(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2018-501539
(P2018-501539A)
(43)公表日　平成30年1月18日(2018.1.18)

(51)Int.Cl.			F I			テーマコード（参考）
G06F	***19/22***	***(2011.01)***	G06F	19/22		
C12N	*15/09*	*(2006.01)*	C12N	15/00	A	

審査請求 未請求　予備審査請求 未請求　（全 23 頁）

(21)出願番号	特願2017-521153 (P2017-521153)
(86)(22)出願日	平成27年10月15日 (2015.10.15)
(85)翻訳文提出日	平成29年5月26日 (2017.5.26)
(86)国際出願番号	PCT/US2015/055807
(87)国際公開番号	WO2016/061396
(87)国際公開日	平成28年4月21日 (2016.4.21)
(31)優先権主張番号	62/064,717
(32)優先日	平成26年10月16日 (2014.10.16)
(33)優先権主張国	米国 (US)

(71)出願人　517131237
カウンシル，　インコーポレイテッド
アメリカ合衆国　カリフォルニア　９４０８０，　サウス　サンフランシスコ，　キンボール　ウェイ　１８０
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100113413
弁理士　森下　夏樹
(74)代理人　100181674
弁理士　飯田　貴敏
(74)代理人　100181641
弁理士　石川　大輔
(74)代理人　230113332
弁護士　山本　健策

最終頁に続く

(54)【発明の名称】バリアントコーラー

(57)【要約】
基準ゲノム配列に対してゲノムサンプルからバリアントを読み取るためのプロセスおよびシステムが、提供される。例示的プロセスは、リードのセットを収集することと、リードからｋ－ｍｅｒグラフを生成することとを含む。例えば、ｋ－ｍｅｒグラフは、収集されたリードの全ての可能なサブストリングを表すように構築されることができる。ｋ－ｍｅｒグラフは、連続的グラフにまとめられ、可能なハプロタイプのセットが、連続的グラフから生成されてもよい。プロセスはさらに、共通シーケンサエラーのためのフィルタを提供するエラーテーブルを生成してもよい。プロセスは、次いで、ハプロタイプのセットおよび生成されたエラーテーブルに基づいて、ディプロタイプのセットを生成し、ディプロタイプのセットをスコアリングし、基準ゲノムからバリアントを識別してもよい。

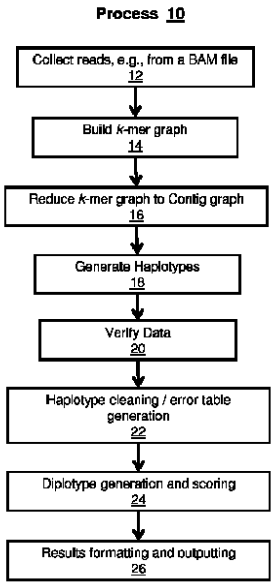

FIG. 1

【特許請求の範囲】
【請求項１】
基準ゲノム配列に対してゲノムサンプルからバリアントを判定するためのコンピュータ実装方法であって、
少なくとも１つのプロセッサおよびメモリを有する電子デバイスにおいて、
先に配列決定されたサンプルからの配列データのエラーテーブルにアクセスするステップと、
ゲノムサンプルから収集されたリードのセットから可能なハプロタイプのセットを判定するステップと、
前記可能なハプロタイプのセットおよび前記エラーテーブルに基づいて、ディプロタイプのセットを生成するステップであって、前記可能なハプロタイプのセットは、前記エラーテーブルによってフィルタリングされる、ステップと、
前記ディプロタイプのセットをスコアリングするステップと、
前記ディプロタイプのセットをスコアリングするステップに基づいて、バリアントを出力するステップと、を含む、方法。
【請求項２】
収集されたリードのセットからｋ－ｍｅｒグラフを生成するステップと、
前記生成されたｋ－ｍｅｒグラフを連続的グラフにまとめるステップと、
前記連続的グラフから前記可能なハプロタイプのセットを生成するステップと、
をさらに含む、請求項１に記載の方法。
【請求項３】
前記ディプロタイプのセットをスコアリングするステップはさらに、ディプロタイプ毎に事後確率を判定することを含む、請求項１に記載の方法。
【請求項４】
前記エラーテーブルを生成するステップをさらに含み、前記エラーテーブルを生成するステップは、
リードを基準サンプルに対してアライメントすることと、
リードが前記基準サンプルとミスマッチを有する部位を判定することと、
ミスマッチを有する部位を前記エラーテーブルに追加することと、を含む、請求項１に記載の方法。
【請求項５】
前記エラーテーブルを生成するステップはさらに、シーケンサエラーと関連付けられない部位を前記エラーテーブルからフィルタリングすることを含む、請求項４に記載の方法。
【請求項６】
前記エラーテーブルを生成するステップはさらに、Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ試験、Ｂａｙｅｓ　Ｆａｃｔｏｒ試験、またはＳｔｒａｎｄ　Ｂｉａｓ試験の１つまたはそれより多くを使用して、閾値に満たない部位を前記エラーテーブルからフィルタリングすることを含む、請求項４に記載の方法。
【請求項７】
配列データのエラーテーブルを生成するためのコンピュータ実装方法であって、
少なくとも１つのプロセッサおよびメモリを有する電子デバイスにおいて、
ゲノムサンプルから収集されたリードのセットから可能なハプロタイプのセットを判定するステップと、
前記収集されたリードのセットを基準サンプルに対してアライメントするステップと、
前記基準サンプルから前記収集されたリードのセットのリードがミスマッチを有する部位を判定するステップと、
ミスマッチを有する部位をエラーテーブルに追加するステップと、を含む、方法。
【請求項８】
前記可能なハプロタイプのセットを判定するステップは、
ｋ－ｍｅｒグラフを前記収集されたリードのセットから生成することと、

