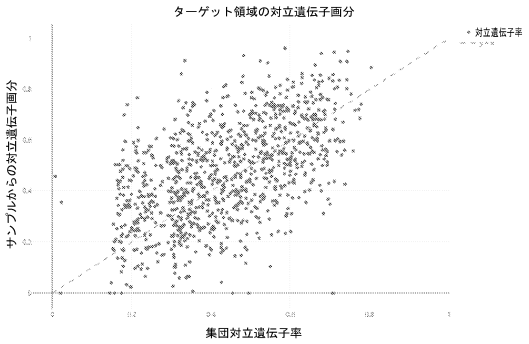


Ｂ

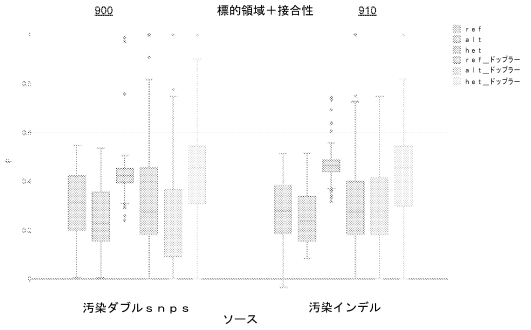
【図７】



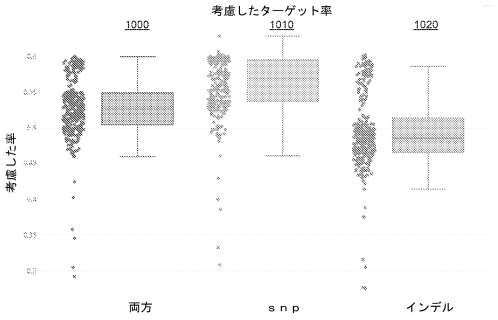
【図８】



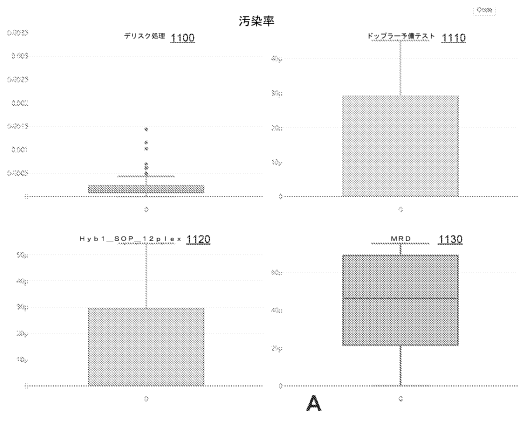
【図９】



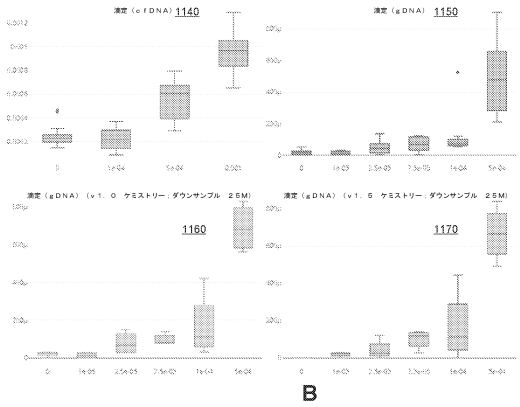
【図１０】



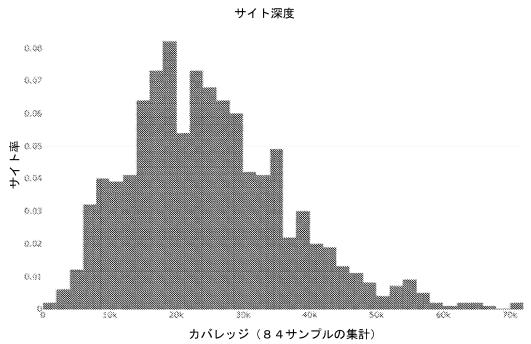
【図１１Ａ】



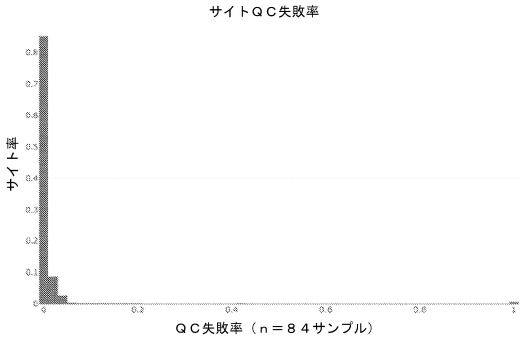
【図１１Ｂ】



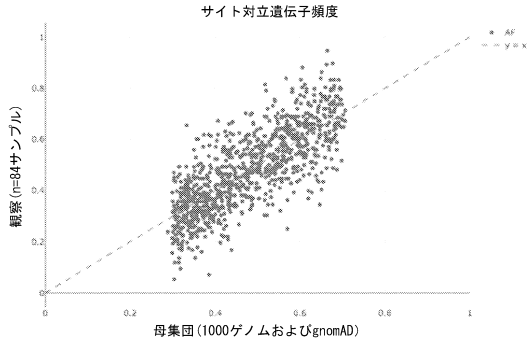
【図１２】



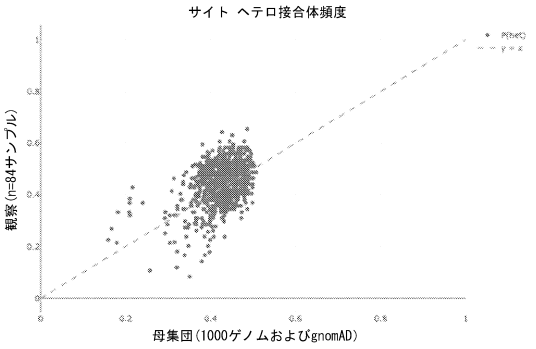
【図１３】



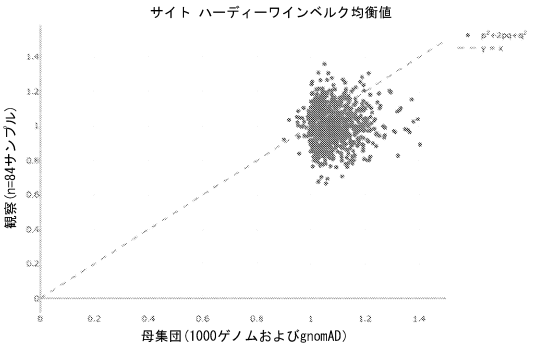
【図１４】



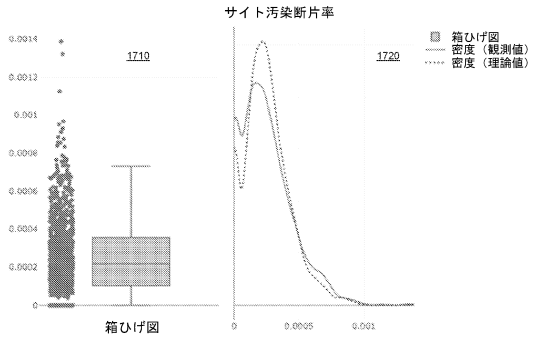
【図１５】



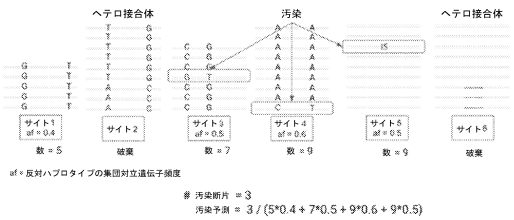
【図１６】



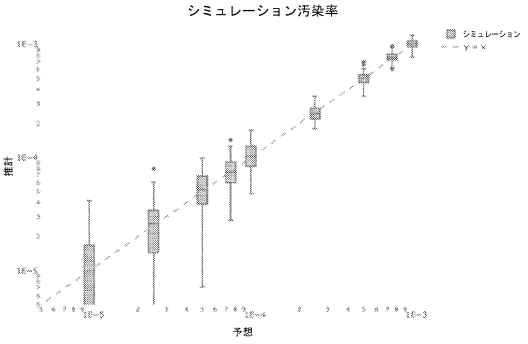
【図１７】



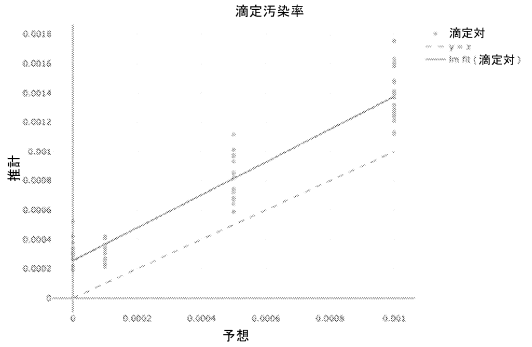
【図１８】



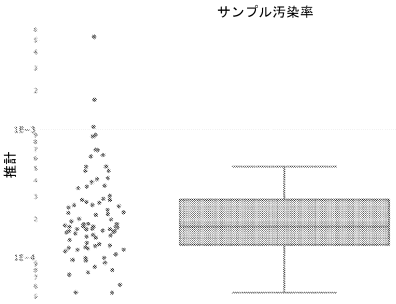
【図１９】



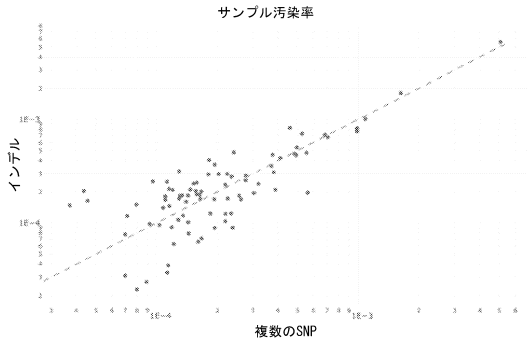
【図２０】



【図２１】



【図２２】



【図２３－１】

参照	開始	終了	メチル化	スコア	間隔	pos_count	neg_count	ソース	説明
chr2	108676624	108676632	Hypo	3.912114	0	0	0	contamination_double_snps	chr2:108676624:A>T,chr2:108676631:C>G
chr2	108676624	108676632	Hyper	3.912114	0	0	0	contamination_double_snps	chr2:108676624:A>T,chr2:108676631:C>G
chr7	154332790	154332799	Hypo	3.869331	0	0	0	contamination_double_snps	chr7:154332790:C>G,chr7:154332798:G>C
chr7	154332790	154332799	Hyper	3.869331	0	0	0	contamination_double_snps	chr7:154332790:C>G,chr7:154332798:G>C
chr20	248925	248935	Hypo	3.857307	0	0	0	contamination_double_snps	chr20:248925:C>A,chr20:248934:A>T
chr20	248925	248935	Hyper	3.857307	0	0	0	contamination_double_snps	chr20:248925:C>A,chr20:248934:A>T
chr10	108512079	108512086	Hypo	3.854286	0	0	0	contamination_double_snps	chr10:108512079:C>G,chr10:108512085:C>A
chr10	108512079	108512086	Hyper	3.854286	0	0	0	contamination_double_snps	chr10:108512079:C>G,chr10:108512085:C>A
chr4	16736064	16736067	Hypo	3.823068	0	0	0	contamination_double_snps	chr4:16736064:T>G,chr4:16736066:G>T
chr4	16736064	16736067	Hyper	3.823068	0	0	0	contamination_double_snps	chr4:16736064:T>G,chr4:16736066:G>T
chr4	72575890	72575901	Hypo	3.815061	0	0	0	contamination_double_snps	chr4:72575890:C>A,chr4:72575900:C>A
chr4	72575890	72575901	Hyper	3.815061	0	0	0	contamination_double_snps	chr4:72575890:C>A,chr4:72575900:C>A
chr6	158434980	158434991	Hypo	3.806239	0	0	0	contamination_double_snps	chr6:158434980:G>C,chr6:158434990:G>C
chr6	158434980	158434991	Hyper	3.806239	0	0	0	contamination_double_snps	chr6:158434980:G>C,chr6:158434990:G>C
chr13	20880885	20880889	Hypo	3.795089	0	0	0	contamination_double_snps	chr13:20880885:A>T,chr13:20880888:A>T
chr13	20880885	20880889	Hyper	3.795089	0	0	0	contamination_double_snps	chr13:20880885:A>T,chr13:20880888:A>T
chr6	21618915	21618925	Hyper	3.767993	0	0	0	contamination_double_snps	chr6:21618915:T>G,chr6:21618924:G>C
chr6	21618915	21618925	Hypo	3.767993	0	0	0	contamination_double_snps	chr6:21618915:T>G,chr6:21618924:G>C
chr5	8827279	8827283	Hypo	3.762267	0	0	0	contamination_double_snps	chr5:8827279:G>C,chr5:8827282:A>T
chr5	8827279	8827283	Hyper	3.762267	0	0	0	contamination_double_snps	chr5:8827279:G>C,chr5:8827282:A>T
chr5	103457798	103457806	Hypo	3.754163	0	0	0	contamination_double_snps	chr5:103457798:C>A,chr5:103457805:G>C
chr5	103457798	103457806	Hyper	3.754163	0	0	0	contamination_double_snps	chr5:103457798:C>A,chr5:103457805:G>C
chr13	91794575	91794584	Hypo	3.753434	0	0	0	contamination_double_snps	chr13:91794575:C>G,chr13:91794583:C>G
chr13	91794575	91794584	Hyper	3.753434	0	0	0	contamination_double_snps	chr13:91794575:C>G,chr13:91794583:C>G
chr8	20489721	20489728	Hyper	3.748329	0	0	0	contamination_double_snps	chr8:20489721:C>A,chr8:20489727:C>A
chr8	20489721	20489728	Hypo	3.748329	0	0	0	contamination_double_snps	chr8:20489721:C>A,chr8:20489727:C>A
chr5	109750753	109750763	Hyper	3.745406	0	0	0	contamination_double_snps	chr5:109750753:T>G,chr5:109750762:C>A
chr5	109750753	109750763	Hypo	3.745406	0	0	0	contamination_double_snps	chr5:109750753:T>G,chr5:109750762:C>A
chr21	41585518	41585523	Hypo	3.740602	0	0	0	contamination_double_snps	chr21:41585518:C>G,chr21:41585522:G>T
chr21	41585518	41585523	Hyper	3.740602	0	0	0	contamination_double_snps	chr21:41585518:C>G,chr21:41585522:G>T
chr22	27741446	27741454	Hypo	3.736007	0	0	0	contamination_double_snps	chr22:27741446:G>T,chr22:27741453:T>G
chr22	27741446	27741454	Hyper	3.736007	0	0	0	contamination_double_snps	chr22:27741446:G>T,chr22:27741453:T>G
chr1	242404448	242404456	Hyper	3.736007	0	0	0	contamination_double_snps	chr1:242404448:G>C,chr1:242404455:A>C
chr1	242404448	242404456	Hypo	3.736007	0	0	0	contamination_double_snps	chr1:242404448:G>C,chr1:242404455:A>C
chr17	61565130	61565133	Hypo	3.732159	0	0	0	contamination_double_snps	chr17:61565130:T>A,chr17:61565132:G>T
chr17	61565130	61565133	Hyper	3.732159	0	0	0	contamination_double_snps	chr17:61565130:T>A,chr17:61565132:G>T
chr14	75482826	75482828	Hyper	3.725651	0	0	0	contamination_double_snps	chr14:75482826:C>A,chr14:75482827:G>T
chr14	75482826	75482828	Hypo	3.725651	0	0	0	contamination_double_snps	chr14:75482826:C>A,chr14:75482827:G>T
chr10	36939305	36939307	Hypo	3.725651	0	0	0	contamination_double_snps	chr10:36939305:A>T,chr10:36939306:A>C
chr10	36939305	36939307	Hyper	3.725651	0	0	0	contamination_double_snps	chr10:36939305:A>T,chr10:36939306:A>C
chr13	88734744	88734755	Hypo	3.723231	0	0	0	contamination_double_snps	chr13:88734744:T>A,chr13:88734754:G>C

【図２３－２】

chr13	88734744	88734755	Hyper	3.723231	0	0	0	contamination_double_snps	chr13:88734744:T>A,chr13:88734754:G>C
chr2	239488791	239488802	Hypo	3.723231	0	0	0	contamination_double_snps	chr2:239488791:C>G,chr2:239488801:C>A
chr2	239488791	239488802	Hyper	3.723231	0	0	0	contamination_double_snps	chr2:239488791:C>G,chr2:239488801:C>A
chr14	100279482	100279492	Hyper	3.719376	0	0	0	contamination_double_snps	chr14:100279482:G>C,chr14:100279491:G>C
chr14	100279482	100279492	Hypo	3.719376	0	0	0	contamination_double_snps	chr14:100279482:G>C,chr14:100279491:G>C
chr14	99262737	99262744	Hyper	3.719295	0	0	0	contamination_double_snps	chr14:99262737:A>T,chr14:99262743:G>T
chr14	99262737	99262744	Hypo	3.719295	0	0	0	contamination_double_snps	chr14:99262737:A>T,chr14:99262743:G>T
chr20	37923956	37923967	Hyper	3.714409	0	0	0	contamination_double_snps	chr20:37923956:C>A,chr20:37923966:G>T
chr20	37923956	37923967	Hypo	3.714409	0	0	0	contamination_double_snps	chr20:37923956:C>A,chr20:37923966:G>T
chr2	45464078	45464089	Hypo	3.71108	0	0	0	contamination_double_snps	chr2:45464078:A>C,chr2:45464088:C>A
chr2	45464078	45464089	Hyper	3.71108	0	0	0	contamination_double_snps	chr2:45464078:A>C,chr2:45464088:C>A
chr4	117904895	117904903	Hypo	3.708048	0	0	0	contamination_double_snps	chr4:117904895:G>C,chr4:117904902:A>C
chr4	117904895	117904903	Hyper	3.708048	0	0	0	contamination_double_snps	chr4:117904895:G>C,chr4:117904902:A>C
chr20	126922	126930	Hyper	3.708048	0	0	0	contamination_double_snps	chr20:126922:C>A,chr20:126929:A>T
chr20	126922	126930	Hypo	3.708048	0	0	0	contamination_double_snps	chr20:126922:C>A,chr20:126929:A>T
chr17	72487747	72487750	Hypo	3.697846	0	0	0	contamination_double_snps	chr17:72487747:T>G,chr17:72487749:A>T
chr17	72487747	72487750	Hyper	3.697846	0	0	0	contamination_double_snps	chr17:72487747:T>G,chr17:72487749:A>T
chr6	11499886	11499889	Hypo	3.697846	0	0	0	contamination_double_snps	chr6:11499886:A>T,chr6:11499888:A>T
chr6	11499886	11499889	Hyper	3.697846	0	0	0	contamination_double_snps	chr6:11499886:A>T,chr6:11499888:A>T
chr3	125991181	125991196	Hyper	3.689892	0	0	0	contamination_double_snps	chr3:125991181:G>C,chr3:125991195:G>C
chr3	125991181	125991196	Hypo	3.689892	0	0	0	contamination_double_snps	chr3:125991181:G>C,chr3:125991195:G>C
chr6	7018489	7018491	Hyper	3.689704	0	0	0	contamination_double_snps	chr6:7018489:G>T,chr6:7018490:A>C
chr6	7018489	7018491	Hypo	3.689704	0	0	0	contamination_double_snps	chr6:7018489:G>T,chr6:7018490:A>C
chr2	165174445	165174447	Hypo	3.689704	0	0	0	contamination_double_snps	chr2:165174445:T>G,chr2:165174446:T>A
chr2	165174445	165174447	Hyper	3.689704	0	0	0	contamination_double_snps	chr2:165174445:T>G,chr2:165174446:T>A
chr5	14982926	14982928	Hypo	3.689704	0	0	0	contamination_double_snps	chr5:14982926:T>A,chr5:14982927:T>G
chr5	14982926	14982928	Hyper	3.689704	0	0	0	contamination_double_snps	chr5:14982926:T>A,chr5:14982927:T>G
chr21	26693258	26693260	Hypo	3.689704	0	0	0	contamination_double_snps	chr21:26693258:G>C,chr21:26693259:G>C
chr21	26693258	26693260	Hyper	3.689704	0	0	0	contamination_double_snps	chr21:26693258:G>C,chr21:26693259:G>C
chr20	5897776	5897778	Hyper	3.689704	0	0	0	contamination_double_snps	chr20:5897776:A>C,chr20:5897777:G>C
chr20	5897776	5897778	Hypo	3.689704	0	0	0	contamination_double_snps	chr20:5897776:A>C,chr20:5897777:G>C
chr7	22310848	22310850	Hypo	3.689704	0	0	0	contamination_double_snps	chr7:22310848:T>G,chr7:22310849:T>A
chr7	22310848	22310850	Hyper	3.689704	0	0	0	contamination_double_snps	chr7:22310848:T>G,chr7:22310849:T>A
chr14	59907091	59907093	Hypo	3.689704	0	0	0	contamination_double_snps	chr14:59907091:A>T,chr14:59907092:T>G
chr14	59907091	59907093	Hyper	3.689704	0	0	0	contamination_double_snps	chr14:59907091:A>T,chr14:59907092:T>G
chr2	31210084	31210086	Hypo	3.689704	0	0	0	contamination_double_snps	chr2:31210084:G>T,chr2:31210085:A>T
chr2	31210084	31210086	Hyper	3.689704	0	0	0	contamination_double_snps	chr2:31210084:G>T,chr2:31210085:A>T
chr18	19489088	19489093	Hyper	3.688722	0	0	0	contamination_double_snps	chr18:19489088:G>T,chr18:19489092:T>G
chr18	19489088	19489093	Hypo	3.688722	0	0	0	contamination_double_snps	chr18:19489088:G>T,chr18:19489092:T>G
chr1	193429848	193429853	Hypo	3.688722	0	0	0	contamination_double_snps	chr1:193429848:A>C,chr1:193429852:T>A
chr1	193429848	193429853	Hyper	3.688722	0	0	0	contamination_double_snps	chr1:193429848:A>C,chr1:193429852:T>A
chr3	161980902	161980908	Hypo	3.683856	0	0	0	contamination_double_snps	chr3:161980902:C>A,chr3:161980907:C>G

【図２３－３】

chr3	161980902	161980908	Hyper	3.683856	0	0	0	contamination_double_snps	chr3:161980902:C>A,chr3:161980907:C>G
chr12	116230638	116230647	Hypo	3.682006	0	0	0	contamination_double_snps	chr12:116230638:T>G,chr12:116230646:C>A
chr12	116230638	116230647	Hyper	3.682006	0	0	0	contamination_double_snps	chr12:116230638:T>G,chr12:116230646:C>A
chr10	90602407	90602417	Hyper	3.676441	0	0	0	contamination_double_snps	chr10:90602407:A>C,chr10:90602416:A>C
chr10	90602407	90602417	Hypo	3.676441	0	0	0	contamination_double_snps	chr10:90602407:A>C,chr10:90602416:A>C
chr21	42806163	42806167	Hyper	3.675311	0	0	0	contamination_double_snps	chr21:42806163:T>A,chr21:42806166:A>C
chr21	42806163	42806167	Hypo	3.675311	0	0	0	contamination_double_snps	chr21:42806163:T>A,chr21:42806166:A>C
chr1	158173394	158173400	Hypo	3.67327	0	0	0	contamination_double_snps	chr1:158173394:A>T,chr1:158173399:C>G
chr1	158173394	158173400	Hyper	3.67327	0	0	0	contamination_double_snps	chr1:158173394:A>T,chr1:158173399:C>G
chr2	119563496	119563502	Hypo	3.67327	0	0	0	contamination_double_snps	chr2:119563496:C>G,chr2:119563501:A>T
chr2	119563496	119563502	Hyper	3.67327	0	0	0	contamination_double_snps	chr2:119563496:C>G,chr2:119563501:A>T
chr2	29223423	29223429	Hyper	3.67327	0	0	0	contamination_double_snps	chr2:29223423:G>C,chr2:29223428:C>G
chr2	29223423	29223429	Hypo	3.67327	0	0	0	contamination_double_snps	chr2:29223423:G>C,chr2:29223428:C>G
chr14	58048443	58048453	Hyper	3.672997	0	0	0	contamination_double_snps	chr14:58048443:C>G,chr14:58048452:C>A
chr14	58048443	58048453	Hypo	3.672997	0	0	0	contamination_double_snps	chr14:58048443:C>G,chr14:58048452:C>A
chr12	66409704	66409715	Hyper	3.672906	0	0	0	contamination_double_snps	chr12:66409704:C>G,chr12:66409714:A>C
chr12	66409704	66409715	Hypo	3.672906	0	0	0	contamination_double_snps	chr12:66409704:C>G,chr12:66409714:A>C
chr8	116302992	116302943	Hypo	3.672906	0	0	0	contamination_double_snps	chr8:116302992:C>G,chr8:116302942:G>T
chr8	116302992	116302943	Hyper	3.672906	0	0	0	contamination_double_snps	chr8:116302992:C>G,chr8:116302942:G>T
chr9	73294481	73294487	Hyper	3.669275	0	0	0	contamination_double_snps	chr9:73294481:G>T,chr9:73294486:T>G
chr9	73294481	73294487	Hypo	3.669275	0	0	0	contamination_double_snps	chr9:73294481:G>T,chr9:73294486:T>G
chr12	103565452	103565455	Hyper	3.668533	0	0	0	contamination_double_snps	chr12:103565452:G>C,chr12:103565454:A>T
chr12	103565452	103565455	Hypo	3.668533	0	0	0	contamination_double_snps	chr12:103565452:G>C,chr12:103565454:A>T
chr3	173228907	173228910	Hyper	3.668533	0	0	0	contamination_double_snps	chr3:173228907:G>T,chr3:173228909:C>A
chr3	173228907	173228910	Hypo	3.668533	0	0	0	contamination_double_snps	chr3:173228907:G>T,chr3:173228909:C>A
chr2	22037574	22037577	Hypo	3.668533	0	0	0	contamination_double_snps	chr2:22037574:A>T,chr2:22037576:A>T
chr2	22037574	22037577	Hyper	3.668533	0	0	0	contamination_double_snps	chr2:22037574:A>T,chr2:22037578:A>T
chr4	189476179	189476182	Hyper	3.668533	0	0	0	contamination_double_snps	chr4:189476179:T>G,chr4:189476181:C>G
chr4	189476179	189476182	Hypo	3.668533	0	0	0	contamination_double_snps	chr4:189476179:T>G,chr4:189476181:C>G
chr8	12909682	12909689	Hyper	3.662226	0	0	0	contamination_double_snps	chr8:12909682:T>A,chr8:12909688:T>A
chr8	12909682	12909689	Hypo	3.662226	0	0	0	contamination_double_snps	chr8:12909682:T>G,chr8:12909688:T>A
chr3	132871829	132871832	Hypo	3.658993	0	0	0	contamination_double_snps	chr3:132871829:A>T,chr3:132871831:G>C
chr3	132871829	132871832	Hyper	3.658993	0	0	0	contamination_double_snps	chr3:132871829:A>T,chr3:132871831:G>C
chr6	68101418	68101423	Hypo	3.657268	0	0	0	contamination_double_snps	chr6:68101418:C>G,chr6:68101422:C>A
chr6	68101418	68101423	Hyper	3.657268	0	0	0	contamination_double_snps	chr6:68101418:C>G,chr6:68101422:C>A
chr9	7602715	7602720	Hyper	3.657268	0	0	0	contamination_double_snps	chr9:7602715:A>T,chr9:7602719:G>T
chr9	7602715	7602720	Hypo	3.657268	0	0	0	contamination_double_snps	chr9:7602715:A>T,chr9:7602719:G>T
chr3	16561448	16561457	Hypo	3.655046	0	0	0	contamination_double_snps	chr3:16561448:A>T,chr3:16561456:C>G
chr3	16561448	16561457	Hyper	3.655046	0	0	0	contamination_double_snps	chr3:16561448:A>T,chr3:16561456:C>G
chr13	27309092	27309101	Hyper	3.655046	0	0	0	contamination_double_snps	chr13:27309092:T>G,chr13:27309100:T>A
chr13	27309092	27309101	Hypo	3.655046	0	0	0	contamination_double_snps	chr13:27309092:T>G,chr13:27309100:T>A
chr15	86925782	86925791	Hypo	3.655046	0	0	0	contamination_double_snps	chr15:86925782:C>A,chr15:86925790:C>A

【図２３－４】

chr15	86925782	86925791	Hyper	3.655046	0	0	0	contamination_double_snps	chr15:86925782:C>A,chr15:86925790:C>A
chr5	7210794	7210803	Hypo	3.655046	0	0	0	contamination_double_snps	chr5:7210794:G>T,chr5:7210802:A>C
chr5	7210794	7210803	Hyper	3.655046	0	0	0	contamination_double_snps	chr5:7210794:G>T,chr5:7210802:A>C
chr5	125184016	125184025	Hyper	3.655046	0	0	0	contamination_double_snps	chr5:125184016:C>A,chr5:125184024:C>A
chr5	125184016	125184025	Hypo	3.655046	0	0	0	contamination_double_snps	chr5:125184016:C>A,chr5:125184024:C>A
chr17	41055530	41055538	Hypo	3.652131	0	0	0	contamination_double_snps	chr17:41055530:A>T,chr17:41055537:T>A
chr17	41055530	41055538	Hyper	3.652131	0	0	0	contamination_double_snps	chr17:41055530:A>T,chr17:41055537:T>A
chr18	65627717	65627726	Hyper	3.651479	0	0	0	contamination_double_snps	chr18:65627717:C>G,chr18:65627725:C>G
chr18	65627717	65627726	Hypo	3.651479	0	0	0	contamination_double_snps	chr18:65627717:C>G,chr18:65627725:C>G
chr14	95933431	95933437	Hypo	3.643074	0	0	0	contamination_double_snps	chr14:95933431:G>T,chr14:95933436:A>C
chr14	95933431	95933437	Hyper	3.643074	0	0	0	contamination_double_snps	chr14:95933431:G>T,chr14:95933436:A>C
chr13	72369155	72369159	Hypo	3.64249	0	0	0	contamination_double_snps	chr13:72369155:A>C,chr13:72369158:G>T
chr13	72369155	72369159	Hyper	3.64249	0	0	0	contamination_double_snps	chr13:72369155:A>C,chr13:72369158:G>T
chr1	74682252	74682259	Hypo	3.642371	0	0	0	contamination_double_snps	chr1:74682252:A>T,chr1:74682258:C>A
chr1	74682252	74682259	Hyper	3.642371	0	0	0	contamination_double_snps	chr1:74682252:A>T,chr1:74682258:C>A
chr12	82080537	82080544	Hypo	3.642371	0	0	0	contamination_double_snps	chr12:82080537:G>C,chr12:82080543:T>A
chr12	82080537	82080544	Hyper	3.642371	0	0	0	contamination_double_snps	chr12:82080537:G>C,chr12:82080543:T>A
chr11	28670116	28670122	Hyper	3.63908	0	0	0	contamination_double_snps	chr11:28670116:G>T,chr11:28670121:T>A
chr11	28670116	28670122	Hypo	3.63908	0	0	0	contamination_double_snps	chr11:28670116:G>T,chr11:28670121:T>A
chr11	108429199	108429204	Hypo	3.636842	0	0	0	contamination_double_snps	chr11:108429199:A>C,chr11:108429203:T>G
chr11	108429199	108429204	Hyper	3.636842	0	0	0	contamination_double_snps	chr11:108429199:A>C,chr11:108429203:T>G
chr14	56694584	56694589	Hyper	3.636842	0	0	0	contamination_double_snps	chr14:56694584:C>G,chr14:56694588:T>G
chr14	56694584	56694589	Hypo	3.636842	0	0	0	contamination_double_snps	chr14:56694584:C>G,chr14:56694588:T>G
chr5	106611440	106611447	Hypo	3.632192	0	0	0	contamination_double_snps	chr5:106611440:G>C,chr5:106611446:G>C
chr5	106611440	106611447	Hyper	3.632192	0	0	0	contamination_double_snps	chr5:106611440:G>C,chr5:106611446:G>C
chr13	109403614	109403625	Hypo	3.631402	0	0	0	contamination_double_snps	chr13:109403614:A>T,chr13:109403624:T>A
chr13	109403614	109403625	Hyper	3.631402	0	0	0	contamination_double_snps	chr13:109403614:A>T,chr13:109403624:T>A
chr10	129764438	129764440	Hypo	3.630413	0	0	0	contamination_double_snps	chr10:129764438:C>A,chr10:129764439:T>G
chr10	129764438	129764440	Hyper	3.630413	0	0	0	contamination_double_snps	chr10:129764438:C>A,chr10:129764439:T>G
chr8	41193964	41193966	Hyper	3.630413	0	0	0	contamination_double_snps	chr8:41193964:T>G,chr8:41193965:T>G
chr8	41193964	41193966	Hypo	3.630413	0	0	0	contamination_double_snps	chr8:41193964:T>G,chr8:41193965:T>G
chr6	138820723	138820725	Hypo	3.630413	0	0	0	contamination_double_snps	chr6:138820723:T>A,chr6:138820724:T>A
chr6	138820723	138820725	Hyper	3.630413	0	0	0	contamination_double_snps	chr6:138820723:T>A,chr6:138820724:T>A
chr1	119078310	119078312	Hyper	3.630413	0	0	0	contamination_double_snps	chr1:119078310:G>C,chr1:119078311:G>T
chr1	119078310	119078312	Hypo	3.630413	0	0	0	contamination_double_snps	chr1:119078310:G>C,chr1:119078311:G>T
chr2	223005124	223005126	Hyper	3.630413	0	0	0	contamination_double_snps	chr2:223005124:T>G,chr2:223005125:G>T
chr2	223005124	223005128	Hypo	3.630413	0	0	0	contamination_double_snps	chr2:223005124:T>G,chr2:223005125:G>T
chr3	183542725	183542727	Hypo	3.630413	0	0	0	contamination_double_snps	chr3:183542725:A>T,chr3:183542726:G>T
chr3	183542725	183542727	Hyper	3.630413	0	0	0	contamination_double_snps	chr3:183542725:A>T,chr3:183542726:G>T
chr15	79481082	79481090	Hyper	3.630275	0	0	0	contamination_double_snps	chr15:79481082:A>C,chr15:79481089:G>C
chr15	79481082	79481090	Hypo	3.630275	0	0	0	contamination_double_snps	chr15:79481082:A>C,chr15:79481089:G>C
chr3	21866735	21866744	Hypo	3.624519	0	0	0	contamination_double_snps	chr3:21866735:C>A,chr3:21866743:A>C

【図２３－５】

chr3	21866735	21866744	Hyper	3.624519	0	0	0	contamination_double_snps	chr3:21866735:C>A,chr3:21866743:A>C
chr3	172122239	172122245	Hypo	3.623465	0	0	0	contamination_double_snps	chr3:172122239:G>C,chr3:172122244:T>G
chr3	172122239	172122245	Hyper	3.623465	0	0	0	contamination_double_snps	chr3:172122239:G>C,chr3:172122244:T>G
chr16	79781589	79781595	Hyper	3.623465	0	0	0	contamination_double_snps	chr16:79781589:T>G,chr16:79781594:T>A
chr16	79781589	79781595	Hypo	3.623465	0	0	0	contamination_double_snps	chr16:79781589:T>G,chr16:79781594:T>A
chr18	50927103	50927107	Hypo	3.621176	0	0	0	contamination_double_snps	chr18:50927103:T>G,chr18:50927106:G>C
chr18	50927103	50927107	Hyper	3.621176	0	0	0	contamination_double_snps	chr18:50927103:T>G,chr18:50927106:G>C
chr14	51839978	51839982	Hyper	3.621176	0	0	0	contamination_double_snps	chr14:51839978:A>T,chr14:51839981:C>G
chr14	51839978	51839982	Hypo	3.621176	0	0	0	contamination_double_snps	chr14:51839978:A>T,chr14:51839983:C>G
chr14	72730027	72730038	Hypo	3.619251	0	0	0	contamination_double_snps	chr14:72730027:G>C,chr14:72730037:T>G
chr14	72730027	72730038	Hyper	3.619251	0	0	0	contamination_double_snps	chr14:72730027:G>C,chr14:72730037:T>G
chr16	28073431	28073439	Hypo	3.615818	0	0	0	contamination_double_snps	chr16:28073431:C>G,chr16:28073438:G>C
chr16	28073431	28073439	Hyper	3.615818	0	0	0	contamination_double_snps	chr16:28073431:C>G,chr16:28073438:G>C
chr1	152442406	152442415	Hypo	3.610577	0	0	0	contamination_double_snps	chr1:152442406:A>C,chr1:152442414:A>C
chr1	152442406	152442415	Hyper	3.610577	0	0	0	contamination_double_snps	chr1:152442406:A>C,chr1:152442414:A>C
chr7	14791571	14791581	Hyper	3.607475	0	0	0	contamination_double_snps	chr7:14791571:G>C,chr7:14791580:T>G
chr7	14791571	14791581	Hypo	3.607475	0	0	0	contamination_double_snps	chr7:14791571:G>C,chr7:14791580:T>G
chr7	156880917	156880920	Hyper	3.606937	0	0	0	contamination_double_snps	chr7:156880917:G>T,chr7:156880919:G>C
chr7	156880917	156880920	Hypo	3.606937	0	0	0	contamination_double_snps	chr7:156880917:G>T,chr7:156880919:G>C
chr1	9402517	9402520	Hypo	3.606937	0	0	0	contamination_double_snps	chr1:9402517:T>G,chr1:9402519:T>G
chr1	9402517	9402520	Hyper	3.606937	0	0	0	contamination_double_snps	chr1:9402517:T>G,chr1:9402519:T>G
chr10	8074010	8074021	Hypo	3.606239	0	0	0	contamination_double_snps	chr10:8074010:T>G,chr10:8074020:A>C
chr10	8074010	8074021	Hyper	3.606239	0	0	0	contamination_double_snps	chr10:8074010:T>G,chr10:8074020:A>C
chr5	1938372	1938380	Hyper	3.606015	0	0	0	contamination_double_snps	chr5:1938372:G>C,chr5:1938379:C>G
chr5	1938372	1938380	Hypo	3.606015	0	0	0	contamination_double_snps	chr5:1938372:G>C,chr5:1938379:C>G
chr4	136942434	136947444	Hypo	3.604032	0	0	0	contamination_double_snps	chr4:136942434:C>G,chr4:136942443:T>G
chr4	136942434	136947444	Hyper	3.604032	0	0	0	contamination_double_snps	chr4:136942434:C>G,chr4:136942443:T>G
chr18	60649125	60649131	Hyper	3.603856	0	0	0	contamination_double_snps	chr18:60649125:C>G,chr18:60649130:C>A
chr18	60649125	60649131	Hypo	3.603856	0	0	0	contamination_double_snps	chr18:60649125:C>G,chr18:60649130:C>A
chr8	13516830	13516838	Hyper	3.602317	0	0	0	contamination_double_snps	chr8:13516830:G>C,chr8:13516837:T>G
chr8	13516830	13516838	Hypo	3.602317	0	0	0	contamination_double_snps	chr8:13516830:G>C,chr8:13516837:T>G
chr5	129519146	129519156	Hypo	3.594905	0	0	0	contamination_double_snps	chr5:129519146:C>G,chr5:129519155:C>A
chr5	129519146	129519156	Hyper	3.594905	0	0	0	contamination_double_snps	chr5:129519146:C>G,chr5:129519155:C>A
chr9	14373430	14373432	Hyper	3.594466	0	0	0	contamination_double_snps	chr9:14373430:T>G,chr9:14373431:G>C
chr6	14373430	14373432	Hypo	3.594466	0	0	0	contamination_double_snps	chr6:14373430:T>G,chr6:14373431:G>C
chr1	181425312	181425314	Hypo	3.594466	0	0	0	contamination_double_snps	chr1:181425312:A>C,chr1:181425313:T>A
chr1	181425312	181425314	Hyper	3.594466	0	0	0	contamination_double_snps	chr1:181425312:A>C,chr1:181425313:T>A
chr13	96135830	96135832	Hypo	3.594466	0	0	0	contamination_double_snps	chr13:96135830:C>G,chr13:96135831:G>C
chr13	96135830	96135832	Hyper	3.594466	0	0	0	contamination_double_snps	chr13:96135830:C>G,chr13:96135831:G>C
chr2	205138871	205138873	Hypo	3.594466	0	0	0	contamination_double_snps	chr2:205138871:G>C,chr2:205138872:G>C
chr2	205138871	205138873	Hyper	3.594466	0	0	0	contamination_double_snps	chr2:205138871:G>C,chr2:205138872:G>C
chr13	36370309	36370311	Hypo	3.594466	0	0	0	contamination_double_snps	chr13:36370309:A>C,chr13:36370310:A>C

【図２３－６】

chr13	36370309	36370311	Hyper	3.594466	0	0	0	contamination_double_snps
chr12	67481176	67481178	Hypo	3.594466	0	0	0	contamination_double_snps
chr12	67481176	67481178	Hyper	3.594466	0	0	0	contamination_double_snps
chr2	34437929	34437931	Hyper	3.594466	0	0	0	contamination_double_snps
chr2	34437929	34437931	Hyper	3.594466	0	0	0	contamination_double_snps
chr16	48704867	48704869	Hyper	3.594466	0	0	0	contamination_double_snps
chr16	48704867	48704869	Hypo	3.594466	0	0	0	contamination_double_snps
chr1	77110658	77110660	Hypo	3.594466	0	0	0	contamination_double_snps
chr1	77110658	77110660	Hyper	3.594466	0	0	0	contamination_double_snps
chr8	60109227	60109229	Hypo	3.594466	0	0	0	contamination_double_snps
chr8	60109227	60109229	Hyper	3.594466	0	0	0	contamination_double_snps
chr18	66252180	66252182	Hyper	3.594466	0	0	0	contamination_double_snps
chr18	66252180	66252182	Hypo	3.594466	0	0	0	contamination_double_snps
chr9	107477916	107477916	Hyper	3.594466	0	0	0	contamination_double_snps
chr9	107477916	107477916	Hypo	3.594466	0	0	0	contamination_double_snps
chr12	107086239	107086245	Hypo	3.59327	0	0	0	contamination_double_snps
chr12	107086239	107086245	Hyper	3.59327	0	0	0	contamination_double_snps
chr11	107121965	107121971	Hyper	3.59327	0	0	0	contamination_double_snps
chr11	107121965	107121971	Hypo	3.59327	0	0	0	contamination_double_snps
chr17	51655065	51655071	Hypo	3.59327	0	0	0	contamination_double_snps
chr17	51655065	51655071	Hyper	3.59327	0	0	0	contamination_double_snps
chr3	30124939	30124950	Hyper	3.589898	0	0	0	contamination_double_snps
chr3	30124939	30124950	Hypo	3.589898	0	0	0	contamination_double_snps
chr7	50878597	50878601	Hypo	3.588354	0	0	0	contamination_double_snps
chr7	50878597	50878601	Hyper	3.588354	0	0	0	contamination_double_snps
chr2	124366185	124366189	Hypo	3.588354	0	0	0	contamination_double_snps
chr2	124366185	124366189	Hyper	3.588354	0	0	0	contamination_double_snps
chr16	58296039	58296047	Hypo	3.587859	0	0	0	contamination_double_snps
chr16	58296039	58296047	Hyper	3.587859	0	0	0	contamination_double_snps
chr1	217912904	217912908	Hypo	3.584012	0	0	0	contamination_double_snps
chr1	217912904	217912908	Hyper	3.584012	0	0	0	contamination_double_snps
chr21	39695308	39695317	Hypo	3.583617	0	0	0	contamination_double_snps
chr21	39695308	39695317	Hyper	3.583617	0	0	0	contamination_double_snps
chr2	202597628	202597637	Hypo	3.583617	0	0	0	contamination_double_snps
chr2	202597628	202597637	Hyper	3.583617	0	0	0	contamination_double_snps
chr2	208495050	208495059	Hypo	3.583617	0	0	0	contamination_double_snps
chr2	208495050	208495059	Hyper	3.583617	0	0	0	contamination_double_snps
chr2	137967024	137967034	Hypo	3.581445	0	0	0	contamination_double_snps
chr2	137967024	137967034	Hyper	3.581445	0	0	0	contamination_double_snps
chr8	76306266	76306266	Hypo	3.581445	0	0	0	contamination_double_snps
chr8	76306256	76306266	Hyper	3.581445	0	0	0	contamination_double_snps
chr3	182670985	182670993	Hypo	3.578056	0	0	0	contamination_double_snps

【図２３－７】

chr3	182670985	182670993	Hyper	3.578056	0	0	0	contamination_double_snps
chr8	37221556	37221564	Hypo	3.578056	0	0	0	contamination_double_snps
chr8	37221556	37221564	Hyper	3.578056	0	0	0	contamination_double_snps
chr10	14832299	14832309	Hypo	3.578001	0	0	0	contamination_double_snps
chr10	14832299	14832309	Hyper	3.578001	0	0	0	contamination_double_snps
chr5	133558682	133558692	Hypo	3.578001	0	0	0	contamination_double_snps
chr5	133558682	133558692	Hyper	3.578001	0	0	0	contamination_double_snps
chr16	83343537	83343542	Hypo	3.573935	0	0	0	contamination_double_snps
chr16	83343537	83343542	Hyper	3.573935	0	0	0	contamination_double_snps
chr11	3193472	3193475	Hypo	3.572624	0	0	0	contamination_double_snps
chr11	3193472	3193475	Hyper	3.572624	0	0	0	contamination_double_snps
chr11	123362757	123362760	Hypo	3.572624	0	0	0	contamination_double_snps
chr11	123362757	123362760	Hyper	3.572624	0	0	0	contamination_double_snps
chr12	52727255	52727258	Hypo	3.572624	0	0	0	contamination_double_snps
chr12	52727255	52727258	Hyper	3.572624	0	0	0	contamination_double_snps
chr17	80807467	80807470	Hypo	3.572624	0	0	0	contamination_double_snps
chr17	80807467	80807470	Hyper	3.572624	0	0	0	contamination_double_snps
chr13	24234368	24234375	Hypo	3.565448	0	0	0	contamination_double_snps
chr13	24234368	24234375	Hyper	3.565448	0	0	0	contamination_double_snps
chr3	72211115	72211126	Hypo	3.564735	0	0	0	contamination_double_snps
chr3	72211115	72211126	Hyper	3.564735	0	0	0	contamination_double_snps
chr4	20238978	20238988	Hypo	3.56454	0	0	0	contamination_double_snps
chr4	20238978	20238988	Hyper	3.56454	0	0	0	contamination_double_snps
chr6	48504480	48504491	Hypo	3.561406	0	0	0	contamination_double_snps
chr6	48504480	48504491	Hyper	3.561406	0	0	0	contamination_double_snps
chr5	84177956	84177964	Hypo	3.5599	0	0	0	contamination_double_snps
chr5	84177956	84177964	Hyper	3.5599	0	0	0	contamination_double_snps
chr2	45209306	45209314	Hypo	3.5599	0	0	0	contamination_double_snps
chr2	45209306	45209314	Hyper	3.5599	0	0	0	contamination_double_snps
chr17	48048713	48048721	Hypo	3.5599	0	0	0	contamination_double_snps
chr17	48048713	48048721	Hyper	3.5599	0	0	0	contamination_double_snps
chr1	177321385	177321393	Hypo	3.5599	0	0	0	contamination_double_snps
chr1	177321385	177321393	Hyper	3.5599	0	0	0	contamination_double_snps
chr20	22014858	22014861	Hypo	3.558519	0	0	0	contamination_double_snps
chr20	22014859	22014861	Hyper	3.558519	0	0	0	contamination_double_snps
chr10	104244321	104244323	Hypo	3.558519	0	0	0	contamination_double_snps
chr10	104244321	104244323	Hyper	3.558519	0	0	0	contamination_double_snps
chr6	45563734	45563736	Hypo	3.558519	0	0	0	contamination_double_snps
chr6	45563734	45563736	Hyper	3.558519	0	0	0	contamination_double_snps
chr1	219925001	219925059	Hypo	3.558519	0	0	0	contamination_double_snps
chr1	219925001	219925003	Hyper	3.558519	0	0	0	contamination_double_snps
chr10	103048085	103048087	Hypo	3.558519	0	0	0	contamination_double_snps

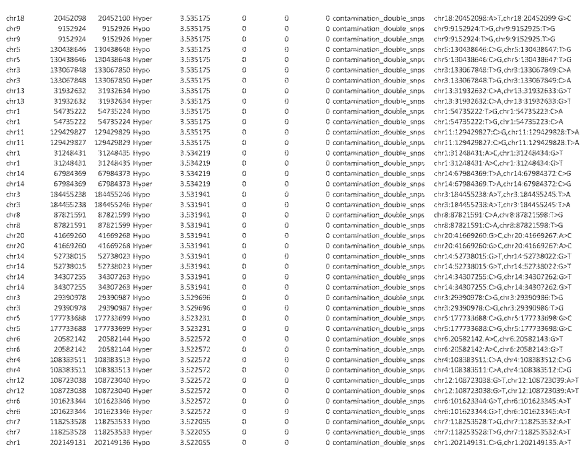
【図２３－８】

chr10	103048085	103048087	Hyper	3.558519	0	0	0	contamination_double_snps
chr4	98803265	98803267	Hypo	3.558519	0	0	0	contamination_double_snps
chr4	98803265	98803267	Hyper	3.558519	0	0	0	contamination_double_snps
chr1	207124929	207124931	Hypo	3.558519	0	0	0	contamination_double_snps
chr1	207124929	207124931	Hyper	3.558519	0	0	0	contamination_double_snps
chr15	71538443	71538445	Hypo	3.558519	0	0	0	contamination_double_snps
chr15	71538443	71538445	Hyper	3.558519	0	0	0	contamination_double_snps
chr10	29016273	29016275	Hypo	3.558519	0	0	0	contamination_double_snps
chr10	29016273	29016275	Hyper	3.558519	0	0	0	contamination_double_snps
chr2	185609754	185609756	Hypo	3.558519	0	0	0	contamination_double_snps
chr2	185609754	185609756	Hyper	3.558519	0	0	0	contamination_double_snps
chr6	145953337	145953339	Hypo	3.558519	0	0	0	contamination_double_snps
chr6	145953337	145953339	Hyper	3.558519	0	0	0	contamination_double_snps
chr1	246161252	246161254	Hypo	3.558519	0	0	0	contamination_double_snps
chr1	246161252	246161254	Hyper	3.558519	0	0	0	contamination_double_snps
chr13	53644106	53644108	Hypo	3.558519	0	0	0	contamination_double_snps
chr13	53644106	53644108	Hyper	3.558519	0	0	0	contamination_double_snps
chr10	78908653	78908655	Hypo	3.558519	0	0	0	contamination_double_snps
chr10	78908653	78908655	Hyper	3.558519	0	0	0	contamination_double_snps
chr7	82954614	82954616	Hypo	3.558519	0	0	0	contamination_double_snps
chr7	82954614	82954616	Hyper	3.558519	0	0	0	contamination_double_snps
chr17	71157069	71157071	Hypo	3.558519	0	0	0	contamination_double_snps
chr17	71157069	71157071	Hyper	3.558519	0	0	0	contamination_double_snps
chr4	91296993	91296995	Hypo	3.558519	0	0	0	contamination_double_snps
chr4	91296993	91296995	Hyper	3.558519	0	0	0	contamination_double_snps
chr7	104017222	104017224	Hypo	3.558519	0	0	0	contamination_double_snps
chr7	104017222	104017224	Hyper	3.558519	0	0	0	contamination_double_snps
chr9	28628805	28628807	Hypo	3.558519	0	0	0	contamination_double_snps
chr9	28628805	28628807	Hyper	3.558519	0	0	0	contamination_double_snps
chr2	6156880	6156882	Hypo	3.558519	0	0	0	contamination_double_snps
chr2	6156880	6156882	Hyper	3.558519	0	0	0	contamination_double_snps
chr14	55905230	55905232	Hypo	3.558519	0	0	0	contamination_double_snps
chr14	55905230	55905232	Hyper	3.558519	0	0	0	contamination_double_snps
chr20	55131786	55131788	Hypo	3.558519	0	0	0	contamination_double_snps
chr20	55131786	55131788	Hyper	3.558519	0	0	0	contamination_double_snps
chr4	36055521	36055523	Hypo	3.558519	0	0	0	contamination_double_snps
chr4	36055521	36055523	Hyper	3.558519	0	0	0	contamination_double_snps
chr16	79212606	79212615	Hypo	3.556657	0	0	0	contamination_double_snps
chr16	79212606	79212615	Hyper	3.556657	0	0	0	contamination_double_snps
chr3	126088329	126088338	Hypo	3.556657	0	0	0	contamination_double_snps
chr3	126088329	126088338	Hyper	3.556657	0	0	0	contamination_double_snps
chr7	147970415	147970424	Hypo	3.556657	0	0	0	contamination_double_snps