前記生成されたｋ－ｍｅｒグラフを連続的グラフにまとめることと、
前記連続的グラフから前記可能なハプロタイプのセットを判定することと、を含む、請求項７に記載の方法。
【請求項９】
非一過性コンピュータ可読記憶媒体であって、
先に配列決定されたサンプルからの配列データのエラーテーブルにアクセスするステップと、
ゲノムサンプルからの収集されたリードのセットから可能なハプロタイプのセットを判定するステップと、
前記可能なハプロタイプのセットおよび前記エラーテーブルに基づいて、ディプロタイプのセットを生成するステップであって、前記可能なハプロタイプのセットは、前記エラーテーブルによってフィルタリングされる、ステップと、
前記ディプロタイプのセットをスコアリングするステップと、
前記ディプロタイプのセットをスコアリングするステップに基づいて、バリアントを出力するステップとのためのコンピュータ実行可能命令を含む、非一過性コンピュータ可読記憶媒体。
【請求項１０】
収集されたリードのセットからｋ－ｍｅｒグラフを生成するステップと、
前記生成されたｋ－ｍｅｒグラフを連続的グラフにまとめるステップと、
前記連続的グラフから前記可能なハプロタイプのセットを生成するステップとをさらに含む、請求項９に記載の非一過性コンピュータ可読記憶媒体。
【請求項１１】
前記ディプロタイプのセットをスコアリングするステップはさらに、ディプロタイプ毎に事後確率を判定することを含む、請求項９に記載の非一過性コンピュータ可読記憶媒体。
【請求項１２】
前記エラーテーブルを生成するステップをさらに含み、前記エラーテーブルを生成するステップは、
リードを基準サンプルに対してアライメントすることと、
リードが前記基準サンプルとミスマッチを有する部位を判定することと、
ミスマッチを有する部位を前記エラーテーブルに追加することとを含む、請求項９に記載の非一過性コンピュータ可読記憶媒体。
【請求項１３】
前記エラーテーブルを生成するステップはさらに、シーケンサエラーと関連付けられない部位を前記エラーテーブルからフィルタリングすることを含む、請求項１２に記載の非一過性コンピュータ可読記憶媒体。
【請求項１４】
前記エラーテーブルを生成するステップはさらに、Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ試験、Ｂａｙｅｓ　Ｆａｃｔｏｒ試験、またはＳｔｒａｎｄ　Ｂｉａｓ試験の１つまたはそれより多くを使用して、閾値に満たない部位を前記エラーテーブルからフィルタリングすることを含む、請求項１２に記載の非一過性コンピュータ可読記憶媒体。
【請求項１５】
システムであって、
１つまたはそれより多くのプロセッサと、
メモリと、
１つまたはそれより多くのプログラムであって、前記１つまたはそれより多くのプログラムは、前記メモリ内に記憶され、前記１つまたはそれより多くのプロセッサによって実行されるように構成され、
先に配列決定されたサンプルからの配列データのエラーテーブルにアクセスするステップと、
ゲノムサンプルから収集されたリードのセットから可能なハプロタイプのセットを判定す

るステップと、
前記可能なハプロタイプのセットおよび前記エラーテーブルに基づいて、ディプロタイプのセットを生成するステップであって、前記可能なハプロタイプのセットは、前記エラーテーブルによってフィルタリングされる、ステップと、
前記ディプロタイプのセットをスコアリングするステップと、
前記ディプロタイプのセットをスコアリングするステップに基づいて、バリアントを出力するステップとのための命令を含む、１つまたはそれより多くのプログラムと、を備える、システム。
【請求項１６】
収集されたリードのセットからｋ－ｍｅｒグラフを生成するステップと、
前記生成されたｋ－ｍｅｒグラフを連続的グラフにまとめるステップと、
前記連続的グラフから前記可能なハプロタイプのセットを生成するステップと、
をさらに含む、請求項９に記載のシステム。
【請求項１７】
前記ディプロタイプのセットをスコアリングするステップはさらに、ディプロタイプ毎に事後確率を判定することを含む、請求項９に記載のシステム。
【請求項１８】
前記エラーテーブルを生成するステップをさらに含み、前記エラーテーブルを生成するステップは、
リードを基準サンプルに対してアライメントすることと、
リードが前記基準サンプルとミスマッチを有する部位を判定することと、
ミスマッチを有する部位を前記エラーテーブルに追加することとを含む、請求項９に記載のシステム。
【請求項１９】
前記エラーテーブルを生成するステップはさらに、シーケンサエラーと関連付けられない部位を前記エラーテーブルからフィルタリングすることを含む、請求項１８に記載のシステム。
【請求項２０】
前記エラーテーブルを生成するステップはさらに、
Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ試験、Ｂａｙｅｓ　Ｆａｃｔｏｒ試験、またはＳｔｒａｎｄ　Ｂｉａｓ試験の１つまたはそれより多くを使用して、閾値に満たない部位を前記エラーテーブルからフィルタリングすることを含む、請求項１８に記載のシステム。

【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願への相互参照）
本出願は、２０１４年１０月１６日に出願された、「ＶＡＲＩＡＮＴ　ＣＡＬＬＥＲ」と題する米国仮出願番号第６２／０６４，７１７号に基づく優先権を主張しており、その内容は、すべての目的のためにその全体が参考として本明細書によって援用される。
【０００２】
（分野）
本願は、概して、ＤＮＡシーケンサリード（ｒｅａｄ）におけるバリアントを識別および定量化するためのプロセスおよびシステムに関し、一実施例では、エラーテーブルの使用を通して基準ゲノム配列からバリアントを識別し、ハプロタイプエラーを除去し、次いで、ディプロタイプ（対のハプロタイプ）を生成およびスコアリングし、バリアントを判定するためのバリアントコーラープロセスおよびシステムに関する。
【背景技術】
【０００３】

（背景）
バリアントコーラー（ｖａｒｉａｎｔ　ｃａｌｌｅｒ）は、概して、基準ゲノム配列に対してＤＮＡ配列リード内のヌクレオチド差異が存在することを判定する。Ｐｌａｔｙｐｕｓ、Ｇｅｎｏｍｅ　Ａｎａｌｙｓｉｓ　Ｔｏｏｌｋｉｔ「ＧＡＴＫ」、およびＦｒｅｅｂａｙｅｓとして知られるものを含む、いくつかの公知のバリアントコーラーが存在する。例えば、Ｐｌａｔｙｐｕｓは、主に、リードの局所再アライメントおよびその局所アセンブリに依拠する、高スループット配列決定データ内のバリアント検出のためのシステムである。Ｐｌａｔｙｐｕｓは、「Ｉｎｔｅｇｒａｔｉｎｇ　ｍａｐｐｉｎｇ－，　ａｓｓｅｍｂｌｙ－　ａｎｄ　ｈａｐｌｏｔｙｐｅ－ｂａｓｅｄ　ａｐｐｒｏａｃｈｅｓ　ｆｏｒ　ｃａｌｌｉｎｇ　ｖａｒｉａｎｔｓ　ｉｎ　ｃｌｉｎｉｃａｌ　ｓｅｑｕｅｎｃｉｎｇ　ａｐｐｌｉｃａｔｉｏｎｓ」においてより詳細に説明されており、参照することによってその全体が本明細書に援用される。