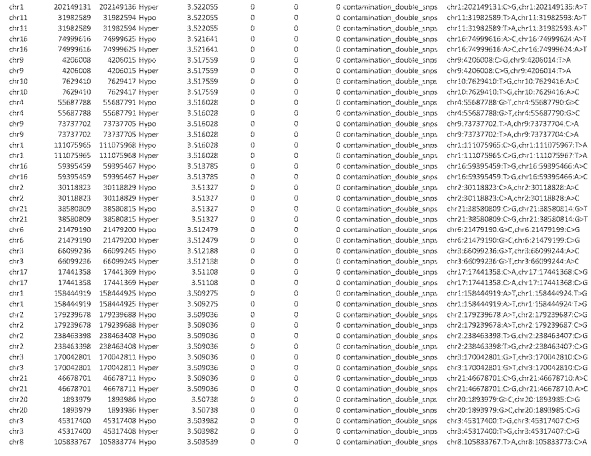
【図２３－９】

chr7	147970415	147970424	Hyper	3.556657	0	0	0	contamination_double_snps
chr2	101341843	101341847	Hypo	3.555533	0	0	0	contamination_double_snps
chr2	101341843	101341847	Hyper	3.555533	0	0	0	contamination_double_snps
chr8	6092094	6092101	Hypo	3.555269	0	0	0	contamination_double_snps
chr8	6092094	6092101	Hyper	3.555269	0	0	0	contamination_double_snps
chr14	21495348	21495355	Hypo	3.555269	0	0	0	contamination_double_snps
chr14	21495348	21495355	Hyper	3.555269	0	0	0	contamination_double_snps
chr17	60831836	60831847	Hypo	3.552584	0	0	0	contamination_double_snps
chr17	60831836	60831847	Hyper	3.552584	0	0	0	contamination_double_snps
chr3	68585765	68585775	Hypo	3.551971	0	0	0	contamination_double_snps
chr3	68585765	68585775	Hyper	3.551971	0	0	0	contamination_double_snps
chr2	109881551	109881557	Hypo	3.543465	0	0	0	contamination_double_snps
chr2	109881551	109881557	Hyper	3.543465	0	0	0	contamination_double_snps
chr4	44908412	44908422	Hypo	3.539572	0	0	0	contamination_double_snps
chr4	44908412	44908423	Hyper	3.539572	0	0	0	contamination_double_snps
chr9	33146392	33146401	Hypo	3.539149	0	0	0	contamination_double_snps
chr9	33146392	33146401	Hyper	3.539149	0	0	0	contamination_double_snps
chr5	67412587	67412596	Hypo	3.539149	0	0	0	contamination_double_snps
chr5	67412587	67412596	Hyper	3.539149	0	0	0	contamination_double_snps
chr1	98918433	98918443	Hypo	3.53851	0	0	0	contamination_double_snps
chr1	98918433	98918443	Hyper	3.53851	0	0	0	contamination_double_snps
chr1	30656037	30656047	Hypo	3.53851	0	0	0	contamination_double_snps
chr1	30656037	30656047	Hyper	3.53851	0	0	0	contamination_double_snps
chr4	44707938	44707941	Hypo	3.538311	0	0	0	contamination_double_snps
chr4	44707938	44707941	Hyper	3.538311	0	0	0	contamination_double_snps
chr7	14032649	14032656	Hypo	3.536414	0	0	0	contamination_double_snps
chr7	14032649	14032656	Hyper	3.536414	0	0	0	contamination_double_snps
chr2	133440237	133440237	Hypo	3.536243	0	0	0	contamination_double_snps
chr2	133440237	133440237	Hyper	3.536243	0	0	0	contamination_double_snps
chr1	214619100	214619102	Hypo	3.535175	0	0	0	contamination_double_snps
chr1	214619100	214619102	Hyper	3.535175	0	0	0	contamination_double_snps
chr12	7297418	7297420	Hypo	3.535175	0	0	0	contamination_double_snps
chr12	7297418	7297420	Hyper	3.535175	0	0	0	contamination_double_snps
chr7	16157196	16157198	Hypo	3.535175	0	0	0	contamination_double_snps
chr7	16157196	16157198	Hyper	3.535175	0	0	0	contamination_double_snps
chr13	64668442	64668444	Hypo	3.535175	0	0	0	contamination_double_snps
chr13	64668442	64668444	Hyper	3.535175	0	0	0	contamination_double_snps
chr7	81789565	81789567	Hypo	3.535175	0	0	0	contamination_double_snps
chr7	81789565	81789567	Hyper	3.535175	0	0	0	contamination_double_snps
chr2	209949181	209949183	Hypo	3.535175	0	0	0	contamination_double_snps
chr2	209949181	209949183	Hyper	3.535175	0	0	0	contamination_double_snps
chr18	20452098	20452100	Hypo	3.535175	0	0	0	contamination_double_snps

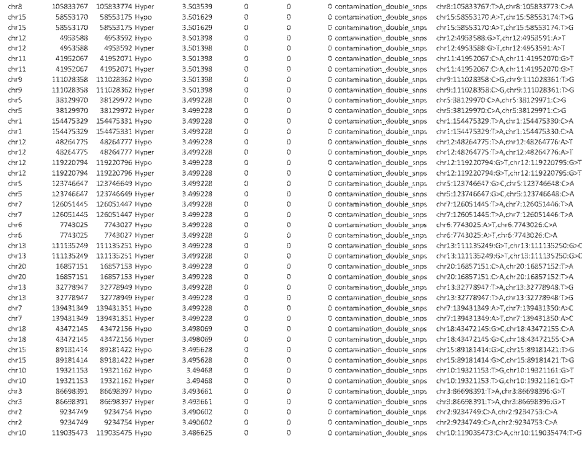
【図２３－１０】



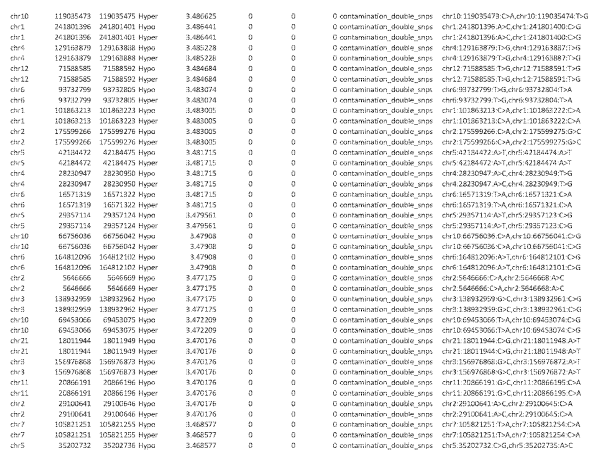
【図２３－１１】



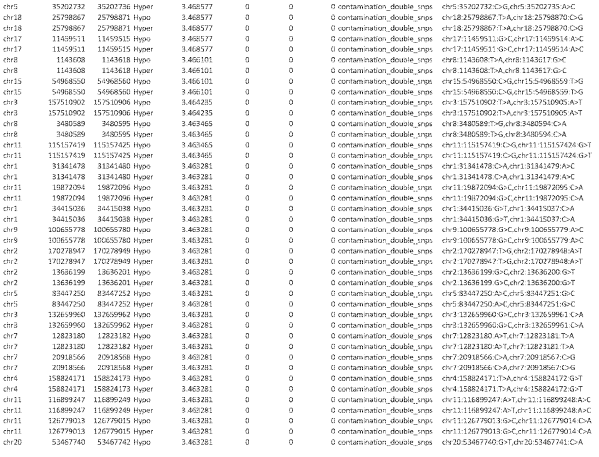
【図２３－１２】



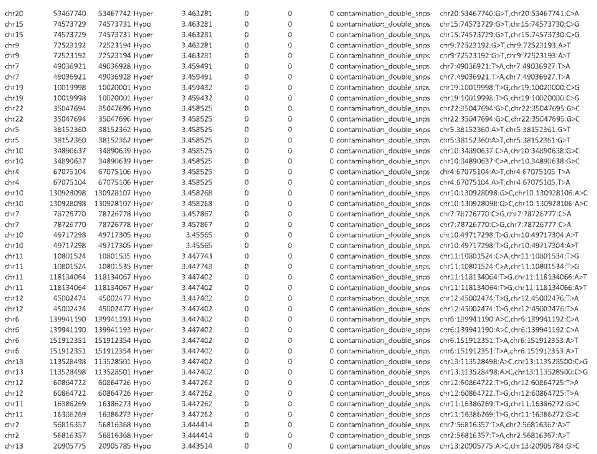
【図２３－１３】



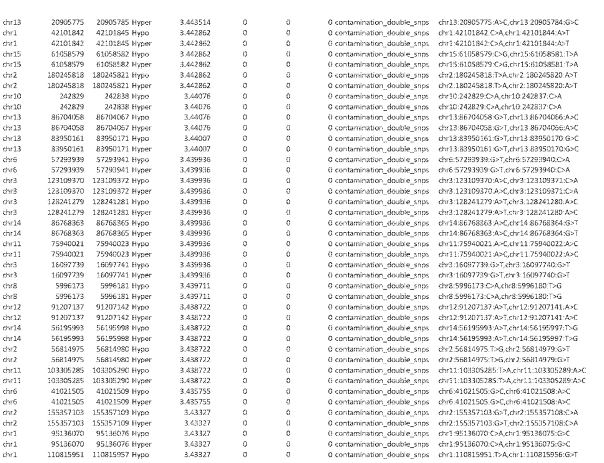
【図２３－１４】



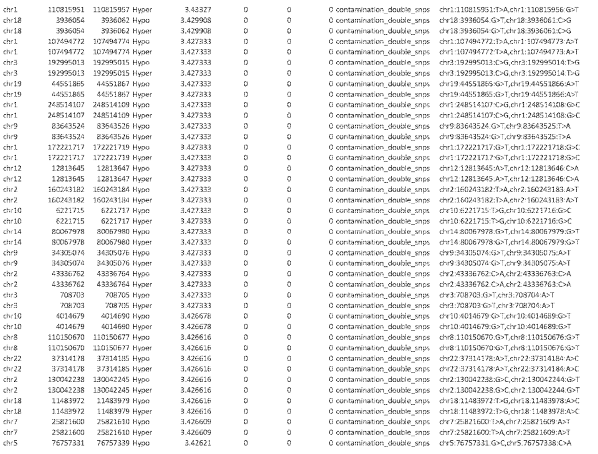
【図２３－１５】



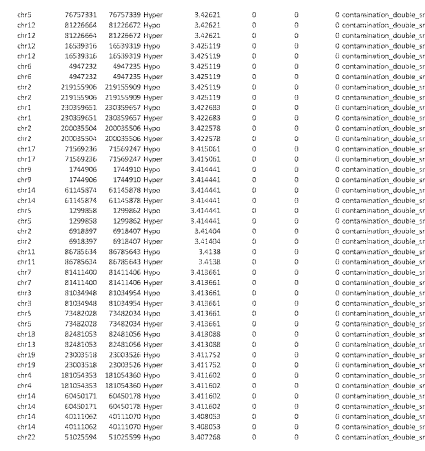
【図２３－１６】



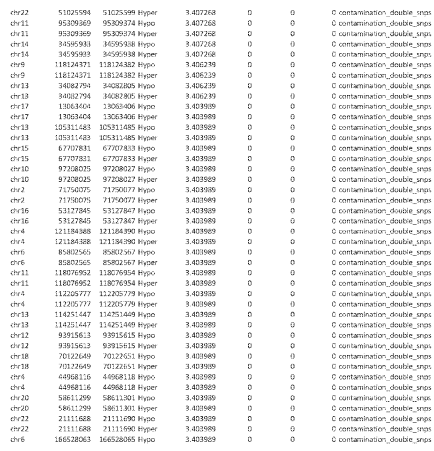
【図２３－１７】



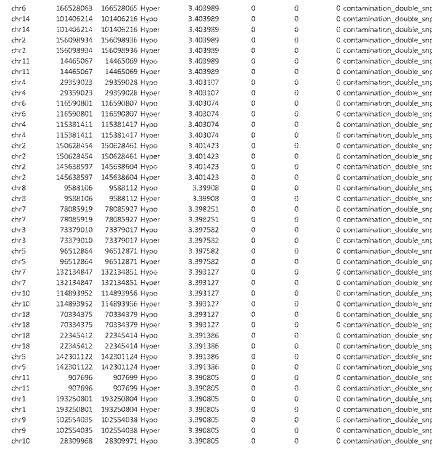
【図２３－１８】



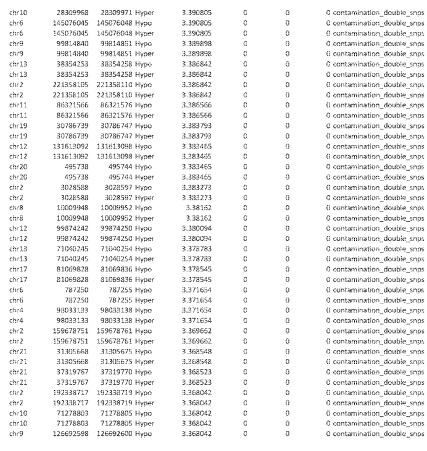
【図２３－１９】



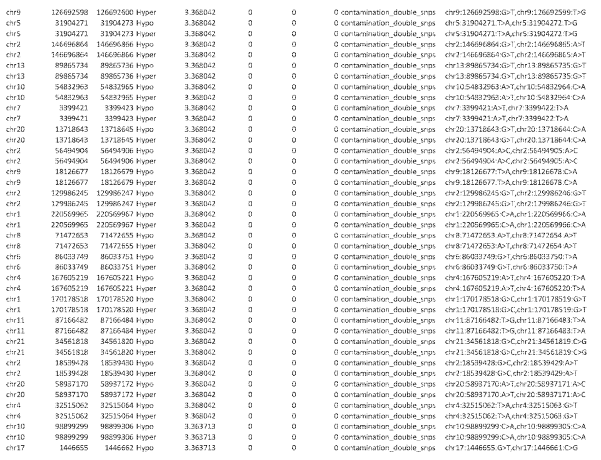
【図２３－２０】



【図２３－２１】

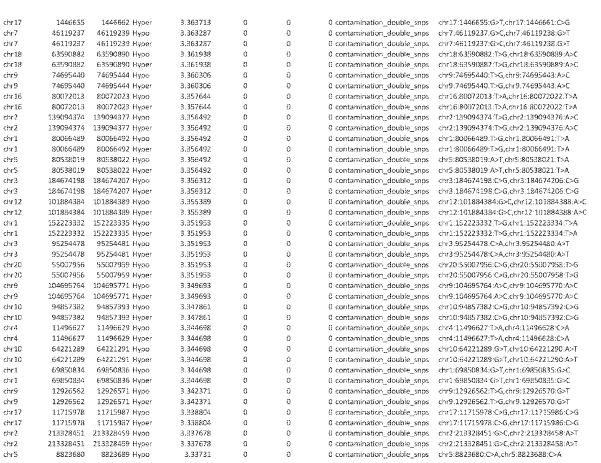


【図２３－２２】



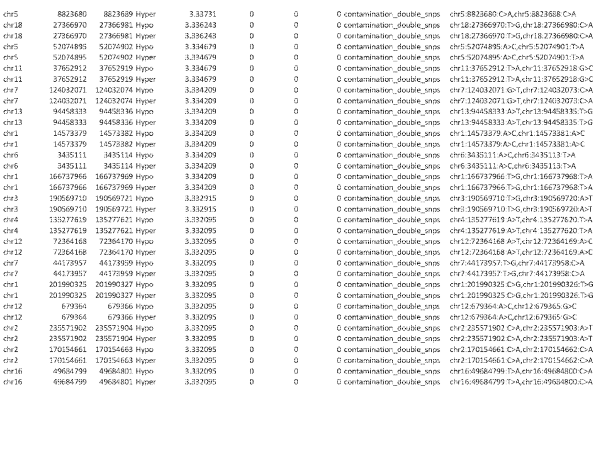
chr9	126692598	126697600	Hyper	3.368042	0	0	0 contamination_double_snps	chr9:126692598:G>T,chr9:126692599:T>G
chr5	31904271	31904273	Hyper	3.368042	0	0	0 contamination_double_snps	chr5:31904271:T>A,chr5:31904272:T>G
chr5	31904271	31904273	Hyper	3.368042	0	0	0 contamination_double_snps	chr5:31904271:T>A,chr5:31904272:T>G
chr2	146696864	146696866	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:146696864:G>T,chr2:146696865:A>T
chr2	146696864	146696866	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:146696864:G>T,chr2:146696865:A>T
chr13	89865734	89865736	Hyper	3.368042	0	0	0 contamination_double_snps	chr13:89865734:G>T,chr13:89865735:G>T
chr13	89865734	89865736	Hyper	3.368042	0	0	0 contamination_double_snps	chr13:89865734:G>T,chr13:89865735:G>T
chr10	54832963	54832965	Hyper	3.368042	0	0	0 contamination_double_snps	chr10:54832963:A>T,chr10:54832964:C>A
chr10	54832963	54832965	Hyper	3.368042	0	0	0 contamination_double_snps	chr10:54832962:A>T,chr10:54832964:C>A
chr7	3399421	3399423	Hyper	3.368042	0	0	0 contamination_double_snps	chr7:3399421:A>T,chr7:3399422:T>A
chr7	3399421	3399423	Hyper	3.368042	0	0	0 contamination_double_snps	chr7:3399421:A>T,chr7:3399422:T>A
chr20	13718643	13718645	Hyper	3.368042	0	0	0 contamination_double_snps	chr20:13718643:G>T,chr20:13718644:C>A
chr20	13718643	13718645	Hyper	3.368042	0	0	0 contamination_double_snps	chr20:13718643:G>T,chr20:13718644:C>A
chr2	56494904	56494906	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:56494904:A>C,chr2:56494905:A>C
chr2	56494904	56494906	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:56494904:A>C,chr2:56494905:A>C
chr8	18126677	18126679	Hyper	3.368042	0	0	0 contamination_double_snps	chr8:18126677:T>A,chr8:18126678:C>A
chr9	18126677	18126679	Hyper	3.368042	0	0	0 contamination_double_snps	chr9:18126677:T>A,chr9:18126678:C>A
chr2	129986245	129986247	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:129986245:G>T,chr2:129986246:G>T
chr2	129986745	129986747	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:129986745:G>T,chr2:129986746:G>T
chr1	220569965	220569967	Hyper	3.368042	0	0	0 contamination_double_snps	chr1:220569965:C>A,chr1:220569966:C>A
chr1	220569965	220569967	Hyper	3.368042	0	0	0 contamination_double_snps	chr1:220569965:C>A,chr1:220569966:C>A
chr8	71472653	71472655	Hyper	3.368042	0	0	0 contamination_double_snps	chr8:71472653:A>T,chr8:71472654:A>T
chr8	71472653	71472655	Hyper	3.368042	0	0	0 contamination_double_snps	chr8:71472653:A>T,chr8:71472654:A>T
chr6	86033749	86033751	Hyper	3.368042	0	0	0 contamination_double_snps	chr6:86033749:G>T,chr6:86033750:T>A
chr6	86033749	86033751	Hyper	3.368042	0	0	0 contamination_double_snps	chr6:86033749:G>T,chr6:86033750:T>A
chr4	167605219	167605221	Hyper	3.368042	0	0	0 contamination_double_snps	chr4:167605219:A>T,chr4:167605220:T>A
chr4	167605219	167605221	Hyper	3.368042	0	0	0 contamination_double_snps	chr4:167605219:A>T,chr4:167605220:T>A
chr1	170178518	170178520	Hyper	3.368042	0	0	0 contamination_double_snps	chr1:170178518:G>C,chr1:170178519:G>T
chr1	170178518	170178520	Hyper	3.368042	0	0	0 contamination_double_snps	chr1:170178518:G>C,chr1:170178519:G>T
chr11	87166482	87166484	Hyper	3.368042	0	0	0 contamination_double_snps	chr11:87166482:T>G,chr11:87166483:T>A
chr11	87166482	87166484	Hyper	3.368042	0	0	0 contamination_double_snps	chr11:87166482:T>G,chr11:87166483:T>A
chr21	34561818	34561820	Hyper	3.368042	0	0	0 contamination_double_snps	chr21:34561818:G>C,chr21:34561819:C>G
chr21	34561818	34561820	Hyper	3.368042	0	0	0 contamination_double_snps	chr21:34561818:G>C,chr21:34561819:C>G
chr2	185394278	185394430	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:185394278:G>C,chr2:185394429:A>T
chr2	185394278	185394430	Hyper	3.368042	0	0	0 contamination_double_snps	chr2:185394278:G>C,chr2:185394429:A>T
chr20	58937170	58937172	Hyper	3.368042	0	0	0 contamination_double_snps	chr20:58937170:A>T,chr20:58937171:A>C
chr20	58937170	58937172	Hyper	3.368042	0	0	0 contamination_double_snps	chr20:58937170:A>T,chr20:58937171:A>C
chr4	32515062	32515064	Hyper	3.368042	0	0	0 contamination_double_snps	chr4:32515062:T>A,chr4:32515063:G>T
chr4	32515062	32515064	Hyper	3.368042	0	0	0 contamination_double_snps	chr4:32515062:T>A,chr4:32515063:G>T
chr10	98899299	98899306	Hyper	3.363713	0	0	0 contamination_double_snps	chr10:98899299:C>A,chr10:98899305:C>A
chr10	98899299	98899306	Hyper	3.363713	0	0	0 contamination_double_snps	chr10:98899299:C>A,chr10:98899305:C>A
chr17	1466655	1466662	Hyper	3.363713	0	0	0 contamination_double_snps	chr17:1466655:G>T,chr17:1466661:C>G

【図２３－２３】



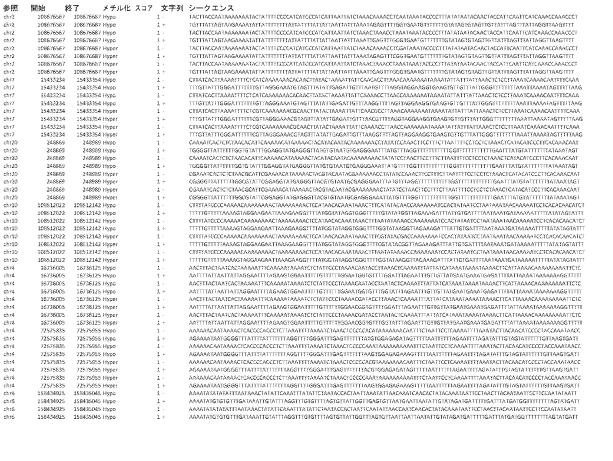
chr17	1466655	1466662	Hyper	3.363713	0	0	0 contamination_double_snps	chr17:1466655:G>T,chr17:1466661:C>G
chr7	46119237	46119239	Hypo	3.363287	0	0	0 contamination_double_snps	chr7:46119237:G>C,chr7:46119238:G>T
chr7	46119237	46119239	Hyper	3.363287	0	0	0 contamination_double_snps	chr7:46119237:G>C,chr7:46119238:G>T
chr18	63590892	63590890	Hyper	3.361938	0	0	0 contamination_double_snps	chr18:63590892:T>G,chr18:63590889:A>C
chr18	63590892	63590890	Hyper	3.361938	0	0	0 contamination_double_snps	chr18:63590887:T>G,chr18:63590889:A>C
chr9	74695440	74695444	Hyper	3.360306	0	0	0 contamination_double_snps	chr9:74695440:T>G,chr9:74695443:A>C
chr9	74695440	74695444	Hyper	3.360306	0	0	0 contamination_double_snps	chr9:74695440:T>G,chr9:74695443:A>C
chr16	80072013	80072023	Hyper	3.357644	0	0	0 contamination_double_snps	chr16:80072013:T>A,chr16:80072022:T>A
chr16	80072013	80072023	Hyper	3.357644	0	0	0 contamination_double_snps	chr16:80072013:T>A,chr16:80072022:T>A
chr7	139094374	139094377	Hyper	3.356492	0	0	0 contamination_double_snps	chr7:139094374:T>G,chr7:139094376:A>C
chr2	139094374	139094377	Hyper	3.356492	0	0	0 contamination_double_snps	chr2:139094374:T>G,chr2:139094376:A>C
chr1	80066489	80066492	Hyper	3.356492	0	0	0 contamination_double_snps	chr1:80066489:T>A,chr1:80066491:T>A
chr1	80066489	80066492	Hyper	3.356492	0	0	0 contamination_double_snps	chr1:80066489:T>A,chr1:80066491:T>A
chr5	80538019	80538022	Hyper	3.356492	0	0	0 contamination_double_snps	chr5:80538019:A>T,chr5:80538021:T>A
chr5	80538019	80538022	Hyper	3.356492	0	0	0 contamination_double_snps	chr5:80538019:A>T,chr5:80538021:T>A
chr3	184674198	184674207	Hyper	3.356312	0	0	0 contamination_double_snps	chr3:184674198:C>G,chr3:184674206:C>G
chr3	184674198	184674207	Hyper	3.356312	0	0	0 contamination_double_snps	chr3:184674198:C>G,chr3:184674205:C>G
chr12	101884384	101884389	Hyper	3.355389	0	0	0 contamination_double_snps	chr12:101884384:G>C,chr12:101884388:A>C
chr12	101884384	101884389	Hyper	3.355389	0	0	0 contamination_double_snps	chr12:101884384:G>C,chr12:101884388:A>C
chr1	152223332	152223335	Hyper	3.351953	0	0	0 contamination_double_snps	chr1:152223332:T>G,chr1:152223334:T>A
chr1	152223332	152223335	Hyper	3.351953	0	0	0 contamination_double_snps	chr1:152223332:T>G,chr1:152223334:T>A
chr3	95254478	95254481	Hyper	3.351953	0	0	0 contamination_double_snps	chr3:95254478:C>A,chr3:95254480:A>T
chr3	95254478	95254481	Hyper	3.351953	0	0	0 contamination_double_snps	chr3:95254478:C>A,chr3:95254480:A>T
chr20	55607956	55607959	Hyper	3.351953	0	0	0 contamination_double_snps	chr20:55607956:C>G,chr20:55607958:T>G
chr20	55607956	55607959	Hyper	3.351953	0	0	0 contamination_double_snps	chr20:55607956:C>G,chr20:55607958:T>G
chr9	104695764	104695771	Hyper	3.348693	0	0	0 contamination_double_snps	chr9:104695764:A>C,chr9:104695770:A>C
chr9	104695764	104695771	Hyper	3.348693	0	0	0 contamination_double_snps	chr9:104695764:A>C,chr9:104695770:A>C
chr10	94857382	94857393	Hyper	3.347861	0	0	0 contamination_double_snps	chr10:94857382:C>G,chr10:94857392:C>G
chr10	94857382	94857393	Hyper	3.347861	0	0	0 contamination_double_snps	chr10:94857382:C>G,chr10:94857392:C>G
chr4	11496627	11496629	Hyper	3.344698	0	0	0 contamination_double_snps	chr4:11496627:T>A,chr4:11496628:C>A
chr4	11496627	11496629	Hyper	3.344698	0	0	0 contamination_double_snps	chr4:11496627:T>A,chr4:11496628:C>A
chr10	64221289	64221291	Hyper	3.344698	0	0	0 contamination_double_snps	chr10:64221289:G>T,chr10:64221290:A>T
chr10	64221289	64221291	Hyper	3.344698	0	0	0 contamination_double_snps	chr10:64221289:G>T,chr10:64221290:A>T
chr1	69850834	69850836	Hyper	3.344698	0	0	0 contamination_double_snps	chr1:69850834:G>T,chr1:69850835:G>C
chr1	69850834	69850836	Hyper	3.344698	0	0	0 contamination_double_snps	chr1:69850834:G>T,chr1:69850835:G>C
chr9	12926562	12926571	Hyper	3.342371	0	0	0 contamination_double_snps	chr9:12926562:T>G,chr9:12926570:G>T
chr9	12926562	12926571	Hyper	3.342371	0	0	0 contamination_double_snps	chr9:12926562:T>G,chr9:12926570:G>T
chr17	11715978	11715987	Hyper	3.338804	0	0	0 contamination_double_snps	chr17:11715978:C>G,chr17:11715986:C>G
chr17	11715978	11715987	Hyper	3.338804	0	0	0 contamination_double_snps	chr17:11715978:C>G,chr17:11715986:C>G
chr2	213328451	213328459	Hyper	3.337678	0	0	0 contamination_double_snps	chr2:213328451:G>C,chr2:213328457:G>C
chr2	213328451	213328459	Hyper	3.337678	0	0	0 contamination_double_snps	chr2:213328451:G>C,chr2:213328458:A>T
chr5	8823680	8823689	Hypo	3.33731	0	0	0 contamination_double_snps	chr5:8823680:C>A,chr5:8823688:C>A

【図２３－２４】



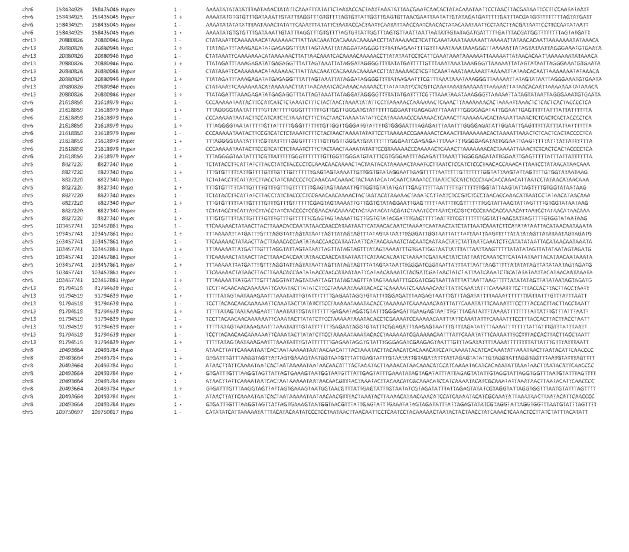
chr5	8823680	8823689	Hyper	3.33731	0	0	0 contamination_double_snps	chr5:8823680:C>A,chr5:8823688:C>A
chr18	27366970	27366981	Hypo	3.336243	0	0	0 contamination_double_snps	chr18:27366970:T>G,chr18:27366980:C>A
chr18	27366970	27366981	Hyper	3.336243	0	0	0 contamination_double_snps	chr18:27366970:T>G,chr18:27366980:C>A
chr5	52074895	52074902	Hypo	3.334679	0	0	0 contamination_double_snps	chr5:52074895:A>C,chr5:52074901:T>A
chr5	52074895	52074902	Hyper	3.334679	0	0	0 contamination_double_snps	chr5:52074895:A>C,chr5:52074901:T>A
chr11	37652912	37652919	Hypo	3.334679	0	0	0 contamination_double_snps	chr11:37652912:T>A,chr11:37652918:G>C
chr11	37652912	37652919	Hyper	3.334679	0	0	0 contamination_double_snps	chr11:37652912:T>A,chr11:37652918:G>C
chr7	124032071	124032074	Hypo	3.334209	0	0	0 contamination_double_snps	chr7:124032071:G>T,chr7:124032073:C>A
chr7	124032071	124032074	Hyper	3.334209	0	0	0 contamination_double_snps	chr7:124032071:G>T,chr7:124032073:C>A
chr13	94458333	94458336	Hypo	3.334209	0	0	0 contamination_double_snps	chr13:94458333:A>T,chr13:94458335:T>G
chr13	94458333	94458336	Hyper	3.334209	0	0	0 contamination_double_snps	chr13:94458333:A>T,chr13:94458335:T>G
chr1	14573379	14573382	Hypo	3.334209	0	0	0 contamination_double_snps	chr1:14573379:A>C,chr1:14573381:A>C
chr1	14573379	14573382	Hyper	3.334209	0	0	0 contamination_double_snps	chr1:14573379:A>C,chr1:14573381:A>C
chr6	3435111	3435114	Hypo	3.334209	0	0	0 contamination_double_snps	chr6:3435111:A>C,chr6:3435113:T>A
chr6	3435111	3435114	Hyper	3.334209	0	0	0 contamination_double_snps	chr6:3435111:A>C,chr6:3435113:T>A
chr1	166737866	166737869	Hypo	3.334209	0	0	0 contamination_double_snps	chr1:166737866:T>G,chr1:166737868:T>A
chr1	166737866	166737869	Hyper	3.334209	0	0	0 contamination_double_snps	chr1:166737866:T>G,chr1:166737868:T>A
chr3	190569710	190569721	Hypo	3.332915	0	0	0 contamination_double_snps	chr3:190569710:T>G,chr3:190569720:A>T
chr3	190569710	190569721	Hyper	3.332915	0	0	0 contamination_double_snps	chr3:190569710:T>G,chr3:190569720:A>T
chr4	135277619	135277621	Hypo	3.332095	0	0	0 contamination_double_snps	chr4:135277619:A>T,chr4:135277620:T>A
chr4	135277619	135277621	Hyper	3.332095	0	0	0 contamination_double_snps	chr4:135277619:A>T,chr4:135277620:T>A
chr12	72364168	72364170	Hyper	3.332095	0	0	0 contamination_double_snps	chr12:72364168:A>T,chr12:72364169:A>C
chr12	72364168	72364170	Hyper	3.332095	0	0	0 contamination_double_snps	chr12:72364168:A>T,chr12:72364169:A>C
chr7	44173957	44173959	Hypo	3.332095	0	0	0 contamination_double_snps	chr7:44173957:T>G,chr7:44173958:C>A
chr7	44173957	44173959	Hyper	3.332095	0	0	0 contamination_double_snps	chr7:44173957:T>G,chr7:44173958:C>A
chr1	201990325	201990327	Hypo	3.332095	0	0	0 contamination_double_snps	chr1:201990325:C>G,chr1:201990326:T>G
chr1	201990325	201990327	Hyper	3.332095	0	0	0 contamination_double_snps	chr1:201990325:C>G,chr1:201990326:T>G
chr12	679364	679366	Hypo	3.332095	0	0	0 contamination_double_snps	chr12:679364:A>C,chr12:679365:G>C
chr12	679364	679366	Hyper	3.332095	0	0	0 contamination_double_snps	chr12:679364:A>C,chr12:679365:G>C
chr2	235571902	235571904	Hypo	3.332095	0	0	0 contamination_double_snps	chr2:235571902:C>A,chr2:235571903:A>T
chr2	235571902	235571904	Hyper	3.332095	0	0	0 contamination_double_snps	chr2:235571902:C>A,chr2:235571903:A>T
chr2	170154661	170154663	Hypo	3.332095	0	0	0 contamination_double_snps	chr2:170154661:C>A,chr2:170154662:C>A
chr2	170154661	170154663	Hyper	3.332095	0	0	0 contamination_double_snps	chr2:170154661:C>A,chr2:170154662:C>A
chr16	49684799	49684801	Hypo	3.332095	0	0	0 contamination_double_snps	chr16:49684798:T>A,chr16:49684800:C>A
chr16	49684799	49684801	Hyper	3.332095	0	0	0 contamination_double_snps	chr16:49684799:T>A,chr16:49684800:C>A

【図２４－１】

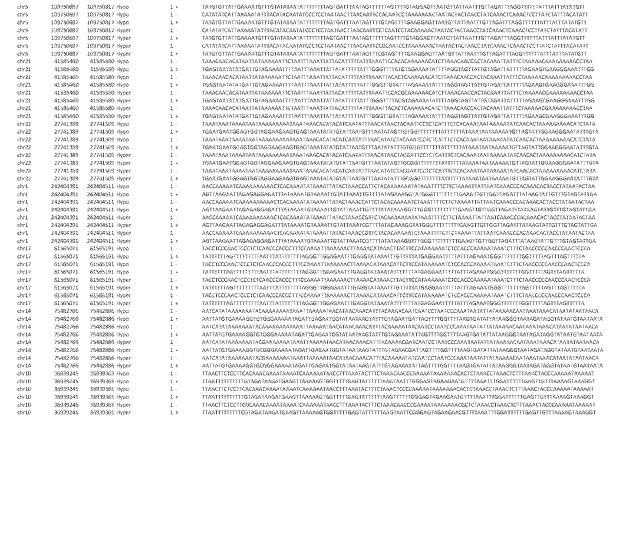


参照	開始	終了	メチル化	スコア	文字列	シーケンス
chr5	10867667	10867667	Hyper		1 +	TACTTGCCATAAAAAATACTATTTTCCCCATCATCCCCATTAATTACTAAAGAAACCTCAATAAATACCCTTATAATAATAGCACTACGATTCAATTCACTAAAACCAAACCCT
chr2	10867667	10867668	Hyper		1 +	TGTTATCTAGTAAGAAAAATATTATTTTTTTTATTATTTATTAAATTAAATAGAGTTTGGTGAATGTTTTTTGTGAGTAGTAGTTTTTTGGTAGTAGTTTTTTTTTTTTTTTTTT
chr2	10867667	10867668	Hyper		1 -	TACTTACCAATAAAAAATACTATTTTCCCCATCATCCCCATTAATTACTAAAGAAACCTCAATAAATACCCTTATAATAATAGCACTACGATTCAATTCACTAAAACCAAACCCT
chr5	10867667	10867668	Hyper		1 -	TGTTATCTAGTAAGAAAAATATTATTTTTTTTATTATTTATTAAATTAAATAGAGTTTGGTGAATGTTTTTTGTGAGTAGTAGTTTTTTGGTAGTAGTTTTTTTTTTTTTTTTT
chr2	10867667	10867667	Hyper		1 -	TACTTACCAATAAAAAATACTATTTTCCCCATCATCCCCATTAATTACTAAAGAAACCTCAATAAATACCCTTATAATAATAGCACTACGATTCAATTCACTAAAACCAAACCCT
chr2	10867667	10867668	Hyper		1 -	TGTTATCTAGTAAGAAAAATATTATTTTTTTTATTATTTATTAAATTAAATAGAGTTTGGTGAATGTTTTTTGGTAGTAGTTTTTTGGTAGTAGTTTTTTTTTTTTTTTTT
chr1	154323234	154323764	Hyper		1 -	CTAACCACTTAAAATTTTCCATCAAAAAAACAGACAACTTATACTAAAATTATTCAAAAGCTTAACCAAAAAA