【発明の概要】
【課題を解決するための手段】
【０００４】
（要旨）
一実施例では、基準ゲノム配列に対してゲノムサンプルからバリアントを読み取るためのコンピュータ実装プロセスが、提供される。このプロセスは、リードのセットを収集することと、リードからｋ－ｍｅｒグラフを生成することとを含む。例えば、ｋ－ｍｅｒグラフは、収集されたリードの全ての可能なサブストリングを表すように構築されることができる。ｋ－ｍｅｒグラフは、連続的グラフにまとめられ、可能なハプロタイプのセットが、連続的グラフから生成されてもよい。プロセスはさらに、エラーテーブルを生成してもよく（例えば、共通シーケンサエラーを識別するための多くの以前のサンプルから）、これは、共通シーケンサエラーのためのフィルタを提供する。プロセスは、次いで、ハプロタイプのセットおよびエラーテーブルに基づいて、ディプロタイプのセットを生成し、ディプロタイプのセットをスコアリングし、基準ゲノムからバリアントを識別してもよい。ディプロタイプをスコアリングすることは、ディプロタイプ毎に事後確率を判定することを含んでもよく、最高スコアリングディプロタイプが結果として報告される。
【０００５】
別の実施例では、配列データのエラーテーブルを生成するためのコンピュータ実装プロセスが、提供される。例示的プロセスは、少なくとも１つのプロセッサおよびメモリを有する電子デバイスにおいて、ゲノムサンプルから収集されたリードのセットから可能なハプロタイプのセットを判定するステップと、収集されたリードのセットを基準サンプルに対してアライメントするステップと、基準サンプルから、収集されたリードのセットのリードがミスマッチを有する部位を判定するステップと、ミスマッチを有する部位をエラーテーブルに追加するステップとを含んでもよい。可能なハプロタイプのセットを判定するステップは、ｋ－ｍｅｒグラフを、収集されたリードのセットから生成することと、生成されたｋ－ｍｅｒグラフを連続的グラフにまとめることと、連続的グラフから可能なハプロタイプのセットを判定することとを含んでもよい。
【０００６】
加えて、バリアントコーラーのためと、エラーテーブルを生成するためのシステム、電子デバイス、グラフィカルユーザインターフェース、および非一過性コンピュータ可読記憶媒体（説明される１つまたはそれより多くのプロセスを実行するためのプログラムおよび命令を含む、記憶媒体）が、説明される。

【図面の簡単な説明】
【０００７】
本願は、同一部分が同一数字によって参照され得る、付随の図面と関連して検討される以下の説明を参照することによって、最良に理解され得る。
【０００８】
【図１】図１は、一実施形態による、例示的コーリングプロセスを図示する。

【０００９】
【図２Ａ】図２Ａ－２Ｃは、図１のプロセスを参照して説明される例示的プロセスを図式的に図示する。
【図２Ｂ】図２Ａ－２Ｃは、図１のプロセスを参照して説明される例示的プロセスを図式的に図示する。
【図２Ｃ】図２Ａ－２Ｃは、図１のプロセスを参照して説明される例示的プロセスを図式的に図示する。
【００１０】
【図３】図３Ａおよび３Ｂは、異なるリードモデルのプロットを図示する。
【００１１】
【図４】図４は、本発明の種々の実施形態が動作し得る、例示的システムおよび環境を図示する。
【００１２】
【図５】図５は、例示的コンピューティングシステムを図示する。
【発明を実施するための形態】
【００１３】
以下の説明は、当業者が、種々の実施形態を作製および使用することが可能となるように提示される。具体的デバイス、技法、および用途の説明は、実施例のみとして提供される。本明細書に説明される実施例の種々の修正は、当業者に容易に明白となり、本明細書に定義される一般的原理は、本技術の精神および範囲から逸脱することなく、他の実施例および用途に適用されてもよい。したがって、開示される技術は、本明細書に説明および示される実施例に限定されず、請求項と一貫した範囲が与えられることが意図される。
【００１４】
本願は、概して、基準ゲノム配列からバリアントを識別するためのバリアントコーラーに関する。一実施例では、バリアントコーラーは、エラーテーブルを生成してエラーをハプロタイプから除去し、ディプロタイプを生成し、ディプロタイプをスコアリングし、基準ゲノム配列からバリアントを識別するためのプロセスを含む。バリアントコーラーの実施例は、Ｐｌａｔｙｐｕｓ、ＧＡＴＫ、Ｆｒｅｅｂａｙｅｓ、およびその他等の公知のコーラーに優るいくつかの進歩を提供し得る。例えば、全ての実施形態または実施例に存在しないが、進歩は、リード内のアライメントの代わりに（例えば、アライメントのためにリードを蓄積し、全リードを使用して、１つのグラフを作成する代わりに）、局所化と、共通シーケンサエラーを防ぐために、エラーテーブルを介して、エラー較正とを含んでもよい。
【００１５】
一実施形態では、バリアントコーラーは、いくつかの処理段階に分割され、各段階は、次の段階への入力としてその出力を提供する。以下の実施例は、配列データを記憶するためのバイナリフォーマットである、Ｂｉｎａｒｙ　Ａｌｉｇｎｍｅｎｔ／Ｍａｐフォーマット「ｂａｍ」または「ＢＡＭ」フォーマットの使用を仮定する。しかしながら、他のデータフォーマット（例えば、Ｓｅｑｕｅｎｃｅ　Ａｌｉｇｎｍｅｎｔ／ＭＡＰフォーマットまたは「ＳＡＭ」フォーマット）も、企図され、そして可能である。一実施例では、各ｂａｍファイル内の各領域の処理は、全ての他の領域およびｂａｍファイルとは完全に別個である。
【００１６】
広義には、かつ一実施例では、ある領域のためのコールを生成するために、図１にプロセス１０として図示される、以下のプロセスが、行われる。プロセス１０の説明と併せて、プロセス１０の種々の側面を図式的に図示する図２Ａ－２Ｃを基準されたい。
【００１７】
最初に、着目配列が、１２において得られる。例えば、リードが、コールの領域と何らかの点で重複するｂａｍファイルから収集されることができる。この処理は、ＢＷＡ、ＢＯＷＴＩＥ、ＭＡＸ等のショートリードアライナを使用して、図２Ａに図式的に図示され

るように、リード２１０をゲノム領域２２０に対してアライメントすることを含んでもよい。収集されたリードは、次いで、その関連付けられたソフトクリッピング情報を使用して、クリッピングされることができる。アライナからの補助情報、例えば、塩基間（ｂａｓｅ－ｔｏ－ｂａｓｅ）アライメント情報は、次いで、破棄されることができ、リードは、単に、塩基の配列となる。（いくつかの実施例では、マッピング品質に基づくフィルタリングが、随意に、行われることができる。）

【００１８】

ｋ－ｍｅｒグラフが、次いで、１４において、収集されたリードから構築され、ｋ－ｍｅｒグラフは、収集されたリードともに含まれる、長さｋの全ての可能なサブストリングを表す。例示的ｋ－ｍｅｒグラフは、図２Ｂに図示され、そこでは、ｋ＝３である（実際は、２０～３０のｋが、ｋ－ｍｅｒが一意であって、例えば、１カ所にのみ生じることを確実にするために使用されてもよい）。例えば、各リードは、ｋ－ｍｅｒおよびｋ－ｍｅｒ遷移を収集するために走査される。各エッジは、その関連付けられた遷移の確率でアノテーションされ、各ｋ－ｍｅｒは、エッジの起点として認められた回数でアノテーションされる。ｋ－ｍｅｒＡとＢとの間の遷移の確率は、ｋ－ｍｅｒＡが認められた合計回数で除算されたｋ－ｍｅｒＡに続くｋ－ｍｅｒＢが認められた回数である。

【００１９】

ｋ－ｍｅｒグラフは、次いで、１６において、処理の単純化のために、連続的（「コンティグ」）グラフにまとめられることができる。コンティググラフは、概して、ゲノム情報の領域をともに形成する、重複セグメントのセットを図示する。例えば、本ステップは、それらが常時、同一経路内で終端する場合、２つのｋ－ｍｅｒを結合することができる。加えて、ｋ－ｍｅｒグラフは、閾値回数未満（例えば、４回未満）認められる任意のｋ－ｍｅｒを破棄し、閾値を下回る（例えば、３％を下回る）確率を有する任意のエッジを破棄することによってフィルタリングされる。いったんｋ－ｍｅｒグラフが作成されると、それは、サイクル、すなわち、それ自体に収束する経路に関してチェックされることができる。グラフがサイクルを有する場合、破棄され、ｋが増加され、グラフが再構築されることができる。したがって、本実施例では、ｋ－ｍｅｒグラフは、サイクルを伴わずに構築される。