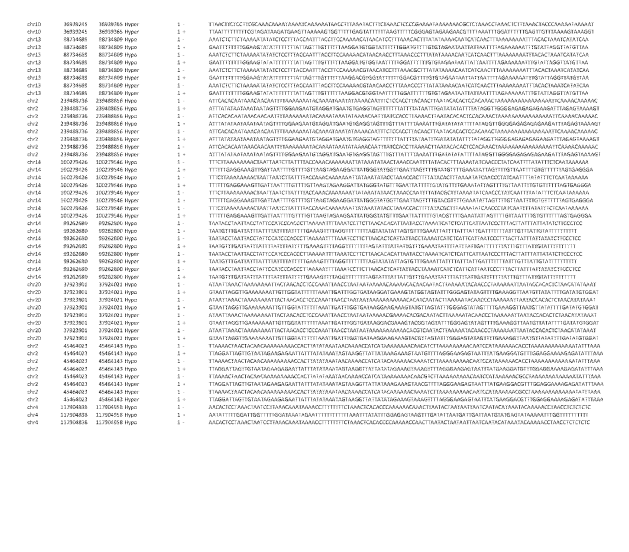
【図２４－２】



【図２４－３】



【図２４－４】



【図２４－５】



【図２４－６】



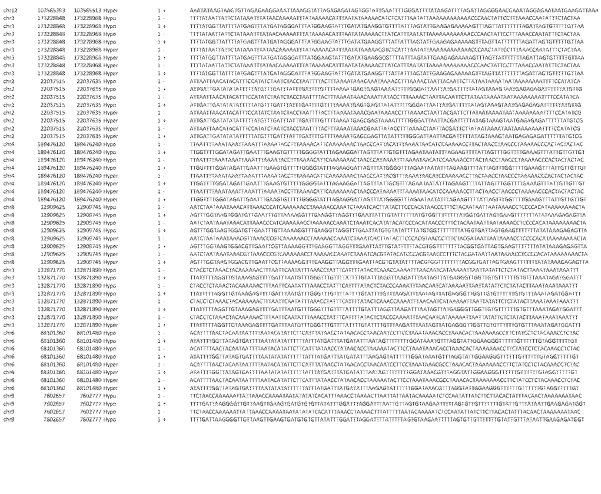
【図２４－７】



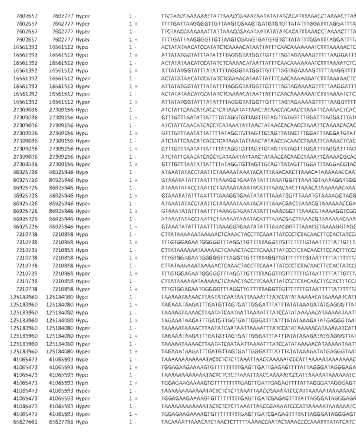
【図２４－８】



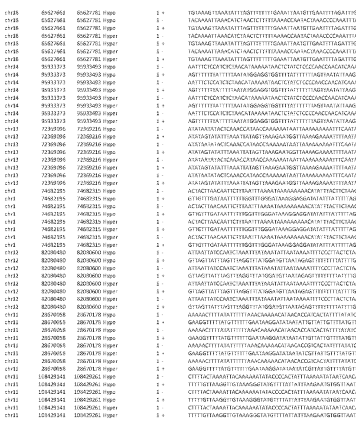
【図２４－９】



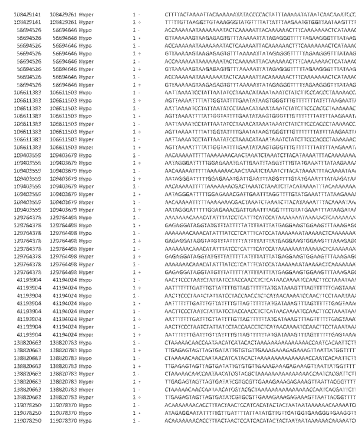
【図２４－１０】



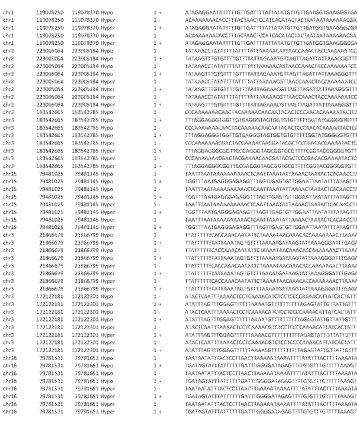
【図２４－１１】



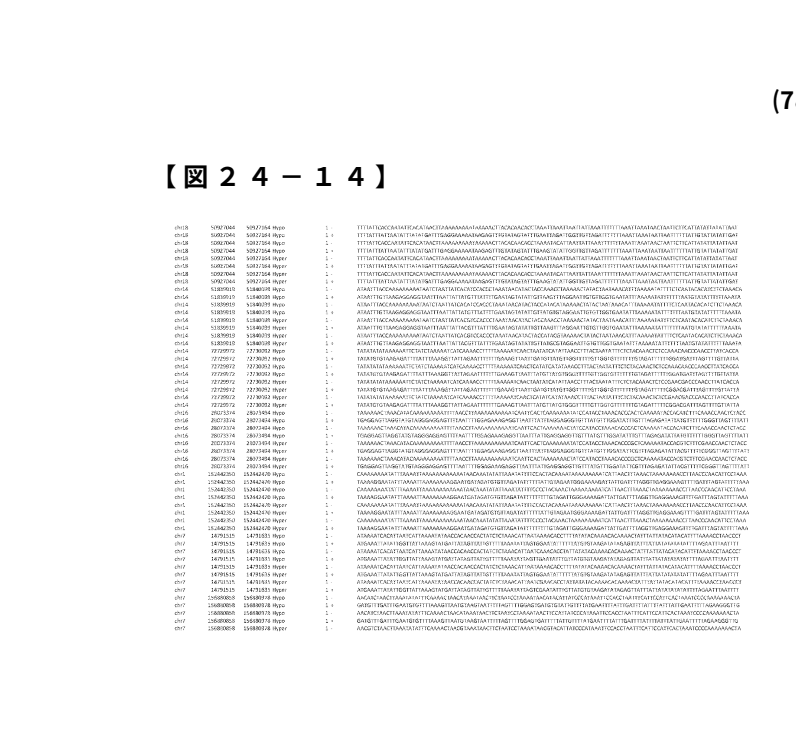
【図２４－１２】



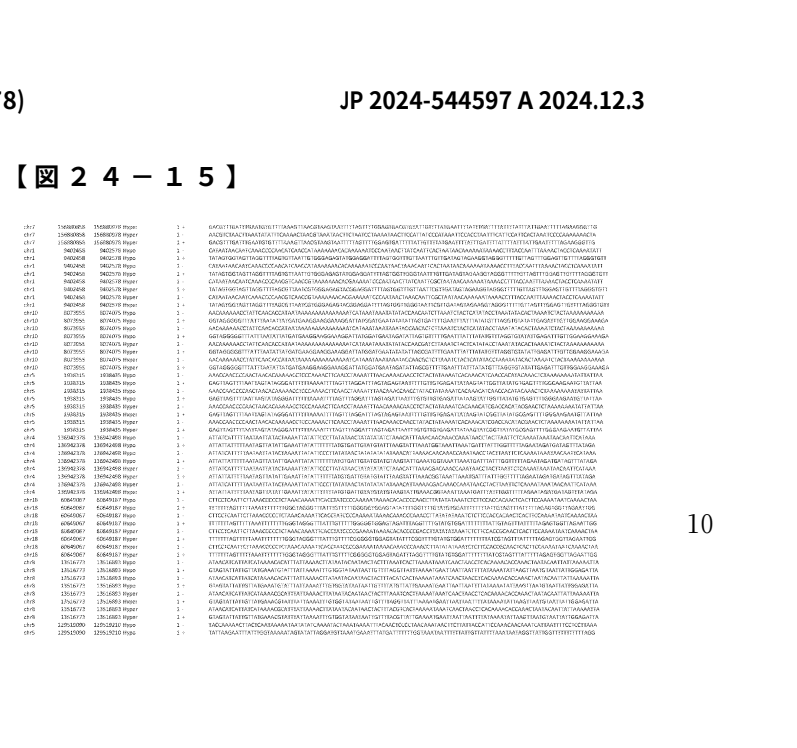
【図２４－１３】



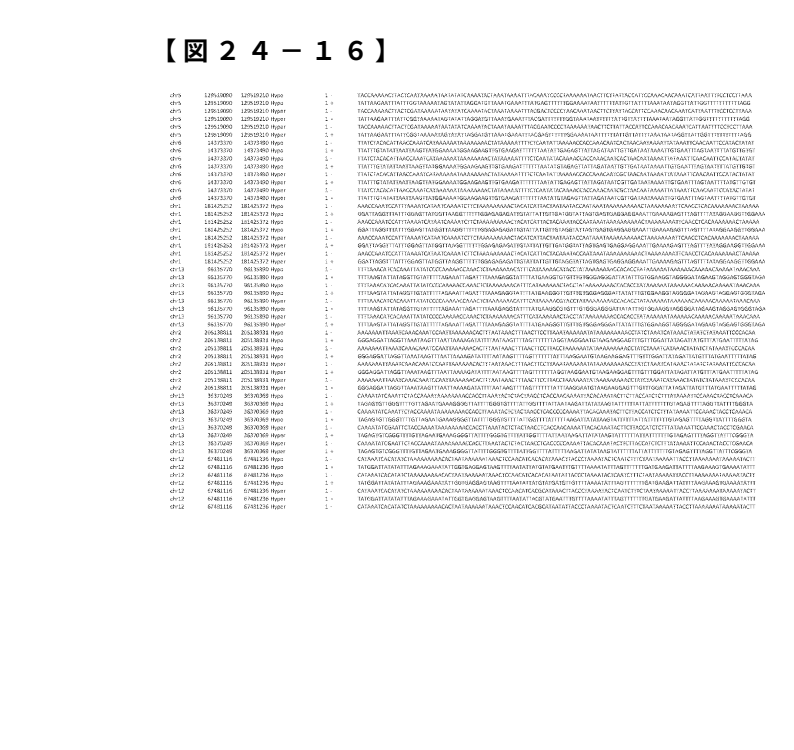
【図２４－１４】



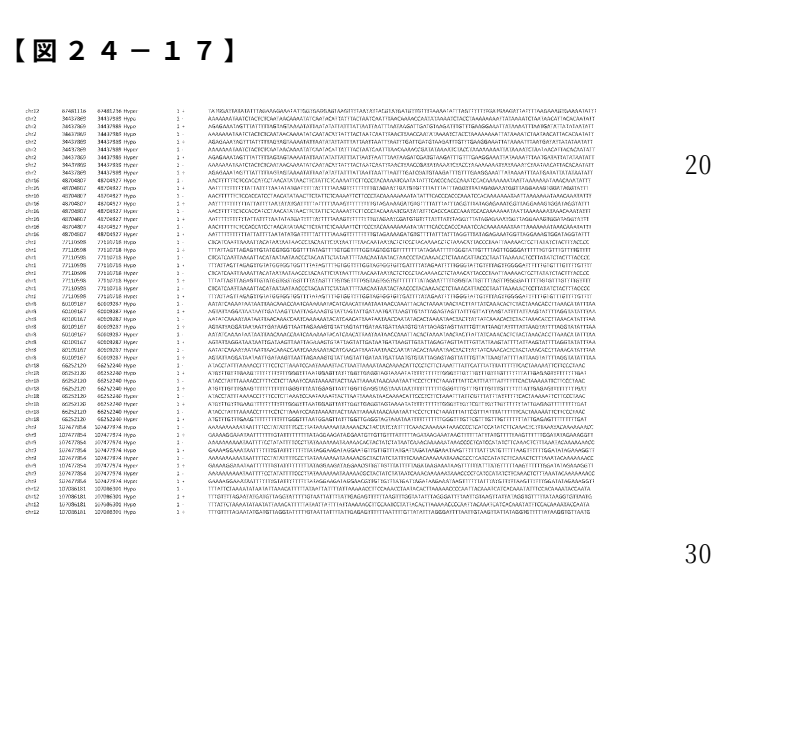
【図２４－１５】



【図２４－１６】



【図２４－１７】



【図２４－１８】



【図２４－１９】



【図２４－２０】



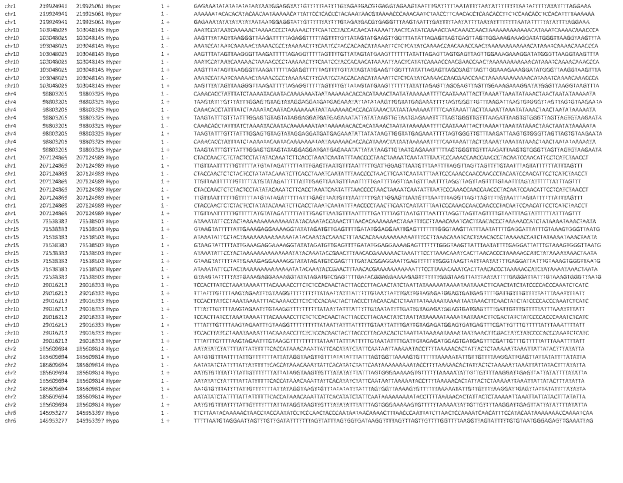
【図２４－２１】



【図２４－２２】



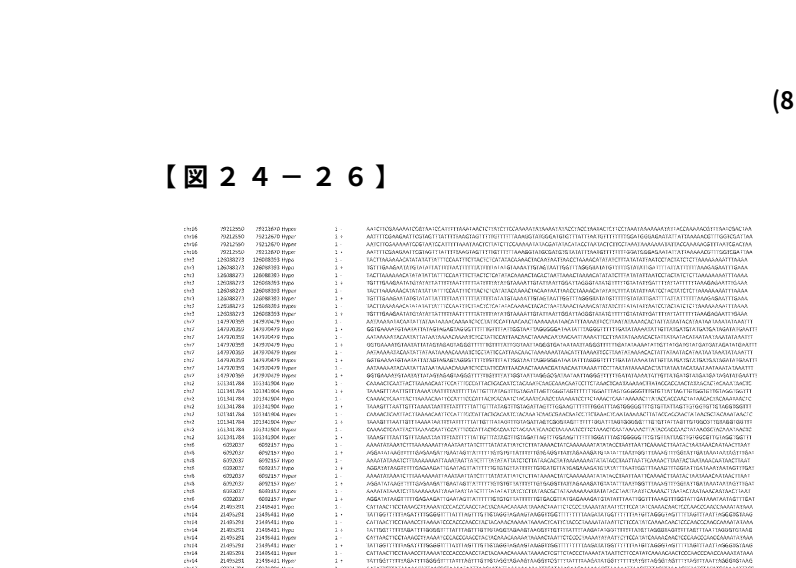
【図２４－２３】



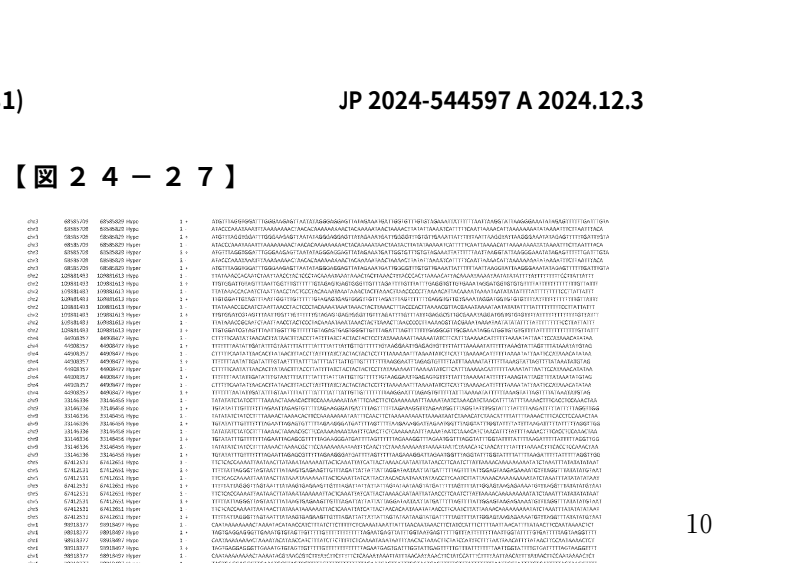
【図２４－２４】

【図２４－２５】

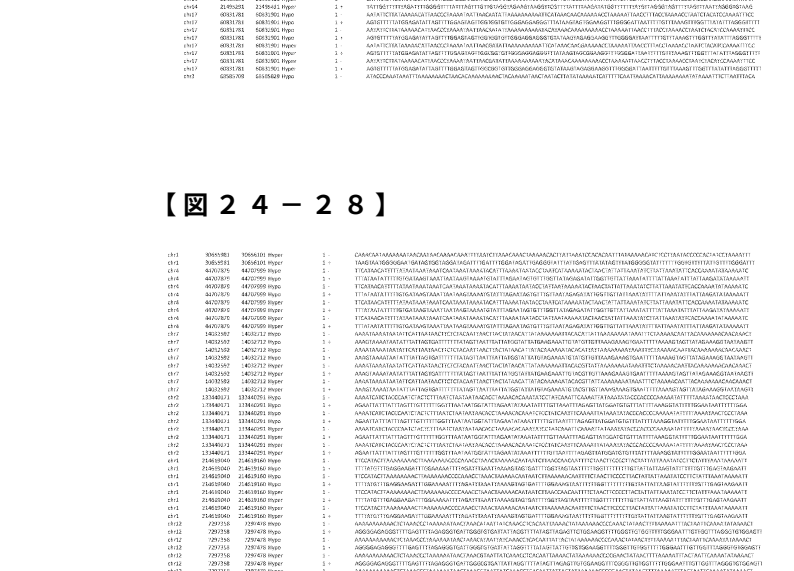
【図２４－２６】



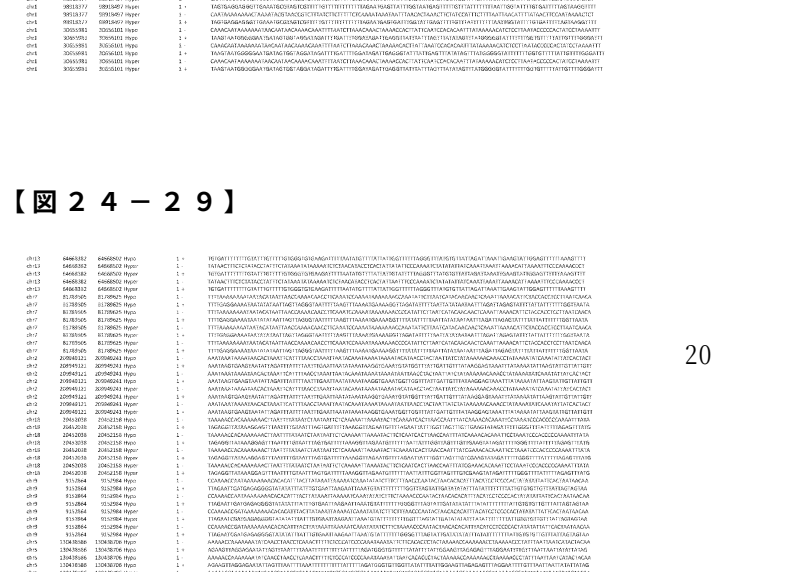
【図２４－２７】



【図２４－２８】



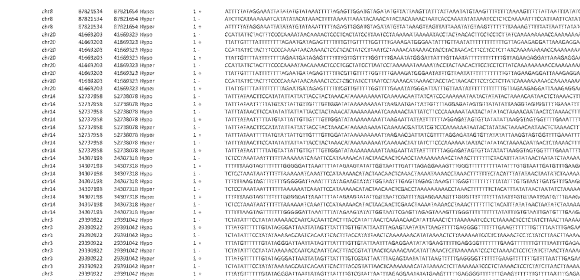
【図２４－２９】



【図２４－３０】



【図２４－３１】



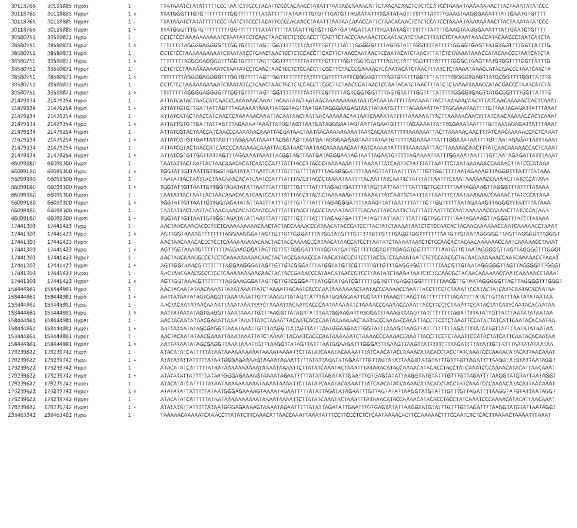
【図２４－３２】



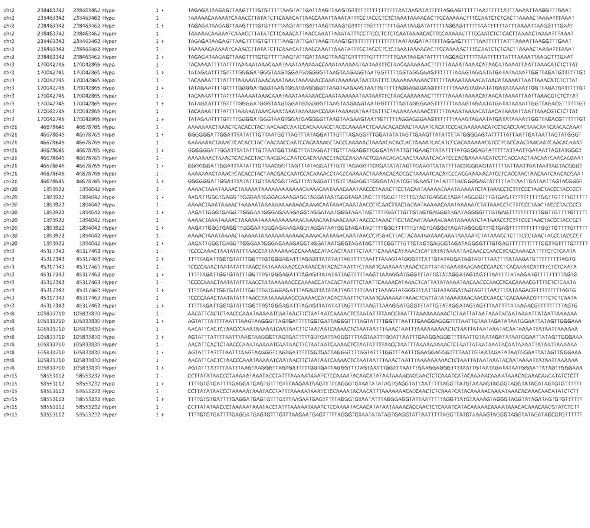
【図２４－３３】



【図２４－３４】



【図２４－３５】



【図２４－３６】



【図２４－３７】



【図２４－３８】



【図２４－３９】



【図２４－４０】



【図２４－４１】



【図２４－４２】



【図２４－４３】



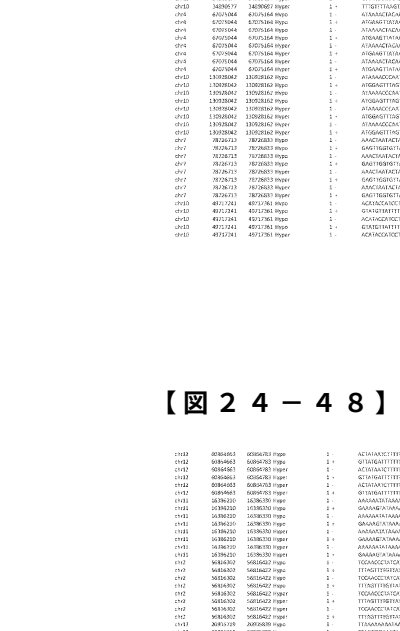
【図２４－４４】



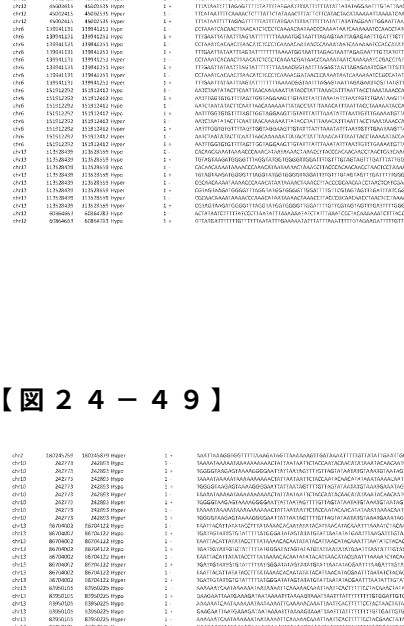
【図２４－４５】



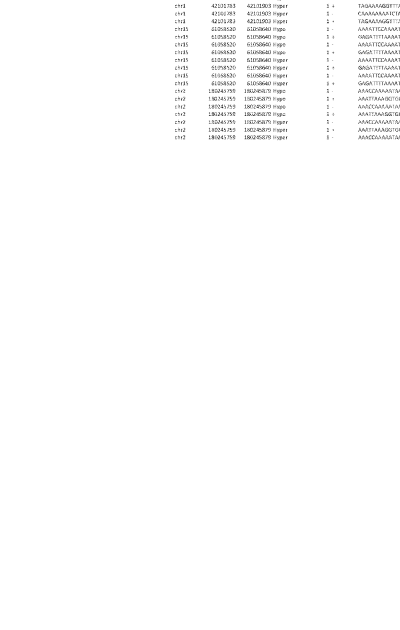
【図２４－４６】



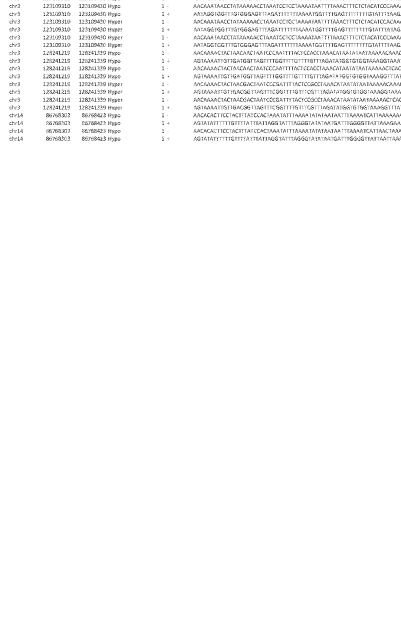
【図２４－４７】



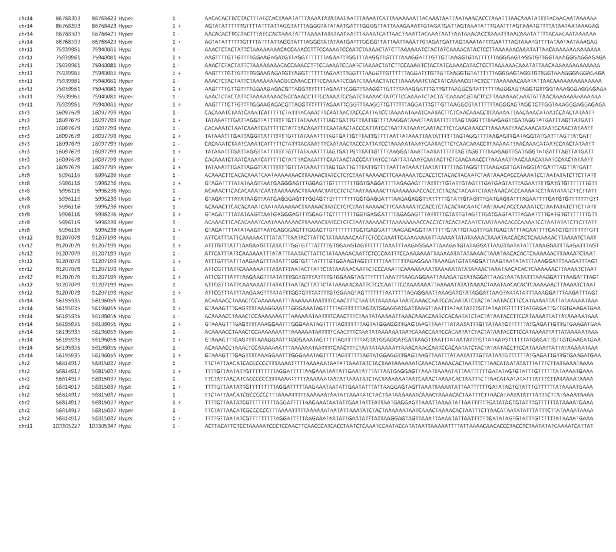
【図２４－４８】



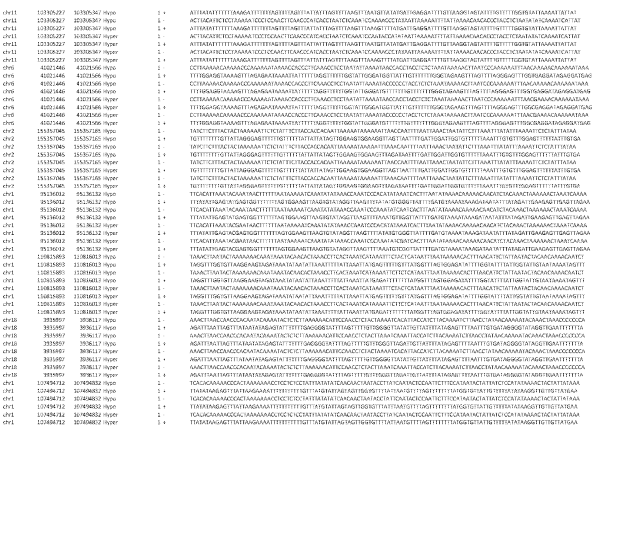
【図２４－４９】



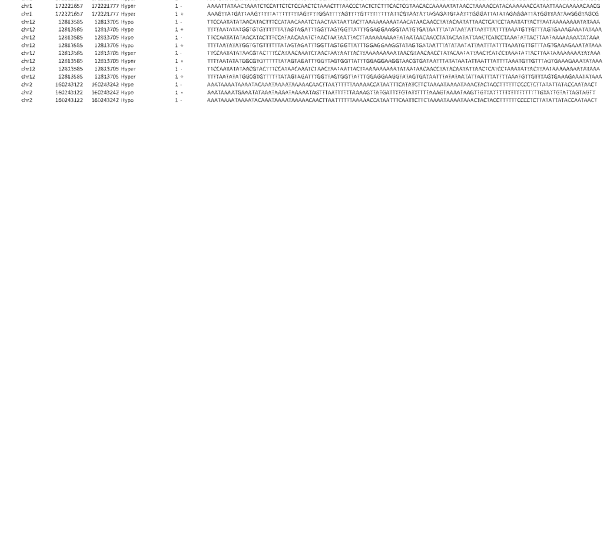
【図２４－５０】



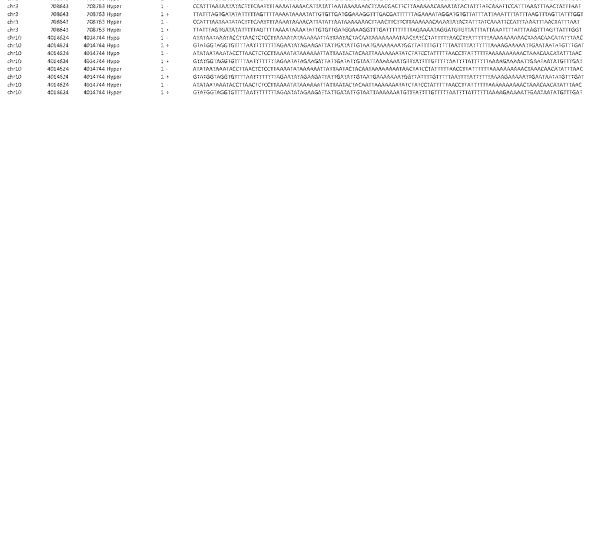
【図２４－５１】



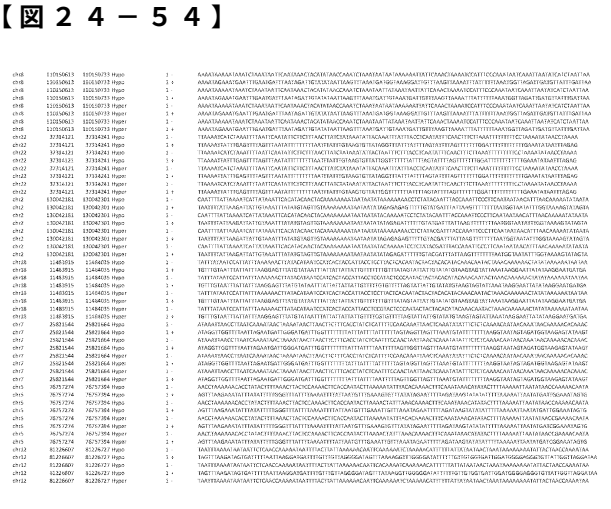
【図２４－５２】



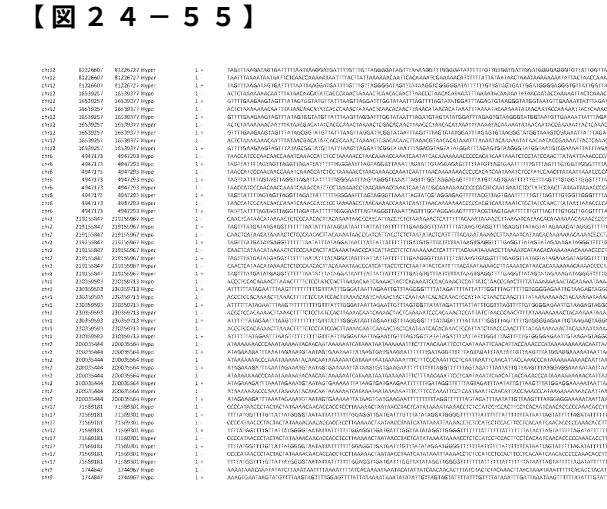
【図２４－５３】



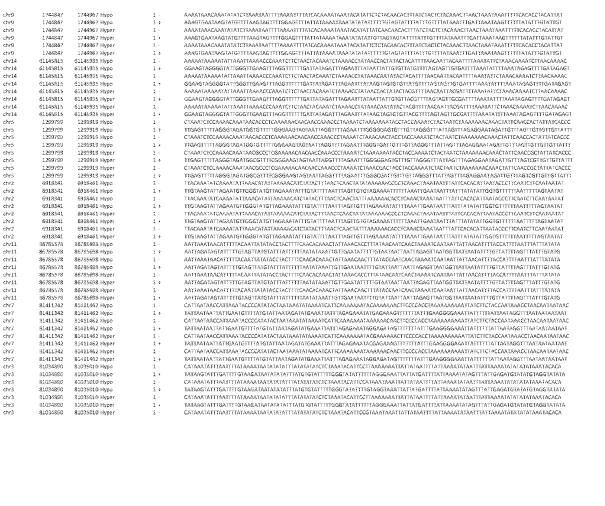
【図２４－５４】



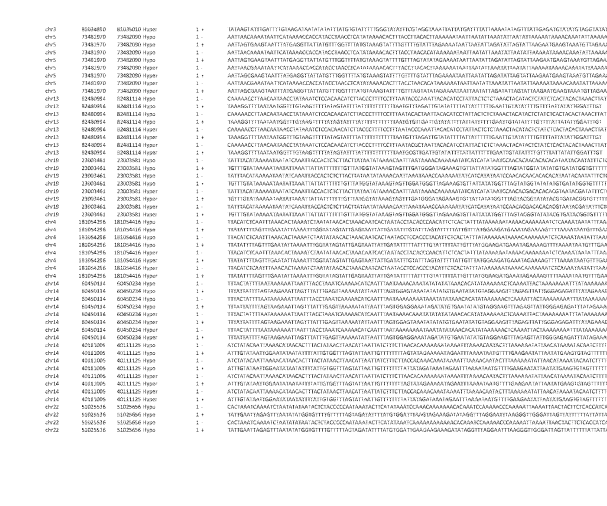
【図２４－５５】



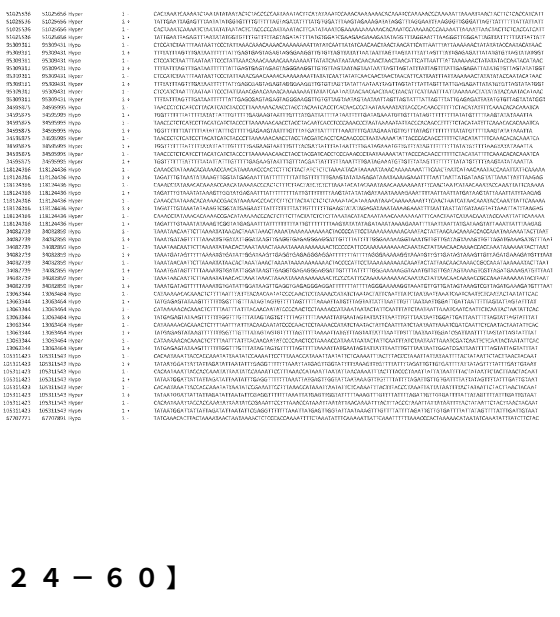
【図２４－５６】



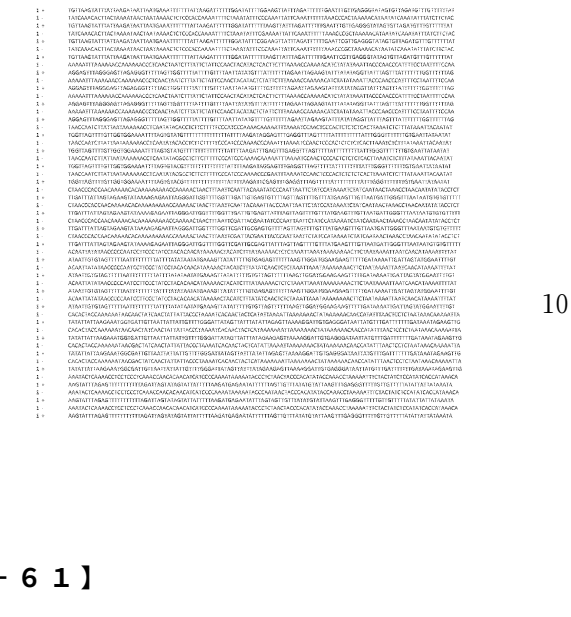
【図２４－５７】



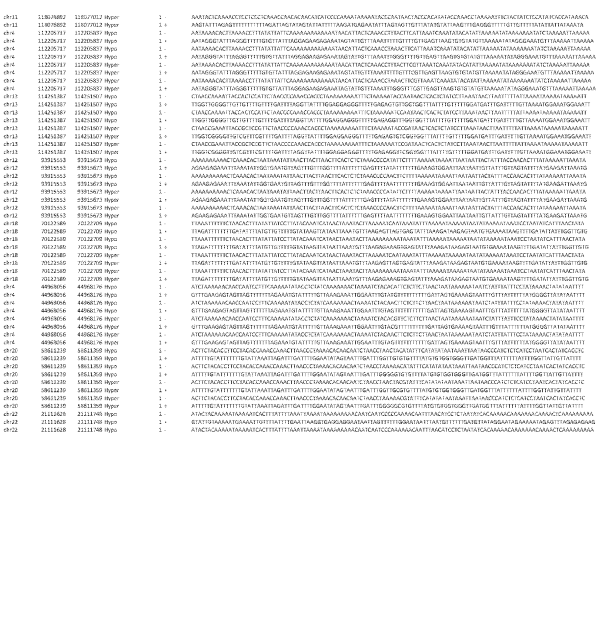
【図２４－５８】



【図２４－５９】



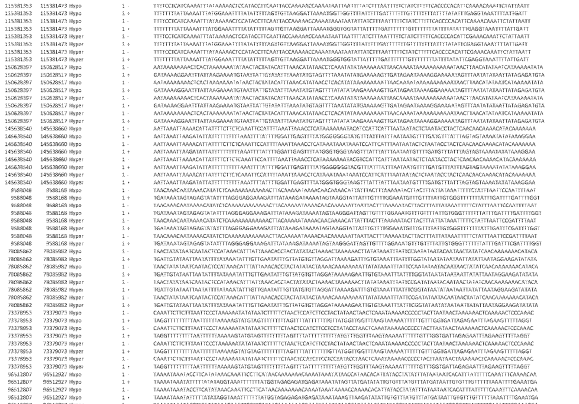
【図２４－６０】



【図２４－６１】



【図２４－６２】



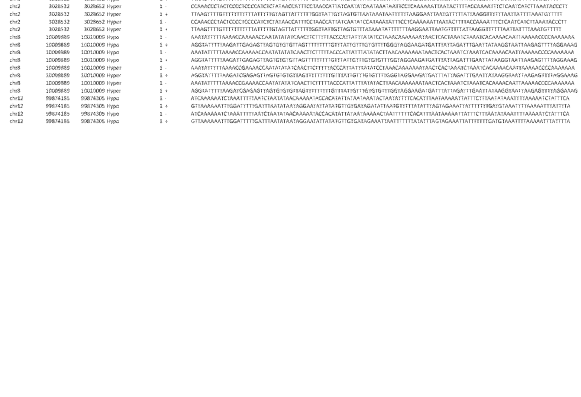
【図２４－６３】



【図２４－６４】



【図２４－６５】



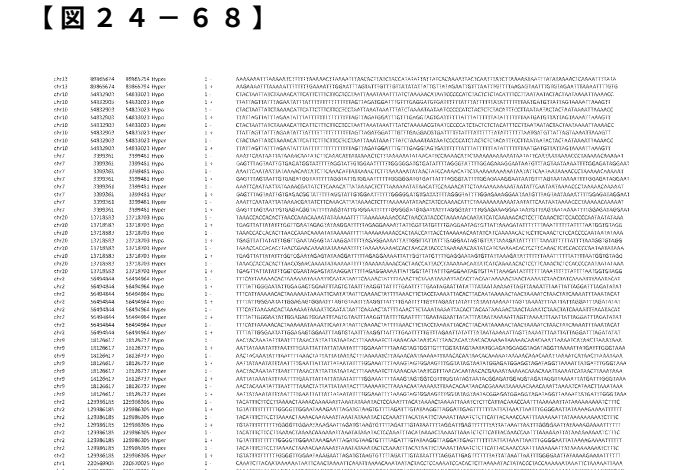
【図２４－６６】



【図２４－６７】



【図２４－６８】



【図２４－６９】



【図２４－７０】



【図２４－７１】

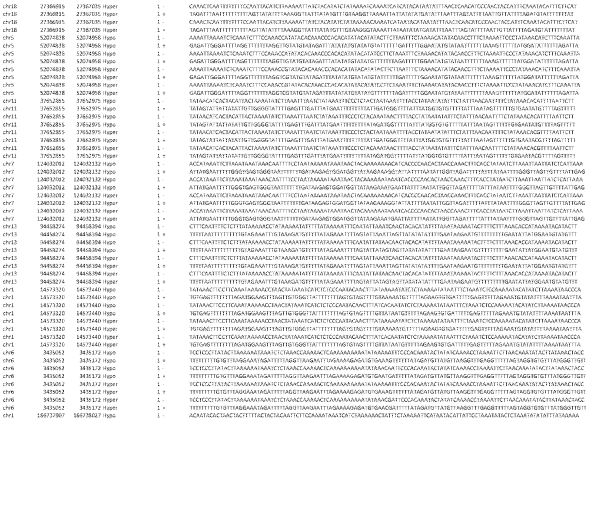


【図２４－７２】

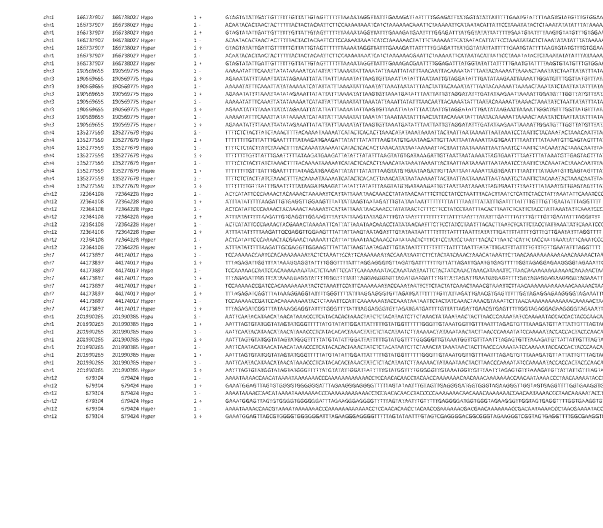
【図２４－７３】



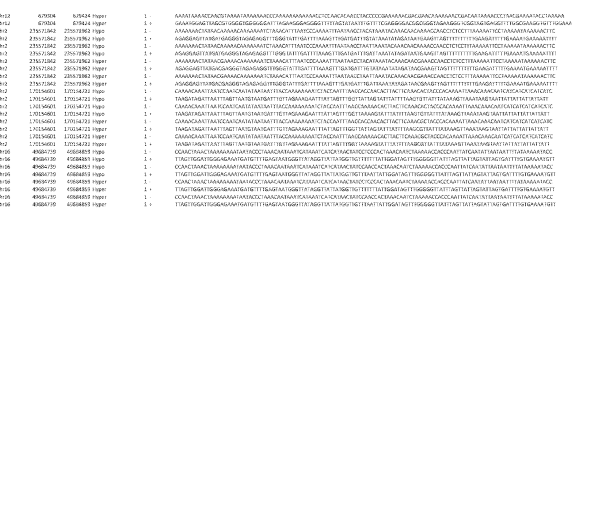
【図２４－７４】



【図２４－７５】



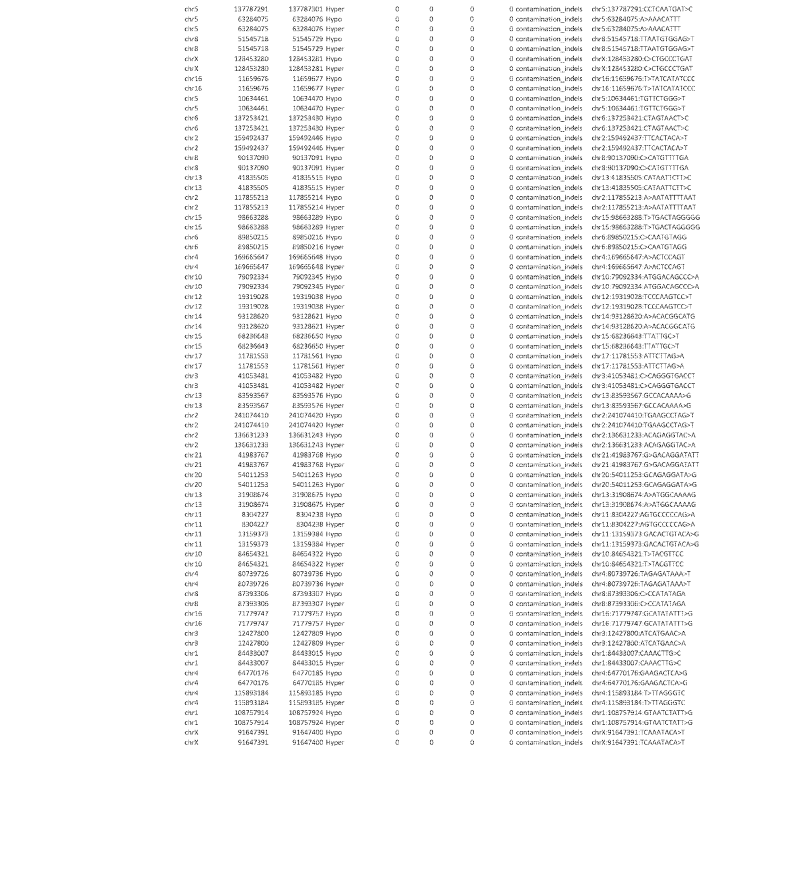
【図２４－７６】



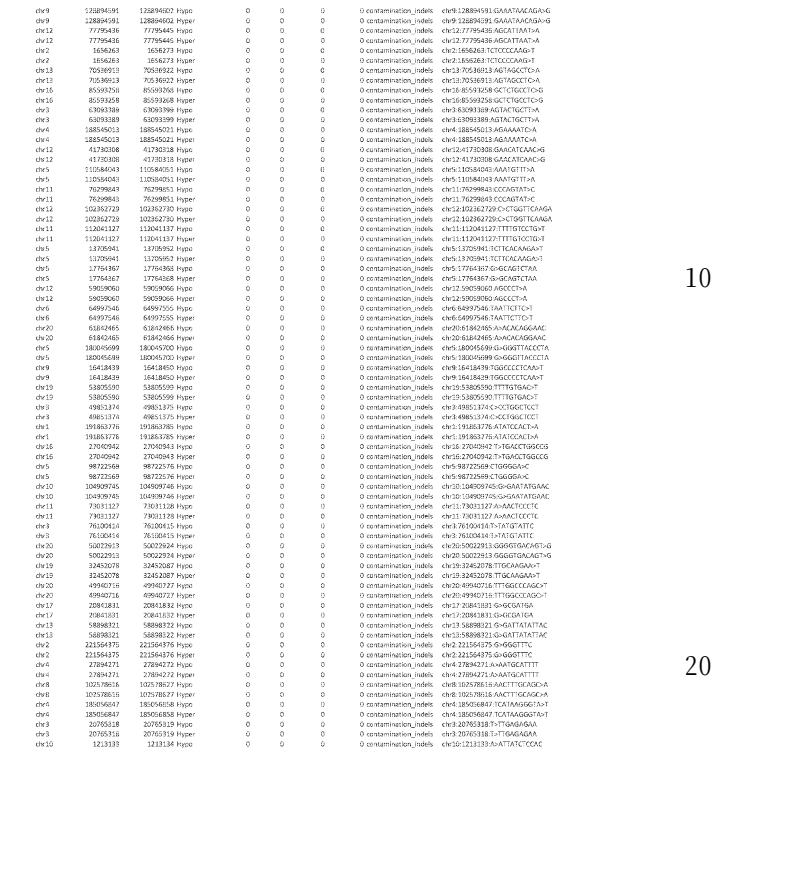
【図２５－１】

参照	開始	終了	メチル化	スコア	閾値	pos_count	neg_count	ソース	説明
chr4	96000607	96000608	Hypo	0	0	0	0	contamination_indels	chr4:96000607-T>TGAAATAATAC
chr4	96000607	96000608	Hyper	0	0	0	0	contamination_indels	chr4:96000607-T>TGAAATAATAC
chr10	57769028	57769029	Hypo	0	0	0	0	contamination_indels	chr10:57769028-A>ACTTCACC
chr10	57769028	57769029	Hyper	0	0	0	0	contamination_indels	chr10:57769028-A>ACTTCACC
chr5	99455945	99455946	Hypo	0	0	0	0	contamination_indels	chr5:99455945-T>TGTATTAAAA
chr5	99455945	99455946	Hyper	0	0	0	0	contamination_indels	chr5:99455945-T>TGTATTAAAA
chr7	12499105	12499107	Hypo	0	0	0	0	contamination_indels	chr7:12499105-A>AATGTGGACAT
chr7	12499105	12499107	Hyper	0	0	0	0	contamination_indels	chr7:12499106-A>AATGTGGACAT
chr15	28267532	28267533	Hypo	0	0	0	0	contamination_indels	chr15:28267532-G>GTGGGTGAGAA
chr15	28267532	28267533	Hyper	0	0	0	0	contamination_indels	chr15:28267532-G>GTGGGTGAGAA
chr8	22949613	22949614	Hypo	0	0	0	0	contamination_indels	chr8:22949613-A>AAAGTACCTG
chr8	22949613	22949614	Hyper	0	0	0	0	contamination_indels	chr8:22949613-A>AAAGTACCTG
chr15	99930624	99930635	Hypo	0	0	0	0	contamination_indels	chr15:99930624-GACTTTTTCC>G
chr15	99930624	99930635	Hyper	0	0	0	0	contamination_indels	chr15:99930624-GACTTTTTCC>G
chr10	119487167	119487168	Hypo	0	0	0	0	contamination_indels	chr10:119487167-C>CTTATAGGA
chr10	119487167	119487168	Hyper	0	0	0	0	contamination_indels	chr10:119487167-C>CTTATAGGA
chr15	46642885	46642895	Hypo	0	0	0	0	contamination_indels	chr15:46642885-TCTATAAGGA>T
chr15	46642885	46642895	Hyper	0	0	0	0	contamination_indels	chr15:46642885-TCTATAAGGA>T
chr20	41072940	41072941	Hypo	0	0	0	0	contamination_indels	chr20:41072940-A>ACAGCAAAG
chr20	41072940	41072941	Hyper	0	0	0	0	contamination_indels	chr20:41072940-A>ACAGCAAAG
chr18	69074008	69074009	Hypo	0	0	0	0	contamination_indels	chr18:69074008-G>GTGGGCTAACA
chr18	69074008	69074009	Hyper	0	0	0	0	contamination_indels	chr18:69074008-G>GTGGGCTAACA
chr17	7401334	7401335	Hypo	0	0	0	0	contamination_indels	chr17:7401334>G>GGCTCTGGGGT
chr17	7401334	7401335	Hyper	0	0	0	0	contamination_indels	chr17:7401334>G>GGCTCTGGGGT
chr7	29793205	29793206	Hypo	0	0	0	0	contamination_indels	chr7:29793205-T>TGGGAGAC
chr7	29793205	29793206	Hyper	0	0	0	0	contamination_indels	chr7:29793205-T>TGGGAGAC
chr3	165404394	165404395	Hypo	0	0	0	0	contamination_indels	chr3:165404394-G>GAGTCTTAGA
chr3	165404394	165404395	Hyper	0	0	0	0	contamination_indels	chr3:165404394-G>GAGTCTTAGA
chr2	36366738	36366749	Hypo	0	0	0	0	contamination_indels	chr2:36366738:CTGCTAGAATA>C
chr2	36366738	36366749	Hyper	0	0	0	0	contamination_indels	chr2:36366738:CTGCTAGAATA>C
chr10	119492705	119492714	Hypo	0	0	0	0	contamination_indels	chr10:119492705:GCCGATATT>G
chr10	119492705	119492714	Hyper	0	0	0	0	contamination_indels	chr10:119492705:GCCGATATT>G
chr6	25408692	25408693	Hypo	0	0	0	0	contamination_indels	chr6:25408692:A>AAGATTAATG
chr6	25408692	25408693	Hyper	0	0	0	0	contamination_indels	chr6:25408692:A>AAGATTAATG
chr10	52475958	52475969	Hypo	0	0	0	0	contamination_indels	chr10:52475958:CACCACTGCAG>C
chr10	52475958	52475969	Hyper	0	0	0	0	contamination_indels	chr10:52475958:CACCACTGCAG>C
chr4	19758127	19758128	Hypo	0	0	0	0	contamination_indels	chr4:19758127:C>CATTTAAAATT
chr4	19758127	19758128	Hyper	0	0	0	0	contamination_indels	chr4:19758127:C>CATTTAAAATT
chr9	113129946	113129947	Hypo	0	0	0	0	contamination_indels	chr9:113129946:A>ACTACAGTT
chr9	113129946	113129947	Hyper	0	0	0	0	contamination_indels	chr9:113129946:A>ACTACAGTT
chr6	62730485	62730495	Hypo	0	0	0	0	contamination_indels	chr6:62730485:GCACTTTGAG>G
chr6	62730485	62730495	Hyper	0	0	0	0	contamination_indels	chr6:62730485:GCACTTTGAG>G
chr6	11692256	11692266	Hypo	0	0	0	0	contamination_indels	chr6:11692256:TTCAGATTCAA>T
chr6	11692256	11692266	Hyper	0	0	0	0	contamination_indels	chr6:11692256:TTCAGATTCAA>T
chr13	78028158	78028169	Hypo	0	0	0	0	contamination_indels	chr13:78028158:GAGTCGATCAC>G
chr13	78028158	78028169	Hyper	0	0	0	0	contamination_indels	chr13:78028158:GAGTCGATCAC>G
chr6	36288878	36288879	Hypo	0	0	0	0	contamination_indels	chr6:36288878:G>GGTCTCAGCT
chr6	36288878	36288879	Hyper	0	0	0	0	contamination_indels	chr6:36288878:G>GGTCTCAGCT
chr4	130620013	130620023	Hypo	0	0	0	0	contamination_indels	chr4:130620013:GCAAAAGAGAA>G
chr4	130620013	130620023	Hyper	0	0	0	0	contamination_indels	chr4:130620013:GCAAAAGAGAA>G
chr9	21162876	21162887	Hypo	0	0	0	0	contamination_indels	chr9:21162876:TAAAAACATAA>T
chr9	21162876	21162887	Hyper	0	0	0	0	contamination_indels	chr9:21162876:TAAAAACATAA>T
chr8	142424471	142424472	Hypo	0	0	0	0	contamination_indels	chr8:142424471:G>GGGTGACCCCA
chr8	142424471	142424472	Hyper	0	0	0	0	contamination_indels	chr8:142424471:G>GGGTGACCCCA
chr6	161638913	161638923	Hypo	0	0	0	0	contamination_indels	chr6:161638913:GGTTTGTTCT>G
chr6	161638913	161638923	Hyper	0	0	0	0	contamination_indels	chr6:161638913:GGTTTGTTCT>G
chr3	17375415	17375425	Hypo	0	0	0	0	contamination_indels	chr3:17375415:AGGGAGAATG>A
chr3	17375415	17375425	Hyper	0	0	0	0	contamination_indels	chr3:17375415:AGGGAGAATG>A
chr12	93046310	93046311	Hypo	0	0	0	0	contamination_indels	chr12:93046310:G>GCCATAATAA
chr12	93046310	93046311	Hyper	0	0	0	0	contamination_indels	chr12:93046310:G>GCCATAATAA
chr2	31142069	31142070	Hypo	0	0	0	0	contamination_indels	chr2:31142069:C>CCTGGAGA
chr2	31142069	31142070	Hyper	0	0	0	0	contamination_indels	chr2:31142069:C>CCTGGAGA
chr15	53945224	53945232	Hypo	0	0	0	0	contamination_indels	chr15:53945224:CCTCTGTA>C
chr15	53945224	53945232	Hyper	0	0	0	0	contamination_indels	chr15:53945224:CCTCTGTA>C
chr8	69536929	69536940	Hypo	0	0	0	0	contamination_indels	chr8:69536929:TACAGCAGCGA>T
chr8	69536929	69536940	Hyper	0	0	0	0	contamination_indels	chr8:69536929:TACAGCAGCGA>T
chr10	120791777	120791785	Hypo	0	0	0	0	contamination_indels	chr10:120791777:TAAATACC>T
chr10	120791777	120791785	Hyper	0	0	0	0	contamination_indels	chr10:120791777:TAAATACC>T
chrX	5996764	5996775	Hypo	0	0	0	0	contamination_indels	chrX:5996764:CCTATTACTCA>C
chrX	5996764	5996775	Hyper	0	0	0	0	contamination_indels	chrX:5996764:CCTATTACTCA>C
chr12	125926855	125926856	Hypo	0	0	0	0	contamination_indels	chr12:125926855:A>AATAGTGATTCC
chr12	125926855	125926856	Hyper	0	0	0	0	contamination_indels	chr12:125926855:A>ATAGTGATTCC
chr5	137787291	137787301	Hypo	0	0	0	0	contamination_indels	chr5:137787291:CCTCAATGAT>C