【００２０】

ハプロタイプ生成が、次いで、１８において、行われることができる。例えば、いったんコンティググラフが構築されると、ハプロタイプ候補のための開始点が、入エッジ（入次数０）を伴わない全コンティグを検索することによって見出されることができる。これらは、領域の開始時のコンティグであるはずであるが、領域の中央におけるコンティグもまた、雑音に起因して作成される場合、本特性を有し得る。次いで、それらのコンティグを開始点として見なし、コンティググラフを通して全ての可能な経路が、列挙され、各経路は、出エッジを伴わないコンティグ（終末）に到達すると終端する。次に進む前に、全経路は、それらのコンティグを結合することによって、ハプロタイプストリングに変換されることができる。簡略化された実施例は、図２Ｃに図示され、開始点は、「１」によって示され、「６」まで続く。各可能な経路は、可能なハプロタイプを生成し、そのうちの１つが、図に示される。

【００２１】

いったん可能なハプロタイプのセットが生成されると、例示的プロセスは、２０において、十分に良好なコールを作成するために十分なデータを有することを検証する（１つまたはそれより多くの仮説形成法（ｈｅｕｒｉｓｔｉｃｓ）を通して）。例えば、プロセスは、所望の領域内の各位置が、十分なｋ－ｍｅｒによって網羅され、領域全体を網羅する、少なくとも１つのハプロタイプが存在することをチェックする。これらのチェックのいずれかが失敗する場合、領域全体に対するコールは、発行され得ない。仮説形成法は、コール内の所望の信頼性に関して調節されることができることを理解されたい。

【００２２】

可能なハプロタイプのセットはさらに、２２において、任意のスコアリングプロセスの

前に、「精緻化」されることができる。コンティググラフから生成されたハプロタイプは、概して、出力またはスコアリングのために好適ではない。故に、一実施例では、スコアリングの前に、それらはいくつかの補正相を受ける。最初に、ハプロタイプは、コーラーが全重複リードを使用し、大部分のハプロタイプが、元々、着目領域のエッジを越えて延在し得るため、この領域にクリッピングされる。一実施例では、ハプロタイプをクリッピングするために、問題の領域に対してアライメントされ、アライメント外の任意の塩基は、破棄される。いったんハプロタイプがクリッピングされると、ハプロタイプ内のエラーが、補正されることができる。例えば、プロセスは、多くのサンプルから、共通シーケンサエラーをリスト化する、エラーテーブル（以下により詳細に説明される）を生成することができ、本エラーテーブルは、可能なハプロタイプのセットからそれらのエラーを除去するために使用されることができる。これらのステップは、複製を含むハプロタイプのセットをもたらし得、これらの複製は、ドロップされることができる。

【００２３】

ディプロタイプは、２４において、ハプロタイプから生成され、そしてスコアリングされることができる。例えば、Ｎ個のハプロタイプのセットが、全ての可能なディプロタイプを生成するために、それ自体と組み合わせられることができる。Ｎ個のハプロタイプに関して、Ｎ（Ｎ＋１）／２個の一意のディプロタイプが存在するであろう。これらのディプロタイプは、次いで、スコアリングされることができ、ディプロタイプのスコアは、その事後確率Ｐ（ｄｉｐｌｏｔｙｐｅ｜ｒｅａｄｓ）に等しい。最高スコアリングディプロタイプは、結果として報告されることができ、その信頼性は、最高確率と次に高い確率との間の比率の対数と等しい。ディプロタイプのスコアリングは、以下により詳細に説明される。