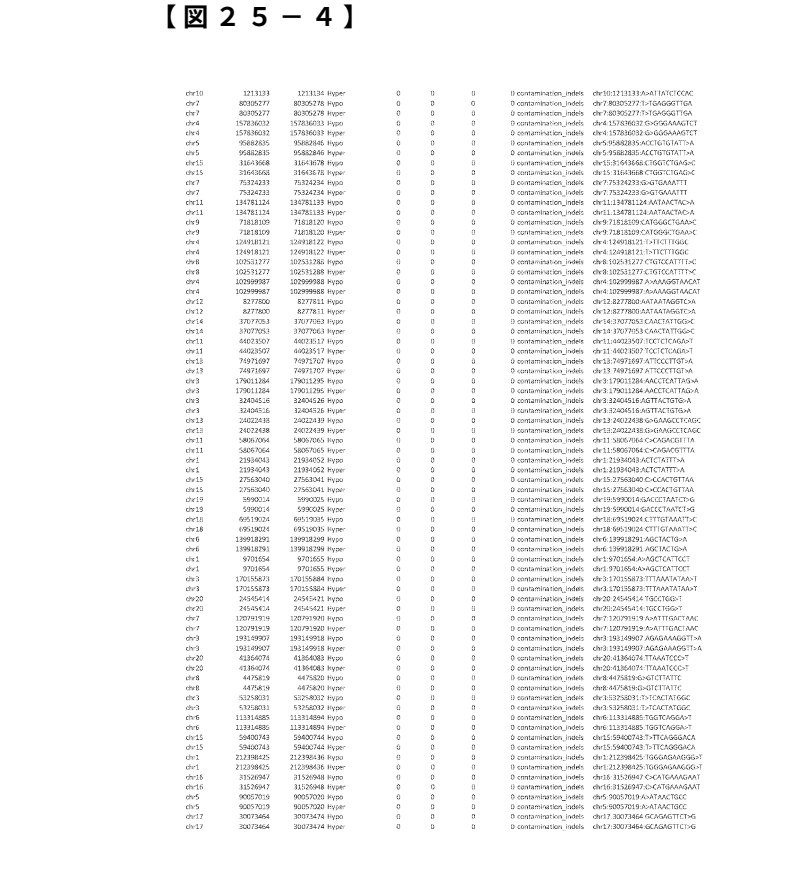
【図２５－２】



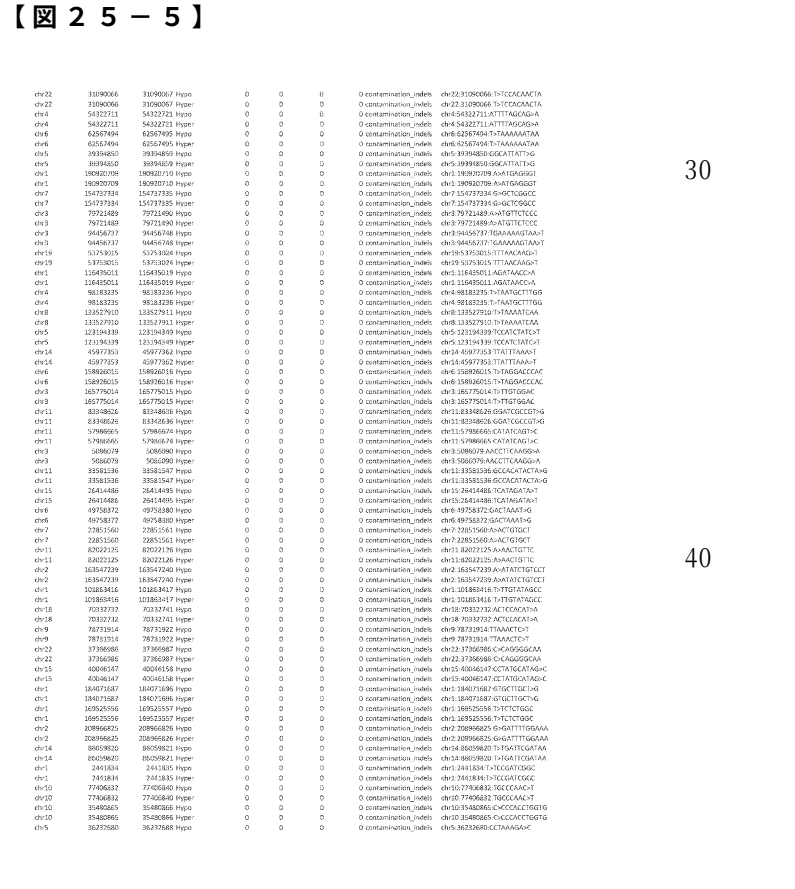
【図２５－３】



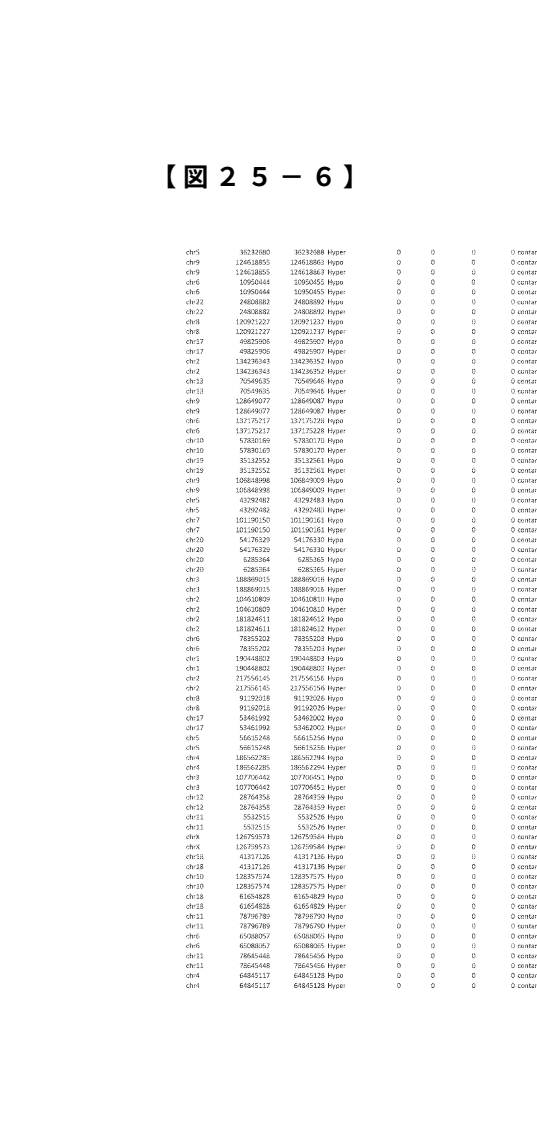
【図２５－４】



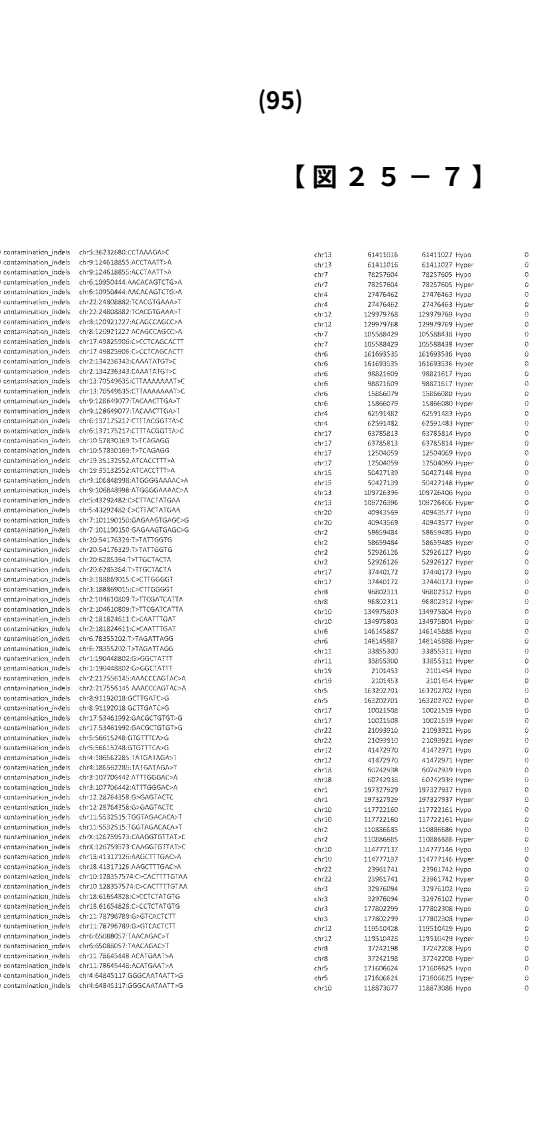
【図２５－５】



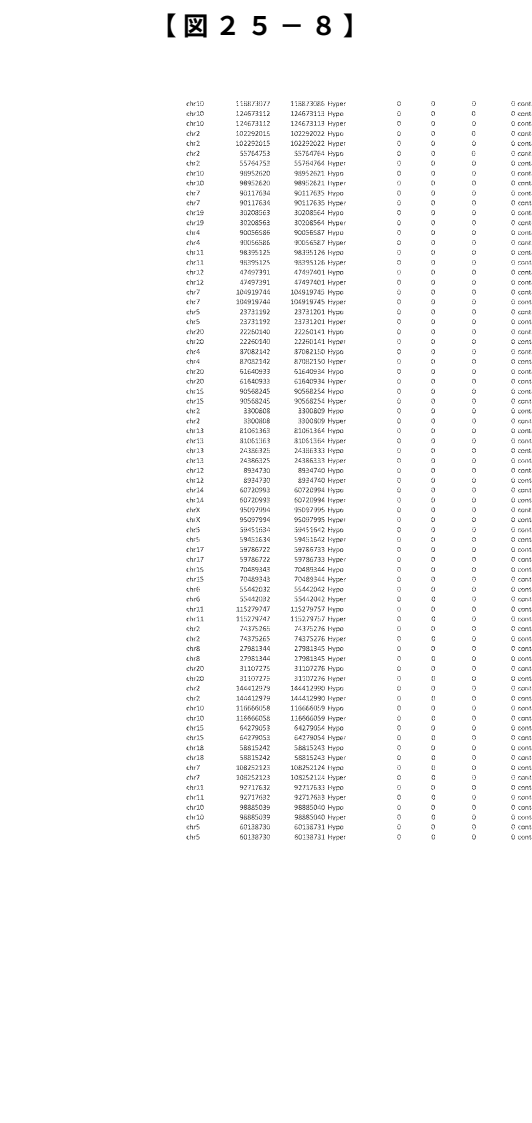
【図２５－６】



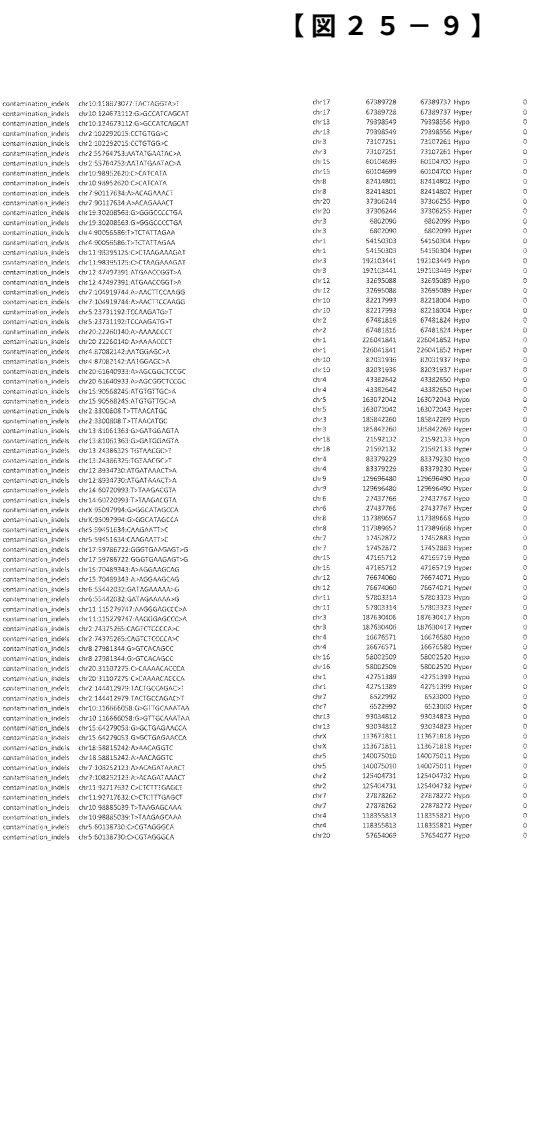
【図２５－７】



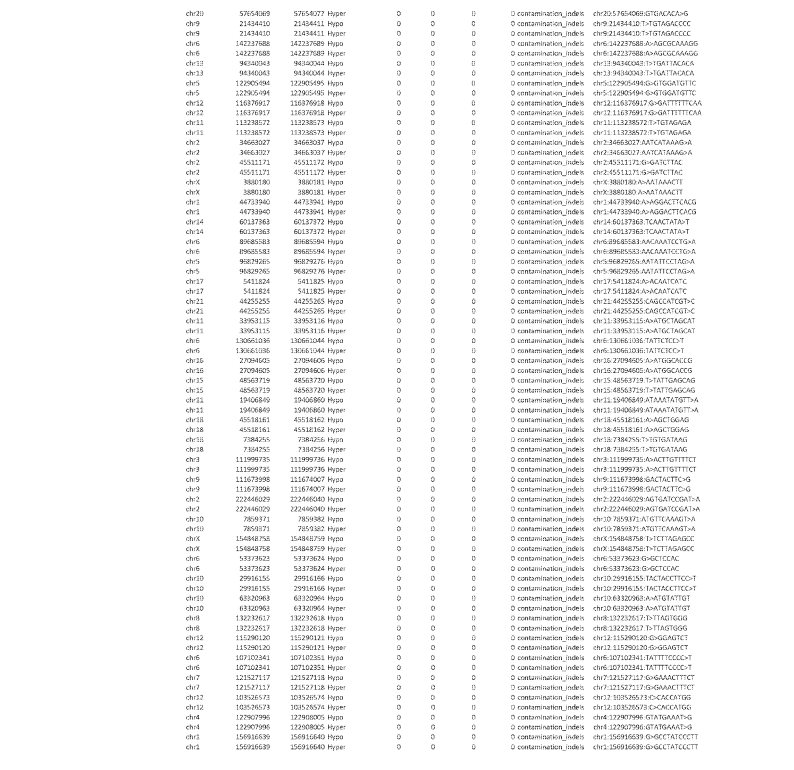
【図２５－８】



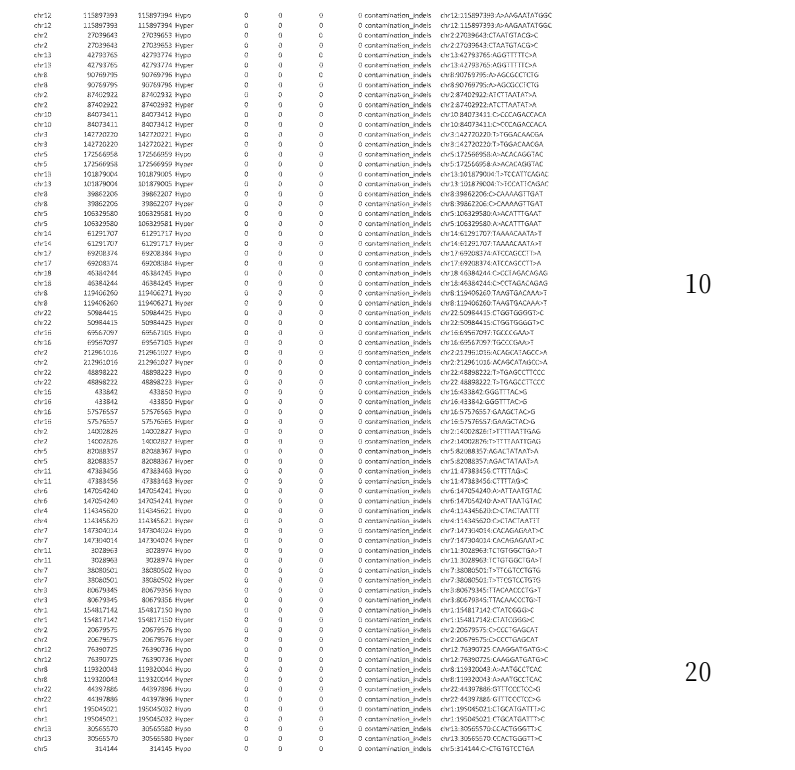
【図２５－９】



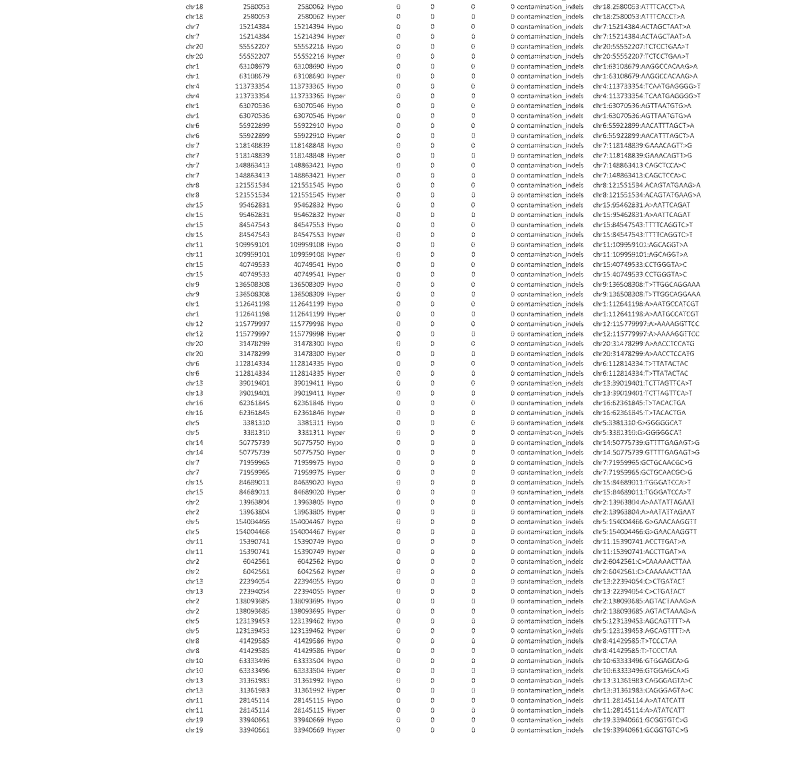
【図２５－１０】



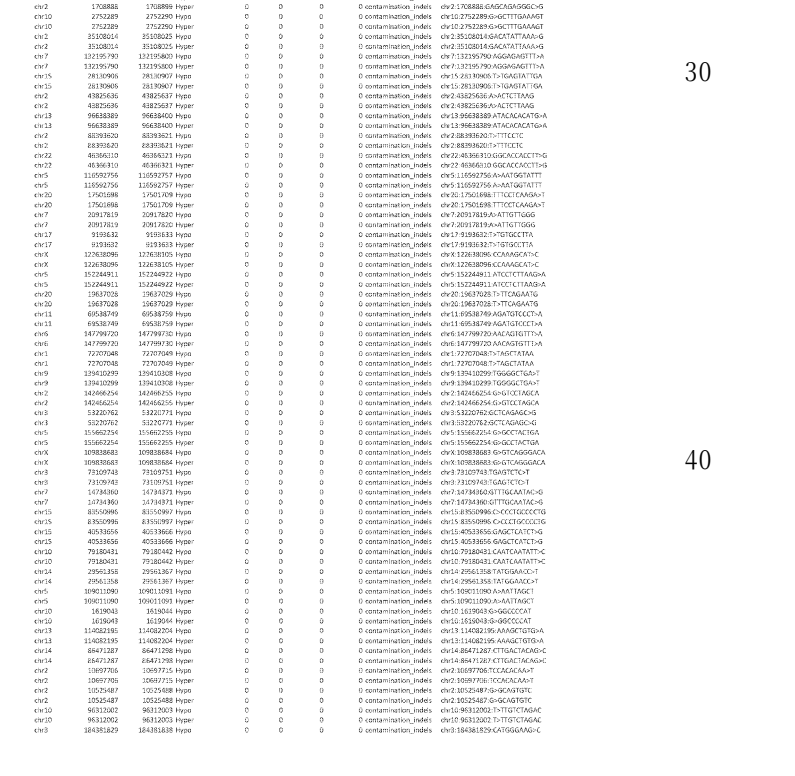
【図２５－１１】



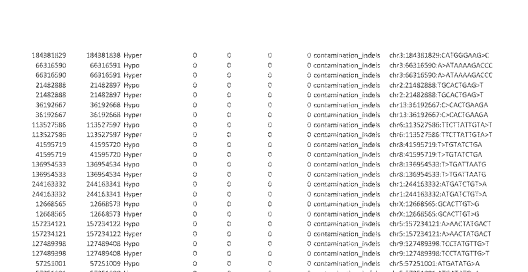
【図２５－１２】



【図２５－１３】



【図２５－１４】



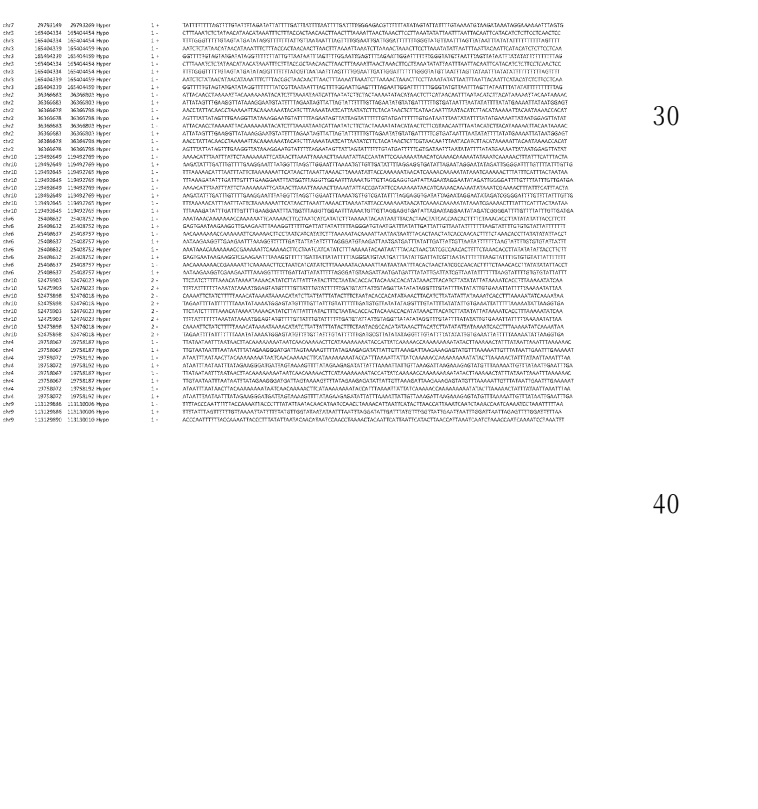
【図２６－１】



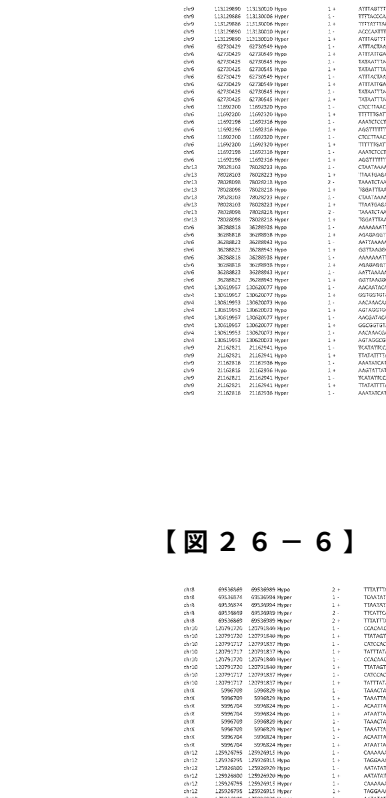
【図２６－２】



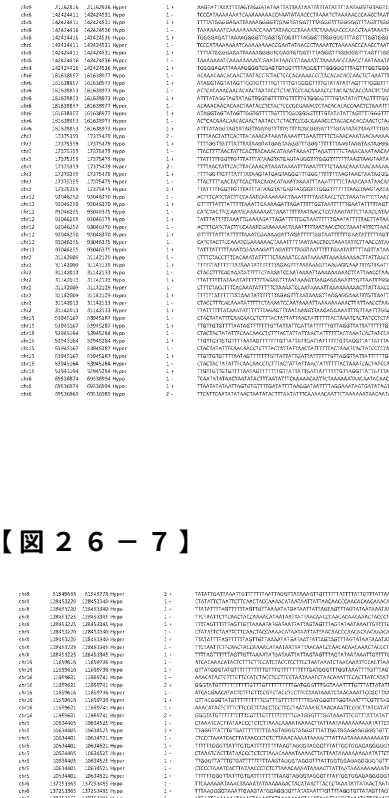
【図２６－３】



【図２６－４】



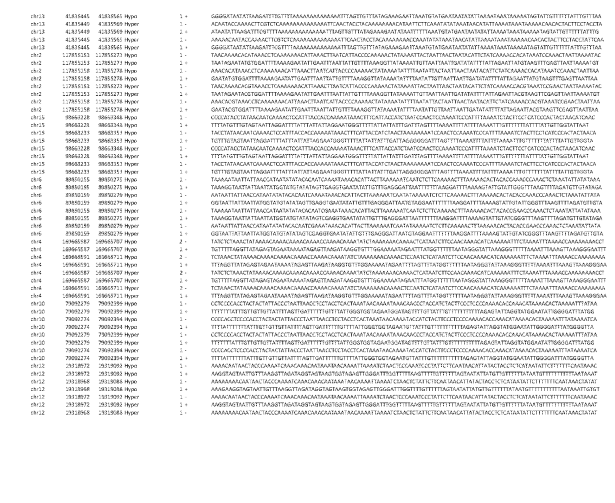
【図２６－５】



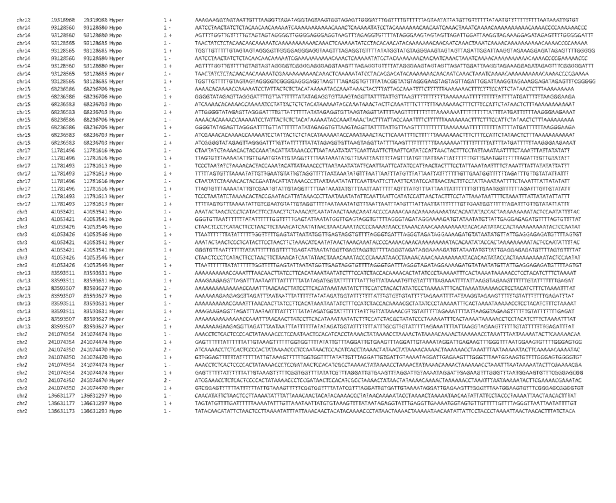
【図２６－６】

【図２６－７】

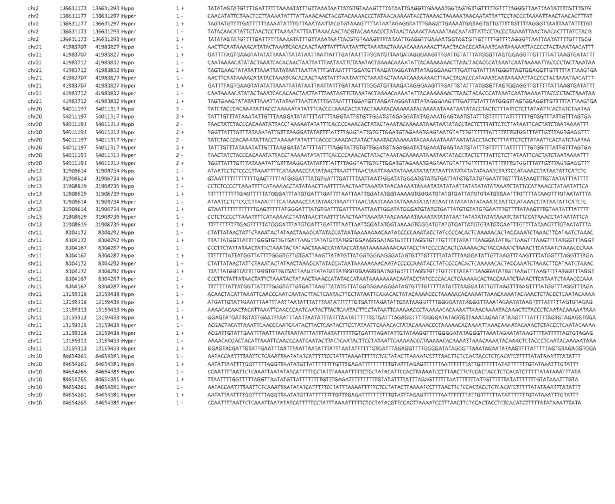
【図２６－８】



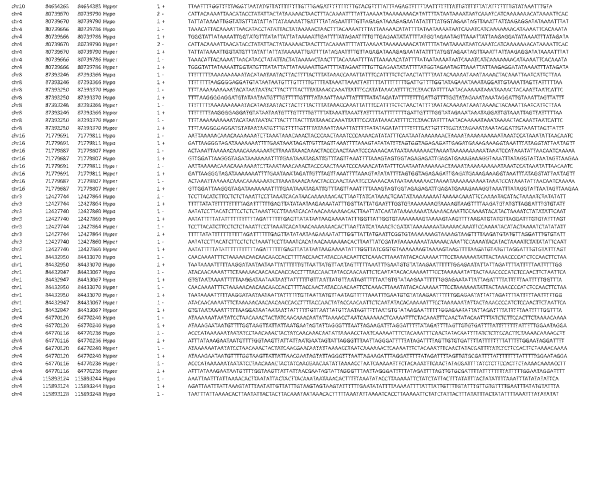
【図２６－９】



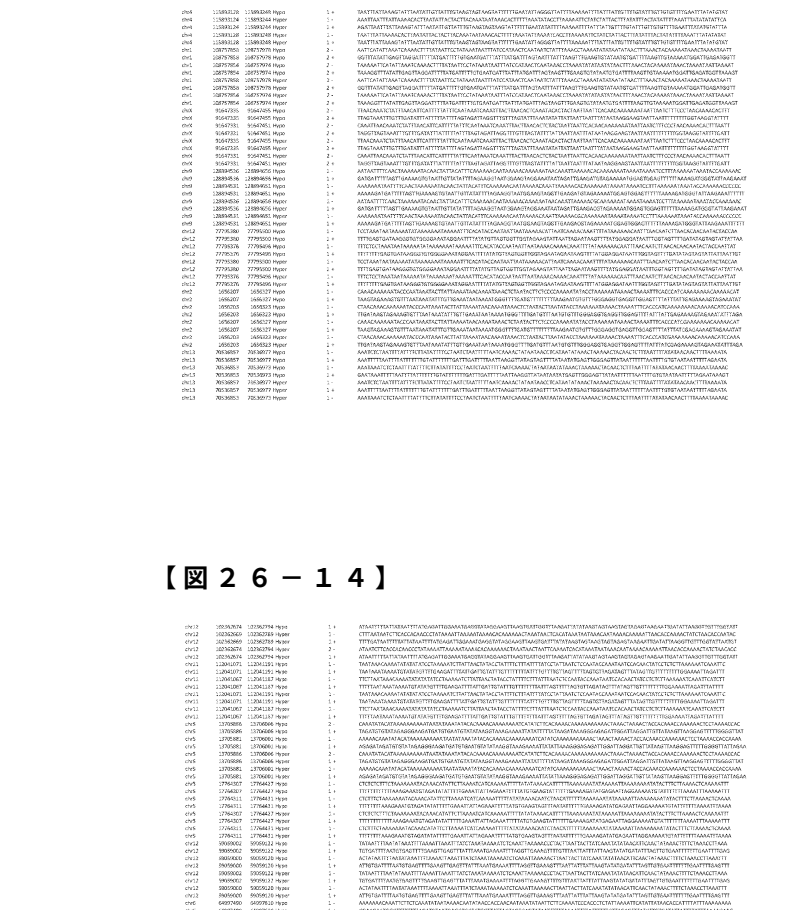
【図２６－１０】



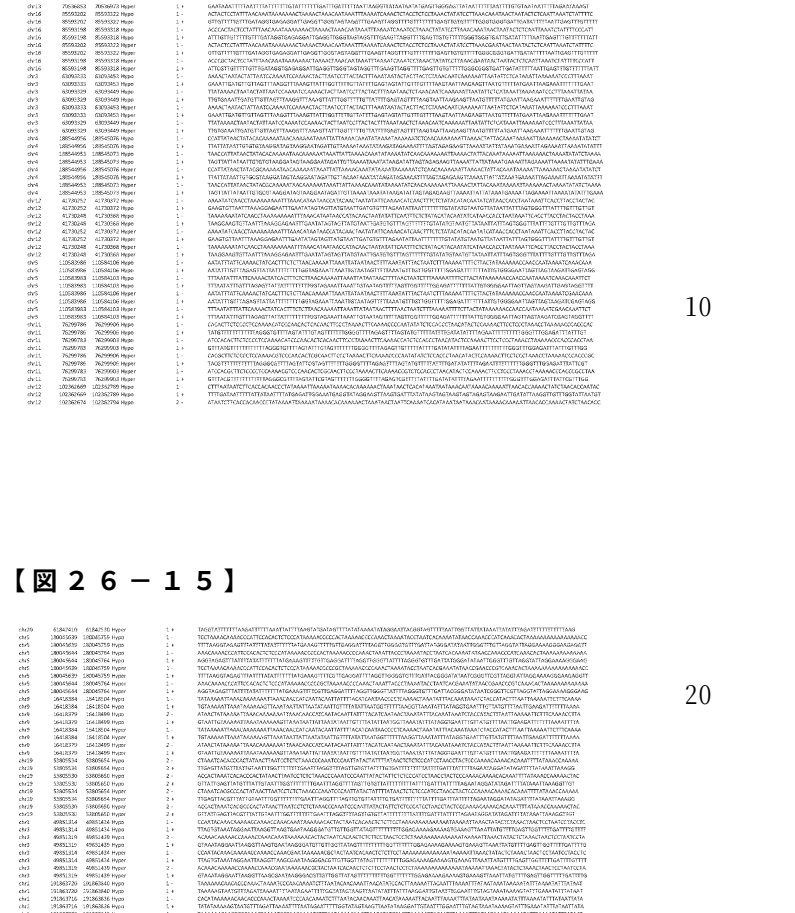
【図２６－１１】



【図２６－１２】



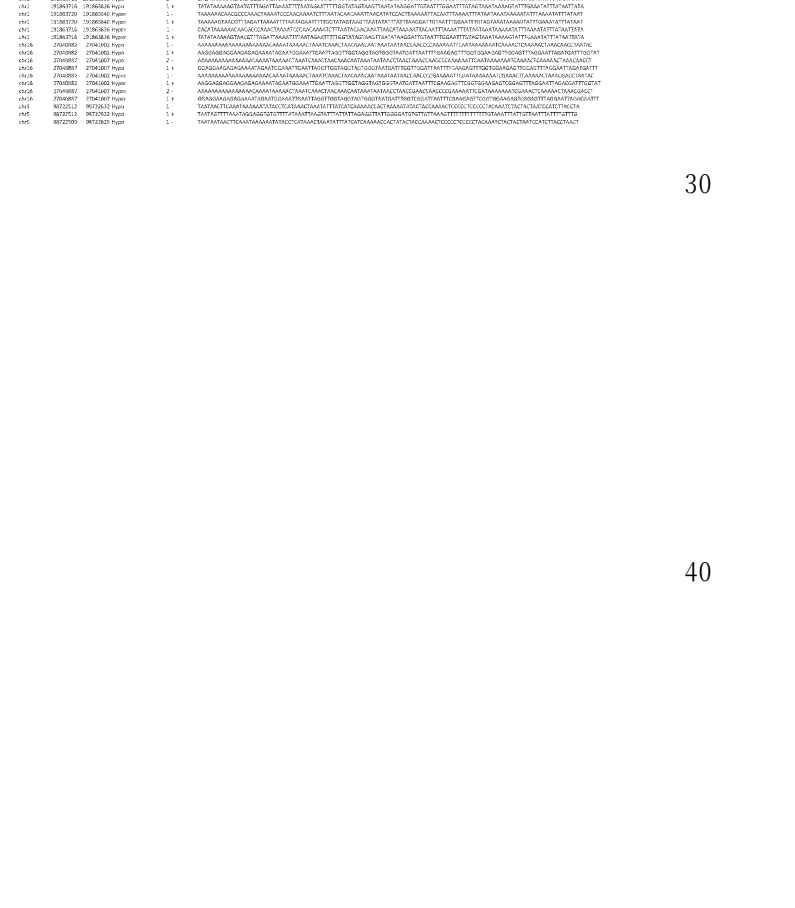
【図２６－１３】



【図２６－１４】



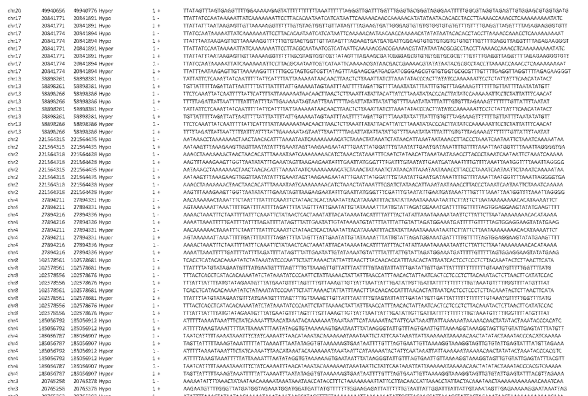
【図２６－１５】



【図２６－１６】



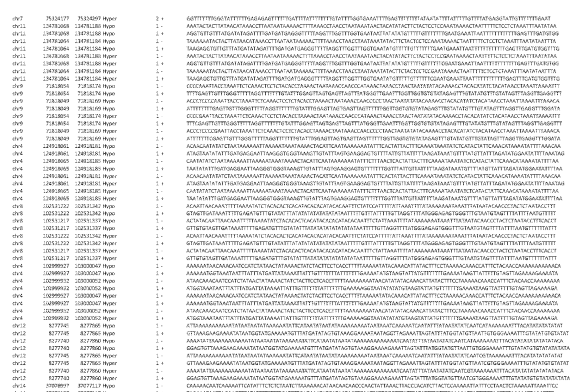
【図２６－１７】



【図２６－１８】



【図２６－１９】



【図２６－２０】



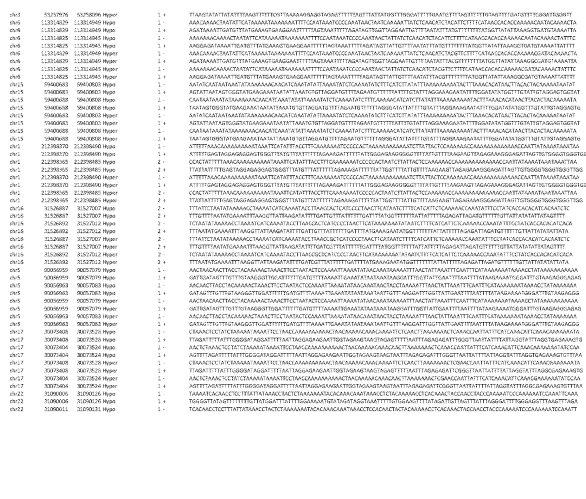
【図２６－２１】



【図２６－２２】



【図２６－２３】



【図２６－２４】



【図２６－２５】



【図２６－２６】



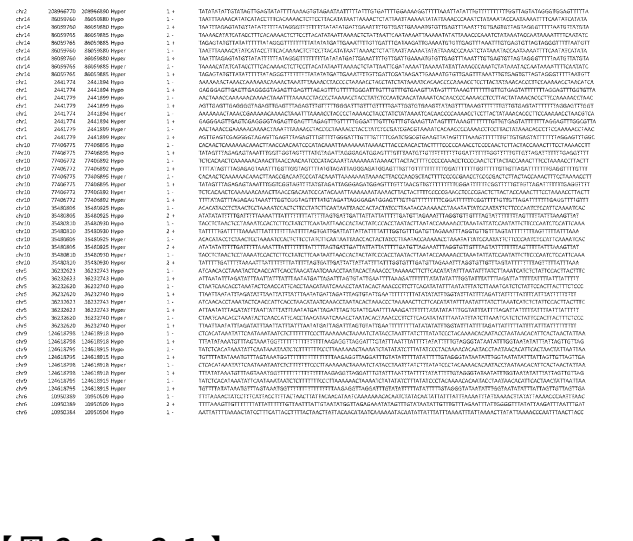
【図２６－２７】



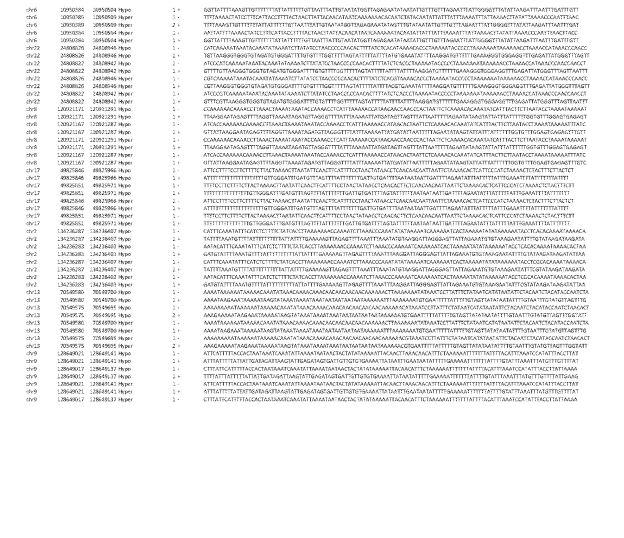
【図２６－２８】



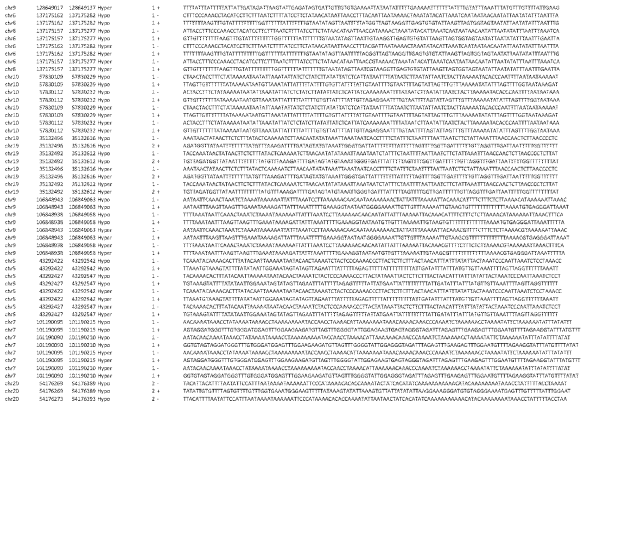
【図２６－２９】



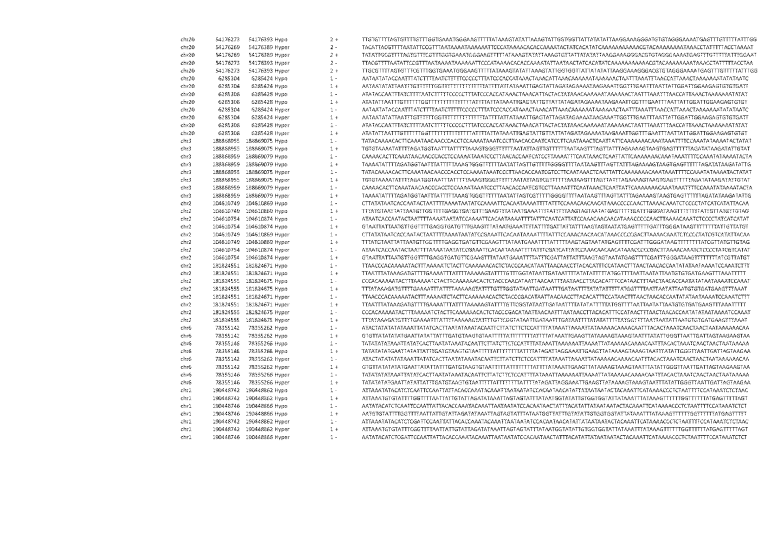
【図２６－３０】



【図２６－３１】



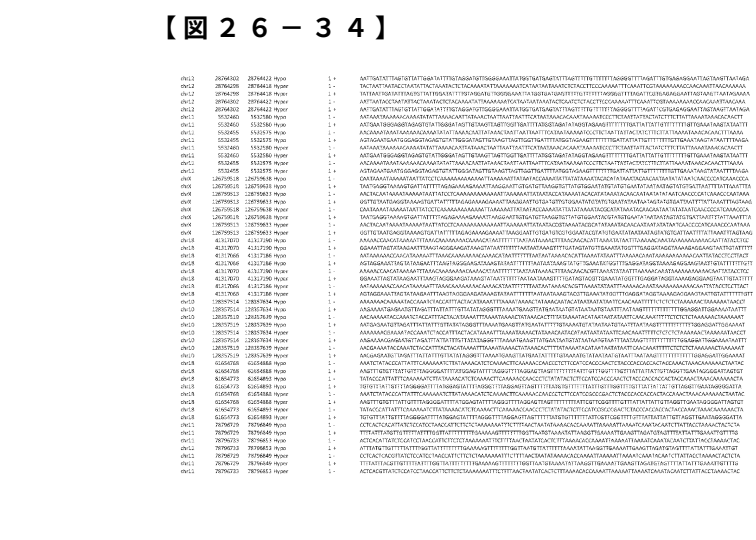
【図２６－３２】



【図２６－３３】



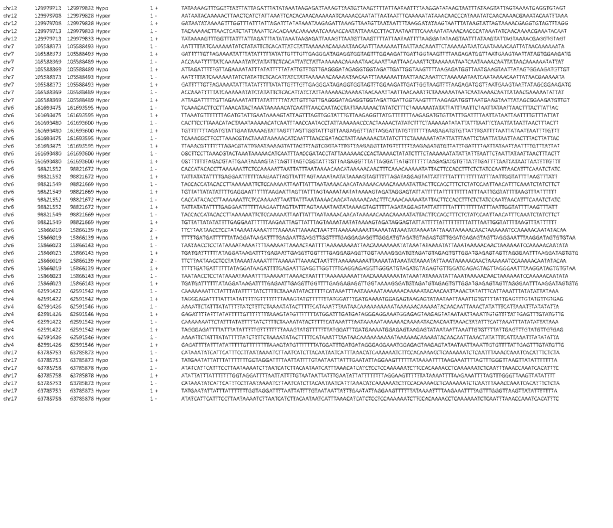
【図２６－３４】



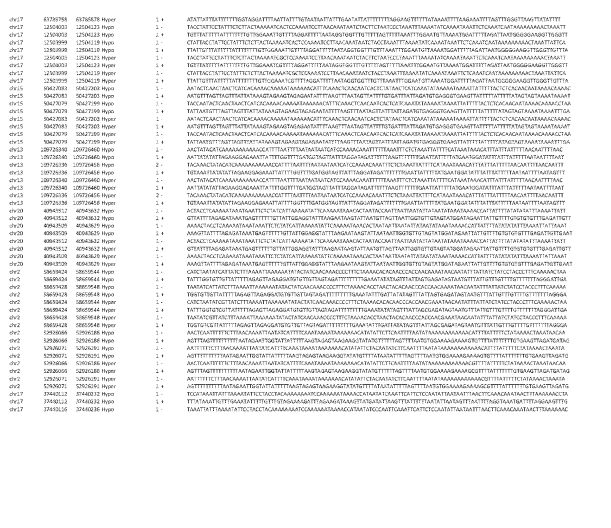
【図２６－３５】



【図２６－３６】



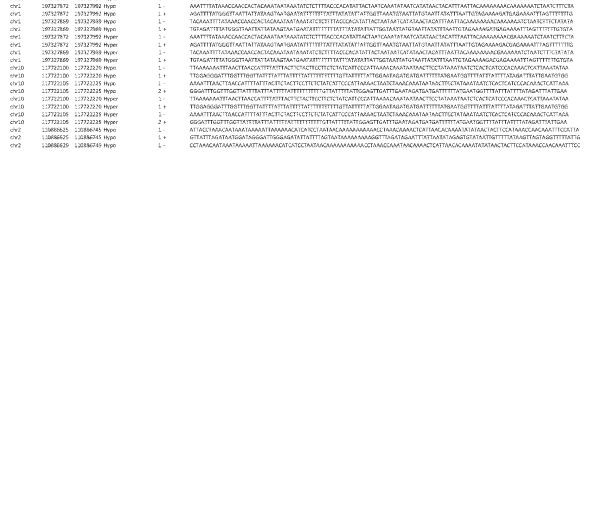
【図２６－３７】



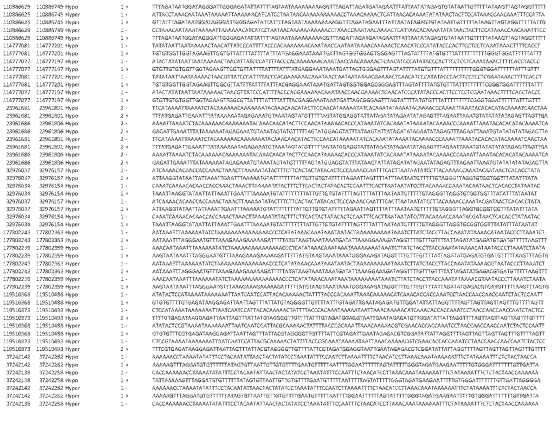
【図２６－３８】



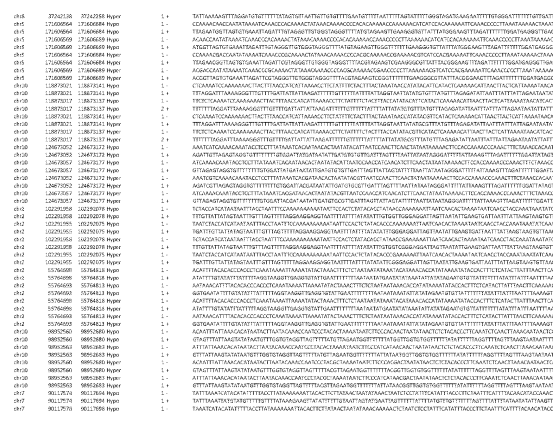
【図２６－３９】



【図２６－４０】



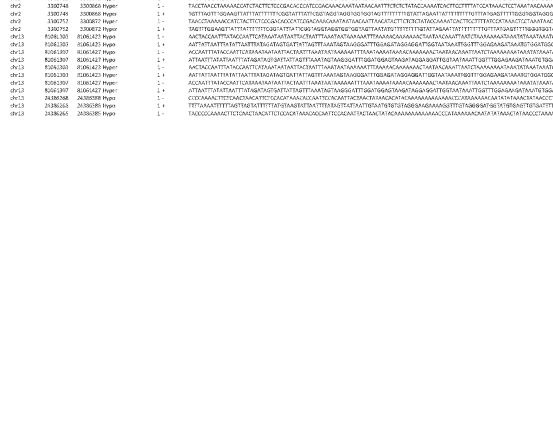
【図２６－４１】



【図２６－４２】



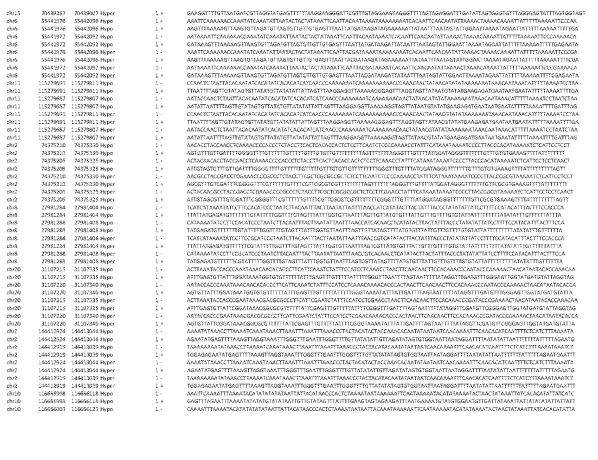
【図２６－４３】



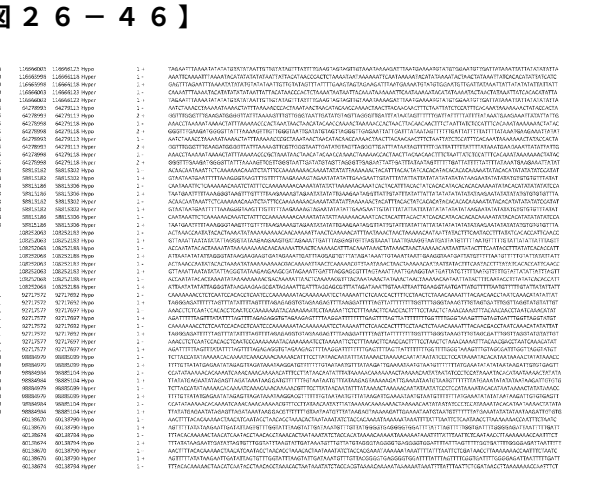
【図２６－４４】



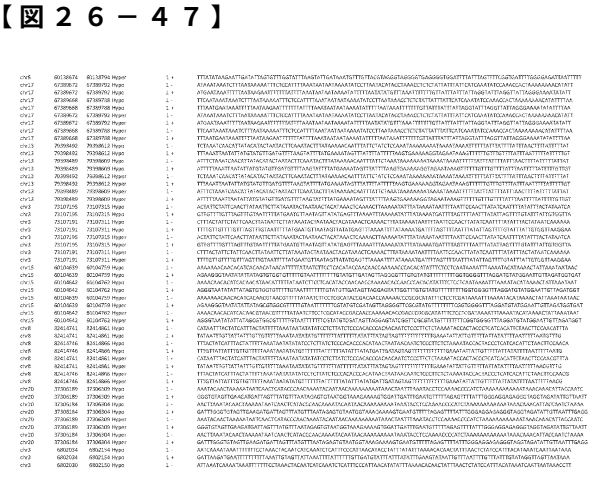
【図２６－４５】



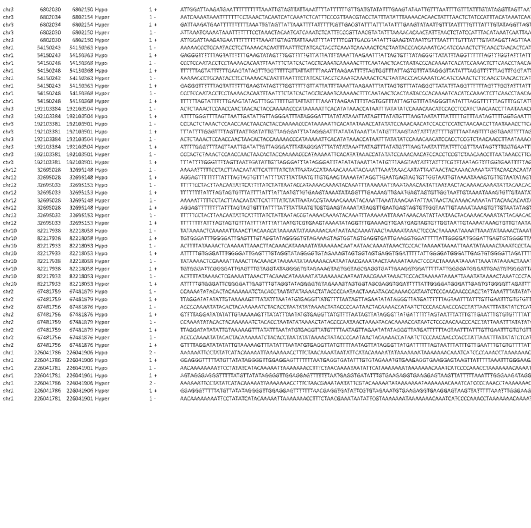
【図２６－４６】



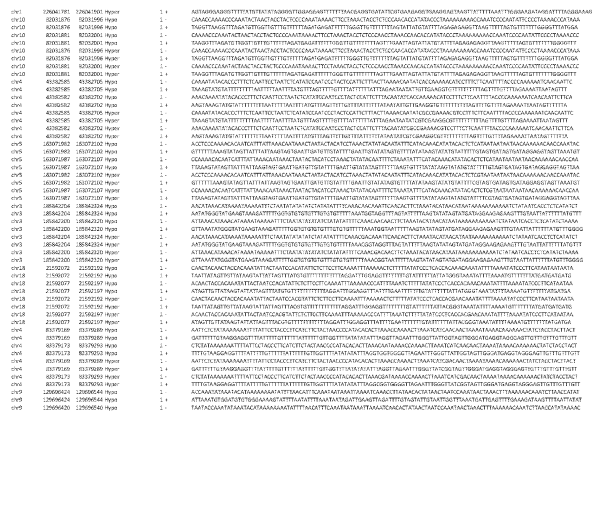
【図２６－４７】



【図２６－４８】



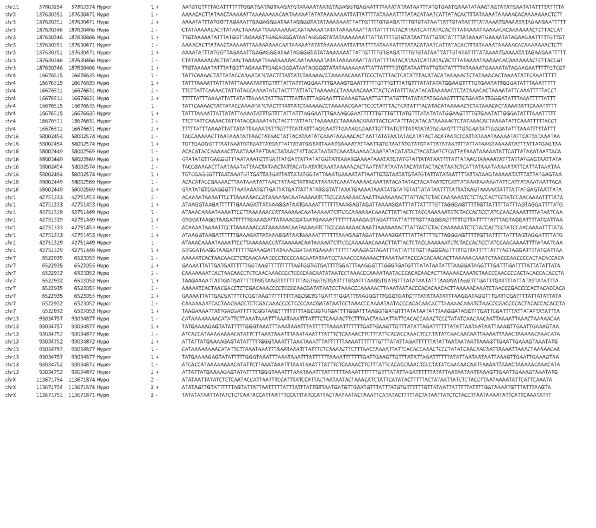
【図２６－４９】



【図２６－５０】



【図２６－５１】



【図２６－５２】



【図２６－５３】



【図２６－５４】



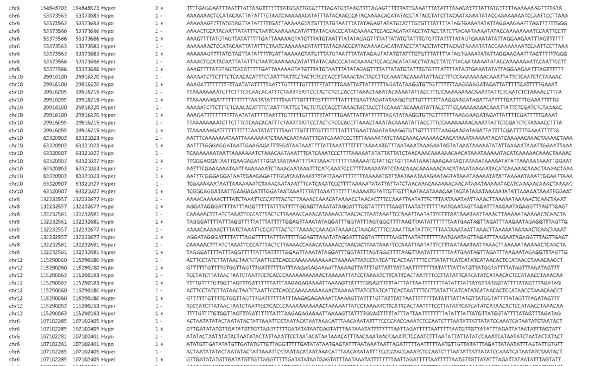
【図２６－５５】



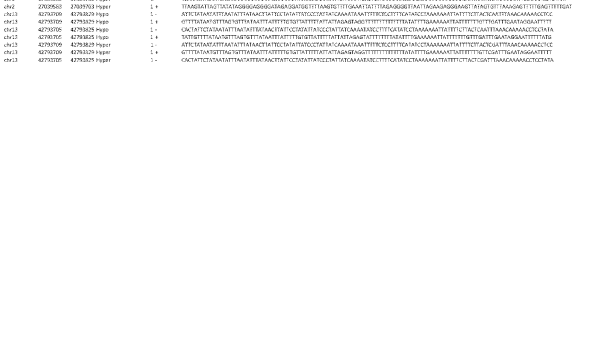
【図２６－５６】



【図２６－５７】



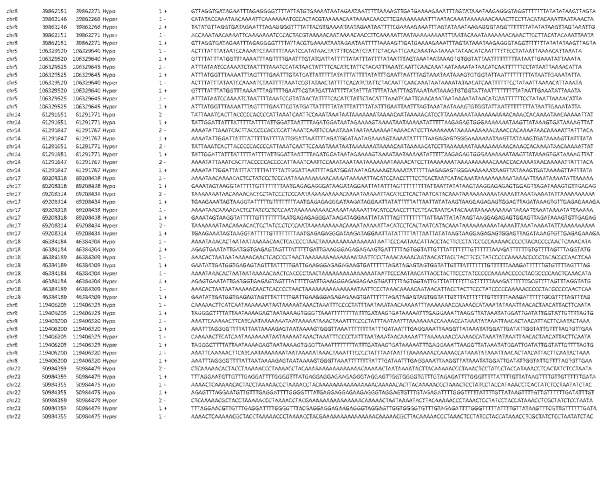
【図２６－５８】



【図２６－５９】



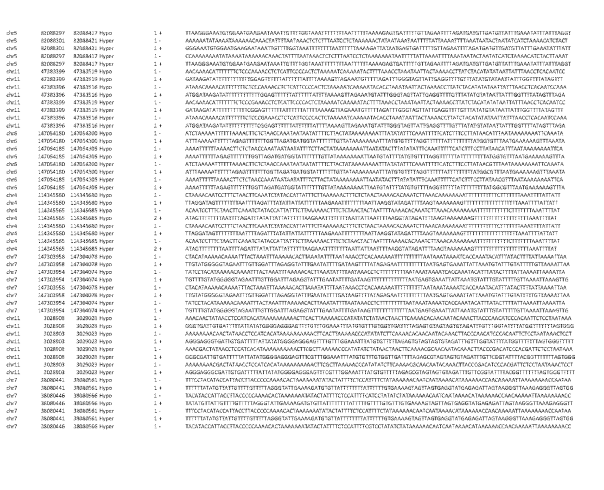
【図２６－６０】



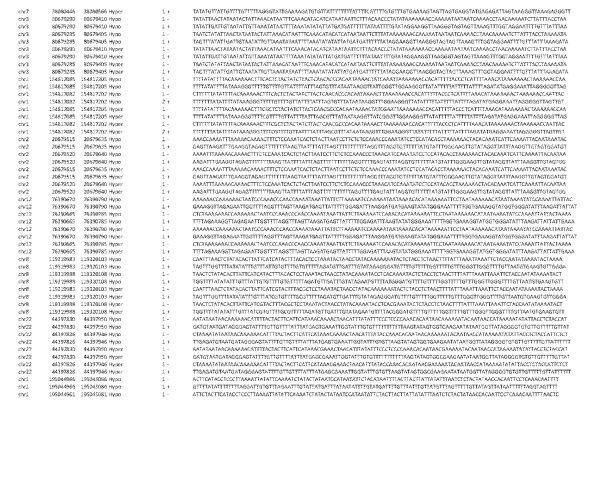
【図２６－６１】



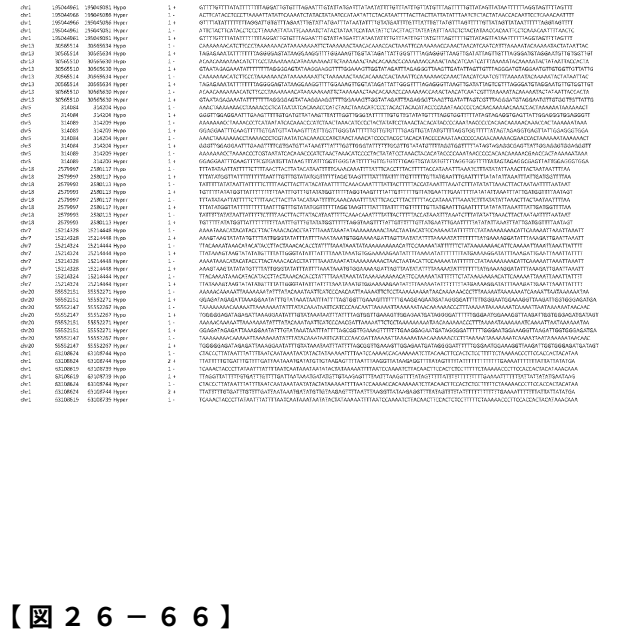
【図２６－６２】



【図２６－６３】



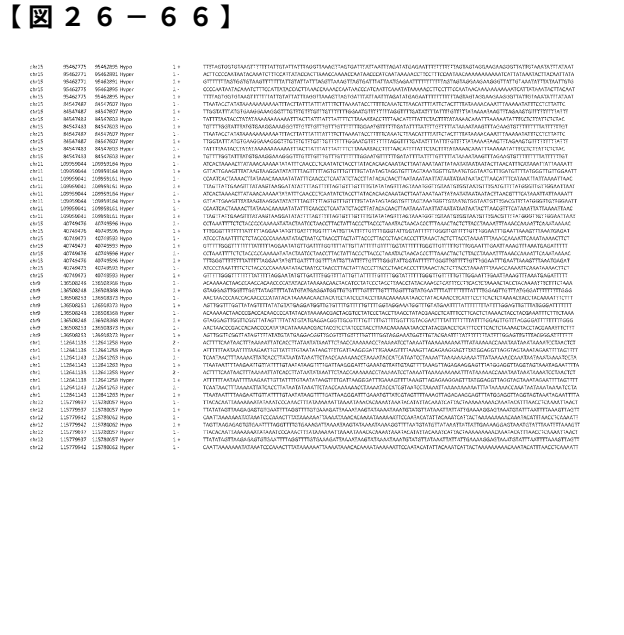
【図２６－６４】



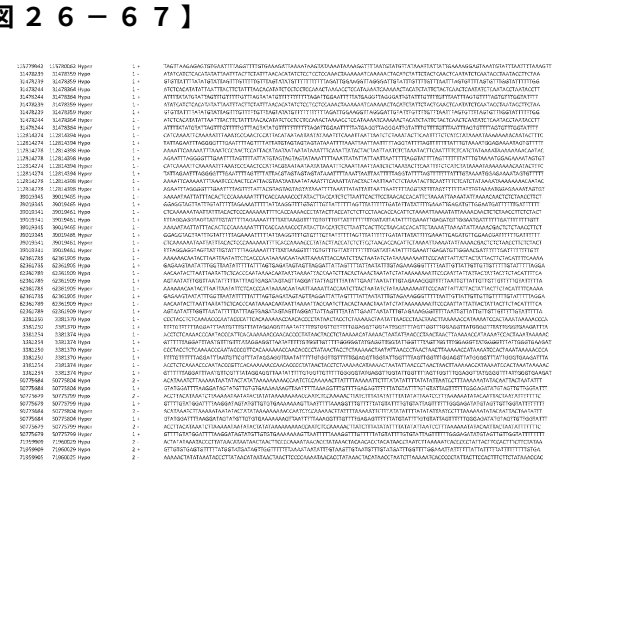
【図２６－６５】



【図２６－６６】



【図２６－６７】



【図２６－６８】



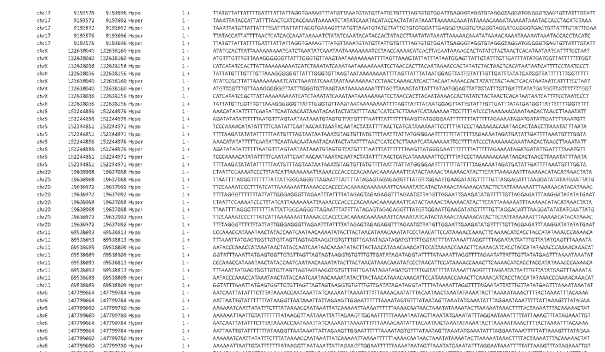
【図２６－６９】



【図２６－７０】

【図２６－７１】

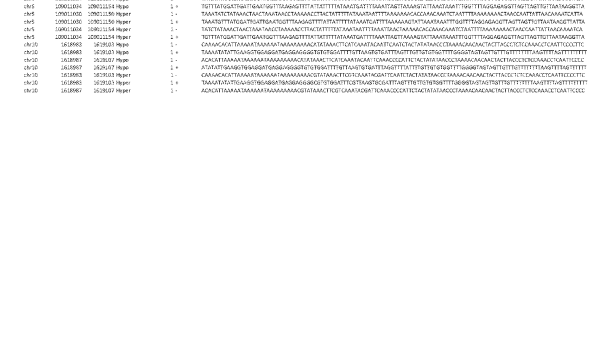
【図２６－７２】



【図２６－７３】



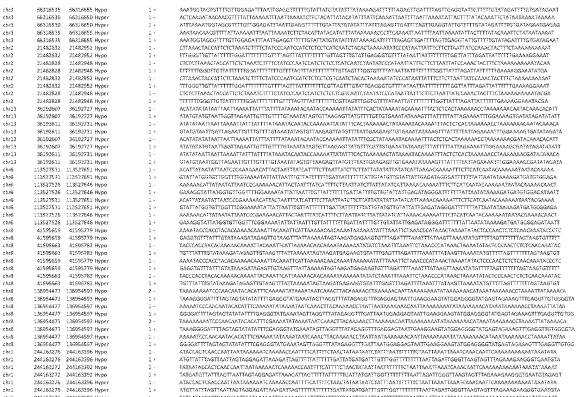
【図２６－７４】



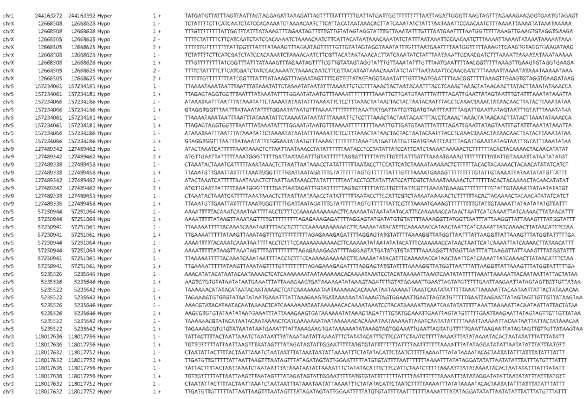
【図２６－７５】



【図２６－７６】



【図２６－７７】



【配列表】

20240544597000001.xml

【誤訳訂正書】
【提出日】令和6年8月5日(2024.8.5)
【誤訳訂正１】
【訂正対象書類名】明細書
【訂正対象項目名】全文
【訂正方法】変更
【訂正の内容】
【発明の詳細な説明】
【技術分野】
【０００１】
[関連出願の相互参照]
本出願は、２０２１年１１月２３日に出願された米国仮出願第６３／２８２，５０９号の利益および優先権を主張し、この仮出願は、参照によりその全体が組み込まれる。
【０００２】
本開示は、がん分類のための汚染片のサンプル汚染検出に関する。
【背景技術】
【０００３】
デオキシリボ核酸（ＤＮＡ）のメチル化は、遺伝子発現の制御において重要な役割を果たしている。異常なＤＮＡメチル化は、がんを含む多くの疾患プロセスに関与している。メチル化配列決定を用いるＤＮＡメチル化プロファイリング（例えば、全ゲノムバイサルファイト配列決定（ＷＧＢＳ））は、がんの検出、診断、および／またはがんのモニタリングのための貴重な診断ツールとしてますます認識されている。例えば、異なるメチル化領域の特異的パターンおよび／または対立遺伝子特異的メチル化パターンは、循環セルフリー（ｃｆ）ＤＮＡを用いた非侵襲的診断の分子マーカーとして有用である。しかし、本技術分野では、サンプルの個体に由来しないｃｆＤＮＡ断片の汚染検出方法の改善に対する必要性が残っている。
【０００４】
本開示は、上記の課題に対処することに向けられている。本明細書において提供する背景説明は、本開示の文脈を一般的に提示するためのものである。本明細書において特に示されない限り、本項に記載する材料は、本出願の特許請求の範囲に対する先行技術ではなく、本項に含めることによって先行技術であること、または先行技術の示唆であることを認めるものではない。
【発明の概要】
【発明が解決しようとする課題】
【０００５】
被験者における疾患状態（がんなど）の早期発見は、早期治療が可能となり、したがって生存の可能性が高まるため重要である。セルフリー（ｃｆ）ＤＮＡサンプル中のＤＮＡ断片の配列決定は、疾患の分類に使用できる特徴を同定するために使用することができる。例えば、がんの評価では、血液サンプルからのセルフリーＤＮＡに基づく特徴（体細胞変異、メチル化状態、または他の遺伝子異常の有無など）は、被験者ががんに罹患している可能性があるかどうかについての洞察、および被験者がどのようながんに罹患している可能性があるかについてのさらなる洞察を提供することができる。この目的のために、本明細書には、被検者が疾患に罹患している可能性を決定するためのセルフリーＤＮＡ（ｃｆＤＮＡ）配列決定データを解析するシステムおよび方法が含まれる。
【０００６】
本開示は、がん分類のための汚染断片のサンプル汚染検出のための改良されたシステムおよび方法を提供することによって、上記で特定された問題に対処する。本システムは、サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから１つ以上の汚染マーカーを同定する。本システムは、同定された汚染マーカーの１つにおいて、それぞれの汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有するサンプル中のあら

ゆるｃｆＤＮＡ断片を、汚染されたｃｆＤＮＡ断片として同定する。システムは、汚染されたｃｆＤＮＡ断片に基づいてサンプルの汚染レベルを推定する。本システムは、がん分類器の訓練に使用される訓練サンプルの両方に対してこの汚染検出を実施することができ、また、がん分類器を展開する際のテストサンプルに対してもこの汚染検出を実施することができる。汚染マーカーには、複数の一塩基多型（ＳＮＰ）部位汚染マーカーおよび／またはインデル部位汚染マーカーが含まれる。いくつかの例において、複数のＳＮＰ部位混入マーカーは、閾値距離内にある少なくとも２つのＳＮＰ部位であって、閾値頻度の範囲内にある集団ハプロタイプ頻度を有する、グアニン－アデニン多型および／もしくはシトシン－チミン多型を除外する、ハーディーワインベルク平衡を確保する、または上記のパラメータの任意の組み合わせである、該少なくとも２つのＳＮＰ部位を含む。いくつかの例では、インデル部位汚染マーカーは、閾値長さ内にあるインデル配列であって、高い複雑性を有する、集団ハプロタイプ頻度が閾値頻度の範囲内にある、ハーディーワインベルク平衡を保証する、または上記のパラメータの任意の組み合わせである、該インデル配列を含む。

【課題を解決するための手段】

【０００７】

第１の態様によれば、検査サンプル中のがんの存在を予測する方法が開示され、該方法は、前記検査サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む、前記検査サンプルを取得するステップと、前記検査サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、前記検査サンプルがホモ接合性ハプロタイプを有する、同定された１つ以上の汚染マーカーのそれぞれについて、同定された汚染マーカーの１つにおいて、それぞれの汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記検査サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、いずれかの同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記汚染レベルが閾値レベル以下であるか否かを判定するステップと、並びに前記汚染レベルが前記閾値レベル以下であると判定されたことに応答して、がん予測を生成するため前記検査サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに対してがん分類を実行するステップと、を備える。

【０００８】

第１の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。

【０００９】

第１の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。

【００１０】

第１の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００１１】

第１の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。

【００１２】

第１の態様の方法において、各複数のＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００１３】

第１の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。

【００１４】

第１の態様の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位

を含む。
【００１５】
第１の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。
【００１６】
第１の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【００１７】
第１の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する。
【００１８】
第１の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【００１９】
第１の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。
【００２０】
第１の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【００２１】
第１の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【００２２】
第１の態様の方法において、前記汚染レベルを推定するステップは、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記検査サンプルのシーケンス深度、前記検査サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる。
【００２３】
第１の態様の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、がんの分類を差し控える。
【００２４】
第１の態様の方法において、前記がん予測は、がんと非がんとの間の二値予測を含む。
【００２５】
第１の態様の方法において、前記がん予測は、複数のがんタイプ間のマルチクラスがん予測を含む。
【００２６】
第１の態様の方法において、前記がんの分類を実行するステップは、さらに、前記検査サンプルのｃｆＤＮＡ断片の初期セットをｐ値フィルタリングでフィルタリングして異常断片のセットを生成するステップであって、前記フィルタリングは、前記初期セットから他の断片に関して閾値ｐ値を下回る断片を除去して前記異常断片のセットを生成するステップを含む。
【００２７】
第１の態様の方法において、前記分類モデルは、機械学習モデルである。
【００２８】
第２の態様によれば、命令がコンピュータプロセッサによって実行されるとき、第１の態様の方法をコンピュータプロセッサに実行させる該命令を記憶する、非一過性のコンピュータ可読記憶媒体を開示する。
【００２９】
第３の態様によれば、コンピュータプロセッサ、および第２の態様の非一過性のコンピュータ可読記憶媒体を備えるシステムを開示する。

【００３０】
第４の態様によれば、検査サンプル中の疾患の存在を予測する方法が開示され、該方法は、前記検査サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む前記検査サンプルを得るステップと、前記検査サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、前記検査サンプルがホモ接合性ハプロタイプを有する１つ以上の汚染マーカーのそれぞれについて、同定された汚染マーカーの１つにおいて、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記検査サンプル中のあらゆる前記ｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、同定された前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、並びに前記汚染レベルが前記閾値レベル以下であると決定することに応答して、疾患予測を生成するため前記検査サンプル中のｃｆＤＮＡ断片の前記配列リードに対して疾患分類を実行するステップと、を備える。
【００３１】
第４の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。
【００３２】
第４の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。
【００３３】
第４の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【００３４】
第４の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。
【００３５】
第４の態様の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【００３６】
第４の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。
【００３７】
第４の態様の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。
【００３８】
第４の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。
【００３９】
第４の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【００４０】
第４の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する。
【００４１】
第４の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【００４２】
第４の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。

【００４３】
第４の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【００４４】
第４の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【００４５】
第４の態様の方法において、前記汚染レベルを推定するステップは、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記検査サンプルのシーケンス深度、前記検査サンプル中のｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる。
【００４６】
第４の態様の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、疾患分類を差し控える。
【００４７】
第４の態様の方法において、前記疾患予測は、疾患と疾患なしとの間の二値予測を備える。
【００４８】
第４の態様の方法において、前記疾患予測は、複数の疾患間のマルチクラスがん予測を含む。
【００４９】
第４の態様の方法において、前記疾患分類を実行するステップは、前記検査サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに基づいて検査特徴ベクターを生成するステップと、および前記検査特徴ベクターを分類モデルに入力し、前記検査サンプルの前記疾患予測を生成するステップと、を含む。
【００５０】
第４の態様の方法において、前記疾患分類を実行するステップは、さらに、異常断片のセット（集合）を生成するために、前記検査サンプルのｃｆＤＮＡ断片の初期セット（集合）をｐ値フィルタリングでフィルタリングするステップを含み、前記フィルタリングするステップは、前記異常断片のセット（集合）を生成するよう、他の断片に関して閾値ｐ値を下回る断片を前記初期セットから除去するステップを含むものであり、前記検査特徴ベクターは、前記異常断片のセットの前記配列リードに基づくものである。
【００５１】
第４の態様の方法において、前記分類モデルは、機械学習モデルである。
【００５２】
第５の態様によれば、命令がコンピュータプロセッサによって実行されると、第４の態様の方法を前記コンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体を開示する。
【００５３】
第６の態様によれば、コンピュータプロセッサと、および第５の態様の前記非一過性のコンピュータ可読記憶媒体とを備えるシステムを開示する。
【００５４】
第７の態様によれば、検査サンプル中の汚染の存在を予測する方法が開示され、該方法は、前記検査サンプル中の複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片に由来する配列リードを取得するステップと、前記配列リードに基づいて、前記検査サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、同定された汚染マーカーの１つにおける、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記検査サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、同定されたあらゆる前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記汚染レベルが

閾値レベル以下であるか否かを決定するステップと、並びに前記汚染レベルが前記閾値レベル以下であると決定することに応答して、前記検査サンプルが汚染されていることを示す通知を生成するステップと、を備える。

【００５５】

第７の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。

【００５６】

第７の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。

【００５７】

第７の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００５８】

第７の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。

【００５９】

第７の態様の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００６０】

第７の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。

【００６１】

第７の態様の方法において、前記複数の汚染マーカーは、表１からの前記複数のＳＮＰ部位を含む。

【００６２】

第７の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。

【００６３】

第７の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。

【００６４】

第７の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する。

【００６５】

第７の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００６６】

第７の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。

【００６７】

第７の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。

【００６８】

第７の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。

【００６９】

第７の態様の方法において、前記汚染レベルを推定するステップは、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記検査サンプルのシーケンシング深度、前記検査サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づ

いて行われる。

【００７０】

第８の態様によれば、命令がコンピュータプロセッサによって実行されると、第７の態様の方法をコンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体を開示する。

【００７１】

第９の態様によれば、コンピュータプロセッサと、および第８の態様の非一過性のコンピュータ可読記憶媒体とを備えるシステムを開示する。

【００７２】

第１０の態様によれば、がん分類モデルを訓練する方法が開示され、該方法は、第１の訓練サンプルを含む複数の訓練サンプルを取得するステップであり、各訓練サンプルは複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片を含むものである、前記訓練サンプルを取得するステップと、各訓練サンプルについて、前記訓練サンプル中の前記ｃｆＤＮＡ断片に由来する配列リードを取得するステップと、前記第１の訓練サンプルについて、前記第１の訓練サンプルの前記配列リードに基づいて、前記第１の訓練サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから１つ以上の複数の汚染マーカーを同定するステップと、前記第１の訓練サンプルについて、同定された汚染マーカーの１つにおける、それぞれの前記汚染マーカーの前記ホモ接合性ハプロタイプとは異なるハプロタイプを有する、前記第１の訓練サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、前記第１の訓練サンプルについて、同定されたあらゆる汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および前記第１の訓練サンプルについて、前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、並びに前記第１の訓練サンプルの前記汚染レベルが前記閾値レベル以上であると決定することに応答して、前記複数の訓練サンプルから前記第１の訓練サンプルを除去するステップと、を備え、前記第１の訓練サンプルを除外する前記複数の訓練サンプルが、前記検査サンプルに対するがん予測を生成するために、前記がん分類モデルを訓練するのに使用される。

【００７３】

第１０の態様の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。

【００７４】

第１０の態様の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。

【００７５】

第１０の態様の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。

【００７６】

第１０の態様の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。

【００７７】

第１０の態様の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。

【００７８】

第１０の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。

【００７９】

第１０の態様の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。

【００８０】

第１０の態様の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部

位を含む。
【００８１】
第１０の態様の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【００８２】
第１０の態様の方法において、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【００８３】
第１０の態様の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【００８４】
第１０の態様の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む。
【００８５】
第１０の態様の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【００８６】
第１０の態様の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【００８７】
第１０の態様の方法において、前記複数の訓練サンプルは、非がんサンプルの第１のコホートおよびがんサンプルの第２のコホートを備え、前記がん分類モデルが、がんの存在の可能性を決定するように訓練される。
【００８８】
第１０の態様の方法において、前記がんサンプルの第２のコホートは、第１のがんタイプを有する１つ以上のサンプルおよび、第２のがんタイプを有する１つ以上の追加のサンプルを含み、ここで前記がん分類モデルは、前記第１のがんタイプの存在の第１の尤度および前記第２のがんタイプの存在の第２の尤度を決定するように訓練される。
【００８９】
第１０の態様の方法において、前記がん分類モデルは、機械学習モデルである。
【００９０】
第１０の態様の方法において、がん分類モデルは、決定木、ニューラルネットワーク、多層パーセプトロン、およびサポートベクターマシンのうちの少なくとも１つである。
【００９１】
第１１の態様によれば、命令がコンピュータプロセッサによって実行されると、第１０の態様の方法をコンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体を開示する。
【００９２】
第１２の態様によれば、コンピュータプロセッサと、および第１１の態様の前記非一過性のコンピュータ可読記憶媒を備えるシステムを開示する。
【００９３】
第１３の態様によれば、非一過性のコンピュータ可読記憶媒体を備えるコンピュータプログラム製品を開示し、前記コンピュータプログラム製品は、訓練された前記がん分類モデルを記憶する前記非一過性のコンピュータ可読記憶媒体を備え、前記コンピュータプログラム製品は、第１１の態様の方法によって作成される。
【００９４】
第１４の態様によれば、個体からの遺伝物質を含む生物学的サンプルを保存するための１つ以上の収集容器と、および複数の汚染マーカーを標的とする複数のプローブであって、表２および表４のうちの少なくとも一方を含むものである、前記複数のプローブと、を備える治療キットを開示する。

【００９５】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む。
【００９６】
第１４の態様の治療キットにおいて、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である。
【００９７】
第１４の態様の治療キットにおいて、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【００９８】
第１４の態様の治療キットにおいて、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する。
【００９９】
第１４の態様の治療キットにおいて、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【０１００】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む。
【０１０１】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む。
【０１０２】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む。
【０１０３】
第１４の態様の治療キットにおいて、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである。
【０１０４】
第１４の態様の治療キットにおいて、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する。
【０１０５】
第１４の態様の治療キットにおいて、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある。
【０１０６】
第１４の態様の治療キットにおいて、複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００のインデル部位を含む。
【０１０７】
第１４の態様の治療キットにおいて、前記複数の汚染マーカーは、表３の前記インデル部位を含む。
【０１０８】
第１４の態様の治療キットにおいて、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む。
【０１０９】
第１４の態様の治療キットにおいて、さらに、生物学的サンプル中の核酸断片を単離するための１つ以上の試薬を備える。
【０１１０】
第１４の態様の治療キットにおいて、さらに、第２の態様の非一過性のコンピュータ可読記憶媒体、第５の態様の非一過性のコンピュータ可読記憶媒体、第８の態様の非一過性

のコンピュータ可読記憶媒体、および第１１の態様の非一過性のコンピュータ可読記憶媒体のうちの１つ以上を含む、前記第１のコンピュータプログラム製品を備える。

【０１１１】

第１４の態様の治療キットは、さらに、第１３の態様のコンピュータプログラム製品を備える。

【図面の簡単な説明】

【０１１２】

【図１】１つ以上の実施形態による、サンプル中の汚染検出のプロセスを説明する例示的なフローチャートである。

【図２Ａ】１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための複数のＳＮＰ部位を同定するプロセスを説明する例示的なフローチャートである。

【図２Ｂ】１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための、インデル部位を同定するプロセスを説明する例示的なフローチャートである。

【図３Ａ】１つ以上の実施形態による、メチル化状態ベクターを得るためにセルフリー（ｃｆ）ＤＮＡの断片を配列決定するプロセスを説明する例示的なフローチャートである。

【図３Ｂ】１つ以上の実施形態による、メチル化状態ベクターを得るためにセルフリー（ｃｆ）ＤＮＡの断片を配列決定する図３Ａのプロセスの例示的な説明図である。

【図４Ａ】１つ以上の実施形態による、健常対照群のデータ構造を生成するプロセスを説明するフローチャートである。

【図４Ｂ】１つ以上の実施形態による、サンプルから異常にメチル化された断片を同定するプロセスを説明する例示的なフローチャートである。

【図５Ａ】１つ以上の実施形態による、がん分類器を訓練するプロセスを説明する例示的なフローチャートである。

【図５Ｂ】１つ以上の実施形態による、がん分類器の訓練に使用される特徴ベクターの生成例を示す図である。

【図６Ａ】１つ以上の実施形態による、核酸サンプルを配列決定するための装置の例示的なフローチャートを示す。

【図６Ｂ】１つ以上の実施形態による、解析システムの例示的なブロック図である。

【図７】所与のサンプルについてホモ接合と判定された各汚染マーカーについて、参照または代替と判定された断片の総数に対する、汚染と判定された断片の数の割合の分布を示す図である。

【図８】例の結果の第１のグループによる、汚染マーカーの対立遺伝子頻度の散布図を示す。

【図９】例の結果の第１のグループによる、汚染マーカーの遺伝子型および接合性を示す２つのグラフを示す図である。

【図１０】第１のグループの例示的な結果に従って、ホモ接合体であり、所与のサンプルについて、そのサンプルの汚染を推定するために有用であると考えられるのに十分な断片が重なる汚染マーカーの割合を示すグラフである。

【図１１Ａ】第１の結果例に従った、異なるバッチのサンプルについての推定汚染レベルのグラフを示す。

【図１１Ｂ】第１の結果例に従った、サンプルの異なるバッチについての推定汚染レベルの追加グラフを示す。

【図１２】第２の結果例群による、表２および表４に列挙された汚染マーカーごとに得られた独特なｃｆＤＮＡ断片の数の分布を示す図である。

【図１３】第２の結果例群による、品質チェックプロセスの結果を示す図である。

【図１４】第２の結果例群による、対立遺伝子頻度の散布図を示す。

【図１５】マーカーがヘテロ接合と呼ばれる頻度の散布図である。

【図１６】第２のグループの結果例による、ハーディーワインベルグ項の値の散布図を示す。

【図１７】第２のグループの結果例に従って、所与のサンプルについてホモ接合と呼ばれ

た各汚染マーカーについて、参照または代替と呼ばれた断片の総数に対する、汚染と分類された断片の数の割合の分布を示す図である。
【図１８】第２のグループの結果例に従って、仮想的な断片セットに対する汚染率を推定するための公式の適用を示す図である。
【図１９】第２のグループの結果例による、シミュレーションデータに汚染割合モデルを適用した結果を示す図である。
【図２０】第２の結果例群に従って、滴定レベルを変化させたｃｆＤＮＡサンプルの４つの滴定対に汚染分画モデルを適用した結果を示す図である。
【図２１】第２の結果例に従った、実験における８４個のｃｆＤＮＡサンプルについての推定汚染割合の分布を示す図である。
【図２２】実験中の８４サンプルについて、複数のＳＮＰマーカーおよびインデルマーカーをそれ自体で別々に考慮した場合に得られた推定値が、第２のグループの例示的な結果に従って、高度に相関しており、有意な系統的バイアスを示さないことを示す図である。
【図２３】１つ以上の実施形態による、複数のＳＮＰ汚染マーカーを含む表１を示す図である。
【図２４】１つ以上の実施形態による、複数のＳＮＰ汚染マーカーのプローブ配列リストを含む表２を示す。
【図２５】１つ以上の実施形態による、インデル汚染マーカーを含む表３を示す。
【図２６】１つ以上の実施形態による、インデル汚染マーカーのプローブ配列リストを含む表４を示す。
【０１１３】
図は、例示のみを目的とした様々な実施形態を描いている。当業者は、以下の考察から、本明細書に記載される原理から逸脱することなく、本明細書に例示される構造および方法の代替の実施形態が採用され得ることを容易に認識するであろう。
【発明を実施するための形態】
【０１１４】
［Ｉ．概要］
＜Ｉ．Ａ．メチル化の概要＞
本明細書に従って、個体からのｃｆＤＮＡ断片を、例えばメチル化されていないシトシンをウラシルに変換するなどの方法で処理し、塩基配列を決定し、配列リードを参照（基準）ゲノムと比較して、ＤＮＡ断片内の特定のＣｐＧ部位のメチル化状態を同定する。各ＣｐＧ部位は、メチル化されている場合とされていない場合がある。健常人と比較して異常なメチル化断片を同定することで、被験者のがんの状態を知ることができる。当技術分野でよく知られているように、ＤＮＡメチル化異常は（健常対照と比較して）様々な影響を引き起こし、それががんの一因となる可能性がある。異常にメチル化されたｃｆＤＮＡ断片の同定には様々な課題が生じる。第一に、ＤＮＡ断片が異常にメチル化されていると判定することは、対照群と比較して重要な意味を持つが、対照群の数が少なければ、対照群のサイズが小さいために統計的なばらつきが生じ、判定が信頼性を失う。さらに、対照群の中でもメチル化の状態は様々であり、被験者のＤＮＡ断片が異常にメチル化されていると判定する際に、これを考慮することは困難である。また、あるＣｐＧ部位のシトシンのメチル化は、その後に続くＣｐＧ部位のメチル化に影響を及ぼす可能性がある。この依存関係をカプセル化することは、それ自体が別の課題となりうる。
【０１１５】
メチル化は、デオキシリボ核酸（ＤＮＡ）において、シトシン塩基のピリミジン環上の水素原子がメチル基に変換され、５－メチルシトシンを形成するときに典型的に起こり得る。特に、メチル化は、本明細書で「ＣｐＧ部位（ＣＰＧ ｓｉｔｅ）」と呼ばれるシトシンとグアニンのジヌクレオチドで起こりうる。他の例では、メチル化は、ＣｐＧ部位の一部ではないシトシン、またはシトシンではない別のヌクレオチドで起こり得るが、これらはまれな現象である。本開示では、わかりやすくするために、メチル化はＣｐＧ部位に関連して議論される。異常なＤＮＡメチル化は、高メチル化または低メチル化として同定するこ

とができ、どちらもがんの状態を示す可能性がある。本開示を通して、ＤＮＡ断片が閾値以上のＣｐＧ部位を含み、それらのＣｐＧ部位の閾値以上の割合がメチル化または非メチル化である場合、ＤＮＡ断片について高メチル化および低メチル化を特徴付けることができる。

【０１１６】

本明細書に記載された原理は、非シトシンメチル化を含む非ＣｐＧコンテクストにおけるメチル化の検出にも同様に適用できる。そのような実施形態において、メチル化を検出するために使用される湿式実験室アッセイは、本明細書に記載されるものとは異なる場合がある。さらに、本明細書で議論されるメチル化状態ベクターは、一般的にメチル化が起こった部位または起こっていない部位（それらの部位が特にＣｐＧ部位でなくても）である要素を含むことができる。この置換により、本明細書に記載される残りのプロセスは同じものとすることができ、その結果、本明細書に記載される本発明の概念は、それらの他の形態のメチル化に適用され得る。

【０１１７】

＜Ｉ．Ｂ．定義＞

「無細胞（セルフリー）核酸（cell free nucleic acid）」または「ｃｆＮＡ」という用語は、個体の体内（例えば血液）を循環する核酸断片であって、１つ以上の健常細胞から、および／または１つ以上の不健全細胞（例えばがん細胞）から起源する核酸断片を指す。「無細胞（セルフリー）ＤＮＡ（cell free DNA）」または「ｃｆＤＮＡ」という用語は、個体の体内（例えば血液）を循環するデオキシリボ核酸断片を指す。さらに、ｃｆＮＡまたは個体の体内のｃｆＤＮＡは、ヒト以外の他の供給源に由来することもある。

【０１１８】

「ゲノム核酸（genomic nucleic acid）」、「ゲノムＤＮＡ（genomic DNA）」、または「ｇＤＮＡ」という用語は、１つ以上の細胞から得られる核酸分子またはデオキシリボ核酸分子を指す。様々な実施形態において、ｇＤＮＡは、健康な細胞（例えば、非腫瘍細胞）または腫瘍細胞（例えば、生検サンプル）から抽出され得る。いくつかの実施形態において、ｇＤＮＡは、白血球などの血球系由来の細胞から抽出され得る。

【０１１９】

「循環腫瘍ＤＮＡ（circulating tumor DNA）」または「ｃｔＤＮＡ」という用語は、腫瘍細胞または他のタイプのがん細胞に由来し、瀕死細胞のアポトーシスまたはネクローシスなどの生物学的プロセスの結果として、または生存腫瘍細胞によって能動的に放出される、個体の体液（例えば、血液、汗、尿、または唾液）中に放出され得る核酸断片を指す。

【０１２０】

「ＤＮＡ断片（DNA fragment）」、「断片（fragment）」、または「ＤＮＡ分子（DNA molecule）」という用語は、一般に、任意のデオキシリボ核酸断片、すなわち、ｃｆＤＮＡ、ｇＤＮＡ、ｃｔＤＮＡなどを指す場合がある。

【０１２１】

「異常な断片（anomalous fragment）」、「異常なメチル化断片（anomalously methylated fragment）」、または「異常なメチル化パターンを有する断片（fragment with an anomalous methylation pattern）」という用語は、ＣｐＧ部位の異常なメチル化を有する断片を指す。断片の異常なメチル化は、対照群における断片のメチル化パターンを観察することの意外性を同定するために、確率モデルを用いて決定することができる。

【０１２２】

「極端なメチル化を伴う異常な断片（unusual fragment with extreme methylation）」または「ＵＦＸＭ」という用語は、低メチル化断片または高メチル化断片を指す。低メチル化断片および高メチル化断片とは、それぞれメチル化または非メチル化の割合が閾値（例えば９０％）を超えるＣｐＧ部位を少なくともいくつか（例えば５つ）有す

る断片を指す。

【０１２３】

「アノマリスコア（anomaly score）」という用語は、そのＣｐＧ部位に重なるサンプルからの異常な断片（または、いくつかの実施形態ではＵＦＸＭ）の数に基づく、ＣｐＧ部位のスコアを指す。異常スコアは、分類のためのサンプルの特徴付けの文脈で使用される。

【０１２４】

本明細書で使用される場合、用語「約（about）」または「おおよそ（approximately）」は、当業者によって決定される特定の値について許容可能な誤差範囲内であることを意味し得、これは、値が測定または決定される方法、例えば測定システムの限界に部分的に依存し得る。例えば、「約（about）」は、当業者の慣行に従って、１標準偏差以内または１標準偏差以上を意味することができる。「約（About）」は、所与の値の±２０％、±１０％、±５％、または±１％の範囲を意味することができる。用語「約（about）」または「おおよそ（approximately）」は、値の１桁以内、５倍以内、または２倍以内を意味し得る。本願明細書および特許請求の範囲に特定の値が記載されている場合、特に断りのない限り、特定の値の許容誤差範囲内を意味する「約（about）」という用語が想定されるべきである。用語「約（about）」は、当業者が一般的に理解する意味を有することができる。用語「約（about）」は、±１０％を指すことができる。「約（about）」という用語は、±５％を指すことができる。

【０１２５】

本明細書で使用される場合、「生物学的サンプル（biological sample）」、「患者サンプル（patient sample）」、または「サンプル（sample）」という用語は、被検者から採取され、被検者に関連する生物学的状態を反映することができ、セルフリー（無細胞）ＤＮＡを含む任意のサンプルを指す。生物学的サンプルの例としては、被検者の血液、全血、血漿、血清、尿、脳脊髄液、糞便、唾液、汗、涙、胸水、心嚢液、または腹膜液が挙げられるが、これらに限定されない。生物学的サンプルには、生死を問わず、被検者由来のあらゆる組織または物質が含まれ得る。生物学的サンプルは、セルフリーサンプルであり得る。生物学的サンプルは、核酸（例えば、ＤＮＡまたはＲＮＡ）またはその断片を含むことができる。核酸（nucleic acid）」という用語は、デオキシリボ核酸（ＤＮＡ）、リボ核酸（ＲＮＡ）、またはそれらのハイブリッドもしくは断片を指すことができる。サンプル中の核酸は、セルフリー核酸であり得る。サンプルは、液体サンプルまたは固体サンプル（例えば、細胞サンプルまたは組織サンプル）であり得る。生物学的サンプルは、体液、例えば、血液、血漿、血清、尿、膣液、（例えば、精巣の）水腫からの液、膣洗浄液、胸水、腹水、脳脊髄液、唾液、汗、涙、喀痰、気管支肺胞洗浄液、乳頭からの排出液、身体の異なる部位（例えば、甲状腺、乳房）からの吸引液などであることができる。生物学的サンプルは便サンプルでもよい。様々な実施形態において、セルフリーＤＮＡについて濃縮された生物学的サンプル（例えば、遠心分離プロトコルを介して得られた血漿サンプル）中のＤＮＡの大部分は、セルフリーであり得る（例えば、ＤＮＡの５０％、６０％、７０％、８０％、９０％、９５％、または９９％より多くがセルフリーであり得る）。生物学的サンプルは、組織または細胞の構造を物理的に破壊するために処理され（例えば、遠心分離および／または細胞溶解）、その結果、細胞内成分を溶液中に放出することができる。この溶液は、さらに酵素、緩衝液、塩、洗浄剤などを含むことができ、この酵素、緩衝液、塩、洗浄剤などは、解析のためにサンプルを調製するために使用することができる。

【０１２６】

本明細書で使用される場合、「対照（control）」、「対照サンプル（control sample）」、「参照（reference）」、「参照サンプル（reference sample）」、「正常（normal）」、および「正常サンプル（normal sample）」という用語は、特定の状態を有しないか、そうでなければ健康である被検者からのサンプルを表す。一例において、本明細書に開示されるような方法は、腫瘍を有する被検者に対して実施され得、ここ

で、参照サンプルは、被検者の健康な組織から採取されたサンプルである。参照サンプルは、から取得することも、データベースから取得することもできる。参照は、例えば、被検者からのサンプルの配列決定から得られた核酸断片配列をマッピングするために使用される参照ゲノムであり得る。参照ゲノムは、生物学的サンプルおよび体質サンプルからの核酸断片配列が整列および比較され得る、ハプロイドまたは二倍体ゲノムを指し得る。体質サンプルの例としては、被験者から得られた白血球のＤＮＡを挙げることができる。ハプロイドゲノムの場合、各遺伝子座には１つのヌクレオチドしか存在しない。２倍体ゲノムの場合、ヘテロ接合性遺伝子座を同定することができる。各ヘテロ接合性遺伝子座は２つの対立遺伝子を持つことができ、どちらの対立遺伝子もその遺伝子座へのアライメントに一致させることができる。

【０１２７】

本明細書において、「がん（cancer）」または「腫瘍（tumor）」という用語は、組織の異常な腫瘤を指し、その腫瘤の成長は正常組織の成長を上回り、正常組織の成長と協調していない。

【０１２８】

本明細書で使用される場合、「健康な（healthy）」という語句は、良好な健康状態を有する被検者を指す。健康な被検者は、悪性疾患または非悪性疾患がないことを示すことができる。「健康な個体（healthy individual）」は、アッセイされる状態とは無関係の、通常「健康（healthy）」とはみなされない他の疾患または状態を有することができる。

【０１２９】

本明細書において、「メチル化（methylation）」という用語は、シトシン塩基のピリミジン環上の水素原子がメチル基に変換され、５－メチルシトシンを形成するデオキシリボ核酸（ＤＮＡ）の修飾を指す。特に、メチル化は、本明細書で「ＣｐＧ部位（CpG sites）」と呼ばれるシトシンとグアニンのジヌクレオチドで起こる傾向がある。ＣｐＧ部位以外のシトシンまたはシトシン以外のヌクレオチドでメチル化が起こることもあるが、これはまれである。ｃｆＤＮＡのメチル化異常は、高メチル化または低メチル化として同定することができ、どちらもがんの状態を示す可能性がある。ＤＮＡメチル化異常は（健常対照と比較して）様々な影響を引き起こす可能性があり、がんの一因となる可能性がある。本明細書に記載された原理は、ＣｐＧコンテクストにおけるメチル化の検出にも、非シトシンメチル化を含む非ＣｐＧコンテクストにおけるメチル化の検出にも同様に適用可能である。さらに、メチル化状態ベクターは、一般にメチル化が起こった部位または起こっていない部位のベクターである要素を含むことができる（それらの部位が特にＣｐＧ部位でなくても）。

【０１３０】

本明細書で互換的に使用されるように、用語「メチル化断片（methylation fragment）」または「核酸メチル化断片（nucleic acid methylation fragment）」は、核酸（例えば、核酸分子および／または核酸断片）のメチル化配列決定によって決定される、複数のＣｐＧ部位における各ＣｐＧ部位のメチル化状態の配列を指す。メチル化断片において、核酸断片中の各ＣｐＧ部位の位置およびメチル化状態は、参照ゲノムに対する配列リード（例えば、核酸の配列決定から得られる）のアライメントに基づいて決定される。核酸メチル化断片は、複数のＣｐＧ部位の各ＣｐＧ部位のメチル化状態（例えば、メチル化状態ベクター）からなり、参照ゲノムにおける核酸断片の位置（例えば、ＣｐＧインデックス、または別の類似の指標を用いて核酸断片中の最初のＣｐＧ部位の位置によって指定される）および核酸断片中のＣｐＧ部位の数を特定する。核酸分子のメチル化配列決定に基づく、参照ゲノムに対する配列リードのアライメントは、ＣｐＧインデックスを用いて実施することができる。本明細書で使用する場合、「ＣｐＧインデックス（CpG index）」という用語は、ヒト参照ゲノムなどの参照ゲノムにおける複数のＣｐＧ部位（例えば、ＣｐＧ　１、ＣｐＧ　２、ＣｐＧ　３など）の各ＣｐＧ部位のリストを意味し、これは電子フォーマットであり得る。ＣｐＧインデックスはさらに、ＣｐＧインデックス

におけるそれぞれのＣｐＧ部位について、対応する参照ゲノムにおける、対応するゲノム上の位置を含んでいる。このようにして、それぞれの核酸メチル化断片中のそれぞれのＣｐＧ部位は、それぞれの参照ゲノム中の特定の位置にインデックス付けされ、ＣｐＧインデックスを用いて決定することができる。

【０１３１】

本明細書で使用される場合、「真陽性（true positive）」（ＴＰ）という用語は、状態を有する被検者に言及する。「真陽性（True positive）」は、腫瘍、がん、前がん状態（例えば、前がん病変）、限局性もしくは転移性がん、または非悪性疾患を有する被検者に言及することができる。「真陽性（True positive）」は、状態を有する被検者に言及することができ、本開示のアッセイまたは方法によって状態を有することが同定される。本明細書で使用される場合、「真陰性（True negative）」（ＴＮ）という用語は、病態を有さないか、または検出可能な病態を有さない被検者に言及する。真陰性は、腫瘍、がん、前がん状態（例えば、前がん病変）、限局性または転移性がん、非悪性疾患、またはその他の健康な被検者などの疾患または検出可能な疾患を有さない被検者に言及することができる。真陰性は、病態を有さないか、検出可能な病態を有さないか、または本開示のアッセイもしくは方法によって病態を有さないと同定される被検者に言及することができる。

【０１３２】

本明細書で使用される場合、「参照ゲノム（reference genome）」という用語は、被験体から同定された配列を参照するために使用され得る、任意の生物またはウイルスの、部分的であるか完全であるかにかかわらず、任意の特定の既知の、配列決定された、または特徴付けられたゲノムを指す。ヒト被験体および他の多くの生物に使用される例示的な参照ゲノムは、アメリカ国立生物工学情報センター（「ＮＣＢＩ」）またはカリフォルニア大学サンタクルーズ校（「ＵＣＳＣ」）がホストするオンラインゲノムブラウザで提供されている。「ゲノム（genome）」とは、核酸配列で表される生物またはウイルスの完全な遺伝情報を指す。本明細書で使用される場合、参照配列または参照ゲノムは、多くの場合、個体または複数の個体からアセンブルされた、または部分的にアセンブルされたゲノム配列である。いくつかの実施形態において、参照ゲノムは、１つ以上のヒト個体からのアセンブルされた、または部分的にアセンブルされたゲノム配列である。参照ゲノムは、ある種の遺伝子セットの代表例とみなすことができる。いくつかの実施形態では、参照ゲノムは染色体に割り当てられた配列からなる。例示的なヒト参照ゲノムとしては、ＮＣＢＩビルド３４（ＵＣＳＣ相当：ｈｇ１６）、ＮＣＢＩビルド３５（ＵＣＳＣ相当：ｈｇ１７）、ＮＣＢＩビルド３６．１（ＵＣＳＣ相当：ｈｇ１８）、ＧＲＣｈ３７（ＵＣＳＣ相当：ｈｇ１９）、およびＧＲＣｈ３８（ＵＣＳＣ相当：ｈｇ３８）が挙げられるが、これらに限定されない。

【０１３３】

本明細書で使用される場合、「配列リード（sequence reads）」または「リード（reads）」という用語は、本明細書に記載されるまたは当該技術分野で公知の任意の配列決定プロセスによって生成されるヌクレオチド配列を指す。リードは、核酸断片の一端から生成することができ（「シングルエンドリード（single-end reads）」）、核酸の両端から生成することもある（例えば、ペアエンドリード（paired-end reads）、ダブルエンドリード（double-end reads））。一部の実施形態では、配列リード（例えば、シングルエンドリードまたはペアエンドリード）は、標的核酸断片の片鎖または両鎖から生成することができる。配列リードの長さは、特定の配列決定技術に関連することが多い。例えば、ハイスループットメソッドは、数十から数百塩基対（ｂｐ）の大きさで変化し得る配列リードを提供する。いくつかの実施形態において、配列リードは、約１５ｂｐ～９００ｂｐの長さの平均、中央値、または平均の長さである（例えば約２０ｂｐ、約２５ｂｐ、約３０ｂｐ、約３５ｂｐ、約４０ｂｐ、約４５ｂｐ、約５０ｂｐ、約５５ｂｐ、約６０ｂｐ、約６５ｂｐ、約７０ｂｐ、約７５ｂｐ、約８０ｂｐ、約８５ｂｐ、約９０ｂｐ、約９５ｂｐ、約１００ｂｐ、約１１０ｂｐ、約１２０ｂｐ、約１３０、約１４０ｂ

ｐ、約１５０ｂｐ、約２００ｂｐ、約４５０ｂｐ、約３００ｂｐ、約３５０ｂｐ、約４００ｂｐ、約４５０ｂｐ、または約５００ｂｐである）。いくつかの実施形態では、配列リードは、約１０００ｂｐ、２０００ｂｐ、５０００ｂｐ、１０，０００ｂｐ、または５０，０００ｂｐ以上の平均、中央値、または平均長である。例えば、ナノポアシーケンシングは、数十から数百から数千塩基対のサイズで変化し得る配列リードを提供することができる。イルミナのパラレル配列では、それほど変化しない配列リードを提供することができ、例えば、配列リードのほとんどは２００ｂｐより小さくなり得る。配列リード（ｓｅｑｕｅｎｃｅ ｒｅａｄ）（または配列リード（ｓｅｑｕｅｎｃｉｎｇ ｒｅａｄ））は、核酸分子（例えば、ヌクレオチドの文字列）に対応する配列情報を指すことができる。例えば、配列リードは、核酸断片の一部からのヌクレオチドの文字列（例えば、約２０～約１５０）に対応し得るものであり、核酸断片の一端または両端のヌクレオチドの文字列に対応し得、または核酸断片全体のヌクレオチドに対応し得る。配列読み取りは、様々な方法、例えば、配列決定技術、またはプローブ、例えば、ハイブリダイゼーションアレイまたはキャプチャープローブ、または増幅技術、例えば、ポリメラーゼ連鎖反応（ＰＣＲ）、または単一プライマーを用いる線形増幅、または等温増幅を用いて得ることができる。

【０１３４】

本明細書で使用される「配列決定／シーケンシング（ｓｅｑｕｅｎｃｉｎｇ）」などの用語は、一般に、核酸またはタンパク質などの生物学的高分子の順序を決定するために使用され得るあらゆる生化学的プロセスを指す。例えば、配列決定データは、ＤＮＡ断片などの核酸分子中のヌクレオチド塩基の全てまたは一部を含み得る。

【０１３５】

本明細書で使用する場合、「シーケンシング深度（ｓｅｑｕｅｎｃｉｎｇ ｄｅｐｔｈ）」という用語は、「カバレッジ（ｃｏｖｅｒａｇｅ）」という用語と互換的に使用され、遺伝子座にアライメントされた独特な核酸標的分子に対応するコンセンサス配列リードによって遺伝子座がカバーされる回数を指し、例えば、シーケンシング深度は、遺伝子座をカバーする独特な核酸標的分子の数に等しい。遺伝子座は、ヌクレオチドのように小さくても、染色体アームのように大きくても、ゲノム全体のように大きくてもよい。シーケンシング深度は、「Ｙｘ」、例えば、５０ｘ、１００ｘなどで表すことができ、ここで「Ｙ」は、核酸標的に対応する配列で遺伝子座がカバーされる回数、例えば、特定の遺伝子座をカバーする独立した配列情報が得られる回数を指す。一部の実施形態では、シーケンシング深度は、配列決定されたゲノムの数に対応する。シーケンシング深度はまた、複数の遺伝子座、または全ゲノムに適用され得、この場合、Ｙは、それぞれ、遺伝子座、ハプロイドゲノム、または全ゲノムが配列決定された平均または平均回数を指し得る。平均深度が引用されている場合、データセットに含まれる異なる遺伝子座の実際の深度は、値の範囲に及ぶ可能性がある。超深度シーケンシングとは、遺伝子座におけるシーケンシング深度が少なくとも１００倍であることを指す。

【０１３６】

本明細書で使用される場合、「感度（ｓｅｎｓｉｔｉｖｉｔｙ）」または「真陽性率（ｔｒｕｅ ｐｏｓｉｔｉｖｅ ｒａｔｅ）」（ＴＰＲ）という用語は、真陽性の数を真陽性と偽陰性の数の合計で割ったものを指す。感度は、ある状態を真に有する集団の割合を正しく同定するアッセイまたは方法の能力を特徴付けることができる。例えば、感度は、集団内のがんを有する被験者の数を正しく同定する方法の能力を特徴付けることができる。別の例では、感度は、がんを示す１つ以上のマーカーを正しく同定する方法の能力を特徴付けることができる。

【０１３７】

本明細書で使用される場合、「特異度（ｓｐｅｃｉｆｉｃｉｔｙ）」または「真陰性率（ｔｒｕｅ ｎｅｇａｔｉｖｅ ｒａｔｅ）」（ＴＮＲ）という用語は、真陰性の数を真陰性と偽陽性の数の合計で割ったものを指す。特異性は、ある状態を真に有していない集団の割合を正しく同定するアッセイまたは方法の能力を特徴付けることができる。例えば、特異性は、ある集団の中でがんでない被験者の数を正しく同定する方法の能力を特徴付けることができる。別の例では、特異性は、がんを示す１つ以上のマーカーを正しく同定する方法の能力を特徴付け

る。
【０１３８】
本明細書で使用する場合、「被験体／被検体（subject）」という用語は、ヒト（例えば、男性ヒト、女性ヒト、胎児、妊娠中の女性、子供など）、非ヒト動物、植物、細菌、真菌または原生生物を含むがこれらに限定されない、任意の生物または非生物を指す。ヒトまたは非ヒト動物であれば、哺乳動物、爬虫類、鳥類、両生類、魚類、偶蹄類、反芻動物、ウシ科（例えば、ウシ）、ウマ科（例えば、ウマ）、ヤギおよびヒツジ（例えば、ヒツジ、ヤギ）、ブタ（例えば、ブタ）、ラクダ科（例えば、ラクダ、ラマ、アルパカ）、サル、類人猿（例えば、ゴリラ、チンパンジー）、クマ科（例えば、クマ）、家禽、イヌ、ネコ、マウス、ラット、魚、イルカ、クジラ、およびサメが被験体になることができる。いくつかの実施形態において、被験体は、あらゆる段階の男性または女性（例えば、男性、女性または子供）である。サンプルが採取される、または本明細書に記載される方法もしくは組成物のいずれかによって処置される被検者は、任意の年齢であり得、成人、幼児または子供であり得る。
【０１３９】
本明細書で使用する場合、「組織（tissue）」という用語は、機能的な単位として一緒になっている細胞群に対応することができる。１つの組織には複数の種類の細胞が存在し得る。異なる種類の組織は、異なる種類の細胞（例えば、肝細胞、肺胞細胞または血球）から構成されることがあるが、異なる生物（母親対胎児）の組織、または健康な細胞対腫瘍細胞に対応することもある。「組織（tissue）」という用語は、一般に、ヒトの体内で見出されるあらゆる細胞群（例えば、心臓組織、肺組織、腎臓組織、鼻咽頭組織、口腔咽頭組織）を指すことができる。いくつかの態様において、「組織（tissue）」または「組織型（tissue type）」という用語は、セルフリー核酸が由来する組織を指すために使用することができる。一例では、ウイルス核酸断片は血液組織に由来し得る。別の例では、ウイルス核酸断片は腫瘍組織に由来し得る。
【０１４０】
本明細書で使用する場合、「ゲノムの（genomic）」という用語は、生物のゲノムの特性を指す。ゲノム特性の例としては、ゲノムの全部または一部の一次核酸配列（例えば、ヌクレオチド多型、インデル、配列再配列、変異頻度などの有無）、ゲノム内の１つ以上の特定のヌクレオチド配列のコピー数（例えば、コピー数、対立遺伝子頻度分率、単一染色体または全ゲノムの倍数性など）、ゲノムの全部または一部のエピジェネティック状態（例えば、メチル化などの共有結合核酸修飾、ヒストン修飾、ヌクレオソームの位置決めなど）、生物のゲノムの発現プロファイル（例えば、遺伝子発現レベル、アイソタイプ発現レベル、遺伝子発現比など）に関するものが挙げられるが、これらに限定されない。
【０１４１】
本明細書で使用する用語は、特定の場合を説明するためのものであり、限定を意図するものではない。本明細書で使用される場合、単数形「a」、「an」および「the」は、文脈上明らかにそうでないことが示されない限り、複数形も含むことが意図される。さらに、「含む（including）」、「含む（includes）」、「有する（having）」、「有する（has）」、「有する（with）」、またはそれらの変形が、詳細な説明および／または特許請求の範囲のいずれかで使用される限り、そのような用語は、用語「備える／含む（comprising）」と同様の方法で包括的であることが意図される。
【０１４２】
＜Ｉ．Ｃ．がん分類の概要＞
一般に、がん分類では、生物学的サンプルが配列決定され、生物学的サンプル中の遺伝物質の配列リードからがんを予測するために解析される。ワークフローは、例えば、医療従事者、配列決定装置、解析システムなどを含む１つ以上の実在物による作用を含むことができる。ワークフローの目的には、個人のがんの検出および／またはモニタリングが含まれる。ヘルスケアの観点からは、ワークフローは他の既存のがん診断ツールを補完する役割を果たすことができる。ワークフローは、がんと診断された個人の治療計画をより良

く知らせるために、がんの早期発見および／または日常的ながんモニタリングを提供する役割を果たすことができる。一般的なワークフローは、代替的に、より一般的に疾患分類に適用することができる。

【０１４３】

医療従事者がサンプル採取を行う。がん分類を受けようとする個人が医療機関を訪れる。医療従事者は、がん分類を実施するためのサンプルを採取する。生物学的サンプルの例としては、被験者の組織生検、血液、全血、血漿、血清、尿、脳脊髄液、糞便、唾液、汗、涙、胸水、心嚢液、または腹膜液が挙げられるが、これらに限定されない。サンプルには、その個人に属する遺伝物質が含まれ、がん分類のために抽出され、塩基配列が決定される場合がある。サンプルが採取されると、サンプルは配列決定装置に提供される。サンプルとともに、医療従事者は、個人に関連する他の情報、例えば、生物学的性別、年齢、民族性、喫煙状態、過去の診断などを収集することができる。１つ以上の実施形態において、医療従事者は、治療キットを利用することができる。治療キットは、１つ以上のサンプル採取容器を含むことができる。治療キットは、試料の処理および解析に使用するための試薬、プローブ、コンピュータプログラム製品、指示書などをさらに含むことができる。

【０１４４】

シーケンシング（配列決定）デバイスは、サンプルに対してサンプルシケンシングを実行する。研究室の臨床医は、シーケンシングの準備として、サンプルに対して１つ以上の処理ステップを実行することができる。臨床医は、試薬、プローブなどを含む処理キットを利用することもできる。準備完了後、臨床医はサンプルを配列装置にロードする。シーケンシングに使用される装置の例を、図６Ａおよび図６Ｂに関連してさらに説明する。配列決定装置は一般に、断片に対応する核酸塩基の配列を決定するために配列決定される核酸の断片を抽出および単離する。配列決定はまた、核物質の増幅を含み得る。様々な配列決定プロセスには、サンガー配列決定、断片解析、および次世代配列決定が含まれる。配列決定は、全ゲノム配列決定または標的パネルを用いた標的配列決定の場合がある。ＤＮＡメチル化の文脈では、バイサルファイト配列決定（例えば、図３Ａおよび図３Ｂにさらに記載）は、ＣｐＧ部位における非メチル化シトシンのバイサルファイト変換を通じてメチル化の状態を決定することができる。サンプル配列決定により、サンプル中の複数の核酸断片の配列が得られる。１つ以上の実施形態において、配列は、メチル化状態ベクターを含み得、各メチル化状態ベクターは、断片上のＣｐＧ部位についてのメチル化状態を記載する。

【０１４５】

次いで、解析システムが配列リードを処理してがん予測を生成する。

【０１４６】

解析システムは、解析前処理を実行することができる。解析前処理には、配列リードの重複除去、カバレッジに関する測定基準の決定、サンプルが汚染されているかどうかの決定、汚染断片の除去、配列決定エラーの呼び出しなどが含まれるが、これらに限定されない。汚染されていると判定されたサンプルは、さらなる解析から差し控えられる場合がある。言い換えれば、解析システムは、汚染されたサンプルに対してさらなる解析（疾患またはがんの分類など）を行うことを差し控えることができる。一部の実施形態では、汚染サンプルは物理的に廃棄される場合がある。

【０１４７】

解析システムは１つ以上の解析を実行する。解析は、統計解析または１つ以上の訓練されたモデルの適用であり、サンプルが由来する個体の少なくともがんの状態を予測する。ＣｐＧ部位のメチル化、一塩基多型（ＳＮＰ）、挿入または欠失（インデル）、他のタイプの遺伝子変異など、様々な遺伝的特徴が評価および考慮され得る。メチル化の文脈では、解析は、異常メチル化同定（例えば、図４Ａおよび図４Ｂにさらに記載）、特徴抽出（例えば、図５Ａおよび図５Ｂにさらに記載）、およびがん予測を決定するためのがん分類器の適用（例えば、図５Ａおよび図５Ｂにさらに記載）を含み得る。がん分類器は、がん