【００２４】

結果は、次いで、２６において、フォーマットされ（必要に応じて）、要求に応じて書き出されることができる。例えば、フォーマットがＪａｖａＳｃｒｉｐｔ Ｏｂｊｅｃｔｉｏｎ Ｎｏｔａｔｉｏｎ（「ｊｓｏｎ」または「ＪＳＯＮ」）またはＶａｒｉａｎｔ Ｃａｌｌ Ｆｏｒｍａｔ（「ｖｃｆ－ｆｕｌｌ」である場合、さらなる処理は、本実施例では、必要とされず、コールは、単に、ディスクに書き出される。しかしながら、結果フォーマットが、Ｖａｒｉａｎｔ Ｃａｌｌ Ｆｏｒｍａｔ－Ｓｉｎｇｌｅ Ｎｕｃｌｅｏｔｉｄｅ Ｐｏｌｙｍｏｒｐｈｉｓｍ（「ｖｃｆ－ｓｎｐ」）である場合、結果は、より小さいコールに分割され、これは、領域をその個々のＳＮＰおよび挿入欠失に分割する。ｖｃｆ－ｓｎｐフォーマットにおける単一コールは、異なるバリアントが相互にある距離（例えば、１０塩基）内にある、全バリアントから成る。

【００２５】

ディプロタイプのスコアリング

【００２６】

一実施例では、前述のＮ個のハプロタイプのセットは、全ての可能なディプロタイプを生成するために、それ自体と組み合わせられることができる。Ｎ個のハプロタイプに関して、Ｎ（Ｎ＋１）／２個の一意のディプロタイプが存在し得る。これらのディプロタイプは、次いで、スコアリングされる；ディプロタイプのスコアは、その事後確率Ｐ（ｄｉｐｌｏｔｙｐｅ｜ｒｅａｄｓ）に等しい。最高スコアリングディプロタイプは、結果として報告されることができ、その信頼性は、最高確率と次に高い確率との間の比率の対数と等しい。

【００２７】

最良ディプロタイプを候補のリストから判定するために使用される例示的確率スコアリングモデルが、ここで説明される。一実施例では、各ディプロタイプに割り当てられるスコアは、ディプロタイプの事後確率Ｐ（ｄｉｐｌｏｔｙｐｅ｜ｒｅａｄｓ）である。スコアリングのために使用される確率は典型的には、小さいため、一実装では、対数確率が使用される。事後確率は、以下のように、尤度および事前確率に分解されることができる。

【数１】

$$P(diplotype \mid reads) = (1/Z)\, P(reads \mid diplotype)\, P(diplotype)$$

式中、Ｚ＝Ｐ（ｒｅａｄｓ）は、ある正規化定数であって、算出されない。Ｚは、ディプロタイプから独立するため、２つのディプロタイプを比較する目的のために無視されることができる。事前確率Ｐ（ｄｉｐｌｏｔｙｐｅ）および尤度Ｐ（ｒｅａｄｓ｜ｄｉｐｌｏｔｙｐｅ）が、次いで、別個に算出されることができる。

【００２８】

事前確率を算出するために、本実施例では、大部分の領域は、基準に類似すると仮定され得る。ディプロタイプの確率は、したがって、ディプロタイプが基準から生物学的変異を介して生成された確率である。本実施例は、これが、単に、基準から生成されているハプロタイプの確率の積と仮定する（選択に起因して、完全に正確ではないが、概して、十分であることを理解されたい）。したがって、ディプロタイプの確率は、以下のように表されることができる。

【数２】

$$P(diplotype) = P(haplotype\_1)\, P(haplotype\_2)$$

【００２９】

生成されているハプロタイプの確率は、それが全ての可能な方法において生成されている確率の総和であって、ハプロタイプと基準の各可能なアライメントは、ハプロタイプを生成する異なる方法に対応する。しかしながら、全アライメントにわたって総和を行うことは、算出上、扱いにくくあり得、したがって、本実施例は、確率質量の大部分が単一アライメント内に含有され、最高確率を有するものと仮定する。したがって、Ｐ（ｈａｐｌｏｔｙｐｅ）を算出するために、プロセスは、ハプロタイプを基準に対してアライメントする。アライメントの間に使用されるマッチ、ミスマッチ、ギャップ開放、およびギャップ伸長パラメータは、生物学的変異に起因して生じるような事象の対数確率に対応する。アライメントは、スコアを最大限にするため、対数確率を最大限にし、したがって、最高確率アライメントをもたらし得る。例えば、１塩基変化は、約１，０００塩基毎に生じ、したがって、ミスマッチパラメータは、ｌｏｇ（１／１０００）となり得る。

【００３０】

尤度Ｐ（ｒｅａｄｓ｜ｄｉｐｌｏｔｙｐｅ）の算出は、類似プロセスを使用する。最初に、実施例は、全リードが独立であると仮定し、これは、尤度が以下のように書き直されることを可能にする。