予測を決定するために抽出された特徴を入力する。がん予測はラベルまたは値となる。ラベルは特定のがん状態を示すことがあり、例えば、二値ラベルはがんの有無を示すことができ、多クラスラベルはスクリーニングされる複数のがん種から１つ以上のがん種を示すことができる。値は、特定のがん状態の可能性、例えばがんの可能性、および／または特定のがん種の可能性を示すことができる。予測値はさらに、がんシグナルの定量化を示す場合があり、これには１つ以上の特定の起源組織のシグナルの定量化が含まれる場合がある。

【０１４８】

解析システムは医療従事者に予測を返す。医療従事者は、がん予測に基づいて治療計画を確立または調整することができる。治療の最適化についてはセクションＶＩ．Ｃ．治療でさらに説明する。

【０１４９】

［ＩＩ．サンプル処理］

＜ＩＩ．Ａ．汚染検出＞

図１は、１つ以上の実施形態による、サンプル中の汚染検出プロセス１００を説明する例示的なフローチャートである。一般に、サンプルは、健康であるか、がんであることが知られているか、がんであることが疑われているか、または事前情報が知られていない個人からのものであってよい。サンプルは、血液、血漿、血清、尿、糞便、および唾液サンプルからなる群から選択され得る。あるいは、サンプルは、全血、血液分画（例えば、白血球（ＷＢＣ））、組織生検、胸水、心嚢液、脳脊髄液、および腹膜液からなる群から選択することもできる。サンプルを用いたがん分類または他の分類／解析に先立ち、汚染検出プロセス１００は試料の汚染レベルを決定する。例えば、プロセス１００は、規則（例えば、閾値汚染レベルまたは間隔）と比較されたときに、サンプルが汚染されているかどうかを決定する汚染推定値および／または信頼区間を出力することができる。汚染検出は、個体のホモ接合性対立遺伝子とは異なる対立遺伝子を有する断片を同定するための汚染マーカーとして、一組の遺伝子配列を利用する。本明細書において、プロセス１００は、解析システム（その例は、図６Ａおよび図６Ｂならびに対応する説明に提供される）によって実行されるものとして記載されるが、ステップの一部または全部は、他の比較可能に記載された配列決定装置および／またはコンピュータプロセッサによって実行されてもよい。

【０１５０】

解析システムは、汚染マーカープローブを含む標的パネルで、サンプル中のｃｆＤＮＡ断片を配列決定する。汚染マーカーは、ヒトゲノム中の遺伝子配列であり、挿入－欠失（インデル）部位および複数の一塩基多型（ＳＮＰ）部位を含み得る。いくつかの実施形態において、汚染マーカーは、表１に列挙された複数のＳＮＰ部位および表３に列挙されたインデル部位の任意の組み合わせからなる。例えば、汚染マーカーは、表１に列挙された複数のＳＮＰ部位のうち少なくとも１００、２００、３００、４００、５００、６００、７００、８００、９００、または１，０００部位および、表３に列挙されたインデル部位のうち少なくとも１００、２００、３００、４００、５００、６００、７００、８００、９００、または１，０００部位を含む。汚染マーカープローブは、ＤＮＡの一本鎖上の汚染マーカー配列またはＤＮＡの両鎖上の汚染マーカー配列を標的とするように設計することができる。メチル化バイサルファイト配列決定の文脈では、プローブは、高メチル化断片、低メチル化断片、またはその両方を標的とするように設計され得る。プローブはまた、汚染マーカー配列の１つ、いくつか、または全てのハプロタイプを標的とするように設計され得る。例えば、２つのＳＮＰ部位からなる複数のＳＮＰ部位（最大４つの潜在的ハプロタイプ）の場合、プローブは、１つ、いくつか、または全てのハプロタイプを標的とし得る。汚染マーカーの潜在的なハプロタイプの全てを標的とするプローブを設計することにより、いずれか１つの特定のハプロタイプへの参照バイアスを回避することができる。いくつかの実施形態において、汚染マーカー用に設計されたプローブは、表２に列挙された複数のＳＮＰ部位用のプローブおよび表４に列挙されたインデル部位用のプローブの

任意の組み合わせからなる。汚染マーカーの選択およびその後のプローブ設計については、図２Ａおよび図２Ｂで後述する。配列決定の結果、解析システムはサンプル中のｃｆＤＮＡ断片の配列リードを有する。

【０１５１】

解析システムは、試料がホモ接合性ハプロタイプを有する複数の汚染マーカーから１２０個の１つ以上の汚染マーカーを同定する。各汚染マーカーにおけるサンプルのハプロタイプを決定するために、解析システムは、その汚染マーカーにおけるｃｆＤＮＡ断片の全配列リードのハプロタイプを評価する。配列リードのハプロタイプは、汚染マーカー部位に対して設計された全てのプローブにリードを再整列させ、最も良好な整列を有するプローブに対応する対立遺伝子を特定することにより、より正確に決定することができる。解析システムが十分な数のリード、例えば少なくとも１０、２０、３０、４０、５０を観察し、１つのハプロタイプを有する配列リードの割合が非常に高い、例えば９０％、９１％、９２％、９３％、９４％、９５％、９６％、９７％、９８％、または９９％を超える場合、解析システムは、サンプルがそのハプロタイプについてホモ接合体である（すなわち、サンプル中の汚染マーカーのコピーが両方とも同じハプロタイプである）と判定する。高いパーセンテージを超えない汚染マーカーは、ホモ接合体ではないと判定され、むしろヘテロ接合体（すなわち、サンプル中の汚染マーカーの２つのコピーは、異なるハプロタイプである）と判定される場合がある。

【０１５２】

解析システムは、同定された汚染マーカーの１つにおいて、それぞれの汚染マーカーのホモ接合体ハプロタイプとは異なるハプロタイプを有するあらゆるｃｆＤＮＡ断片を汚染断片として同定する。同定された汚染マーカーごとに、解析システムは、同定された汚染マーカーにおいてサンプルのホモ接合性ハプロタイプとは異なるハプロタイプを有する場合、ｃｆＤＮＡ断片を汚染（「汚染断片（複数可）（contaminated fragment(s)）」とも呼ばれる）として標識またはその他の方法で同定する（ステップ１３０）。利用される複数の汚染マーカーのうち、任意のサンプルは、最初の複数のサブセット（部分集合）についてホモ接合のハプロタイプを有する可能性が高い。

【０１５３】

解析システムは、特定された汚染断片に基づいて、汚染レベルを推定する（ステップ１４０）。例えば、同定された汚染断片がない場合、汚染レベルは０または０％となる。解析システムは、サンプルのシーケンシング深度、サンプル中のｃｆＤＮＡ断片の総数、実施された汚染マーカーの数、同定された汚染断片の総数、またはそれらの組み合わせに基づいて、汚染レベルをさらに推定することができる。いくつかの例では、解析システムは、ある信頼区間内の汚染を推定するために、異なるハプロタイプを有するｃｆＤＮＡ断片をカウントする。

【０１５４】

解析システムは、汚染レベルを閾値と比較することにより、サンプルが汚染されているか否かを判断する（ステップ１５０）。いくつかの例では、解析システムは推定汚染レベルを閾値汚染レベル、限界値、または間隔と比較する。閾値汚染レベルまたは間隔は、用途に応じて調整または変化させることができる。例えば、機密性の高いアプリケーションでは、非常に低い汚染閾値が必要となる場合がある。単なる例として、閾値汚染レベルは０．１～１．０％または０．０１～１．０％の間の値または間隔とすることができる。いくつかの例では、閾値汚染限界は０．１％である。いくつかの例では、閾値汚染限界は０．０１％である。一部の実施例では、サンプルの推定汚染レベルが閾値汚染限界値を超える場合、解析システムは、がん分類結果の生成を拒否し（例えば、サンプルをがんまたは非がんと呼ぶことを拒否する）、分類器を訓練するための訓練セットにサンプルを含めることを拒否し（例えば、がん／非がんまたはがんの様々な起源組織を呼ぶための二値またはマルチクラス分類器）、および／または汚染サンプルに基づく疾患または非疾患分類結果を防止することができる。追加的および／または代替的に、解析システムは比較を使用して、サンプルが汚染されているか否かを判定する。例えば、サンプルが閾値未満である

場合、解析システムはサンプルが汚染されていないと判定することができる。逆に、試料が閾値以上である場合、解析システムは試料が汚染されていると判定することができる。

【０１５５】

図２Ａは、１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための複数のＳＮＰ部位を同定するプロセス２００を説明する例示的なフローチャートである。プロセス２００は、本明細書では、解析システムによって実行されるものとして記載されるが、ステップの一部または全部は、他の比較可能に記載された配列決定デバイスおよび／またはコンピュータプロセッサによって実行されてもよい。

【０１５６】

解析システムは、閾値距離内にある複数のＳＮＰ部位を特定する（ステップ２０５）。複数のＳＮＰ部位は、少なくとも２つのＳＮＰ部位を含む遺伝子配列である。１つ以上の実施形態において、複数のＳＮＰ部位は、２、３、４、または５つのＳＮＰ部位を含む。解析システムは、閾値距離、例えば、５塩基対（ｂｐ）、１０ｂｐ、１５ｂｐ、２０ｂｐ、または２５ｂｐを設定する。ＳＮＰ部位が近ければ近いほど、単一断片上に存在する確率が高くなるが、閾値距離が小さいほど、実行可能な複数のＳＮＰ部位の数も制限される。解析システムは、上記の考慮事項および／またはターゲットアッセイパネルに含めることができる汚染マーカープローブの予算を考慮して、閾値距離を調整することができる。

【０１５７】

解析システムは、閾値範囲内の集団ハプロタイプ頻度を有するハプロタイプを有する複数のＳＮＰ部位を含む（ステップ２１０）。２－ＳＮＰ部位では、それぞれ２つのバリアント（変異型）対立遺伝子を有する２つのＳＮＰが存在する。つまり、両方のＳＮＰ部位が置換を持たない［０、０］の第１のハプロタイプ、両方のＳＮＰ部位が置換を持つ［１、１］の第２のハプロタイプ、第２のＳＮＰ部位だけが置換を持つ［０、１］の第３のハプロタイプ、および第１のＳＮＰ部位だけが置換を持つ［１、０］の第４のハプロタイプである。解析システムは、ステップ２１０に含まれる複数のＳＮＰ部位から、集団ハプロタイプ頻度が約５０％の閾値範囲内にある２つのハプロタイプ（２－ＳＮＰ部位における４つの潜在的ハプロタイプのうちの）を有するものを特定する。集団ハプロタイプ頻度は、ゲノムデータベースから得るか、集団を代表するサンプルセット（集合）を用いて決定することができる。閾値範囲は、５０％の±１％、±２％、±３％、±４％、±５％、±６％、±７％、±８％、±９％、または±１０％とすることができる。例えば、解析システムは、ハプロタイプ［０、０］および［１、１］が閾値範囲内の実質的な集団ハプロタイプ頻度を有する一方、［１、０］および［０、１］が小さな集団ハプロタイプ頻度を有する２－ＳＮＰ部位を含む。別の例では、解析システムは、ハプロタイプ［１、０］および［０、１］が閾値範囲内の実質的な集団ハプロタイプ頻度を有する一方、［０、０］および［１、１］が小さな集団ハプロタイプ頻度を有する２－ＳＮＰ部位を含む。

【０１５８】

解析システムは、グアニン－アデニン多型およびシトシン－チミン多型を有する複数のＳＮＰ部位を除外する（ステップ２１５）。解析システムは、例えばメチル化配列決定のためのバイサルファイト配列決定での問題を回避するために、このような複数のＳＮＰ部位を除外する。他のタイプの配列決定に依存する実施形態では、解析システムはステップ２１５をスキップすることができる。

【０１５９】

解析システムは、各複数ＳＮＰ部位におけるハプロタイプ、がハーディーワインベルク平衡にあることを保証する（ステップ２２０）。複数のＳＮＰ部位の各ＳＮＰ部位について、解析システムはＳＮＰ部位の各対立遺伝子の対立遺伝子頻度を計算する。対立遺伝子頻度は、遺伝子データベースから得られたデータ、または集団を代表するサンプルセットを用いて計算することができる。次に、解析システムは、各ＳＮＰ部位の対立遺伝子頻度がハーディーワインベルグ平衡にあるかどうかをチェックする。

【数１】

$$p^2 + 2pq + q^2 = 1$$

ここで、ｐはＳＮＰ部位の第１の対立遺伝子の対立遺伝子頻度であり、ｑはＳＮＰ部位の第２の対立遺伝子の対立遺伝子頻度である。複数のＳＮＰ部位の各ＳＮＰ部位がハーディーワインベルグ平衡にあるかどうかをチェックすることにより、進化または変異している可能性のあるＳＮＰ部位、すなわち平衡にないＳＮＰ部位から生じるアーチファクトを確実に防止することができる。解析システムは、ハーディーワインベルグ平衡の計算にある程度の許容誤差を組み込むことができる。

【０１６０】

解析システムは、ステップ２０５、２１０、２１５、および２２０のうちの１つ以上を省略することができる。前述のように、バイサルファイト配列決定を行わない実施形態では、ステップ２１５を省略することができる。他の実施形態では、ステップ２１０および／またはステップ２２０が省略され得る。

【０１６１】

この時点で、解析システムは、ステップ２０５、２１０、２１５、および２２０の１つ、いくつか、または全ての後に、複数のＳＮＰ部位を汚染マーカーとして使用可能であると同定した。解析システムは、配列決定パネルに実装され得る汚染マーカーの予算に基づいて、実行可能な複数のＳＮＰ部位の数をさらに切り詰めることができる。解析システムは、ゲノム全体にわたる多重ＳＮＰ部位汚染マーカーの分布を最適化することができる。解析システムはまた、ステップ２０５、２１０、２１５、および２２０における様々なパラメータを調整して、汚染マーカーとして選択される複数のＳＮＰ部位の数を最適化することができる。例えば、解析システムは、ステップ２０５において閾値距離を増加させ、ステップ２１０、２１５、および２２０において考慮され得る複数のＳＮＰ部位の数を増加させることができる。別の例として、解析システムは、ステップ２１０において閾値範囲を減少させて、ステップ２１５および２２０において考慮され得る複数のＳＮＰ部位の数を減少させることができる。表１は、例示的な実施態様に従って、汚染マーカーとして使用するために選択された複数のＳＮＰ部位のリストを含む。１つ以上の実施態様において、複数のＳＮＰ部位は、基準のリストに従ってランク付けされ得る。基準には、（１）ＳＮＰ位置を取り囲む配列の複雑さ（ｋ－ｍｅｒエントロピー）、（２）設計されたプローブとゲノム内の他の領域との類似性、（３）理想値０．５からの集団ハプロタイプ頻度の偏差、（４）実際の配列決定されたサンプルで観察される、その部位での読み取り重複率、などが含まれ得る。

【０１６２】

解析システムは、複数のＳＮＰ部位汚染マーカーの各ハプロタイプを標的とする汚染マーカープローブを設計する（ステップ２２５）。各複数のＳＮＰ部位汚染マーカーについて、ステップ２１０でどの２つのハプロタイプが考慮されるかに応じて、解析システムは、２つのハプロタイプの各々を標的とするプローブを設計する。解析システムは、各ハプロタイプの両方のＤＮＡ鎖を標的とするプローブを設計することもできる。汚染マーカーの各ハプロタイプを標的とするプローブを設計することにより、配列決定における参照または代替バイアスを回避することができる。別の実施形態では、解析システムは、各複数ＳＮＰ部位汚染マーカーの参照配列を標的とする単一のプローブを設計する。表２は、例示的な実施態様に従って、表１の複数のＳＮＰ部位汚染マーカーに対して設計されたプローブのリストを含む。

【０１６３】

図２Ｂは、１つ以上の実施形態による、汚染検出における汚染マーカーとして使用するための、インデル部位を同定するステップ２３０を説明する例示的なフローチャートであ

る。ステップ２３０は、本明細書では、解析システムによって実行されるものとして記載されるが、ステップの一部または全部は、他の比較可能に記載された配列決定デバイスおよび／またはコンピュータプロセッサによって実行されてもよい。

【０１６４】

解析システムは、長さの範囲内でインデル部位を同定する（ステップ２３５）。インデル部位とは、個体のゲノムから挿入または欠失された遺伝子配列のことである。長さの範囲は、例えば、５～１０ｂｐ、５～１５ｂｐ、５～２０ｂｐ、５～２５ｂｐ、５～５０ｂｐ、５～１００ｂｐ、１０～１５ｂｐ、１０～２０ｂｐ、１０～２５ｂｐ、１０～５０ｂｐ、１０～１００ｂｐ、１５～２０ｂｐ、１５～２５ｂｐ、１５～５０ｂｐ、または１５～１００ｂｐであり得る。

【０１６５】

解析システムには、複雑性の高いインデル部位が含まれる（ステップ２４０）。複雑度の低いインデル部位には、ホモポリマーおよび単純なタンデムリピートが含まれる。ホモポリマーは、１つの反復ヌクレオチドの文字列であり、例えば、ＡＣＧＴＴＴＴＴＴＴＴＴＴＴＴＴＴＴＡＣＧは、１５個のチミンのホモポリマーを含む。ホモポリマーについては、低複雑性とみなされるための閾値反復数が存在する可能性があり、例えば、５回以上の反復は低複雑性とみなされる。単純タンデムリピートは、繰り返されるヌクレオチドタンデムの文字列であり、例えば、ＡＣＧＴＣＡＴＣＡＴＣＡＴＣＡＴＣＡＴＣＡＴＣＡＴＡＣＧＴは、ヌクレオチドタンデムＣＡＴの７つの繰り返される実例を含む。単純なタンデム反復の場合、低複雑性とみなされる反復の閾値数が存在する可能性があり、例えば、低複雑性とみなされる反復は３回以上である。高複雑性配列には、ホモポリマーまたは単純タンデム反復を含まない、より長い配列が含まれる。例えば、高複雑性配列コンテクストは、特に長いインデルまたは特に特異的な長い挿入、例えば、ＡＣＧＴＡＣＣＧＧＧＴＴＴＴＣＡ（ここで、「ＡＣＣＧＧＧＴＴＴＴ」は挿入配列である）を含み得る。高複雑度インデルの配列は、例えばポリメラーゼ連鎖反応（ＰＣＲ）または配列決定などのサンプル処理によってもたらされるエラーとは対照的に、汚染された断片に対するスクリーニングを確実にする。

【０１６６】

解析システムは、閾値範囲内の集団対立遺伝子頻度を有するインデル部位を含む（ステップ２４５）。集団対立遺伝子頻度は、ゲノムデータベースから得るか、集団を代表するサンプルセットを用いて決定することができる。閾値範囲は、５０％の±１％、±２％、±３％、±４％、±５％、±６％、±７％、±８％、±９％、または±１０％とすることができる。挿入配列の場合、第１の対立遺伝子は挿入配列を含まないが、第２の対立遺伝子は挿入配列を含む。欠失配列の場合、第１の対立遺伝子は欠失配列を含むが、第２の対立遺伝子は欠失配列を含まない。

【０１６７】

解析システムは、各インデル部位の対立遺伝子がハーディーワインベルク平衡にあることを保証する（ステップ２５０）。各インデル部位について、解析システムはインデル部位の各対立遺伝子の対立遺伝子頻度を計算する。対立遺伝子頻度は、遺伝子データベースから得られたデータ、または母集団を代表するサンプルセットを用いて計算することができる。次に、解析システムは、各インデル部位の対立遺伝子頻度がハーディーワインベルク平衡にあるかどうかをチェックする。

【数２】

$$p^2 + 2pq + q^2 = 1$$

ここで、ｐはインデル部位の第１の対立遺伝子の対立遺伝子頻度であり、ｑはインデル部位の第２の対立遺伝子の対立遺伝子頻度である。各インデル部位がハーディーワインベルグ平衡にあるかどうかをチェックすることにより、進化または変異している可能性のあるインデル部位、すなわち平衡にないインデル部位から生じるアーチファクトを確実に防ぐことができる。解析システムはハーディーワインベルグ平衡の計算にある程度の許容誤差を組み込むことができる。

【０１６８】

解析システムは、ステップ２３５、２４０、２４５、および２５０の１つ以上を省略することができる。

【０１６９】

この時点で、解析システムは、ステップ２３５、２４０、２４５、および２５０の１つ、いくつか、または全ての後に、汚染マーカーとして使用可能なインデル部位を同定している。解析システムは、配列決定パネルに実装され得る汚染マーカーの予算に基づいて、実行可能なインデル部位の数をさらに削減することができる。解析システムは、ゲノム全体にわたるインデル部位汚染マーカーの分布を最適化することができる。解析システムはまた、ステップ２３５、２４０、２４５、および２５０における様々なパラメータを調整して、汚染マーカーとして選択されるインデル部位の数を最適化することができる。例えば、解析システムはステップ２３５で長さの範囲を広げ、ステップ２４０、２４５、２５０で考慮される可能性のあるインデル部位の数を増やすことができる。表３は、例示的な実施態様に従って、汚染マーカーとして使用するために選択されたインデル部位のリストを含む。１つ以上の実施態様において、インデル部位は１つ以上の基準に従ってランク付けすることができる。基準には、（１）設計されたプローブとゲノムの他の領域との類似性、（２）集団ハプロタイプ頻度の理想値０．５からの偏差、（３）実際の配列決定されたサンプルで観察される、その部位での読み取り重複率、他の基準、またはそれらの組み合わせが含まれる。インデル部位は、ランキングに従って、またインデル部位の予算に従って選択することができる。

【０１７０】

解析システムは、インデル部位汚染マーカーの各対立遺伝子を標的とする２５５個の汚染マーカープローブを設計する。解析システムは、２つの対立遺伝子のそれぞれを標的とするプローブを設計する。解析システムはまた、各対立遺伝子の両方のＤＮＡ鎖を標的とするプローブを設計してもよい。汚染マーカーの各対立遺伝子を標的とするプローブを設計することにより、配列決定における参照または代替バイアスを回避することができる。別の実施形態では、解析システムは、各インデル部位汚染マーカーの参照配列を標的とする単一のプローブを設計する。表４は、例示的な実施態様に従って、表３のインデル部位汚染マーカー用に設計されたプローブのリストを含む。

【０１７１】

＜ＩＩ．Ｂ．ＤＮＡ断片のメチル化状態ベクターの生成＞

図３Ａは、１つ以上の実施形態による、メチル化状態ベクターを得るためにｃｆＤＮＡの断片を配列決定するプロセス３００を説明する例示的なフローチャートである。ＤＮＡメチル化を解析するために、解析システムはまず、複数のｃｆＤＮＡ分子からなる個人からのサンプルを取得する（ステップ３１０）。さらなる実施形態において、プロセス３００は、他のタイプのＤＮＡ分子の配列決定にも適用され得る。

【０１７２】

サンプルから、解析システムは各ｃｆＤＮＡ分子を単離することができる。ｃｆＤＮＡ分子は、非メチル化シトシンをウラシルに変換するために処理することができる。一実施形態では、この方法は、メチル化されたシトシンを変換することなく、メチル化されていないシトシンをウラシルに変換するＤＮＡの重亜硫酸塩処理を使用する。例えば、ＥＺ ＤＮＡ Ｍｅｔｈｙｌａｔｉｏｎ（商標）－Ｇｏｌｄ、ＥＺ ＤＮＡ Ｍｅｔｈｙｌａｔｉｏｎ（商標）－ＤｉｒｅｃｔまたはＥＺ ＤＮＡ Ｍｅｔｈｙｌａｔｉｏｎ（商標）－Ｌｉｇｈｔｎｉｎｇキット（Ｚｙｍｏ Ｒｅｓｅａｒｃｈ Ｃｏｒｐから入手可能、カリ

フォルニア州アーバイン）などの市販のキットを、重亜硫酸塩変換に使用する。別の実施形態では、非メチル化シトシンのウラシルへの変換は、酵素反応を用いて達成される。例えば、ＡＰＯＢＥＣ－Ｓｅｑ（ＮＥＢｉｏｌａｂｓ、マサチューセッツ州イプスウィッチ）のような、非メチル化シトシンのウラシルへの変換のための市販のキットを使用することができる。

【０１７３】

変換されたｃｆＤＮＡ分子から、配列決定ライブラリーを調製することができる（ステップ３３０）。ライブラリー調製中、アダプターライゲーションにより、核酸分子（例えば、ＤＮＡ分子）に一意的分子識別子（ＵＭＩ）を付加することができる。ＵＭＩは、アダプターライゲーション中にＤＮＡ断片（例えば、物理的剪断、酵素消化、および／または化学的断片化によって断片化されたＤＮＡ分子）の末端に付加される短い核酸配列（例えば、４～１０塩基対）であり得る。ＵＭＩは、特定のＤＮＡ断片に由来する配列リードを同定するために使用できる独特なタグとして機能する縮退塩基対であり得る。アダプターライゲーション後のＰＣＲ増幅において、ＵＭＩは結合したＤＮＡ断片と共に複製される。これにより、下流の解析において、同じオリジナル断片に由来する配列リードを同定することができる。

【０１７４】

任意に、配列決定ライブラリーは、複数のハイブリダイゼーションプローブを用いて、がんの状態について有益なｃｆＤＮＡ分子またはゲノム領域について濃縮され得る。ハイブリダイゼーションプローブは、特に特定されたｃｆＤＮＡ分子または標的領域にハイブリダイズすることができる短いオリゴヌクレオチドであり、その後の配列決定および解析のために、それらの断片または領域を濃縮する。ハイブリダイゼーションプローブは、研究者の関心のある特定のＣｐＧ部位のセットをターゲットとした高深度解析を行うために使用することができる。ハイブリダイゼーションプローブは、１倍、２倍、３倍、４倍、５倍、６倍、７倍、８倍、９倍、１０倍、または１０倍以上のカバレッジであり、１つ以上の標的配列にわたってタイリングすることができる。例えば、２倍のカバレッジでタイリングされたハイブリダイゼーションプローブは、標的配列の各部分が２つの独立したプローブにハイブリダイズするようなオーバーラッププローブからなる。ハイブリダイゼーションプローブは、１倍未満のカバレッジで、１つ以上の標的配列にわたってタイリングすることができる。

【０１７５】

一実施形態では、ハイブリダイゼーションプローブは、非メチル化シトシンをウラシルに変換するために（例えば、重亜硫酸塩を使用して）処理されたＤＮＡ分子を濃縮するように設計される。濃縮中、ハイブリダイゼーションプローブ（本明細書では「プローブ（probes）」とも呼ぶ）は、がん（または疾患）の有無、がんの状態、またはがんの分類（例えば、がんクラスまたは起源組織）について情報提供する核酸断片を標的としてプルダウンするために使用することができる。プローブは、ＤＮＡの標的鎖（相補的）にアニール（またはハイブリダイズ）するように設計することができる。標的鎖は、「ポジティブ（positive）」鎖（例えば、ｍＲＮＡに転写され、その後タンパク質に翻訳される鎖）または相補的な「ネガティブ（negative）」鎖である。プローブの長さは１０塩基対、１００塩基対、１０００塩基対のいずれでもよい。プローブは、メチル化部位パネルに基づいて設計することができる。プローブは、特定の変異または、特定のがんまたは他のタイプの疾患に対応すると疑われるゲノム（例えば、ヒトや他の生物）の標的領域を解析するために、標的遺伝子パネルに基づいて設計することができる。さらに、プローブは標的領域の重複部分をカバーすることもできる。

【０１７６】

一旦調製されると、配列決定ライブラリーまたはその一部は、複数の配列リードを得るために配列決定され得る。配列リードは、コンピュータソフトウェアによる処理および解釈のために、コンピュータ読み取り可能なデジタル形式であってもよい。配列リードは、アライメント位置情報を決定するために参照ゲノムにアライメントされ得る。アライメン

ト位置情報は、所定の配列リードの開始ヌクレオチド塩基および終了ヌクレオチド塩基に対応する参照ゲノム中の領域の開始位置および終了位置を示す場合がある。アライメント位置情報は、開始位置および終了位置から決定され得る配列リード長も含み得る。参照ゲノム中の領域は、遺伝子または遺伝子のセグメントに関連する場合がある。配列リードは、R_1およびR_2と表記されるリードペアで構成される。例えば、第１リードR_1は核酸断片の第１末端から配列決定され、第２リードR_2は核酸断片の第２末端から配列決定される。したがって、第１のリードR_1および第２のリードR_2のヌクレオチド塩基対は、参照ゲノムのヌクレオチド塩基と一貫して（例えば、反対の向きで）整列される場合がある。リードペアR_1およびR_2から得られるアライメント位置情報には、第１のリード（例えば、R_1）の末端に対応する参照ゲノム内の開始位置および、第２のリード（例えば、R_2）の末端に対応する参照ゲノム内の終了位置が含まれる場合がある。言い換えれば、参照ゲノムにおける開始位置と終了位置は、核酸断片が対応する参照ゲノム内の可能性の高い位置を表すことができる。ＳＡＭ（配列アライメントマップ）形式またはＢＡＭ（二値）形式の出力ファイルが生成され、メチル化状態の決定などのさらなる解析のために出力される。

【０１７７】

解析システムは、配列リードから、参照ゲノムとのアライメントに基づいて、各ＣｐＧ部位の位置およびメチル化状態を決定する（ステップ３５０）。解析システムは、参照ゲノムにおける断片の位置（例えば、各断片の最初のＣｐＧ部位の位置で指定されたもの、または他の類似の指標）、断片内のＣｐＧ部位の数、および断片内の各ＣｐＧ部位のメチル化（例えば、Ｍと表記）、非メチル化（例えば、Ｕと表記）、または不確定（例えば、Ｉと表記）のいずれかのメチル化状態を指定する、各断片のメチル化状態ベクターを生成する（ステップ３６０）。観察される状態には、メチル化された状態およびメチル化されていない状態があるが、観察されない状態は不確定である。不確定なメチル化状態は、塩基配列の誤りおよび／またはＤＮＡ断片の相補鎖のメチル化状態の不一致に由来する可能性がある。メチル化状態ベクターは、後で使用および処理するために、一時的または永続的なコンピュータメモリに保存することができる。さらに、解析システムは、単一のサンプルから重複するリードまたは重複するメチル化状態ベクターを除去することができる。解析システムは、１つ以上のＣｐＧ部位を有する特定の断片が、閾値数またはパーセンテージを超える不確定メチル化状態を有することを決定してもよく、そのような断片を除外してもよく、またはそのような断片を選択的に含んでもよいが、そのような不確定メチル化状態を考慮したモデルを構築してもよく、そのようなモデルの１つについては、図４と関連して後述する。

【０１７８】

図３Ｂは、１つ以上の実施形態による、メチル化状態ベクターを得るためにｃｆＤＮＡ分子を配列決定する図３Ａのプロセス３００の例示的な図である。一例として、解析システムは、この例では３つのＣｐＧ部位を含むｃｆＤＮＡ分子３１２を受け取る。図示のように、ｃｆＤＮＡ分子３１２の第１および第３のＣｐＧ部位はメチル化３１４されている。処理ステップ３２０中に、ｃｆＤＮＡ分子３１２は変換されて、変換ｃｆＤＮＡ分子３２２を生成する。処理ステップ３２０中に、メチル化されていなかった第２のＣｐＧ部位は、そのシトシンがウラシルに変換される。しかし、第１および第３のＣｐＧ部位は変換されなかった。

【０１７９】

変換後、配列決定ライブラリーが調製され（ステップ３３０）、配列決定が行われ（ステップ３４０）、配列リード３４２が生成される。解析システムは、配列リード３４２を参照ゲノム３４４にアライメントする（ステップ３５０）。参照ゲノム３４４は、断片ｃｆＤＮＡがヒトゲノムのどの位置に由来するかについてのコンテクストを提供する。この単純化された例では、解析システムは、３つのＣｐＧ部位がＣｐＧ部位２３、２４および２５（説明の便宜のために使用される任意の参照識別子）に相関するように、配列リード３４２をアライメントする（ステップ３５０）。このようにして、解析システムは、ｃｆ

ＤＮＡ分子３１２上の全てのＣｐＧ部位のメチル化状態、およびＣｐＧ部位がマッピングされるヒトゲノム内の位置の両方の情報を生成することができる。図示のように、メチル化されている配列リード３４２上のＣｐＧ部位は、シトシンとして読み取られる。この例では、シトシンは配列リード３４２の１番目および３番目のＣｐＧ部位にのみ現れているため、元のｃｆＤＮＡ分子の１番目および３番目のＣｐＧ部位がメチル化されていると推測できる。一方、第２のＣｐＧ部位はチミンとして読み取れる（Ｕは配列決定過程でＴに変換される）ので、第２のＣｐＧ部位は元のｃｆＤＮＡ分子ではメチル化されていないと推測できる。これら２つの情報であるメチル化の状態および位置を用いて、解析システムは断片ｃｆＤＮＡ３１２のメチル化状態ベクター３５２を生成する（ステップ３６０）。この例では、得られたメチル化状態ベクター３５２は＜M_{23}、U_{24}、M_{25}＞であり、Ｍはメチル化ＣｐＧ部位に対応し、Ｕは非メチル化ＣｐＧ部位に対応し、添え字番号は参照ゲノムにおける各ＣｐＧ部位の位置に対応する。

【０１８０】

生物学的サンプル中の核酸から配列リードを得るために、1つ以上の代替的な配列決定方法を使用することができる。１つ以上の配列決定方法は、核酸（例えば、セルフリー核酸）から測定された配列リードの数を得るために使用され得る任意の形態の配列決定を含み得、これには、Ｒｏｃｈｅ　４５４プラットフォーム、Ａｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓ　ＳＯＬＩＤプラットフォーム、Ｈｅｌｉｃｏｓ　Ｔｒｕｅ　Ｓｉｎｇｌｅ　Ｍｏｌｅｃｕｌｅ　ＤＮＡ配列決定技術、Ａｆｆｙｍｅｔｒｉｘ社のシーケンシング・バイ・ハイブリダイゼーション・プラットフォーム、Ｐａｃｉｆｉｃ　Ｂｉｏｓｃｉｅｎｃｅｓの１分子リアルタイム（ＳＭＲＴ）技術、４５４　Ｌｉｆｅ　Ｓｃｉｅｎｃｅｓ、Ｉｌｌｕｍｉｎａ／Ｓｏｌｅｘａ、Ｈｅｌｉｃｏｓ　Ｂｉｏｓｃｉｅｎｃｅｓのシーケンシング・バイ・シンセシス・プラットフォーム、Ａｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓのシーケンシング・バイ・ライゲーション・プラットフォームなどがある。Ｌｉｆｅ　ｔｅｃｈｎｏｌｏｇｉｅｓのＩＯＮ　ＴＯＲＲＥＮＴテクノロジーおよびＮａｎｏｐｏｒｅシーケンシングも、生物学的サンプル中の核酸（例えば、セルフリー核酸）から配列リードを得るために使用できる。シーケンシング・バイ・シンセシスおよび可逆ターミネーターベースの配列（例えば、ＩｌｌｕｍｉｎａのＧｅｎｏｍｅ　Ａｎａｌｙｚｅｒ、Ｇｅｎｏｍｅ　Ａｎａｌｙｚｅｒ　ＩＩ、ＨＩＳＥＱ　２０００、ＨＩＳＥＱ　４５００（Ｉｌｌｕｍｉｎａ、カリフォルニア州サンディエゴ））を使用して、遺伝子型データセットを形成するために、訓練被験者の生物学的サンプルから得られたセルフリー核酸から配列リードを得ることができる。何百万ものセルフリー核酸（例えば、ＤＮＡ）断片を並行して配列決定することができる。この種の配列決定技術の一例では、光学的に透明なスライドを含むフローセルが使用され、その表面には８つの個別レーンがあり、オリゴヌクレオチドアンカー（例えば、アダプタープライマー）が結合している。セルフリー核酸サンプルは、検出を容易にするシグナルまたはタグを含むことができる。生物学的サンプルから得られたセルフリー核酸からの配列リードの取得は、例えば、フローサイトメトリー、定量的ポリメラーゼ連鎖反応（ｑＰＣＲ）、ゲル電気泳動、遺伝子チップ分析、マイクロアレイ、質量分析、サイトフルオリメトリー分析、蛍光顕微鏡、共焦点レーザー走査顕微鏡、レーザー走査型サイトメトリー、アフィニティークロマトグラフィー、手動バッチモード分離、電場懸濁、配列決定、およびこれらの組み合わせなどの様々な技術を用いて、シグナルまたはタグの定量情報を得ることを含むことができる。

【０１８１】

１つ以上の配列決定（シーケンシング）方法は、全ゲノム配列決定アッセイを含み得る。全ゲノム配列決定アッセイは、コピー数変異またはコピー数異常のような大きな変異を決定するために使用され得る全ゲノムまたは全ゲノムの実質的部分についての配列リードを生成する物理的アッセイを含み得る。このような物理的アッセイは、全ゲノム配列決定技術または全エクソーム配列決定技術を用いることができる。全ゲノム配列決定アッセイは、被検者のゲノム全体にわたって、少なくとも１倍、２倍、３倍、４倍、５倍、６倍、７倍、８倍、９倍、１０倍、少なくとも２０倍、少なくとも３０倍、または少なくとも４

０倍の平均シーケンシング深度を有し得る。いくつかの実施形態では、シーケンシング深度は約３０，０００倍である。１つ以上の配列決定方法は、標的化パネル配列決定アッセイを含み得る。標的化パネル配列決定アッセイは、遺伝子の標的化パネルについて、少なくとも５０，０００倍、少なくとも５５，０００倍、少なくとも６０，０００倍、または少なくとも７０，０００倍の平均シーケンシング深度を有し得る。遺伝子の標的化パネルは、４５０～５００の遺伝子を含み得る。遺伝子の標的化パネルは、５００±５遺伝子の範囲、５００±１０遺伝子の範囲、または５００±２５遺伝子の範囲を含み得る。

【０１８２】

１つ以上の配列決定方法は、ペアエンドシーケンシングを含み得る。１つ以上の配列決定方法は、複数の配列リードを生成し得る。複数の配列リードは、１０～７００、５０～４００、または１００～３００の範囲の平均長を有し得る。１つ以上の配列決定方法は、メチル化配列決定アッセイを含み得る。メチル化配列決定は、ｉ）全ゲノムメチル化配列決定、または、ｉｉ）複数の核酸プローブを使用する標的ＤＮＡメチル化配列決定であり得る。例えば、メチル化配列決定は、全ゲノムバイサルファイト配列決定（例えば、ＷＧＢＳ）である。メチル化配列決定は、メチロームの最も情報量の多い領域を標的とする複数の核酸プローブ、独自のメチル化データベース、および先行するプロトタイプの全ゲノムおよび標的配列決定アッセイを用いる標的ＤＮＡメチル化配列決定であり得る。

【０１８３】

メチル化配列決定は、それぞれの核酸メチル化断片中の１つ以上の５－メチルシトシン（５ｍＣ）および／または５－ヒドロキシメチルシトシン（５ｈｍＣ）を検出し得る。メチル化配列決定は、それぞれの核酸メチル化断片中の、１つ以上の非メチル化シトシンまたは１つ以上のメチル化シトシンの、対応する１つ以上のウラシルへの変換を含み得る。１つ以上のウラシルは、メチル化配列決定中に、１つ以上の対応するチミンとして検出され得る。１つ以上の非メチル化シトシンまたは１つ以上のメチル化シトシンの変換は、化学的変換、酵素的変換、またはそれらの組み合わせからなり得る。

【０１８４】

例えば、バイサルファイト変換は、メチル化されたシトシン（例えば、５－メチルシトシンまたは５－ｍＣ）をそのまま残して、シトシンをウラシルに変換することを含む。ＤＮＡによっては、約９５％のシトシンがメチル化されていない場合があり、得られるＤＮＡ断片には、チミンで表される多くのウラシルが含まれる場合がある。配列決定の前に核酸を処理するために酵素的変換プロセスを使用することができ、これは様々な方法で行うことができる。重亜硫酸塩を含まない変換の一例は、非修飾シトシンに影響を与えることなく５－メチルシトシンおよび５－ヒドロキシメチルシトシンを非破壊的かつ直接検出するための、重亜硫酸塩を含まない塩基分解能配列決定法、ＴＥＴ－アシステッドピリジンボラン配列決定（ＴＡＰＳ）からなる。それぞれの核酸メチル化断片中の対応する複数のＣｐＧ部位におけるＣｐＧ部位のメチル化状態は、メチル化配列決定によってＣｐＧ部位がメチル化されていると判定された場合にはメチル化され、メチル化配列決定によってＣｐＧ部位がメチル化されていないと判定された場合にはメチル化されていないことがある。

【０１８５】

メチル化配列決定アッセイ（例えば、ＷＧＢＳおよび／または標的メチル化配列決定）は、約１，０００倍、２，０００倍、３，０００倍、５，０００倍、１０，０００倍、１５，０００倍、２０，０００倍、または３０，０００倍までを含むが、これらに限定されない平均シーケンシング深度を有し得る。メチル化配列決定は、３０，０００倍を超えるシーケンシング深度、例えば、少なくとも４０，０００倍または５０，０００倍を有することができる。全ゲノムバイサルファイト配列決定法は、２０倍～５０倍の平均シーケンシング深度を有し得、標的化メチル化配列決定法は、１００倍～１０００倍の平均有効深度を有し得、ここで、有効深度は、標的化メチル化配列決定によって得られる同じ数の配列リードを得るための等価全ゲノムバイサルファイト配列決定カバレッジであり得る。

【０１８６】

メチル化配列決定（例えば、ＷＧＢＳおよび／または標的化メチル化配列決定）に関するさらなる詳細については、例えば、２０１８年３月１３日に出願された「メチル化断片異常検出（Methylation Fragment Anomaly Detection）」と題する米国特許出願第６２／６４２，４８０号、および２０１９年１２月１８日に出願された「メチル化情報の使用により細胞源分画を推定するシステムおよび方法（Systems and Methods for Estimating Cell Source Fractions Using Methylation Information）」と題する米国特許出願第１６／７１９，９０２号を参照されたく、これらの各々は参照により本明細書に組み込まれる。断片メチル化パターンを得るために、本明細書に開示されたもの、および／またはその任意の改変、置換、または組み合わせを含む、メチル化配列決定のための他の方法を使用することができる。メチル化配列決定は、例えば、２０１９年３月１３日に出願された「異常断片の検出および分類（Anomalous Fragment Detection and Classification）」と題する米国特許出願第１６／３５２，６０２号に記載されるように、または２０２０年５月１３日に出願された「モデルベースの特徴化および分類（Model-Based Featurization and Classification）」と題する米国特許出願第１５／９３１，０２２号に開示される技術のいずれかに従って、１つ以上のメチル化状態ベクターを同定するために使用され得、これらの各々は、参照により本明細書に組み込まれる。

【０１８７】

核酸のメチル化配列決定および得られた１つ以上のメチル化状態ベクターを使用して、複数の核酸メチル化断片を得ることができる。それぞれの対応する複数の核酸メチル化断片（例えば、それぞれの遺伝子型データセットについて）は、１００個以上の核酸メチル化断片から構成され得る。各対応する複数の核酸メチル化断片にわたる核酸メチル化断片の平均数は、１０００個以上の核酸メチル化断片、５０００個以上の核酸メチル化断片、１０，０００個以上の核酸メチル化断片、２０，０００個以上の核酸メチル化断片、または３０，０００個以上の核酸メチル化断片からなり得る。各対応する複数の核酸メチル化断片にわたる核酸メチル化断片の平均数は、１０，０００個～５０，０００個の核酸メチル化断片であり得る。対応する複数の核酸メチル化断片は、１０００個以上、１万個以上、１０万個以上、１００万個以上、１０００万個以上、１億個以上、５億個以上、１０億個以上、２０億個以上、３０億個以上、４０億個以上、５０億個以上、６０億個以上、７０億個以上、８０億個以上、９０億個以上、または１００億個以上の核酸メチル化断片からなり得る。対応する複数の核酸メチル化断片の平均長さは、１４０～４８０ヌクレオチドであり得る。

【０１８８】

核酸を配列決定する方法およびメチル化配列決定データに関するさらなる詳細は、２０２０年３月４日に出願された「オートエンコーダを用いたがん状態判定システムおよび方法（Systems and Methods for Cancer Condition Determination Using Autoencoders）」と題する米国仮特許出願第６２／９８５，２５８号に開示されており、その全体が参照により本明細書に組み込まれる。

【０１８９】

＜ＩＩ．Ｃ．異常断片の特定＞

解析システムは、サンプルのメチル化状態ベクターを使用して、サンプルの異常断片を決定することができる。解析システムは、サンプル中の各断片について、その断片に対応するメチル化状態ベクターを用いて、その断片が異常断片であるかどうかを判定することができる。いくつかの実施形態では、解析システムは、各メチル化状態ベクターについて、そのメチル化状態ベクターまたは健常対照群においてさらに確率の低い他のメチル化状態ベクターを観察する確率を記述するｐ値スコアを計算する。ｐ値スコアを計算するプロセスについては、セクションＩＩ．Ｃ．ｉ．のＰ値フィルタリングでさらに後述する。解析システムは、閾値ｐ値スコアを下回るメチル化状態ベクターを有する断片を異常断片として決定することができる。一部の実施形態では、解析システムはさらに、メチル化または非メチル化の割合がある閾値を超える少なくともいくつかのＣｐＧ部位を有する断片を

、それぞれ高メチル化および低メチル化断片として標識する。高メチル化断片または低メチル化断片は、極端なメチル化を伴う異常な断片（ＵＦＸＭ）とも呼ばれることがある。他の実施形態では、解析システムは、異常断片を決定するための様々な他の確率モデルを実装することができる。他の確率モデルの例には、混合モデル、深層確率モデルなどが含まれる。いくつかの実施形態では、解析システムは、異常な断片を識別するために、以下に説明するプロセスの任意の組み合わせを使用することができる。同定された異常断片を用いて、解析システムは、他の処理、例えば、がん分類器の訓練および展開における使用のために、サンプルのメチル化状態ベクターのセットをフィルタリングすることができる。

【０１９０】

＜ＩＩ．Ｃ．Ｉ．Ｐ値フィルタリング＞

いくつかの実施形態では、解析システムは、各メチル化状態ベクターについて、健常対照群の断片からのメチル化状態ベクターと比較したｐ値スコアを計算する。ｐ値スコアは、そのメチル化状態ベクターに一致するメチル化状態を観察する確率、または健常対照群においてさらに確率の低い他のメチル化状態ベクターを観察する確率を表すことができる。ＤＮＡ断片が異常にメチル化されていると判定するために、解析システムは、正常にメチル化されている断片が大多数を占める健常対照群を使用することができる。異常な断片を判定するためにこの確率的解析を行う場合、健常対照群を構成する対照被験者群と比較して、判定に重みを持たせることができる。健常対照群における頑健性を確保するために、解析システムは、ＤＮＡ断片を含むサンプルを供給するために、ある閾値数の健常個体を選択することができる。以下の図４Ａは、解析システムがｐ値スコアを計算し得る健常対照群のデータ構造を生成する方法を説明する。図４Ｂは、生成されたデータ構造の使用によりｐ値スコアを計算する方法を説明する。

【０１９１】

図４Ａは、実施形態による、健常対照群のデータ構造を生成するプロセス４００を説明するフローチャートである。健常対照群のデータ構造を作成するために、解析システムは、複数の健常個体から複数のＤＮＡ断片（例えば、ｃｆＤＮＡ）を受け取ることができる。メチル化状態ベクターは、例えば、プロセス３００を介して、各断片について同定することができる。

【０１９２】

各断片のメチル化状態ベクターを用いて、解析システムはメチル化状態ベクターをＣｐＧ部位の文字列に細分化することができる。いくつかの実施形態では、解析システムは、得られた文字列が全て所定の長さ未満になるように、メチル化状態ベクターを細分化する（ステップ４０５）。例えば、長さ１１のメチル化状態ベクターを長さ３以下の文字列に細分化すると、長さ３の文字列が９個、長さ２の文字列が１０個、長さ１の文字列が１１個となる。別の例では、長さ７のメチル化状態ベクターを長さ４以下の文字列に細分化すると、長さ４の文字列が４個、長さ３の文字列が５個、長さ２の文字列が６個、長さ１の文字列が７個となる。メチル化状態ベクターが、指定された文字列の長さより短いか、同じ長さである場合、メチル化状態ベクターは、そのベクターの全てのＣｐＧ部位を含む単一の文字列に変換されてもよい。

【０１９３】

解析システムは、ベクター内の可能性のあるＣｐＧ部位およびメチル化状態の可能性ごとに、指定されたＣｐＧ部位を文字列の最初のＣｐＧ部位として有し、そのメチル化状態の可能性を有する対照群に存在する文字列の数をカウントすることによって、文字列を集計する（ステップ４１０）。例えば、あるＣｐＧ部位において、文字列の長さが３であるとすると、２＾３、または８通りの文字列構成があり得る。その与えられたＣｐＧ部位で、８つの可能性のある文字列構成のそれぞれについて、解析システムは、各メチル化状態のベクターの可能性が対照群にいくつ出現するかを集計する（ステップ４１０）。この例を続けると、これは、＜M_x、M_{x+1}、M_{x+2}＞、＜M_x、M_{x+1}、U_{x+2}＞、．．．、＜U_x、U_{x+1}、U_{x+2}＞を参照ゲノム中の各開始ＣｐＧ部位ｘについて計算

する量を集計することを含み得る。解析システムは、各開始ＣｐＧ部位と文字列の可能性について集計されたカウントを格納するデータ構造を作成する（ステップ４１５）。

【０１９４】

文字列の長さに上限を設定することには、いくつかの利点がある。第一に、文字列の最大長によっては、解析システムによって作成されるデータ構造のサイズが劇的に増大する可能性がある。例えば、文字列の最大長が４である場合、長さ４の文字列を集計するために、全てのＣｐＧ部位が少なくとも２＾４の数を持つことを意味する。文字列の最大長を５に増やすと、全てのＣｐＧ部位が集計するためにさらに２＾４または１６の数を持つことを意味し、以前の文字列長に比べて集計する数（および必要なコンピュータ・メモリ）が２倍になる。文字列サイズを小さくすることにより、データ構造の作成と性能（例えば、後述するように後でアクセスするために使用する）を、計算とストレージの点で合理的に保つことができる。第二に、最大文字列長を制限する統計的な考慮点として、文字列カウントを使用する下流モデルのオーバーフィッティングを避けることができる。ＣｐＧ部位の長い文字列が生物学的に結果に強い影響を与えない場合（例えば、がんの存在を予測する異常性の予測）、ＣｐＧ部位の大きな文字列に基づいて確率を計算することは、利用可能でない可能性のあるかなりの量のデータを使用するため問題となる可能性があり、したがって、モデルが適切に実行するには疎らになりすぎる可能性がある。例えば、事前１００個のＣｐＧ部位を条件として異常／がんの確率を計算する場合、長さ１００のデータ構造中の文字列のカウントを使用することができ、理想的には事前１００個のメチル化状態に正確に一致するものがある。長さ１００の文字列が疎らにしかカウントされていない場合、検査サンプルに含まれる長さ１００の文字列が異常か否かを判定するのに十分なデータがない可能性がある。

【０１９５】

図４Ｂは、１つ以上の実施形態による、サンプルから異常にメチル化された断片を同定するためのステップ４２０を説明するフローチャートである。ステップ４２０において、解析システムは、被検者のｃｆＤＮＡ断片からメチル化状態ベクターを生成する（ステップ３００）。解析システムは、各メチル化状態ベクターを以下のように処理することができる。

【０１９６】

所与のメチル化状態ベクターについて、解析システムは、メチル化状態ベクター内の同じ開始ＣｐＧ部位および同じ長さ（すなわち、ＣｐＧ部位のセット）を有するメチル化状態ベクターの全ての可能性を列挙する（ステップ４３０）。各メチル化状態は一般にメチル化または非メチル化のいずれかであるため、各ＣｐＧ部位には事実上２つの可能性があり、したがってメチル化状態ベクターの異なる可能性の数は２のべき乗に依存し、長さｎのメチル化状態ベクターは2^nの可能性のあるメチル化状態ベクターと関連することになる。メチル化状態ベクターが１つ以上のＣｐＧ部位の不確定状態を含む場合、解析システムは、観察された状態を有するＣｐＧ部位のみを考慮して、メチル化状態ベクターの可能性を列挙することができる（ステップ４３０）。

【０１９７】

解析システムは、識別された開始ＣｐＧ部位およびメチル化状態ベクター長について、健常対照群データ構造にアクセスすることにより、メチル化状態ベクターの各可能性を観察する確率を計算する（ステップ４４０）。いくつかの実施形態において、所与の可能性を観察する確率を計算することは、共同確率計算をモデル化するためにマルコフ連鎖確率を使用する。マルコフモデルは、少なくとも部分的に、対応する複数のＣｐＧ部位を有する健康な非がんコホートデータセット中のそれらの核酸メチル化断片にわたる、それぞれの断片（例えば、核酸メチル化断片）の対応する複数のＣｐＧ部位における各ＣｐＧ部位のメチル化状態の評価に基づいて、訓練され得る。例えば、マルコフモデル（例えば、隠れマルコフモデルまたはＨＭＭ）を用いて、複数の核酸メチル化断片中の核酸メチル化断片について、配列中の各状態について、配列中の次の状態を観察する可能性を決定する確率のセットが与えられた場合に、メチル化状態の配列（例えば、「Ｍ」または「Ｕ」から

なる）が観察され得る確率を決定する。確率のセットは、ＨＭＭを訓練することによって得ることができる。このような訓練は、観察されたメチル化状態配列（例えば、メチル化パターン）の初期訓練データセットが与えられた場合に、統計的パラメータ（例えば、第１の状態が第２の状態に遷移し得る確率（遷移確率）および／または所定のメチル化状態がそれぞれのＣｐＧ部位について観察され得る確率（放出確率））を計算することを含み得る。ＨＭＭは、教師あり訓練（例えば、基礎となる配列ならびに観察された状態が既知であるサンプルを使用する）および／または教師なし訓練（例えば、ビタビ学習、最尤推定、期待値最大化訓練、および／またはバウム－ウェルチ訓練）を使用して訓練することができる。他の実施形態において、マルコフ連鎖確率以外の計算方法が、メチル化状態ベクターの各可能性を観察する確率を決定するために使用される。例えば、そのような計算方法は、学習された表現を含み得る。ｐ値閾値は、０．０１～０．１０、または０．０３～０．０６であり得る。ｐ値閾値は０．０５とすることができる。ｐ値の閾値は０．０１未満、０．００１未満、または０．０００１未満とすることができる。

【０１９８】

解析システムは、各可能性について計算された確率を使用して、メチル化状態ベクターのｐ値スコアを計算する（ステップ４５０）。いくつかの実施形態において、これは、問題のメチル化状態ベクターに一致する可能性に対応する計算された確率を特定することを含む。具体的には、これは、メチル化状態ベクターと同じＣｐＧ部位のセット、または同様に同じ開始ＣｐＧ部位および長さを有する可能性であり得る。解析システムは、特定された確率以下の確率を有するあらゆる可能性の計算された確率を合計して、ｐ値スコアを生成することができる。

【０１９９】

このｐ値は、断片のメチル化状態ベクター、または健常対照群でさらに確率の低い他のメチル化状態ベクターを観察する確率を表すことができる。低いｐ値スコアは、一般に、健常対照群に対して、健常個体ではまれであり、断片が異常にメチル化されたラベルを付けられるメチル化状態ベクターに対応し得る。高いｐ値スコアは、一般に、健康な個体において相対的な意味で存在すると予想されるメチル化状態ベクターに関連し得る。例えば、健常対照群ががんでない群である場合、低いｐ値は、がんでない群と比較して断片が異常にメチル化されていることを示し、したがって、被検者におけるがんの存在を示す可能性がある。

【０２００】

上記のように、解析システムは、複数のメチル化状態ベクターのそれぞれについてｐ値スコアを計算することができ、それぞれが検査サンプル中のｃｆＤＮＡ断片を表す。どの断片が異常にメチル化されているかを同定するために、解析システムは、ｐ値スコアに基づいてメチル化状態ベクターのセットをフィルタリングすることができる（ステップ４６０）。いくつかの実施形態では、ｐ値スコアを閾値と比較し、閾値以下の断片のみを残すことによってフィルタリングが実行される。この閾値ｐ値スコアは、０．１、０．０１、０．００１、０．０００１などのオーダーであり得る。

【０２０１】

プロセス４００の結果例によると、解析システムは、訓練においてがんでない参加者について、異常なメチル化パターンを有する断片の中央値（範囲）２，８００（１，５００～１２，０００）を得ることができ、訓練においてがんである参加者について、異常なメチル化パターンを有する断片の中央値（範囲）３，０００（１，２００～４２０，０００）を得ることができる。これらの異常なメチル化パターンを持つ断片のフィルターセットは、セクションＩＩＩで後述するように、下流の解析に使用することができる。

【０２０２】

いくつかの実施形態では、解析システムは、メチル化状態ベクターの可能性を決定し、ｐ値を計算するためにスライディングウィンドウを使用する（ステップ４５５）。可能性を列挙し、メチル化状態ベクター全体についてｐ値を計算するのではなく、解析システムは、可能性を列挙し、連続するＣｐＧ部位のウィンドウのみについてｐ値を計算すること

ができ、ウィンドウは、少なくともいくつかの断片よりも（ＣｐＧ部位の）長さが短い（そうでなければ、ウィンドウは目的を果たさない）。ウィンドウの長さは、静的、使用者による決定、動的、またはその他の方法で選択することができる。

【０２０３】

ウィンドウより大きいメチル化状態ベクターについてｐ値を計算する場合、ウィンドウは、ベクター内の最初のＣｐＧ部位から始まるウィンドウ内のベクターからのＣｐＧ部位の連続セットを識別することができる。解析システムは、最初のＣｐＧ部位を含むウィンドウのｐ値スコアを計算することができる。解析システムは次に、ウィンドウをベクター内の第２のＣｐＧ部位まで「スライド（slide）」させ、第２のウィンドウについて別のｐ値スコアを計算することができる。このように、ウィンドウサイズｌとメチル化ベクター長ｍに対して、各メチル化状態ベクターはｍ－ｌ＋１のｐ値スコアを生成することができる。ベクターの各部分についてｐ値計算を完了した後、全てのスライディングウィンドウから最も低いｐ値スコアを、メチル化状態ベクターの全体的なｐ値スコアとすることができる。他の実施形態では、解析システムは、メチル化状態ベクターのｐ値スコアを集約して、総合ｐ値スコアを生成する。

【０２０４】

スライディングウィンドウを使用することで、メチル化状態ベクターの列挙された可能性および、それに対応する確率計算の数を減らすことができる。現実的な例を挙げると、５４個以上のＣｐＧ部位を持つ断片がある。単一のｐスコアを生成するために２＾５４（～１．８×１０＾１６）の可能性について確率を計算する代わりに、分析システムは、その断片のメチル化状態ベクターの５０のウィンドウのそれぞれについて５０のｐ値計算をもたらすサイズ５のウィンドウ（例えば）を使用することができる。５０の計算のそれぞれは、２＾５（３２）のメチル化状態ベクターの可能性を列挙することができ、合計で５０×２＾５（１．６×１０＾３）の確率計算になる。その結果、実行すべき計算が大幅に削減され、異常断片の正確な同定には意味がなくなる。

【０２０５】

不確定状態を有する実施形態では、解析システムは、断片のメチル化状態ベクター中の不確定状態を有するＣｐＧ部位を合計したｐ値スコアを計算することができる。解析システムは、不確定状態を除くメチル化状態ベクターの全てのメチル化状態とコンセンサスを有する全ての可能性を同定することができる。解析システムは、特定された可能性の確率の合計として、メチル化状態ベクターに確率を割り当てることができる。一例として、解析システムは、ＣｐＧ部位１および３に対するメチル化状態が観察され、ＣｐＧ部位１および３における断片のメチル化状態とコンセンサスが得られているので、＜M_1、M_2、U_3＞および＜M_1、U_2、U_3＞のメチル化状態ベクターの可能性の確率の合計として、＜M_1、I_2、U_3＞のメチル化状態ベクターの確率を計算することができる。不確定状態を有するＣｐＧ部位を合計するこの方法は、２＾ｉまでの可能性の確率の計算を使用することができ、ここでｉは、メチル化状態ベクター中の不確定状態の数を示す。追加の実施形態において、動的プログラミングアルゴリズムが、１つ以上の不確定状態を有するメチル化状態ベクターの確率を計算するために実施され得る。

【０２０６】

いくつかの実施形態では、確率および／またはｐ値スコアを計算する計算負荷は、少なくとも一部の計算をキャッシュすることによってさらに低減され得る。例えば、解析システムは、メチル化状態ベクター（またはそのウィンドウ）の可能性に関する確率の計算を、一時的または永続的なメモリにキャッシュすることができる。他の断片が同じＣｐＧ部位を有する場合、可能性確率をキャッシュすることにより、基礎となる可能性確率を再計算する必要なく、ｐスコア値を効率的に計算することができる。同様に、解析システムは、ベクター（またはそのウィンドウ）からのＣｐＧ部位のセットに関連するメチル化状態ベクターの可能性の各々についてｐ値スコアを計算することができる。解析システムは、同じＣｐＧ部位を含む他の断片のｐ値スコアを決定する際に使用するために、ｐ値スコアをキャッシュしてもよい。一般に、同じＣｐＧ部位を有するメチル化状態ベクターの可能

性のｐ値スコアは、同じＣｐＧ部位のセットからの可能性の異なるもののｐ値スコアを決定するために使用され得る。

【０２０７】

１つ以上の核酸メチル化断片は、領域モデルまたはがん分類器を訓練する前にフィルタリングすることができる。核酸メチル化断片をフィルタリングすることは、対応する複数の核酸メチル化断片から、１つ以上の選択基準（例えば、１つの選択基準を下回るまたは上回る）を満たさないそれぞれの核酸メチル化断片を除去することを含み得る。１つ以上の選択基準は、ｐ値閾値を含み得る。それぞれの核酸メチル化断片の出力ｐ値は、それぞれの核酸メチル化断片の対応するメチル化パターンと、それぞれの核酸メチル化断片の対応する複数のＣｐＧ部位を有する健康な非がんコホートデータセット中のそれらの核酸メチル化断片のメチル化パターンの対応する分布との比較に基づいて、少なくとも部分的に決定され得る。

【０２０８】

複数の核酸メチル化断片をフィルタリングすることは、ｐ値閾値を満足しないそれぞれの核酸メチル化断片を除去することを含み得る。フィルターは、第１の複数の核酸メチル化断片にわたって観察されたメチル化パターンを用いて、それぞれの核酸メチル化断片のそれぞれのメチル化パターンに適用することができる。それぞれの核酸メチル化断片（例えば、断片１、．．．、断片Ｎ）のそれぞれのメチル化パターンは、メチル化部位の識別子と対応するメチル化パターンで識別され、１と０の配列で表され、対応する１つ以上のメチル化部位（例えば、ＣｐＧ部位）を含み、ここで、各「１」は、１つ以上のＣｐＧ部位におけるメチル化ＣｐＧ部位を表し、各「０」は、１つ以上のＣｐＧ部位における非メチル化ＣｐＧ部位を表す。第１の複数の核酸メチル化断片にわたって観察されたメチル化パターンは、第１の複数の核酸メチル化断片によって集合的に表されるＣｐＧ部位状態（例えば、ＣｐＧ部位Ａ、ＣｐＧ部位Ｂ、．．．、ＣｐＧ部位ＺＺＺ）についてのメチル化状態分布を構築するために使用することができる。核酸メチル化断片の処理に関するさらなる詳細は、２０２０年３月４日に出願された「オートエンコーダを用いたがん状態判定システムおよび方法（Systems and Methods for Cancer Condition Determination Using Autoencoders）」と題する米国仮特許出願第６２／９８５，２５８号に開示されており、その全体が参照により本明細書に組み込まれる。

【０２０９】

それぞれの核酸メチル化断片が、異常メチル化スコア閾値未満の異常メチル化スコアを有する場合、それぞれの核酸メチル化断片は、１つ以上の選択基準における選択基準を満足しないことがある。この状況において、異常メチル化スコアは、混合モデルによって決定され得る。例えば、混合モデルは、同じ長さで同じ対応するゲノム位置の可能なメチル化状態ベクターの数に基づいて、それぞれの核酸メチル化断片のメチル化状態ベクター（例えば、メチル化パターン）の尤度を決定することによって、核酸メチル化断片の異常なメチル化パターンを検出することができる。これは、参照ゲノムの各ゲノム位置において、指定された長さのベクターについて複数の可能なメチル化状態を生成することによって実行することができる。複数の可能性のあるメチル化状態を用いて、可能性のあるメチル化状態の総数と、それに続くゲノム位置における各予測メチル化状態の確率を決定することができる。次に、参照ゲノム内のゲノム位置に対応するサンプル核酸メチル化断片の尤度を、サンプル核酸メチル化断片を予測された（例えば、可能性のある）メチル化状態にマッチングさせ、予測されたメチル化状態の計算された確率を検索することによって決定することができる。その後、サンプル核酸メチル化断片の確率に基づいて異常メチル化スコアを計算することができる。

【０２１０】

それぞれの核酸メチル化断片は、それぞれの核酸メチル化断片の残基数が閾値未満である場合、１つ以上の選択基準における選択基準を満たすことができない。残基の閾値数は、１０～５０、５０～１００、１００～１５０、または１５０以上であり得る。残基数の閾値は、２０～９０の固定値とすることができる。それぞれの核酸メチル化断片が閾値未

満のＣｐＧ部位を有する場合、それぞれの核酸メチル化断片は、１つ以上の選択基準における選択基準を満たさないことがある。ＣｐＧ部位の閾値数は、４、５、６、７、８、９、または１０であり得る。それぞれの核酸メチル化断片は、それぞれの核酸メチル化断片のゲノム開始位置およびゲノム終了位置が、それぞれの核酸メチル化断片がヒトゲノム参照配列において閾値未満のヌクレオチド数を表すことを示す場合、１つ以上の選択基準における選択基準を満たすことができない。

【０２１１】

フィルタリングは、対応する複数の核酸メチル化断片中の核酸メチル化断片のうち、対応する複数の核酸メチル化断片中の別の核酸メチル化断片と同じ対応するメチル化パターンを有し、同じ対応するゲノム開始位置およびゲノム終了位置を有する核酸メチル化断片を除去することができる。このフィルタリングステップは、場合によってはＰＣＲ重複を含む、完全な重複である冗長な断片を除去することができる。このフィルタリングにより、対応する複数の核酸メチル化断片中の別の核酸メチル化断片と、対応するゲノム開始位置およびゲノム終了位置が同じで、異なるメチル化状態の数が閾値未満である核酸メチル化断片を除去することができる。核酸メチル化断片の保持に使用される異なるメチル化状態の閾値数は、１、２、３、４、５、または５以上であり得る。例えば、第２の核酸メチル化断片と同じ対応するゲノムの開始および終了位置を有するが、それぞれのＣｐＧ部位（例えば、参照ゲノムにアライメントされる）において少なくとも１、少なくとも２、少なくとも３、少なくとも４、または少なくとも５の異なるメチル化状態を有する第１の核酸メチル化断片が保持される。別の例として、同じメチル化状態ベクターを有するが（例えば、メチル化パターン）、第２の核酸メチル化断片と対応するゲノムの開始および終了位置が異なる第１の核酸メチル化断片も保持される。

【０２１２】

フィルタリングは、複数の核酸メチル化断片中のアッセイアーチファクトを除去することができる。アッセイアーチファクトの除去は、配列決定されたハイブリダイゼーションプローブから得られた配列リードおよび／またはバイサルファイト変換中に変換を受けなかった配列から得られた配列リードを除去することを含み得る。フィルタリングは、汚染物質を除去し得る（例えば、配列決定、核酸単離、および／またはサンプル調製に起因する）。

【０２１３】

フィルタリングは、複数の訓練被験者にわたるがん状態に対するそれぞれのメチル化断片の相互情報フィルタリングに基づいて、複数のメチル化断片からメチル化断片のサブセットを除去することができる。例えば、相互情報は、同時にサンプリングされた関心のある２つの状態間の相互依存性の尺度を提供することができる。相互情報は、１つ以上のデータセットからＣｐＧ部位の独立したセット（例えば、核酸メチル化断片の全部または一部内）を選択し、２つのサンプルグループ間（例えば、遺伝子型データセットのサブセットおよび／またはグループ、生物学的サンプル、および／または被験者）でＣｐＧ部位のセットに対するメチル化状態の確率を比較することによって決定することができる。相互情報スコアは、スライディングウィンドウのそれぞれのフレームにおけるそれぞれの領域における第１の条件対第２の条件のメチル化パターンの確率を示すことができ、したがって、それぞれの領域の識別力を示す。相互情報スコアは、選択されたＣｐＧ部位のセットおよび／または選択されたゲノム領域を横切って進行するにつれて、スライディングウィンドウの各フレームの各領域について同様に計算され得る。相互情報フィルタリングに関するさらなる詳細は、２０２０年１２月１１日に出願された「パッチ畳み込みニューラルネットワークを用いたがん分類（Cancer Classification using Patch Convolutional Neural Networks）」と題する米国特許出願１７／１１９，６０６号に開示されており、その全体が参照により本明細書に組み込まれる。

【０２１４】

＜ＩＩ．Ｃ．ＩＩ．過剰メチル化断片および過剰メチル化断片＞

いくつかの実施形態では、解析システムは、ＣｐＧ部位の数が閾値以上であり、ＣｐＧ

部位のメチル化率が閾値以上であるか、またはＣｐＧ部位の非メチル化率が閾値以上である断片として異常断片を決定する。断片（またはＣｐＧ部位）の長さの閾値の例としては、３以上、４以上、５以上、６以上、７以上、８以上、９以上、１０以上などがある。メチル化または非メチル化のパーセンテージの閾値の例としては、８０％以上、８５％以上、９０％以上、９５％以上、または５０％～１００％の範囲内の他の割合が挙げられる。

【０２１５】

＜ＩＩ．Ｄ．例示的解析システム＞

図６Ａは、１つ以上の実施形態による核酸サンプルを配列決定するための装置の例示的なフローチャートである。この例示的なフローチャートは、シーケンサー６２０および解析システム６００などのデバイスを含む。シーケンサー６２０および解析システム６００は、図３Ａのプロセス３００、図４Ａのプロセス４００、図４Ｂのステップ４２０、および本明細書に記載される他のステップにおける１つ以上のステップを実行するために、連動して動作し得る。

【０２１６】

様々な実施形態において、シーケンサー６２０は、濃縮核酸サンプル６１０を受け取る。図６Ａに示すように、シーケンサー６２０は、特定のタスク（例えば、シーケンシングの開始またはシーケンシングの終了）とのユーザー対話を可能にするグラフィカルユーザーインターフェース６２５、ならびに濃縮断片サンプルを含むシーケンシングカートリッジをロードするため、および／またはシーケンシングアッセイを実行するために必要なバッファーをロードするための１つ以上のロードステーション６３０を含むことができる。したがって、シーケンサー６２０のユーザーがシーケンサー６２０のローディングステーション６３０に必要な試薬およびシーケンシングカートリッジを準備すると、ユーザーはシーケンサー６２０のグラフィカルユーザーインターフェース６２５と対話することによってシーケンシングを開始することができる。一旦開始されると、シーケンサー６２０はシーケンシングを実行し、核酸サンプル６１０からの濃縮断片の配列リードを出力する。

【０２１７】

いくつかの実施形態では、シーケンサー６２０は、解析システム６００と通信可能に結合される。解析システム６００は、１つ以上のＣｐＧ部位におけるメチル化状態の評価、バリアントコールまたは品質管理などの様々な用途のために配列リードを処理するために使用されるいくつかのコンピューティングデバイスを含む。シーケンサー６２０は、配列リードをＢＡＭファイル形式で解析システム６００に提供することができる。分析システム６００は、無線、有線、または無線および有線の組み合わせの通信技術を通じてシーケンサー６２０に通信可能に結合され得る。一般に、解析システム６００は、プロセッサおよび、プロセッサによって実行されると、プロセッサに配列リードを処理させるか、または本明細書に開示される方法もしくはプロセスのいずれかの１つ以上のステップを実行させるコンピュータ命令を記憶する非一過性のコンピュータ可読記憶媒体を備えるように構成される。

【０２１８】

いくつかの実施形態では、配列リードは、例えば、図３Ａのプロセス３００のうちのステップ３４０を介して、アライメント位置情報を決定するために、当該分野で公知の方法を使用して参照ゲノムにアライメントされ得る。アライメント位置は、一般に、所定の配列リードの開始ヌクレオチド塩基および終了ヌクレオチド塩基に対応する参照ゲノム内の領域の開始位置および終了位置を記述し得る。メチル化配列決定に対応して、アライメント位置情報は、参照ゲノムへのアライメントに従って配列リードに含まれる最初のＣｐＧ部位および最後のＣｐＧ部位を示すように一般化されてもよい。アライメント位置情報はさらに、所定の配列リード内の全てのＣｐＧ部位のメチル化状態および位置を示す場合がある。参照ゲノム内の領域は、遺伝子または遺伝子のセグメントと関連付けられることがあり、そのようなものとして、解析システム６００は、配列リードに、その配列リードにアライメントする１つ以上の遺伝子をラベル付けすることがある。一実施形態では、断片の長さ（またはサイズ）は、開始位置および終了位置から決定される。

【０２１９】
様々な実施形態において、例えば、ペアエンド配列決定プロセスが使用される場合、配列リードは、Ｒ＿１およびＲ＿２と示されるリードペアで構成される。例えば、第１のリードＲ＿１は、二本鎖ＤＮＡ（ｄｓＤＮＡ）分子の第１の末端から配列決定される場合があり、一方、第２のリードＲ＿２は、二本鎖ＤＮＡ（ｄｓＤＮＡ）の第２の末端から配列決定される場合がある。したがって、第１のリードＲ＿１および第２のリードＲ＿２のヌクレオチド塩基対は、参照ゲノムのヌクレオチド塩基と一貫して（例えば、反対方向に）整列される可能性がある。リードペアＲ＿１およびＲ＿２から得られるアライメント位置情報には、第１のリード（例えば、Ｒ＿１）の末端に対応する参照ゲノム内の開始位置と、第２のリード（例えば、Ｒ＿２）の末端に対応する参照ゲノム内の終了位置とが含まれる場合がある。言い換えれば、参照ゲノムにおける開始位置と終了位置は、核酸断片が対応する参照ゲノム内の可能性の高い位置を表すことができる。ＳＡＭ（配列アライメントマップ）形式またはＢＡＭ（二値）形式の出力ファイルが生成され、さらなる解析のために出力される。

【０２２０】
次に図６Ｂを参照すると、図６Ｂは、一実施形態によるＤＮＡサンプルを処理するための解析システム６００のブロック図である。解析システムは、ＤＮＡサンプルの解析に使用するための１つ以上のコンピューティングデバイスを実装する。解析システム６００は、配列プロセッサ６４０、配列データベース６４５、モデルデータベース６５５、モデル６５０、パラメータデータベース６６５、およびスコアエンジン６６０を含む。いくつかの実施形態において、解析システム６００は、図３Ａのプロセス３００および図４Ａのプロセス４００の一部または全部を実行する。

【０２２１】
配列プロセッサ６４０は、サンプルからの断片についてメチル化状態ベクターを生成する。断片上の各ＣｐＧ部位において、配列プロセッサ６４０は、図３Ａのプロセス３００を介して、参照ゲノムにおける断片の位置、断片中のＣｐＧ部位の数、および断片中の各ＣｐＧ部位のメチル化、非メチル化、または不確定のいずれかのメチル化状態を指定する、各断片のメチル化状態ベクターを生成する。配列プロセッサ６４０は、配列データベース６４５に断片のメチル化状態ベクターを格納することができる。配列データベース６４５のデータは、サンプルからのメチル化状態ベクターが互いに関連付けられるように組織化されてもよい。

【０２２２】
さらに、複数の異なるモデル６５０が、モデルデータベース６５５に保存されるか、またはテストサンプルと共に使用するために検索されることがある。一実施例では、モデルは、異常断片から導出された特徴ベクターを使用して、テストサンプルのがん予測を決定するための訓練されたがん分類器である。がん分類器の訓練と使用については、「セクションＩＩＩ．がんを判定するためのがん分類器」に関連してさらに説明する。解析システム６００は、１つ以上のモデル６５０を訓練し、様々な訓練されたパラメータをパラメータデータベース６６５に格納することができる。解析システム６００は、モデル６５０を機能と共にモデルデータベース６５５に格納する。

【０２２３】
推論中、スコアエンジン６６０は、出力を返すために１つ以上のモデル６５０を使用する。スコアエンジン６６０は、パラメータデータベース６６５からの訓練されたパラメータと共に、モデルデータベース６５５内のモデル６５０にアクセスする。各モデルに従って、スコアエンジンはモデルのための適切な入力を受信し、受信した入力、パラメータ、および入力と出力とを関連付ける各モデルの関数に基づいて出力を計算する。一部の使用例では、スコアエンジン６６０はさらに、モデルから計算された出力に対する信頼度に関連するメトリクスを計算する。他の使用例では、スコアエンジン６６０はモデルで使用するための他の中間値を計算する。

【０２２４】

［ＩＩＩ．がんを判定するためのがん分類器］
＜ＩＩＩ．Ａ．概要＞
がん分類器は、検査サンプルの特徴ベクターを受け取り、検査サンプルががん、より具体的には特定のがんタイプを有する検査の被検者からのものであるか否かを決定するように訓練することができる。がん分類器は、複数の分類パラメータおよび、入力としての特徴ベクターと、分類パラメータを用いて入力特徴ベクター上で動作する関数によって決定される出力としてのがん予測との間の関係を表す関数とを含むことができる。一部の実施形態では、がん分類器に入力する特徴ベクターは、検査サンプルから決定された異常断片のセットに基づく。異常断片は、図４Ｂのステップ４２０を介して決定されてもよく、より具体的には、ステップ４２０のステップ４７０を介して決定される高メチル化断片および低メチル化断片、または何らかの他のステップに従って決定される異常断片であってもよい。がん分類器を配備する前に、解析システムはがん分類器を訓練することができる。
【０２２５】
＜ＩＩＩ．Ｂ．がん分類器の訓練＞
図５Ａは、実施形態による、がん分類器を訓練するプロセス５００を説明するフローチャートである。解析システムは、異常断片のセットおよびがんタイプのラベルを各々が有する複数の訓練サンプルを取得する（ステップ５１０）。複数の訓練サンプルは、「非がん（non-cancer）」という一般的なラベルを有する健常人からのサンプル、「がん（cancer）」という一般的なラベルを有する被験者からのサンプル、または特定のラベル（例えば、「乳がん（breast cancer）」、「肺がん（lung cancer）」など）を有する被験者からのサンプルの任意の組み合わせを含むことができる。あるがん種の被験者からの訓練サンプルは、そのがん種のコホートまたはがん種コホートと呼ばれることがある。解析システムは、がん分類器の訓練に使用される訓練サンプルが汚染されていないことを保証することができる。訓練サンプルが汚染されているかどうかを決定するために、解析システムは図１のプロセス１００を実行することができる。
【０２２６】
解析システムは、各訓練サンプルについて、訓練サンプルの異常断片セットに基づく特徴ベクターを決定する（ステップ５２０）。解析システムは、ＣｐＧ部位の初期セット内の各ＣｐＧ部位について異常スコアを計算することができる。ＣｐＧ部位の初期セットは、ヒトゲノム中の全てのＣｐＧ部位またはその一部であってもよく、10^4、10^5、10^6、10^7、10^8などのオーダーであってもよい。一実施形態では、解析システムは、異常断片セット中にＣｐＧ部位を包含する異常断片が存在するか否かに基づいて、二値スコアリングで特徴ベクターの異常スコアを定義する。別の実施形態では、解析システムは、ＣｐＧ部位と重なる異常断片のカウントに基づいて異常スコアを定義する。一例では、解析システムは、異常断片が存在しない場合の第１のスコア、少数の異常断片が存在する場合の第２のスコア、および少数以上の異常断片が存在する場合の第３のスコアを割り当てる３値スコアリングを使用することができる。例えば、解析システムは、ＣｐＧ部位と重なるサンプル中の異常断片を５個カウントし、５個のカウントに基づいて異常スコアを算出する。
【０２２７】
一旦全ての異常スコアが訓練サンプルについて決定されると、分析システムは、各要素について、初期セット内のＣｐＧ部位の１つに関連する異常スコアの１つを含む要素のベクターとして、特徴ベクターを決定することができる。解析システムは、サンプルのカバレッジに基づいて、特徴ベクターの異常スコアを正規化することができる。ここで、カバレッジは、分類器に使用されるＣｐＧ部位の初期セットによってカバーされる全てのＣｐＧ部位にわたる、または所与の訓練サンプルについての異常断片セットに基づく、シーケンシング深度の中央値または平均値を指すことができる。
【０２２８】
例として、訓練特徴ベクターの行列５２２を示す図５Ｂを参照する。この例では、解析システムは、がん分類器のための特徴ベクターを生成する際に考慮するＣｐＧ部位［Ｋ］

５２６を同定した。解析システムは、訓練サンプル［Ｎ］５２４を選択する。解析システムは、訓練サンプル［ｎ１］の特徴ベクターにおいて使用される第１の任意のＣｐＧ部位［ｋ１］の第１の異常スコア５２８を決定する。解析システムは、異常断片のセット内の各異常断片をチェックする。解析システムが、第１のＣｐＧ部位を含む少なくとも１つの異常断片を識別した場合、解析システムは、図５Ｂに示されるように、第１のＣｐＧ部位に対する第１の異常スコア５２８を１として決定する。第２の任意のＣｐＧ部位［ｋ２］を考慮すると、解析システムは同様に、第２のＣｐＧ部位［ｋ２］を含む少なくとも１つの異常断片のセットをチェックする。解析システムが、第２のＣｐＧ部位を含むそのような異常断片を発見しない場合、解析システムは、図５Ｂに示されるように、第２のＣｐＧ部位［ｋ２］に対する第２の異常スコア５２９を０に決定する。解析システムがＣｐＧ部位の初期セットに対する全ての異常スコアを決定すると、解析システムは、第１のＣｐＧ部位［ｋ１］に対する１の第１の異常スコア５２８、第２のＣｐＧ部位［ｋ２］に対する０の第２の異常スコア５２９、およびそれ以降の異常スコアとを含む異常スコアを含む第１のトレーニングサンプル［ｎ１］に対する特徴ベクターを決定し、これにより特徴ベクター［１、０、．．．］を形成する。

【０２２９】

サンプルの特徴づけに対するその他のアプローチは、以下に記載されている。「モデルベースのフィーチャライゼーションおよび分類（Model-Based Featurization and Classification）」と題する米国出願第１５／９３１，０２２号、「ターゲット配列決定のための混合モデル（Mixture Model for Targeted Sequencing）」と題する米国出願第１６／５７９，８０５号、「異常断片の検出および分類（Anomalous Fragment Detection and Classification）」と題する米国出願第１６／３５２，６０２号、および「セルフリーＤＮＡサンプルのメチル化断片に基づく起源判定（Source of Origin Deconvolution Based on Methylation Fragments in Cell-Free DNA Samples）」と題する米国出願第１６／７２３，７１６号は、全て参照によりその全体が組み込まれる。

【０２３０】

解析システムは、がん分類器において使用するために考慮されるＣｐＧ部位をさらに制限することができる。解析システムは、ＣｐＧ部位の初期セット内の各ＣｐＧ部位について、訓練サンプルの特徴ベクターに基づいて情報利得を計算する（ステップ５３０）。ステップ５２０から、各訓練サンプルは、ＣｐＧ部位の初期セットにおける全てのＣｐＧ部位の異常スコアを含む特徴ベクターを有する。しかし、ＣｐＧ部位の初期セット中の一部のＣｐＧ部位は、がん種を区別する上で他のＣｐＧ部位ほど有益でない可能性があり、あるいは他のＣｐＧ部位と重複している可能性がある。

【０２３１】

一実施形態では、解析システムは、がんの種類ごとに、また初期セット内のＣｐＧ部位ごとに情報利得を計算し、そのＣｐＧ部位を分類器に含めるかどうかを決定する（ステップ５３０）。情報利得は、所与のがん種を有する訓練サンプルについて、他の全てのサンプルと比較して計算される。例えば、２つの確率変数「異常断片（anomalous fragment）」（「ＡＦ」）および「がんタイプ（cancer type）」（「ＣＴ」）が使用される。一実施形態では、ＡＦは、上記の異常スコア／特徴ベクターについて決定されたように、所与のサンプルにおいて所与のＣｐＧ部位に重なる異常断片が存在するか否かを示す二値変数である。ＣＴは、がんが特定のタイプであるかどうかを示すランダム変数である。解析システムは、ＡＦが与えられたＣＴに関する相互情報を計算する。すなわち、特定のＣｐＧ部位に重なる異常断片があるかどうかがわかれば、がんのタイプについて何ビットの情報が得られるかを計算する。実際には、第１のがん種について、解析システムは他の各がん種に対する一対の相互情報利得を計算し、また他の全てのがん種にわたる相互情報利得を合計する。

【０２３２】

所与のがん型について、解析システムはこの情報を使用して、ＣｐＧ部位がどの程度が

んに特異的であるかに基づいてＣｐＧ部位をランク付けすることができる。この手順は、検討中の全てのがん種について繰り返すことができる。ある特定の領域が、あるがんの訓練用サンプルでは一般的に異常メチル化されているが、他のがん型の訓練用サンプルまたは健康な訓練用サンプルではメチル化されていない場合、それらの異常断片と重なるＣｐＧ部位は、あるがん型に対して高い情報利得を持つ可能性がある。各がん種のランク付けされたＣｐＧ部位は、がん分類器で使用するために、そのランクに基づいて選択されたＣｐＧ部位セットに貪欲に追加（選択）することができる（ステップ５４０）。

【０２３３】

追加の実施形態では、解析システムは、がん分類器で使用される有益なＣｐＧ部位を選択するための他の選択基準を考慮することができる。１つの選択基準は、選択されたＣｐＧ部位が他の選択されたＣｐＧ部位から閾値以上の分離を有することである。例えば、選択されたＣｐＧ部位は、他の選択されたＣｐＧ部位から閾値以上の塩基対数（例えば、１００塩基対）離れていることであり、閾値以内の分離であるＣｐＧ部位は、がん分類器における検討のために両方とも選択されないようにする。

【０２３４】

一実施形態では、初期セットから選択されたＣｐＧ部位のセットに従って、解析システムは、必要に応じて、訓練サンプルの特徴ベクターを修正することができる（ステップ５５０）。例えば、解析システムは、選択されたＣｐＧ部位のセットに含まれないＣｐＧ部位に対応する異常スコアを除去するために、特徴ベクターを切り捨てることができる。

【０２３５】

解析システムは、訓練サンプルの特徴ベクターを用いて、多数の方法のいずれかでがん分類器を訓練することができる。特徴ベクターは、ステップ５２０からのＣｐＧ部位の初期セット、またはステップ５５０からのＣｐＧ部位の選択セットに対応することができる。一実施形態では、解析システムは、訓練サンプルの特徴ベクターに基づいてがんと非がんを区別するために、二値がん分類器を訓練する（ステップ５６０）。このように、解析システムは、健常人からの非がんサンプルおよび被験者からのがんサンプルの両方を含む訓練サンプルを使用する。各訓練サンプルは、「がん（cancer）」または「非がん（non-cancer）」の２つのラベルのいずれかを有することができる。この実施形態では、分類器はがんの存在または非存在の可能性を示すがん予測を出力する。

【０２３６】

別の実施形態では、解析システムは、多くのがん種（原発組織（ＴＯＯ）ラベルとも呼ばれる）を区別するために、マルチクラスがん分類器を訓練する（ステップ５７０）。がんタイプは、１つ以上のがんを含むことができ、非がんタイプを含むこともできる（任意の追加の他の疾患または遺伝的障害なども含むことができる）。そうするために、解析システムは、がん型コホートを使用することができ、また、非がん型コホートを含んでも含まなくてもよい。この多がんの実施形態では、がん分類器は、分類されるがん型の各々についての予測値からなるがん予測（より具体的には、ＴＯＯ予測）を決定するように訓練される。予測値は、所与の訓練サンプル（および推論中のテストサンプル）が各がん種を有する尤度に対応し得る。ある実施態様では、予測値は０～１００でスコア化され、予測値の累積は１００に等しい。例えば、がん分類器は、乳がん、肺がん、非がんの予測値を含むがん予測を返す。例えば、分類器は、検査サンプルが乳がんの可能性６５％、肺がんの可能性２５％、非がんの可能性１０％であるというがん予測を返すことができる。解析システムは、さらに予測値を評価して、１つ以上のＴＯＯラベル、例えば、最も高い予測値を有する第１のＴＯＯラベル、２番目に高い予測値を有する第２のＴＯＯラベルなどを示すＴＯＯ予測とも呼ばれ得る、サンプル中の１つ以上のがんの存在予測を生成することができる。上記の例を継続し、割合が与えられた場合、この例では、システムは、乳がんが最も高い可能性を有することから、サンプルが乳がんに罹患していると判定することができる。

【０２３７】

いずれの実施形態においても、解析システムは、特徴ベクターを有する訓練サンプルの

セットをがん分類器に入力し、分類器の関数が訓練特徴ベクターを対応するラベルに正確に関連付けるように分類パラメータを調整することによって、がん分類器を訓練する。解析システムは、がん分類器の反復バッチ訓練のために、訓練サンプルを１つ以上の訓練サンプルのセットにグループ化することができる。訓練特徴ベクターを含む訓練サンプルの全てのセットを入力し、分類パラメータを調整した後、がん分類器を十分に訓練して、ある程度の誤差の範囲内でその特徴ベクターに従ってテストサンプルをラベル付けすることができる。解析システムは、多数の方法のいずれか１つに従ってがん分類器を訓練することができる。一例として、二値がん分類器は、対数損失関数を用いて訓練されるＬ２－正則化ロジスティック回帰分類器であってもよい。別の例として、多がん分類器は多項ロジスティック回帰であってもよい。実際には、いずれのタイプのがん分類器も、他の技術を用いて訓練することができる。これらの技法は、カーネル法、ランダムフォレスト分類器、混合モデル、オートエンコーダモデル、多層ニューラルネットワークなどの機械学習アルゴリズムなどを使用する可能性があるなど、多数ある。

【０２３８】

分類器には、ロジスティック回帰アルゴリズム、ニューラルネットワークアルゴリズム、サポートベクターマシンアルゴリズム、ナイーブベイズアルゴリズム、最近傍アルゴリズム、ブースティング木アルゴリズム、ランダムフォレストアルゴリズム、決定木アルゴリズム、多項ロジスティック回帰アルゴリズム、線形モデル、または線形回帰アルゴリズムを含めることができる。

【０２３９】

＜ＩＩＩ．Ｃ．がん分類器の展開＞

がん分類装置の使用中、解析システムは、がんの種類が未知の被験者から検査サンプルを得ることができる。解析システムは、プロセス３００、４００、および４２０の任意の組み合わせでＤＮＡ分子から構成される検査サンプルを処理して、異常断片のセットを得ることができる。解析システムは、プロセス５００で議論された同様の原理に従って、がん分類器で使用するための検査特徴ベクターを決定することができる。解析システムは、がん分類器によって使用される複数のＣｐＧ部位中の各ＣｐＧ部位について異常スコアを計算することができる。例えば、がん分類器は、１，０００個の選択されたＣｐＧ部位の異常スコアを含む特徴ベクターを入力として受け取る。したがって、解析システムは、異常断片のセットに基づいて、１，０００個の選択されたＣｐＧ部位の異常スコアを含むテスト特徴ベクターを決定することができる。解析システムは、訓練サンプルと同じ方法で異常スコアを計算することができる。一部の実施形態では、解析システムは、異常断片のセット中にＣｐＧ部位を包含する高メチル化断片または低メチル化断片が存在するか否かに基づいて、異常スコアを二値スコアとして定義する。

【０２４０】

分析システムは次に、検査特徴ベクターをがん分類器に入力することができる。その後、がん分類器の機能は、プロセス５００で訓練された分類パラメータおよび検査特徴ベクターに基づいてがん予測を生成することができる。第１の態様において、がん予測は二値であり、「がん（cancer）」または「非がん（non-cancer）」からなる群から選択され得、第２の態様において、がん予測は多くのがん種および「非がん（non-cancer）」からなる群から選択される。付加的な実施形態において、がん予測は、多数のがんタイプの各々についての予測値を有する。さらに、解析システムは、検査サンプルががんの種類の１つである可能性が最も高いと判定することができる。検査サンプルのがん予測が、乳がんの可能性６５％、肺がんの可能性２５％、非がんの可能性１０％である上記の例に従うと、解析システムは検査サンプルが乳がんに罹患している可能性が最も高いと判定することができる。別の例では、がん予測が、がんでない可能性６０％、がんの可能性４０％という二値である場合、解析システムは、検査サンプルががんでない可能性が最も高いと判定する。追加の実施形態において、最も高い尤度を有するがん予測は、被検者をそのがん種を有すると呼ぶために、やはり閾値（例えば、４０％、５０％、６０％、７０％）と比較され得る。最も高い尤度を有するがん予測がその閾値を超えない場合、解析システ

ムは決定的でない結果を返すことがある。

【０２４１】

追加の実施形態では、解析システムは、プロセス５００のステップ５６０で訓練されたがん分類器を、プロセス５００のステップ５７０またはプロセス５００で訓練された別のがん分類器と連鎖させる。解析システムは、プロセス５００のステップ５６０において二値分類器として訓練されたがん分類器に検査特徴ベクターを入力することができる。解析システムはがん予測の出力を受け取ることができる。がん予測は、被検者ががんに罹患している可能性が高いか、またはがんに罹患していない可能性が高いかについての二値であってもよい。他の実施態様では、がん予測は、がんの可能性および非がんである可能性を記述する予測値を含む。例えば、がん予測は８５％のがん予測値および１５％の非がん予測値を有する。解析システムは、被験者ががんである可能性が高いと判定することができる。解析システムが、被験者ががんである可能性が高いと判定すると、解析システムは、異なるがんタイプを区別するように訓練された多クラスがん分類器に検査特徴ベクターを入力することができる。多クラスがん分類器は、検査特徴ベクターを受け取り、複数のがんタイプのがんタイプのがん予測を返すことができる。例えば、多クラスがん分類器は、被検者が卵巣がんである可能性が最も高いことを指定するがん予測を提供する。別の実施態様では、多クラスがん分類器は、複数のがん種の各がん種の予測値を提供する。例えば、がん予測値は、４０％の乳がんタイプ予測値、１５％の大腸がんタイプ予測値、および４５％の肝臓がん予測値を含むことができる。

【０２４２】

二値がん分類の一般化された実施形態によれば、解析システムは、検査サンプルの配列決定データ（例えば、メチル化配列決定データ、ＳＮＰ配列決定データ、他のＤＮＡ配列決定データ、ＲＮＡ配列決定データなど）に基づいて、検査サンプルのがんスコアを決定することができる。解析システムは、サンプルががんである可能性が高いかどうかを予測するための二値閾値カットオフに対して、検査サンプルのがんスコアを比較することができる。二値閾値カットオフは、１つ以上のＴＯＯサブタイプクラスに基づくＴＯＯ閾値化を用いて調整することができる。解析システムはさらに、１つ以上の可能性の高いがん種を示すがん予測を決定するために、多クラスがん分類器において使用するための検査サンプルの特徴ベクターを生成することができる。

【０２４３】

この分類器は、被検者、例えば疾患状態が未知の被検者の疾患状態を決定するために使用され得る。本方法は、被検者から得られた生物学的サンプル中の対応する複数の核酸断片の複数のゲノム特性中の各ゲノム特性についての値を含む、検査ゲノムデータ構築物（例えば、単一時点検査データ）を電子的形態で取得することを含み得る。次いで、本方法は、検査ゲノムデータ構成体を検査分類器に適用し、それにより被検者における疾患状態の状態を決定することを含み得る。被検者は、以前に疾患状態と診断されたことがない場合がある。

【０２４４】

分類器は、少なくとも（ｉ）第１の時点で被検者から取得された第１の生物学的サンプルから生成された第１の検査ゲノムデータ構成体、および（ｉｉ）第２の時点で被検者から取得された第２の生物学的試料から生成された第２の検査ゲノムデータ構成体を使用する時間的分類器とすることができる。

【０２４５】

訓練された分類器は、被検者、例えば、疾患状態が未知の被検者の疾患状態を決定するために使用することができる。この場合、本方法は、被検者についての検査時系列データセットを電子形式で取得することを含むことができ、検査時系列データセットには、複数の時点におけるそれぞれの時点について、それぞれの時点において被検者から得られた対応する生物学的サンプル中の対応する複数の核酸断片の複数の遺伝子型特性についての値を含む対応する検査遺伝子型データ構築物および、複数の時点における連続する時点のそれぞれの対について、連続する時点のそれぞれの対の間の時間の長さの表示を含む。次い

で、本方法は、検査遺伝子型データ構成体を検査分類器に適用し、それにより被検者における疾患状態の状態を決定することを含むことができる。被検者は、以前に疾患状態と診断されたことがない場合がある。

【０２４６】

＜ＩＶ．応用＞

いくつかの実施形態において、本発明の方法、解析システムおよび／または分類器は、がんの存在を検出するため、がんの進行または再発をモニタリングするため、治療反応または有効性をモニタリングするため、最小残存病変（ＭＲＤ）の存在を決定またはモニタリングするため、またはそれらの任意の組み合わせに使用することができる。例えば、本明細書に記載されるように、分類器を使用して、検査特徴ベクターががんを有する被検者からのものである可能性を記述する確率スコア（例えば、０～１００）を生成することができる。一部の実施形態では、確率スコアは、被検者ががんであるか否かを判定するために閾値確率と比較される。他の実施形態では、尤度スコアまたは確率スコアは、疾患の進行を監視するため、または治療の有効性（例えば、治療効果）を監視するために、複数の異なる時点（例えば、治療前または治療後）で評価することができる。さらに他の実施形態では、尤度スコアまたは確率スコアは、臨床的決定（例えば、がんの診断、治療の選択、治療効果の評価など）を行うため、または臨床的決定に影響を及ぼすために使用することができる。例えば、一実施形態では、確率スコアが閾値を超えた場合、医師は適切な治療を処方することができる。

【０２４７】

＜ＩＶ．Ａ．がんの早期発見＞

いくつかの実施形態では、本発明の方法および／または分類器は、がんを有する疑いのある被検者におけるがんの有無を検出するために使用される。例えば、分類器（例えば、セクションＩＩＩで上述し、セクションＶで検討する）を使用して、検査特徴ベクターががんを有する被検者からのものである可能性を記述するがん予測を決定することができる。

【０２４８】

一実施形態では、がん予測は、検査サンプルががんに罹患しているかどうかの尤度（例えば、０～１００でスコア化される）である（すなわち、二値分類）。したがって、解析システムは、被検者ががんであるかどうかを決定するための閾値を決定することができる。例えば、６０以上または等しいがん予測は、被検者ががんに罹患していることを示すことができる。さらに他の実施形態では、６５以上または等しい、７０以上または等しい、７５以上または等しい、８０以上または等しい、８５以上または等しい、９０以上または等しい、または９５以上または等しいがん予測は、被検者ががんに罹患していることを示す。他の実施形態では、がん予測は疾患の重症度を示すことができる。例えば、８０のがん予測は、８０未満のがん予測（例えば、７０の確率スコア）と比較して、より重篤な形態、またはより後期のがんを示すことがある。同様に、経時的ながん予測値の増加（例えば、２つ以上の時点で採取された同一被験者からの複数のサンプルからの検査特徴ベクターを分類することによって決定される）は、疾患の進行を示すことがあり、または経時的ながん予測値の減少は、治療の成功を示すことがある。

【０２４９】

別の実施形態では、がん予測は多数の予測値から構成され、分類され（すなわち、多クラス分類）、複数のがんタイプの各々が予測値を有する（例えば、０～１００でスコア化される）。予測値は、所与の訓練サンプル（および推論中、訓練サンプル）ががんタイプの各々を有する尤度に対応し得る。解析システムは、最も高い予測値を有するがん種を特定し、被検者がそのがん種を有する可能性が高いことを示すことができる。他の実施形態では、解析システムはさらに、最高予測値を閾値（例えば、５０、５５、６０、６５、７０、７５、８０、８５など）と比較して、被検者がそのがん種を有する可能性が高いと判定する。他の実施形態では、予測値は疾患の重症度も示すことができる。例えば、８０を超える予測値は、６０の予測値と比較して、がんのより重篤な形態、またはより後期の段

階を示すことがある。同様に、経時的な予測値の増加（例えば、２つ以上の時点で採取された同一被験者からの複数のサンプルからの検査特徴ベクターを分類することによって決定される）は、疾患の進行を示すことができ、または経時的な予測値の減少は、治療の成功を示すことができる。

【０２５０】

本発明の態様によれば、本発明の方法およびシステムは、複数のがん徴候を検出または分類するように訓練することができる。例えば、本発明の方法、システムおよび分類器は、１個以上、２個以上、３個以上、５個以上、１０個以上、１５個以上、または２０個以上の異なるタイプのがんの存在を検出するために使用することができる。

【０２５１】

本発明の方法、システムおよび分類器を使用して検出できるがんの例には、がん腫、リンパ腫、芽細胞腫、肉腫、および白血病またはリンパ系悪性腫瘍が含まれる。このようながんのより特定の例としては、扁平上皮がん（例えば、上皮性扁平上皮がん）、皮膚がん、悪性黒色腫、小細胞肺がんを含む肺がん、非小細胞肺がん（「ＮＳＣＬＣ」）、肺腺がんおよび肺扁平上皮がんを含む肺がん、腹膜のがん、胃腸がんを含む胃がんまたは胃腸がん、膵臓がん（例えば、膵管腺がん）、子宮頸がん、卵巣がん（例えば、高悪性度漿液性卵巣がん）、肝臓がん（例えば、肝細胞がん（ＨＣＣ））、肝細胞がん、肝がん、膀胱がん（例えば、尿路上皮性膀胱がん）、精巣（胚細胞腫瘍）がん、乳がん（例えば、ＨＥＲ２陽性乳がん、ＨＥＲ２陰性乳がん、トリプルネガティブ乳がん）、脳腫瘍（例えば、星細胞腫、神経膠腫（例えば、膠芽腫））、結腸がん、直腸がん、大腸がん、子宮内膜がんまたは子宮体がん、唾液腺がん、腎臓がんまたは腎臓がん（例えば、腎細胞がん、腎芽細胞腫またはウィルムス腫瘍）、前立腺がん、外陰がん、甲状腺がん、肛門がん、陰茎がん、頭頸部がん、食道がん、および上咽頭がん（ＮＰＣ）が含まれるが、これらに限定されない。がんの追加的な例としては、網膜芽細胞腫、卵胞膜細胞腫、男化腫瘍、非ホジキンリンパ腫（ＮＨＬ）、多発性骨髄腫および急性血液悪性腫瘍を含むがこれらに限定されない血液悪性腫瘍、子宮内膜症、線維肉腫、絨毛がん、喉頭がん、カポジ肉腫、神経鞘腫、乏突起膠腫、神経芽腫、横紋筋肉腫、骨原性肉腫、平滑筋肉腫、尿路がんが含まれるが、これらに限定されない。

【０２５２】

いくつかの実施形態において、がんは、肛門がん、膀胱がん、乳がん、子宮頸がん、大腸がん、食道がん、胃がん、頭頸部がん、肝胆道がん、白血病、肺がん、リンパ腫、悪性黒色腫、多発性骨髄腫、卵巣がん、膵臓がん、前立腺がん、腎臓がん、甲状腺がん、子宮がん、またはそれらの任意の組み合わせのうちの１つ以上である。

【０２５３】

いくつかの実施形態において、１つ以上のがんは、肛門がん、大腸がん、食道がん、頭頸部がん、肝胆道がん、肺がん、卵巣がん、膵臓がん、ならびにリンパ腫および多発性骨髄腫などの「高シグナル（high-signal）」がん（５年がん特異的死亡率が５０％を超えるがんと定義される）であり得る。高シグナルがんはより侵攻性の傾向があり、一般的に患者から採取した検査サンプル中のセルフリー核酸濃度は平均以上である。

【０２５４】

＜ＩＶ．Ｂ．がんおよび治療のモニタリング＞

いくつかの実施形態において、がん予測は、疾患の進行をモニタリングするため、または治療の有効性（例えば、治療効果）をモニタリングするために、複数の異なる時点（例えば、または治療の前後）で評価することができる。例えば、本発明は、第１の時点でがん患者から第１のサンプル（例えば、第１の血漿ｃｆＤＮＡサンプル）を取得し、そこから（本明細書に記載されるように）第１のがん予測を決定し、第２の時点でがん患者から第２の検査サンプル（例えば、第２の血漿ｃｆＤＮＡサンプル）を取得し、そこから（本明細書に記載されるように）第２のがん予測を決定することを含む方法を含む。

【０２５５】

特定の実施形態では、第１の時点はがん治療前（例えば、切除手術または治療的介入前

）であり、第２の時点はがん治療後（例えば、切除手術または治療的介入後）であり、分類器は治療の有効性を監視するために利用される。例えば、２回目のがん予測値が１回目のがん予測値に比べて減少した場合、治療は成功したとみなされる。しかし、第２のがん予測値が第１のがん予測値に比べて増加した場合は、治療は成功しなかったとみなされる。他の実施形態では、第１および第２の時点の両方が、がん治療の前（例えば、切除手術または治療介入の前）である。さらに他の実施形態では、第１の時点および第２の時点の両方が、がん治療後（例えば、切除手術後または治療介入後）である。さらに他の実施形態において、ｃｆＤＮＡサンプルは、第１および第２の時点でがん患者から取得され、分析され得る。例えば、がんの進行をモニタリングするため、がんが寛解状態にあるかどうかを決定するため（例えば、治療後に）、残存疾患または疾患の再発をモニタリングまたは検出するため、または治療（例えば、治療上の）効果をモニタリングするためである。

【０２５６】

当業者であれば、患者におけるがんの状態をモニタリングするために、任意の所望の時点のセットにわたってがん患者から検査サンプルを取得し、本発明の方法に従って分析することができることを容易に理解するであろう。いくつかの実施形態において、第１および第２の時点は、約１５分～約３０年の時間量、例えば、約１、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、または約２４時間、例えば約１、２、３、４、５、１０、１５、２０、２５、または約５０日、または例えば約１、２、３、４、５、６、７、８、９、１０、１１、または約１２ヶ月、または例えば約１、１．５、２、２．５、３、３．５、４、４．５、５、５．５、６、６．５、７、７．５、８、８．５、９、９．５、１０、１０．５、１１、１１．５、１２、１２．５、１３、１３．５、１４、１４．５、１５、１５．５、１６、１６．５、１７、１７．５、１８、１８．５、１９、１９．５、２０、２０．５、２１、２１．５、２２、２２．５、２３、２３．５、２４、２４．５、２５、２５．５、２６、２６．５、２７、２７．５、２８、２８．５、２９、２９．５または約３０年である。他の実施形態では、少なくとも５ヶ月に１回、少なくとも６ヶ月に１回、少なくとも１年に１回、少なくとも２年に１回、少なくとも３年に１回、少なくとも４年に１回、または少なくとも５年に１回、患者から検査サンプルを得ることができる。

【０２５７】

＜ＩＶ．Ｃ．治療＞

さらに別の実施形態では、がん予測は、臨床的決定（例えば、がんの診断、治療選択、治療効果の評価など）を行うため、または臨床的決定に影響を及ぼすために使用することができる。例えば、一実施形態では、がん予測（例えば、がんについて、または特定のがん種について）が閾値を超えた場合、医師は適切な治療（例えば、切除手術、放射線療法、化学療法、および／または免疫療法）を処方することができる。

【０２５８】

分類器（本明細書に記載）は、サンプル特徴ベクターががんを有する被検者からのものであるというがん予測を決定するために使用することができる。一実施形態では、がん予測値が閾値を超えると、適切な治療（例えば、切除手術または治療）が処方される。例えば、一実施形態では、がん予測値が６０以上または等しい場合、１つ以上の適切な治療法が処方される。別の実施形態では、がん予測値が６５以上または等しい、７０以上または等しい、７５以上または等しい、８０以上または等しい、８５以上または等しい、９０以上または等しい、または９５以上または等しい場合、１つ以上の適切な治療法が処方される。他の実施形態では、がん予測は疾患の重症度を示すことができる。次いで、疾患の重症度に適合する適切な治療が処方され得る。

【０２５９】

いくつかの実施形態において、治療は、化学療法剤、標的化がん治療剤、分化療法剤、ホルモン療法剤、および免疫療法剤からなる群から選択される１つ以上のがん治療剤である。例えば、治療は、アルキル化剤、代謝拮抗剤、アントラサイクリン、抗腫瘍抗生物質、細胞骨格破壊剤（タキサン）、トポイソメラーゼ阻害剤、分裂阻害剤、コルチコステロ

イド、キナーゼ阻害剤、ヌクレオチドアナログ、白金系薬剤、およびそれらの任意の組み合わせからなる群から選択される１つ以上の化学療法剤であり得る。いくつかの実施形態において、治療は、シグナル伝達阻害剤（例えば、チロシンキナーゼおよび成長因子受容体阻害剤）、ヒストン脱アセチル化酵素（ＨＤＡＣ）阻害剤、レチノイン受容体作動薬、プロテオソーム阻害剤、血管新生阻害剤、およびモノクローナル抗体結合体からなる群から選択される１つ以上の標的化がん治療剤である。いくつかの実施形態では、治療は、トレチノイン、アリトレチノイン、ベキサロテンなどのレチノイドを含む１つ以上の分化療法剤である。いくつかの実施形態では、治療は、抗エストロゲン剤、アロマターゼ阻害剤、黄体ホルモン剤、エストロゲン剤、抗アンドロゲン剤、およびＧｎＲＨアゴニストまたはアナログからなる群から選択される１つ以上のホルモン療法剤である。一実施形態において、治療は、リツキシマブ（ＲＩＴＵＸＡＮ）およびアレムツズマブ（ＣＡＭＰＡＴＨ）のようなモノクローナル抗体療法、ＢＣＧ、インターロイキン－２（ＩＬ－２）およびインターフェロン－アルファのような非特異的免疫療法およびアジュバント、免疫調節薬、例えばサリドマイドおよびレナリドミド（ＲＥＶＬＩＭＩＤ）からなる群から選択される１つ以上の免疫療法剤である。腫瘍の種類、がんの病期、以前のがん治療または治療薬への曝露およびがんの他の特徴などの特徴に基づいて、適切ながん治療薬を選択することは、熟練した医師またはがん専門医の能力の範囲内である。

【０２６０】

＜Ｖ．キットの実施＞

また、本明細書には、がん分類器に関する方法を含む上述の方法を実施するためのキットも開示する。キットは、遺伝物質を含む個体からサンプルを収集するための１つ以上の収集容器を含み得る。サンプルは、血液、血漿、血清、尿、糞便、唾液、他のタイプの体液、またはそれらの任意の組み合わせを含むことができる。このようなキットには、サンプルから核酸を単離するための試薬を含めることができる。試薬は、緩衝液および検出剤を含む、核酸を配列決定するための試薬をさらに含み得る。１つ以上の実施形態において、キットは、特定のゲノム領域、特定の変異、特定の遺伝子変異、またはそれらの組み合わせを標的とするためのプローブを含む１つ以上の配列決定パネルを含み得る。１つ以上の実施形態において、キットは、例えば、表２、表４、またはそれらのいくつかの組み合わせからの、汚染を標的とするプローブを含む少なくとも１つのパネルを含む。他の実施形態において、キットを介して収集されたサンプルは、サンプル中の核酸を配列決定するために配列決定パネルを使用し得る配列決定研究所に提供される。

【０２６１】

キットはさらに、キットに含まれる試薬の使用説明書を含むことができる。例えば、キットはサンプルを採取し、検査サンプルから核酸を抽出するための説明書を含むことができる。説明書の例としては、試薬を添加する順序、被験試料から核酸を単離するために使用する遠心速度、核酸の増幅方法、核酸の配列決定方法、またはそれらの組み合わせを挙げることができる。説明書はさらに、記載された方法のいずれかのステップを実行する目的で、分析システム２００としてのコンピューティングデバイスを操作する方法を明らかにすることができる。

【０２６２】

上記の構成要素に加えて、キットは、本開示全体を通して記載される様々な方法を実行するためのコンピュータソフトウェアを記憶したコンピュータ読み取り可能な記憶媒体を含むことができる。これらの説明書が存在し得る１つの形態は、適切な媒体または基材に印刷された情報、例えば、情報が印刷された紙片または複数の紙片、キットの包装中、パッケージ挿入物中に印刷された情報としてである。さらに別の手段としては、コンピュータ可読媒体、例えばディスケット、ＣＤ、ハードドライブ、ネットワークデータ記憶装置であり、その上にコンピュータコードの形で命令が保存されている。さらに別の手段としては、インターネットを経由して、削除されたサイトの情報にアクセスできるウェブサイトのアドレスがある。

【０２６３】

［ＶＩ．結果例］
＜ＶＩ．Ａ．サンプルの収集および処理＞
研究デザインおよびサンプルであるＣＣＧＡ（ＮＣＴ０２８８９９７８）は、長期追跡調査を伴う、他施設共同研究、症例対照研究、観察研究である。３４２施設から約１５，０００人の参加者から非同定生物検体が収集された。サンプルはトレーニングセット（１，７８５例）とテストセット（１，０１５例）に分けられ、サンプルは各コホートにおける部位間のがん種および非がん種の分布が事前に規定された通りになるように選択され、がんおよび非がんのサンプルは性別で年齢を頻度的に一致させた。
【０２６４】
全ゲノムバイサルファイト配列決定は、血漿からｃｆＤＮＡを単離し、ｃｆＤＮＡの解析に全ゲノムバイサルファイト配列決定（ＷＧＢＳ；深さ３０倍）を採用した。ｃｆＤＮＡを、修正ＱＩＡａｍｐ循環核酸キット（Ｑｉａｇｅｎ、メリーランド州ジャーマンタウン）を用いて、患者１人につき２本の血漿チューブ（合計量１０ｍｌまで）から抽出した。最大７５ｎｇの血漿ｃｆＤＮＡを、ＥＺ－９６　ＤＮＡメチル化キット（Ｚｙｍｏ　Ｒｅｓｅａｒｃｈ社、Ｄ５００３）を用いて重亜硫酸塩に変換した。変換されたｃｆＤＮＡは、Ａｃｃｅｌ－ＮＧＳ　Ｍｅｔｈｙｌ－Ｓｅｑ　ＤＮＡライブラリー調製キット（Ｓｗｉｆｔ　ＢｉｏＳｃｉｅｎｃｅｓ、ミシガン州アナーバー）の使用によりデュアルインデックスシーケンシングライブラリーを調製するために使用され、構築されたライブラリーはＫＡＰＡ　イルミナプラットフォーム用ライブラリー定量キット（Ｋａｐａ　Ｂｉｏｓｙｓｔｅｍｓ、マサチューセッツ州ウィルミントン）を用いて定量された。１０％のＰｈｉＸ　ｖ３ライブラリー（Ｉｌｌｕｍｉｎａ、ＦＣ－１１０－３００１）とともに４つのライブラリーをプールし、Ｉｌｌｕｍｉｎａ　ＮｏｖａＳｅｑ　７０００　Ｓ２フローセル上でクラスター化した後、１５０ｂｐペアエンドシーケンシング（３０倍）を行った。
【０２６５】
各サンプルについて、ＷＧＢＳ断片セットを、異常なメチル化パターンを有する断片の小さなサブセットに縮小した。さらに、高メチル化または低メチル化ｃｆＤＮＡ断片が選択された。異常なメチル化パターンを有し、高メチル化または低メチル化、すなわちＵＦＸＭを有するｃｆＤＮＡ断片が選択された。がんでない個体で高頻度に生じる断片、または不安定なメチル化を有する断片は、がんの状態を分類するための高度に識別的な特徴を生み出す可能性は低い。そこで、ＣＣＧＡ研究から得られた、がんのない非喫煙者１０８人（年齢：５８±１４歳、女性７９人［７３％］）からなる独立した参照セット（すなわち、参照ゲノム）を用いて、統計モデルと典型的な断片のデータ構造を作成した。これらのサンプルは、セクションＩＩ．Ｃ．で前述したように、断片内のＣｐＧメチル化状態の所与の配列の尤度を推定するマルコフ連鎖モデル（次数３）の訓練に使用された。このモデルは、正常な断片範囲（ｐ値＞０．００１）内で較正されることが実証され、マルコフモデルからのｐ値が＞０．００１である断片を、異常性が不十分であるとして拒絶するために使用された。
【０２６６】
上述のように、さらなるデータ削減ステップでは、少なくとも５つのＣｐＧがカバーされ、平均メチル化が０．９以上（高メチル化）または０．１未満（低メチル化）の断片のみを選択した。この手順により、訓練においてがんを発症していない参加者のＵＦＸＭ断片の中央値（範囲）は２，８００（１，５００～１２，０００）、訓練においてがんを発症した参加者のＵＦＸＭ断片の中央値（範囲）は３，０００（１，２００～４２０，０００）となった。このデータ削減手順は参照セットデータのみを使用するため、この段階は各サンプルに１回のみ適用する必要があった。
【０２６７】
＜ＶＩ．Ｂ．汚染検出結果＞
結果例の最初のグループとして、表２および表４の汚染マーカープローブの性能を調べるために、コードベースの開発にさらに力を入れる前に、意図的な汚染の有無にかかわらず、異なる内部研究からの血漿ｃｆＤＮＡサンプルおよびｇＤＮＡサンプルの複数のバッ

チを予備分析に使用した。

【０２６８】

図７は、所与のサンプルについてホモ接合と呼ばれた各汚染マーカーについて、基準または代替と呼ばれた断片の総数に対する、汚染と分類された断片の数の割合の分布を、第１の結果例群に従って示したものである。分析システムは、表２および表４の汚染マーカープローブを利用して、サンプル中の汚染断片を同定した。汚染断片の割合は、汚染マーカープローブと重なる各サンプル中の断片の総数に対する同定された汚染断片の比率として計算した。グラフ７００に示すように、各サンプルの汚染断片の割合をプロットし、Ｙ軸は汚染断片の割合を示した。０．００１を超えるサンプルもいくつかあったが、０．００１以下のサンプルが圧倒的に多かった。グラフ７００の箱ひげ図では、下位四分位値、中央値、上位四分位値の全てが０．０００５を大きく下回っている。グラフ７１０は、Ｘ軸に汚染片の割合を、Ｙ軸にサンプル数をプロットしたもので、同じ結果を示している。第２グループの結果例に対応する図については、図１７を参照されたい。

【０２６９】

図８は、第１の結果例に従った、汚染マーカーの対立遺伝子頻度の散布図である。解析システムは、ＣＣＧＡ研究とそのフォローアップ研究からのサンプルを用いて、汚染マーカーの対立遺伝子頻度を計算した。解析システムはまた、ゲノムデータベースＧｅｎｏｍｅｓ　ＰｒｏｊｅｃｔおよびｇｎｏｍＡＤから集団対立遺伝子分率を得た（グラフ１０００）。各汚染マーカーは、ｘ軸が集団対立遺伝子分率を示し、ｙ軸がサンプルから計算した対立遺伝子分率を示してプロットしている。第２グループの結果例に対応する図については、図１４を参照されたい。

【０２７０】

図９は、第１のグループの例示的な結果に従って、ＳＮＰ汚染マーカー（グラフ９００）およびインデル汚染マーカー（グラフ９１０）の遺伝子型および接合性を示す２つのグラフを示す。複数のＳＮＰ汚染マーカーについてのグラフ９００において、左側の３つの箱ひげ図は、サンプルの第１のセット（基準ホモ接合体、代替ホモ接合体、およびヘテロ接合体）の接合性の比例内訳を表し、右側の３つの箱ひげ図は、サンプルの第２のセット（基準ホモ接合体、代替ホモ接合体、およびヘテロ接合体）の接合性の比例内訳を表す。同様に、インデル汚染マーカーのグラフ９１０も、２セットのサンプル間で分割されたインデル汚染メーカーの接合性の比例内訳を表す。図１５および図１６を参照されたい。

【０２７１】

図１０は、第１の結果例群に従って、所定のサンプルについて、ホモ接合性であり、そのサンプルの汚染を推定するのに有用であると考えられるのに十分な断片が重複している汚染マーカーの割合を示すグラフである。分析システムは、各サンプルの汚染マーカーの接合性を決定した。次に、分析システムは、サンプルにおいてホモ接合であることが判明した汚染マーカーの割合を計算した。サンプル中のホモ接合体であることが判明した汚染マーカーの数が多いほど、汚染検出能力が高くなる。さらに、これらのホモ接合部位は、サンプルの汚染を推定するのに有用と考えられるだけのサンプル断片が重なるなど、様々な品質チェック基準を経て、所定のサンプルに有用な部位の最終的な割合が得られる。その割合は、多重ＳＮＰ部位汚染マーカー、インデル部位汚染マーカー、多重ＳＮＰ部位汚染マーカーおよびインデル部位汚染マーカーの両方について、特にプロットした。最初の箱ひげ図１０００では、割合の中央値は約０．５２５であった。複数のＳＮＰ部位汚染マーカーに関する２番目の箱ひげ図１０１０では、中央値の割合は約０．５７５であった。３番目の箱ひげ図１０２０では、インデル部位汚染マーカーであった。中央値は約０．４７５であった。図１５および図１３を参照されたい。

【０２７２】

図１１Ａおよび１１Ｂは、第１の結果例群に従った、サンプルの異なるバッチについての推定汚染レベル検出のグラフを示す。グラフ１１００は、サンプルの「脱リスク（Derisking）」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１１０は、サンプルの「ｄｏｐｐｌｅｒ＿ｐｒｅｌｉｍ＿ｔｅｓｔ」バッチからの推定サンプル汚染レ

ベルの分布を示す。グラフ１１２０は、サンプルの「Ｈｙｂ１＿ＳＯＰ＿１２ｐｌｅｘ」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１３０は、サンプルの「ＭＲＤ」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１４０は、「ｃｆＤＮＡ滴定（cfDNA titration）」バッチサンプルからの推定サンプル汚染レベルの分布を示す。グラフ１１５０は、サンプルの「ｇＤＮＡ滴定（gDNA titration）」バッチからの推定サンプル汚染レベルの分布を示す。グラフ１１６０は、サンプルの「ダウンサンプリングｇＤＮＡ滴定（downsampled gDNA titration）（ｖ１．０化学）」バッチからの推定サンプル汚染レベルの分布を示す。また、グラフ１１７０は、「ダウンサンプルｇＤＮＡ滴定（downsampled gDNA titration）（ｖ１．５化学）」バッチ試料からの推定サンプル汚染レベルの分布を示す。第２グループの結果例に対応する図については、図２０および図２１を参照されたい。

【０２７３】

第２の結果例群として、表２および表４の汚染マーカープローブの性能を調べるために、意図的な汚染のない８４人の一個人からの血漿ｃｆＤＮＡサンプルのコホートを、改訂され成熟したコードベースを用いた最終分析に用いた。

【０２７４】

図１２は、実験中の８４サンプルについて集計した、表２および表４に列挙した汚染マーカーごとに得られた特異なｃｆＤＮＡ断片の数の分布を示す。ｘ軸は、汚染マーカーを標的とするように設計されたプローブと重複する特異なｃｆＤＮＡ断片分子の数を示し、全８４サンプルにわたって集計し、ｙ軸は、ｘ軸によって示されるビンに該当する汚染マーカーの割合を示す。

【０２７５】

表２および表４に記載されているような断片マーカーは、分析システムのステップ２２０および２５０なしで選択した。ハーディーワインベルグ平衡にないマーカーは、参照ハプロタイプおよび代替ハプロタイプの集団データベース頻度に基づいてフィルタリングした。この場合、ハーディーワインベルグ項（ｐ２＋２ｐｑ＋ｑ２）の値は［０．９、１．５］の範囲であった（１より高い値への偏りは、集団の非ランダム交配条件において予想される）。表２および表４の１０００マーカーのうち、この条件では４マーカーが不合格であった。

【０２７６】

各汚染マーカーは、そのマーカーのプローブと重複するｃｆＤＮＡ断片について、個々のサンプルに関してさらなる品質チェック（ＱＣ）を受けた。特定のサンプルについて品質チェック条件に合格しなかったマーカーは、汚染割合の推定を含む、特定のマーカーについてのゲノムバリアントコールに依存するあらゆる分析について、そのサンプルについて後に廃棄された。

【０２７７】

実験の８４サンプルからの汚染マーカーおよび１サンプルの各ペアについて、複数のＳＮＰマーカーおよびインデルマーカーのハプロタイプ、および対立遺伝子をそれぞれ参照または代替として確実に呼び出すために使用することができる、ゲノム変異体の位置範囲に重複する特異なｃｆＤＮＡ断片分子の最小数、この場合２０が、一定の割合（例えば、複数のＳＮＰマーカーでは０．７、インデルマーカーでは０．５）以上のｃｆＤＮＡ断片について、参照または代替マーカーとして使用することができる。後者の条件は、マーカーおよびサンプルのペアに期待される理論的な二項確率分布（ヘテロ接合と呼ばれるマーカーの平均は０．５、またはホモ接合と呼ばれるサンプル中の全マーカーの総汚染率に依存する平均）に従って、参照ハプロタイプと呼ばれるｃｆＤＮＡ断片の数と代替ハプロタイプと呼ばれるｃｆＤＮＡ断片の数の比が、非常にあり得ないものであってはならないこと（例えば、ｐ値＜１０－５）をチェックするために拡張する。

【０２７８】

図１３は、この品質チェックプロセスの結果を示している。前述の基準に基づいて、８４サンプル全てについて８５１マーカーが合格し、８３サンプルについて６８マーカーが

合格し、８２サンプルについて２０マーカーが合格した。不合格マーカーは、８４サンプル全てで７マーカーであった（ハーディーワインベルグ平衡の合格条件を満たさなかった４個を含む）。残りのマーカーの不合格率は、この両極端の間でばらつきがあった。

【０２７９】

図１４～１６は、８４サンプル中少なくとも８２サンプルについて品質チェック基準に合格した汚染マーカーについて、表２および表４の汚染マーカーの基となるゲノム変異体の特徴を示す図である。図１４は、これらの汚染マーカーを選択するための対立遺伝子頻度範囲である［０．３、０．７］の範囲にある対立遺伝子頻度の散布図を示す。図１５は、マーカーがヘテロ接合と呼ばれる頻度の散布図を示す。図１６は、ハーディーワインベルグ項の値の散布図を示す。これらのグラフのｘ軸は、集団データベース１０００ Ｇｅｎｏｍｅｓ（複数のＳＮＰマーカー用）またはｇｎｏｍＡＤ（インデルマーカー用）で利用可能なデータから得られた値または計算された値を表し、ｙ軸は、我々の実験における８４個のｃｆＤＮＡサンプルから計算された値を表す。全てのポイントは、ｙ＝ｘ線（グラフでは破線で表現）の周りに多少のノイズを伴って横たわることが予想された。

【０２８０】

ｃｆＤＮＡ断片のセットにおける汚染率を推定するために、まず、汚染源に由来するｃｆＤＮＡ断片を同定した。マーカーと呼ばれるホモ接合型バリアントと反対のバリアントとして呼ばれるｃｆＤＮＡ断片は、汚染断片として分類した。断片が汚染マーカー部位と重ならない、またはヘテロ接合体として遺伝子型決定された汚染マーカー部位と重なる断片のサブセットは、汚染断片の分類および汚染率を推定する目的で無視した。汚染率を推定する能力は、少なくとも１つの汚染断片を見つけることを条件とした。汚染断片は分数化できないため、検出可能な汚染断片の下限は、検討対象となる断片の総数に反比例した。

【０２８１】

図１７は、全８４検体にわたる各マーカーについて集計した、所与の検体についてホモ接合体と呼ばれた各汚染マーカーについて、参照または代替と呼ばれた断片の総数に対する、汚染である分類された断片の数の割合の分布を示す。グラフ１７１０は、ジッター散布図および、分布の極値および四分位数を示すこれらの値の箱ひげ図を示す。これらの値の算術平均は2×10^{-4}である。検出下限より低い値は０に抑圧され、その結果、０に小さなバンプが生じる。グラフ１７２０は、同じデータの平滑化密度線（実線）、および二項パラメータ$p = 2 \times 10^{-4}$、サンプルサイズは特定の汚染マーカーに対するｃｆＤＮＡ断片の集合数として、データに二項分布フィットを用いて得られたシミュレーション値（破線）を示す。全体として、０と最頻値付近で若干のずれはあるものの、よく適合している。

【０２８２】

汚染断片の概念を発展させ、あるｃｆＤＮＡサンプルの汚染率を、外部ソースからのｃｆＤＮＡ汚染の根本的プロセスをモデル化することによって推定した。サンプルに対してホモ接合体として呼ばれるマーカーだけを考慮すると、汚染率がｃｆで、サンプルに対して呼ばれるホモ接合体とは反対のバリアントの集団対立遺伝子頻度がａｆｉである場合、マーカーに対する汚染断片が観察される確率はｃｆｘａｆｉである。このマーカーでの断片の総数がｎｉであれば、このマーカーにおける汚染断片の期待数は、ｃｆｘａｆｉｘｎｉである。確率論の期待値の直線性の性質を用いると、全てのマーカーについてこれらの期待数を合計して、全てのマーカーにわたる汚染断片の期待総数の式、$nc = \Sigma_i (cf \times af_i \times n_i)$を得ることができる。全ての既知の変数を片側に持ってくるように再配列すると、汚染割合の推定値は、$cf = nc / (\Sigma_i (af_i \times n_i))$となる。

【０２８３】

図１８は、仮想的な断片セットに対する汚染率を推定するための式の適用を示す図である。この例では、４つの複数のＳＮＰおよび２つのインデル汚染マーカーがあり、そのうち３つの複数のＳＮＰおよび１つのインデルマーカーだけがホモ接合と呼ばれる。３つの汚染断片の同定、各マーカーの断片の総数、反対ハプロタイプの各マーカーでの対立遺伝

子頻度を考慮すると、サンプル全体の汚染断片の推定値が得られる。

【０２８４】

図１９は、シミュレーションデータに汚染割合モデルを適用した結果を示している。各ポイントについて、表２および表４の各汚染マーカーについて、それらの集団頻度に基づいて、バックグラウンドゲノムおよび汚染源ゲノムの両方についてバリアントコール（ホモ接合参照、ホモ接合代替、またはヘテロ接合）をサンプリングすることによってデータを生成した。次に、参照および代替ハプロタイプ断片のカウントを、実験から得られた８４サンプルのうち、このマーカーについて観察されたｃｆＤＮＡ断片の平均数をサイズとする二項分布からサンプリングし、確率を２つのバリアントコールおよびシミュレーションパラメータとして入力された汚染割合の関数とした。推定値のばらつきを捉えるため、各シミュレーションは１００回繰り返した。グラフのｘ軸はシミュレーションのパラメータとして使用された汚染率を表し、ｙ軸はモデルによって推定された汚染率の値を表す。破線はｙ＝ｘ線を表す。各箱ひげ図は、指定された汚染率パラメータで実行された１００のシミュレーションの推定汚染率の範囲を示す。この範囲は、中央値が期待値に非常に近く、汚染度が高くなるにつれて四分位範囲が小さくなることを示す。これらの結果は、汚染率を推定するために本明細書で説明したモデルを検証するものである。

【０２８５】

図２０は、それぞれ滴定レベルが異なるｃｆＤＮＡサンプルの４つの滴定対に、汚染率モデルを適用した結果を示す。滴定対におけるドナーサンプルは汚染として扱い、その滴定レベルは汚染率として扱う。汚染率の推定値が確実に得られていることを確認するため、各組の各滴定レベルも、インプット材料の入手可能性に応じて複数回繰り返した。Ｘ軸は滴定レベル、Ｙ軸はそのレベルで滴定した各組の複製に対する推定汚染率を示す。実線は全観測値に対する線形モデルフィットを示し、破線はｙ＝ｘ線を示す。このグラフから、線形モデルフィットはｙ＝ｘ線とほぼ同じ傾きを有するが、２本の線の切片は異なっており、全てのサンプルに約2.5×10^{-4}という一定の意図しないバックグラウンド汚染があることがわかる。これらのサンプル内の汚染断片を詳しく調べると、この一定の残留汚染は実在し、バリアントコーリングなどの技術的ノイズによるものではないことが明らかになった（ここでは図示していない）。これらの結果は、汚染率を推定するための本明細書に記載のモデルをさらに検証するものである。

【０２８６】

図２１は、実験における８４個のｃｆＤＮＡサンプルの推定汚染率の分布を示す。このグラフは、ジッター散布図ならびに推定汚染率値の箱ひげ図を示す。中央値は1.8×10^{-4}、平均値は3.2×10^{-4}である。この平均値は、同じサンプルセットで得られた汚染断片の割合の平均値（2×10^{-4}）とは異なり、高いことに注意すべきであり、これは、汚染断片モデルが、マーカーが両方のゲノムで同じハプロタイプを持つ場合に同定できなかった汚染断片も考慮するためであると予想される。

【０２８７】

図２２は、複数のＳＮＰマーカーおよびインデルマーカーを別々に考慮した場合、実験における８４サンプルについて得られた推定値が高い相関を示し、有意な系統的バイアスを示さないことを示す。このことは、両タイプの汚染マーカーの性能が同等であることを証明している。

【０２８８】

＜ＶＩＩ．追加考察＞

以上の実施形態の詳細な説明は、本開示の特定の実施形態を示す添付の図面を参照する。異なる構造および動作を有する他の実施形態は、本開示の範囲から逸脱しない。「本発明（the invention）」等の用語は、本明細書に記載される出願人の発明の多くの代替的な態様または実施形態の特定の具体例を参照して使用され、その使用またはその不存在は、出願人の発明の範囲または特許請求の範囲を限定することを意図するものではない。

【０２８９】

本発明の実施形態は、本明細書における操作を実行するための装置に関することもでき

る。この装置は、必要な目的のために特別に構成されてもよく、および／または、コンピュータに格納されたコンピュータプログラムによって選択的に起動または再構成される汎用のコンピューティングデバイスで構成されてもよい。このようなコンピュータプログラムは、非一過性の、有形のコンピュータ可読記憶媒体、または、コンピュータシステムバスに結合され得る、電子命令を記憶するのに適した任意のタイプの媒体に記憶され得る。さらに、本明細書で言及されるコンピューティングシステムは、単一のプロセッサを含んでもよいし、コンピューティング能力を高めるために複数のプロセッサ設計を採用する構造であってもよい。

【０２９０】

分析システムによって実行されるものとして本明細書で説明されるステップ、操作、またはプロセスのいずれかは、単独で、または他のコンピューティングデバイスと組み合わせて、装置の１つ以上のハードウェアまたはソフトウェアモジュールを用いて実行または実装され得る。一実施形態では、ソフトウェアモジュールは、プロセスのいずれかまたは全てを実行するためにコンピュータプロセッサによって実行され得る、コンピュータプログラムコードを含むコンピュータ可読媒体を含むコンピュータプログラム製品、操作、または説明されたステップ、で実装される。

【誤訳訂正２】

【訂正対象書類名】特許請求の範囲

【訂正対象項目名】全文

【訂正方法】変更

【訂正の内容】

【特許請求の範囲】

【請求項１】

検査サンプル中のがんの存在を予測する方法であって、

前記検査サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む、前記検査サンプルを取得するステップと、

前記検査サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、

同定された汚染マーカーの１つにおける、それぞれの汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記検査サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、

同定されたあらゆる前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および

前記汚染レベルが閾値レベル以下であるか否かを判定するステップと、並びに

前記汚染レベルが前記閾値レベル以下であると判定したことに応答して、がん予測を生成するために前記検査サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに対してがん分類を実行するステップと、

を備える、方法。

【請求項２】

請求項１に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。

【請求項３】

請求項２に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。

【請求項４】

請求項２～３のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。

【請求項５】

請求項２～４のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。

【請求項６】
請求項２～５のいずれか一項に記載の方法において、各複数のＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項７】
請求項２～６のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項８】
請求項２～７のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、方法。
【請求項９】
請求項１～８のいずれか一項に記載の方法において、前記複数の汚染マーカーは挿入－欠失（インデル）部位を含む、方法。
【請求項１０】
請求項９に記載の方法において、前記インデル部位は５ｂｐｓ～１０ｂｐｓである、方法。
【請求項１１】
請求項９～１０のいずれか一項に記載の方法において、前記インデル部位は４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する、方法。
【請求項１２】
請求項９～１１のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項１３】
請求項９～１２のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項１４】
請求項９～１３のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項１５】
請求項１～１４のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項１６】
請求項１～１５のいずれか一項に記載の方法において、前記汚染レベルを推定するステップは、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記検査サンプルのシーケンシング深度、前記検査サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項１７】
請求項１～１６のいずれか一項に記載の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、がんの分類を差し控える、方法。
【請求項１８】
請求項１～１７のいずれか一項に記載の方法において、前記がん予測は、がんと非がんとの間の二値予測を含む、方法。
【請求項１９】
請求項１～１８のいずれか一項に記載の方法において、前記がん予測は、複数のがんタイプ間のマルチクラスがん予測を含む、方法。
【請求項２０】
請求項１～１９のいずれか一項に記載の方法において、前記がんの分類を実行するステップは、
前記検査サンプル中のｃｆＤＮＡ断片の配列リードに基づいて検査特徴ベクターを生成

するステップと、および
前記検査特徴ベクターを分類モデルに入力し、前記検査サンプルの前記がん予測を生成するステップと、
を含む、方法。
【請求項２１】
請求項２０に記載の方法において、前記がんの分類を実行するステップは、さらに、
異常断片のセットを生成するために前記検査サンプルのｃｆＤＮＡ断片の初期セットをｐ値フィルタリングでフィルタリングするステップを含み、前記フィルタリングするステップは、前記異常断片のセットを生成するよう、前記初期セットから他の断片に関して閾値ｐ値を下回る断片を除去するステップを含むものであり、
前記検査特徴ベクターは、前記異常断片のセットの前記配列リードに基づくものである、
方法。
【請求項２２】
請求項２０～２１のいずれか一項に記載の方法において、前記分類モデルは機械学習モデルである、方法。
【請求項２３】
命令がコンピュータプロセッサによって実行されると、請求項１～２２のいずれか一項に記載の方法を前記コンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項２４】
システムであって、
コンピュータプロセッサと、および
請求項２３に記載の前記非一過性のコンピュータ可読記憶媒体と
を備える、システム。
【請求項２５】
検査サンプル中の疾患の存在を予測する方法であって、
前記検査サンプル中のセルフリーＤＮＡ（ｃｆＤＮＡ）断片についての複数の配列リードを含む前記検査サンプルを取得するステップと、
前記検査サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、
同定された汚染マーカーの１つにおける、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記検査サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
同定されたあらゆる前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および
前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、並びに
前記汚染レベルが前記閾値レベル以下であると決定したことに応答して、疾患予測を生成するために前記検査サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに対して疾患分類を実行するステップと
を備える、方法。
【請求項２６】
請求項２５に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項２７】
請求項２６に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項２８】
請求項２６～２７のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。

【請求項２９】
請求項２６～２８のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項３０】
請求項２６～２９のいずれか一項に記載の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項３１】
請求項２６～３０のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項３２】
請求項２６～３１のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、方法。
【請求項３３】
請求項２５～３２のいずれか一項に記載の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、方法。
【請求項３４】
請求項３３に記載の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、方法。
【請求項３５】
請求項３３～３４のいずれか一項に記載の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する、方法。
【請求項３６】
請求項３３～３５のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項３７】
請求項３３～３６のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項３８】
請求項３３～３７のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項３９】
請求項２５～３８のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項４０】
請求項２５～３９のいずれか一項に記載の方法において、前記汚染レベルを推定するステップは、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記検査サンプルのシーケンシング深度、前記検査サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項４１】
請求項２５～４０のいずれか一項に記載の方法において、前記汚染レベルが前記閾値レベル以上であると決定することに応答して、疾患分類を差し控える、方法。
【請求項４２】
請求項２５～４１のいずれか一項に記載の方法において、前記疾患予測は、疾患と疾患なしとの間の二値予測を備える、方法。
【請求項４３】
請求項２５～４２のいずれか一項に記載の方法において、前記疾患予測は、複数の疾患間のマルチクラスがん予測を含む、方法。
【請求項４４】

請求項２５～４３のいずれか一項に記載の方法において、前記疾患分類を実行するステップは、
前記検査サンプル中の前記ｃｆＤＮＡ断片の前記配列リードに基づいて検査特徴ベクターを生成するステップと、および
前記検査特徴ベクターを分類モデルに入力し、前記検査サンプルの前記疾患予測を生成するステップと
を含む、方法。
【請求項４５】
請求項４４に記載の方法において、前記疾患分類を実行するステップは、さらに、
異常断片のセットを生成するために、前記検査サンプルのｃｆＤＮＡ断片の初期セットをｐ値フィルタリングでフィルタリングするステップを含み、前記フィルタリングするステップは、前記異常断片のセットを生成するよう、他の断片に関して閾値ｐ値を下回る断片を前記初期セットから除去するステップを含むものであり、
前記検査特徴ベクターは、前記異常断片のセットの前記配列リードに基づくものである、
方法。
【請求項４６】
請求項４４～４５のいずれか一項に記載の方法において、前記分類モデルは機械学習モデルである、方法。
【請求項４７】
命令がコンピュータプロセッサによって実行されると、請求項２５～４６のいずれか一項に記載の方法をコンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項４８】
システムであって、
コンピュータプロセッサと、および
請求項４７に記載の前記非一過性のコンピュータ可読記憶媒体と
を備える、システム。
【請求項４９】
検査サンプル中の汚染の有無を予測する方法であって、
前記検査サンプル中の複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片に由来する配列リードを取得するステップと、
前記配列リードに基づいて、前記検査サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから、１つ以上の汚染マーカーを同定するステップと、
同定された汚染マーカーの１つにおける、それぞれの前記汚染マーカーのホモ接合性ハプロタイプとは異なるハプロタイプを有する前記検査サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
同定されたあらゆる前記汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および
前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、並びに
前記汚染レベルが前記閾値レベル以上であると決定することに応答して、前記検査サンプルが汚染されていることを示す通知を生成するステップと
を備える、方法。
【請求項５０】
請求項４９に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項５１】
請求項４９に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項５２】

請求項５０～５１のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項５３】
請求項５０～５２のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項５４】
請求項５０～５３のいずれか一項に記載の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項５５】
請求項５０～５４のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項５６】
請求項５０～５５のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１からの前記複数のＳＮＰ部位を含む、方法。
【請求項５７】
請求項４９～５６のいずれか一項に記載の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、方法。
【請求項５８】
請求項５７に記載の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、方法。
【請求項５９】
請求項５７～５８のいずれか一項に記載の方法において、前記インデル部位は、４５％～５５％の範囲内の前記集団ハプロタイプ頻度を有する、方法。
【請求項６０】
請求項５７～５９のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項６１】
請求項５７～６０のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項６２】
請求項５７～６１のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項６３】
請求項４９～６２のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項６４】
請求項４９～６３のいずれか一項に記載の方法において、前記汚染レベルを推定するステップは、さらに、前記同定された汚染ｃｆＤＮＡ断片の数、前記検査サンプルのシーケンシング深度、前記検査サンプル中の前記ｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項６５】
命令がコンピュータプロセッサによって実行されると、請求項４９～６４のいずれか一項に記載の方法をコンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項６６】
システムであって、
コンピュータプロセッサと、および
請求項６５に記載の前記非一過性のコンピュータ可読記憶媒体と

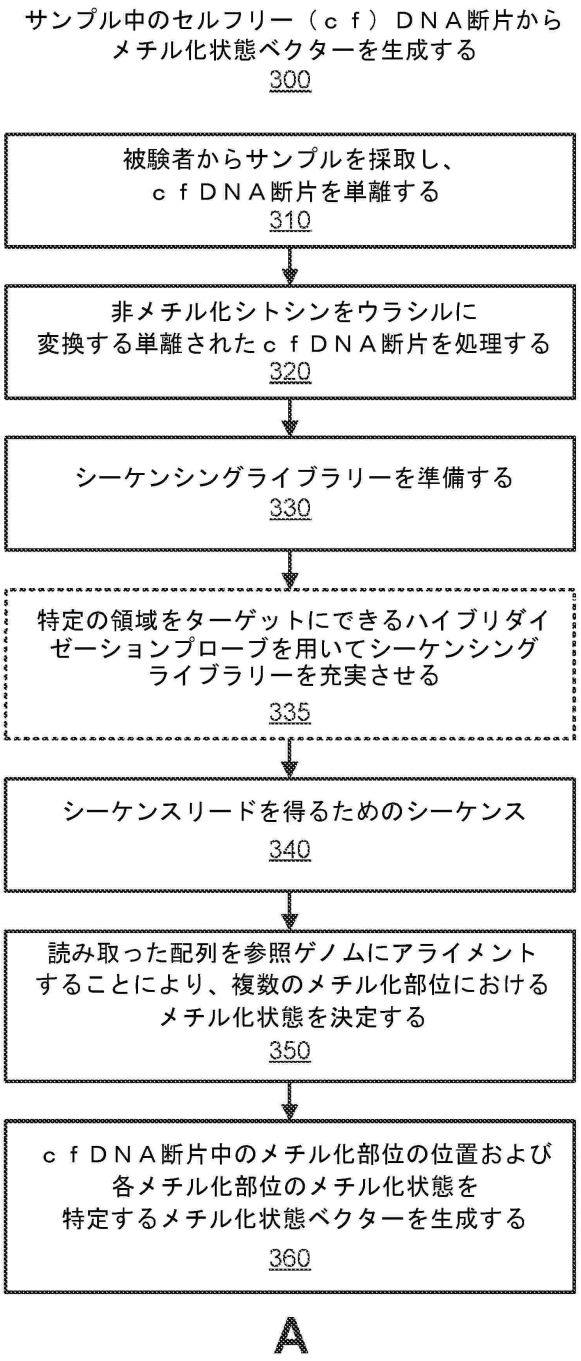
を備える、システム。
【請求項６７】
がん分類モデルを訓練する方法であって、
第１の訓練サンプルを含む複数の訓練サンプルを取得するステップであり、各訓練サンプルは複数のセルフリーＤＮＡ（ｃｆＤＮＡ）断片を含むものである、前記訓練サンプルを取得するステップと、
各訓練サンプルについて、前記訓練サンプル中の前記ｃｆＤＮＡ断片に由来する配列リードを取得するステップと、
前記第１の訓練サンプルについて、前記第１の訓練サンプルの前記配列リードに基づいて、前記第１の訓練サンプルがホモ接合性ハプロタイプを有する複数の汚染マーカーから１つ以上の汚染マーカーを同定するステップと、
前記第１の訓練サンプルについて、同定された汚染マーカーの１つにおける、それぞれの前記汚染マーカーの前記ホモ接合性ハプロタイプとは異なるハプロタイプを有する、前記第１の訓練サンプル中のあらゆるｃｆＤＮＡ断片を、汚染ｃｆＤＮＡ断片として同定するステップと、
前記第１の訓練サンプルについて、同定されたあらゆる汚染ｃｆＤＮＡ断片に基づいて汚染レベルを推定するステップと、および
前記第１の訓練サンプルについて、前記汚染レベルが閾値レベル以下であるか否かを決定するステップと、並びに
前記第１の訓練サンプルの前記汚染レベルが前記閾値レベル以上であると決定することに応答して、前記複数の訓練サンプルから前記第１の訓練サンプルを除去するステップと、
を備え、前記第１の訓練サンプルを除外する前記複数の訓練サンプルが、検査サンプルに対するがん予測を生成するために、前記がん分類モデルを訓練するのに使用される、方法。
【請求項６８】
請求項６７に記載の方法において、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、方法。
【請求項６９】
請求項６８に記載の方法において、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、方法。
【請求項７０】
請求項６８～６９のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項７１】
請求項６８～７０のいずれか一項に記載の方法において、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、方法。
【請求項７２】
請求項６８～７１のいずれか一項に記載の方法において、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項７３】
請求項６８～７２のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、方法。
【請求項７４】
請求項６８～７３のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、方法。
【請求項７５】
請求項６７～７４のいずれか一項に記載の方法において、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、方法。

【請求項７６】
請求項７５に記載の方法において、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、方法。
【請求項７７】
請求項７５～７６のいずれか一項に記載の方法において、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、方法。
【請求項７８】
請求項７５～７７のいずれか一項に記載の方法において、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、方法。
【請求項７９】
請求項７５～７８のいずれか一項に記載の方法において、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、方法。
【請求項８０】
請求項７５～７９のいずれか一項に記載の方法において、前記複数の汚染マーカーは、表３の前記インデル部位を含む、方法。
【請求項８１】
請求項６７～８０のいずれか一項に記載の方法において、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、方法。
【請求項８２】
請求項６７～８１のいずれか一項に記載の方法において、汚染レベルを推定するステップは、さらに、前記第１の訓練サンプル中の同定された汚染ｃｆＤＮＡ断片の数、前記第１の訓練サンプルのシーケンシング深度、前記第１の訓練サンプル中のｃｆＤＮＡ断片の数、および前記汚染マーカーの数のうちの１つ以上に基づいて行われる、方法。
【請求項８３】
請求項６７～８２のいずれか一項に記載の方法において、前記複数の訓練サンプルは、非がんサンプルの第１のコホートおよびがんサンプルの第２のコホートを備え、前記がん分類モデルが、がんの存在の可能性を決定するために訓練される、方法。
【請求項８４】
請求項８３に記載の方法において、前記がんサンプルの第２のコホートは、第１のがんタイプを有する１つ以上のサンプルおよび、第２のがんタイプを有する１つ以上の追加サンプルを含み、ここで前記がん分類モデルは、前記第１のがんタイプの存在の第１の尤度および前記第２のがんタイプの存在の第２の尤度を決定するように訓練される、方法。
【請求項８５】
請求項６７～８４のいずれか一項に記載の方法において、前記がん分類モデルは、機械学習モデルである、方法。
【請求項８６】
請求項８５に記載の方法において、がん分類モデルは、決定木、ニューラルネットワーク、多層パーセプトロン、およびサポートベクターマシンのうちの少なくとも１つである、方法。
【請求項８７】
命令がコンピュータプロセッサによって実行されると、請求項６７～８６のいずれか一項に記載の方法をコンピュータプロセッサに実行させる前記命令を記憶した、非一過性のコンピュータ可読記憶媒体。
【請求項８８】
システムにおいて、
コンピュータプロセッサと、および
請求項８７に記載の前記非一過性のコンピュータ可読記憶媒体と
を備える、システム。
【請求項８９】

コンピュータプログラム製品において、
訓練された前記がん分類モデルを記憶する前記非一過性のコンピュータ可読記憶媒体を備え、前記コンピュータプログラム製品は、請求項６７～８６のいずれか一項に記載の方法により作成される、コンピュータプログラム製品。
【請求項９０】
治療キットにおいて、
個体からの遺伝物質を含む生物学的サンプルを保存するための１つ以上の収集容器と、および
複数の汚染マーカーを標的とする複数のプローブであって、表２および表４のうちの少なくとも一方を含むものである、前記複数のプローブと
を備える、治療キット。
【請求項９１】
請求項９０に記載の治療キットにおいて、前記複数の汚染マーカーは、複数の一塩基多型（複数のＳＮＰ）部位を含む、治療キット。
【請求項９２】
請求項９１に記載の治療キットにおいて、前記複数のＳＮＰ部位は、１０塩基対（ｂｐｓ）以内である、治療キット。
【請求項９３】
請求項９１～９２のいずれか一項に記載の治療キットにおいて、前記複数のＳＮＰ部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、治療キット。
【請求項９４】
請求項９１～９３のいずれか一項に記載の治療キットにおいて、前記複数のＳＮＰ部位は、グアニン－アデニン多型およびシトシン－チミン多型を除外する、治療キット。
【請求項９５】
請求項９１～９４のいずれか一項に記載の治療キットにおいて、各複数ＳＮＰ部位の前記ハプロタイプはハーディーワインベルグ平衡にある、治療キット。
【請求項９６】
請求項９１～９５のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記複数のＳＮＰ部位を含む、治療キット。
【請求項９７】
請求項９１～９６のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、表１の前記複数のＳＮＰ部位を含む、治療キット。
【請求項９８】
請求項９１～９７のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、挿入－欠失（インデル）部位を含む、治療キット。
【請求項９９】
請求項９８に記載の治療キットにおいて、前記インデル部位は、５ｂｐｓ～１０ｂｐｓである、治療キット。
【請求項１００】
請求項９８～９９のいずれか一項に記載の治療キットにおいて、前記インデル部位は、４５％～５５％の範囲内の集団ハプロタイプ頻度を有する、治療キット。
【請求項１０１】
請求項９８～１００のいずれか一項に記載の治療キットにおいて、各インデル部位の前記ハプロタイプはハーディーワインベルク平衡にある、治療キット。
【請求項１０２】
請求項９８～１０１のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、少なくとも５００、少なくとも１，０００、少なくとも１，５００、または少なくとも２，０００の前記インデル部位を含む、治療キット。
【請求項１０３】

請求項９８～１０２のいずれか一項に記載の治療キットにおいて、前記複数の汚染マーカーは、表３の前記インデル部位を含む、治療キット。
【請求項１０４】
請求項９０～１０３のいずれか一項に記載の治療キットにおいて、各汚染マーカーは、前記汚染マーカーの各ハプロタイプを標的とするように設計されたプローブを含む、治療キット。
【請求項１０５】
請求項９０～１０４のいずれか一項に記載の治療キットにおいて、さらに、
生物学的サンプル中の核酸断片を単離するための１つ以上の試薬を備える、治療キット。
【請求項１０６】
請求項９０～１０５のいずれか一項に記載の治療キットにおいて、さらに、
第１のコンピュータプログラム製品であって、
請求項２３に記載の、非一過性のコンピュータ可読記憶媒体、
請求項４７に記載の、非一過性のコンピュータ可読記憶媒体、
請求項６５に記載の、非一過性のコンピュータ可読記憶媒体、および
請求項８７に記載の、非一過性のコンピュータ可読記憶媒体
のうちの１つ以上を含む、前記第１のコンピュータプログラム製品を備える、治療キット。
【請求項１０７】
請求項９０～１０６のいずれか一項に記載の治療キットにおいて、さらに、
請求項８９に記載のコンピュータプログラム製品を備える、治療キット。
【誤訳訂正３】
【訂正対象書類名】図面
【訂正対象項目名】図３Ａ
【訂正方法】変更
【訂正の内容】

【図３Ａ】


サンプル中のセルフリー（ｃｆ）ＤＮＡ断片から
メチル化状態ベクターを生成する
300
被験者からサンプルを採取し、
ｃｆＤＮＡ断片を単離する
310
非メチル化シトシンをウラシルに
変換する単離されたｃｆＤＮＡ断片を処理する
320
シーケンシングライブラリーを準備する
330
特定の領域をターゲットにできるハイブリダイ
ゼーションプローブを用いてシーケンシング
ライブラリーを充実させる
335
シーケンスリードを得るためのシーケンス
340
読み取った配列を参照ゲノムにアライメント
することにより、複数のメチル化部位における
メチル化状態を決定する
350
ｃｆＤＮＡ断片中のメチル化部位の位置および
各メチル化部位のメチル化状態を
特定するメチル化状態ベクターを生成する
360
A

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/080431

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/6827
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, BIOSIS, Sequence Search, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LIU M.C. ET AL: "Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA", ANNALS OF ONCOLOGY, vol. 31, no. 6, 1 June 2020 (2020-06-01), pages 745-759, XP055809431, NL ISSN: 0923-7534, DOI: 10.1016/j.annonc.2020.02.011 the whole document ----- -/--	1,2,5, 18,25, 26,42, 43,49, 50,67

[x] Further documents are listed in the continuation of Box C. [x] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 April 2023	04/05/2023

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Botz, Jürgen

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/080431

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KLEIN E.A. ET AL: "Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set", ANNALS OF ONCOLOGY, vol. 32, no. 9, 1 September 2021 (2021-09-01), pages 1167-1177, XP93040807, NL ISSN: 0923-7534, DOI: 10.1016/j.annonc.2021.05.806 the whole document	1-107
A	ARAVANIS ALEXANDER A ET AL: "Development of plasma cell-free DNA (cfDNA) assays for early cancer detection: first insights from the Circulating Cell-Free Genome Atlas Study (CCGA)", CANCER RESEARCH, vol. 78, no. 13, Suppl. S, July 2018 (2018-07), pages LB-343, XP009543919, & ANNUAL MEETING OF THE AMERICAN-ASSOCIATION-FOR-CANCER-RESEARCH (AACR); CHICAGO, IL, USA; APRIL 14 -18, 2018 the whole document	1-107
X	WO 2020/232109 A1 (GRAIL INC [US]) 19 November 2020 (2020-11-19) the whole document	1,2,5, 18,25, 26,42, 43,49, 50,67
X	WO 2019/178277 A1 (GRAIL INC [US]) 19 September 2019 (2019-09-19) the whole document	1,2,5, 18,25, 26,42, 43,49, 50,67
X	WO 2020/132572 A1 (GRAIL INC [US]) 25 June 2020 (2020-06-25) the whole document	1,2,5, 18,25, 26,42, 43,49, 50,67
8 X	WO 2021/202423 A1 (GRAIL INC [US]) 7 October 2021 (2021-10-07) the whole document	1,2,5, 18,25, 26,42, 43,49, 50,67

1 -/--

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2022/080431

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2021/202424 A1 (GRAIL INC [US]) 7 October 2021 (2021-10-07) the whole document -----	1,2,5, 18,25, 26,42, 43,49, 50,67

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2022/080431

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2020232109 A1	19-11-2020	AU 2020274348 A1	09-12-2021
		CA 3136204 A1	19-11-2020
		CN 113826167 A	21-12-2021
		EP 3969622 A1	23-03-2022
		IL 286874 A	31-10-2021
		JP 2022532892 A	20-07-2022
		TW 202108774 A	01-03-2021
		US 2020365229 A1	19-11-2020
		WO 2020232109 A1	19-11-2020
WO 2019178277 A1	19-09-2019	AU 2019234843 A1	24-09-2020
		CA 3092998 A1	19-09-2019
		CN 111989407 A	24-11-2020
		EP 3765637 A1	20-01-2021
		TW 201938798 A	01-10-2019
		US 2019287652 A1	19-09-2019
		WO 2019178277 A1	19-09-2019
WO 2020132572 A1	25-06-2020	EP 3899953 A1	27-10-2021
		US 2020239965 A1	30-07-2020
		WO 2020132572 A1	25-06-2020
WO 2021202423 A1	07-10-2021	AU 2021248552 A1	03-11-2022
		CA 3169914 A1	07-10-2021
		CN 115335533 A	11-11-2022
		EP 4127231 A1	08-02-2023
		US 2021313006 A1	07-10-2021
		WO 2021202423 A1	07-10-2021
WO 2021202424 A1	07-10-2021	AU 2021248502 A1	03-11-2022
		CA 3170277 A1	07-10-2021
		CN 115461472 A	09-12-2022
		EP 4127232 A1	08-02-2023
		US 2021310075 A1	07-10-2021
		WO 2021202424 A1	07-10-2021

Form PCT/ISA/210 (patent family annex) (April 2005)

フロントページの続き

(51)国際特許分類	ＦＩ	テーマコード（参考）
	Ｇ１６Ｂ　40/00	

MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CV,CV,CZ,DE,DJ,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IQ,IR,IS,IT,JM,JO,JP,KE,KG,KH,KN,KP,KR,KW,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,WS,ZA,ZM,ZW

(72)発明者　シッダールタ バガリア
アメリカ合衆国 カリフォルニア州 ９４０２５ メンロー パーク オブライエン ドライブ １５２５

Ｆターム（参考）　4B029　AA07 BB11 BB20 FA15
4B063　QA01 QA13 QA19 QQ03 QQ42 QR32 QR55 QR62 QS32 QS36