【数３】

$$P(reads \mid diplotype) = product\_i\, \{\, P(read\_i \mid diplotype)\, \}$$

【００３１】

次いで、実施例は、リードが、２つのハプロタイプのディプロタイプから生じ得る（等確率を伴う）か、またはゲノム内の別の場所から無作為に生成され得る（非常に低確率を伴う）かのいずれかであると仮定する。後者の場合は、アライナエラーおよび稀な外れ値を効果的にモデル化する。したがって、リードの確率は、以下のように表されることができる。

【数４】

$$P(read \mid diplotype) = epsilon\, P(read\ is\ random) + (0.5 - epsilon)\, P(read \mid haplotype\_1) + (0.5 - epsilon)\, P(read \mid haplotype\_2)$$

【００３２】

無作為に生成されたリードの確率は、４つの等しい可能性の塩基が存在するため、生成された各塩基に等しい。

【数５】

$$P(\text{read is random}) \sim= 0.25 \wedge \text{len(read)}$$

【００３３】

ハプロタイプが与えられたリードの確率は、アライメントを使用して見出されることができる。本実施例は、ハプロタイプが、基となるゲノムの真の配列であって、リードが、エラーを含んだ配列決定プロセスを使用して、この配列から生成されると仮定する。したがって、アライメントパラメータは、シーケンサエラーの率であるはずである。例えば、ミスマッチパラメータは、任意の塩基においてシーケンサが１つの塩基変化を作成する確率の対数であるはずである。事前確率と同様に、プロセスは、最良アライメントを算出し、スコアを確率として使用する。

【００３４】

他のスコアリングプロセスが、ここで説明されるものの代わりに、またはそれに加えて、使用されてもよく、例えば、他のパラメータ、値、仮定、および算出プロセスを含むことが、当業者によって理解されるはずである。

【００３５】

エラーテーブル生成

【００３６】

一般に、そして一実施例では、エラーテーブルは、共通シーケンサエラーを防ぐためのフィルタのように作用し、これは、いくつかの領域を別様にコーリングすることを非常に困難にし得る。一実施例では、エラーテーブルを生成するために、同一領域に対してデータを含む、数百（例えば、１００～３００またはそれを上回る）サンプルが、使用される。本実施例では、所与の領域に対するエラーテーブル生成は、以下のステップを受ける。

１．　サンプル毎に、リードを基準に対してアライメントする。基準内の塩基毎に、異なるバリアントがそこで認められる回数をカウントする（バリアントは、４つの塩基、異なる長さ欠失、および異なる挿入である）。本プロセスは、フォワードリードおよびバックワードリードに関して別個に行われることができる。

２．　ある閾値を上回るバリアントが存在する、すなわち、ある閾値％のリードが非基準対立遺伝子を有する、部位を見出す。例えば、閾値は、１％であることができる。これらの部位は、エラーテーブルに入る候補部位である。

３．　次に、エラーテーブル部位が、フィルタリングされる。フィルタリングにおける例示的ステップは、以下の次の節においてより詳細に説明される。

４．　フィルタは、部位のいくつかをエラーテーブルから除去する。フィルタリング後、部位は、Ｓｉｎｇｌｅ　Ｎｕｃｌｅｏｔｉｄｅ　Ｐｏｌｙｍｏｒｐｈｉｓｍ　Ｄａｔａｂａｓｅ「ｄｂＳＮＰ」（および潜在的に、複数のｄｂＳＮＰ　Ｖａｒｉａｎｔ　Ｃａｌｌｅｒ　Ｆｏｒｍａｔｓ「ＶＣＦ」）と比較される。ｄｂＳＮＰ内に生じ、共通である、任意の部位は、エラーテーブルから除去されることができる。

５．　エラーテーブルは、大容量ＪＳＯＮファイルとしてディスクに書き込まれ、部位毎の記録は、基準塩基および各代替塩基の頻度を示す。例えば、１％を上回る頻度を伴う任意の代替塩基は、フィルタリングされてもよい。フィルタリングのためのカットオフは、システム自体内で構成可能であることができ、したがって、エラーテーブル内にあっても、フィルタリングを保証するために十分ではない。しかしながら、カットオフは、非常に類似する。例えば、プロセスは、エラーテーブル内にある１．５％を上回る頻度を伴う任意のものをフィルタリングすることができる。

【００３７】

エラーテーブルは、着目領域毎に１回生成され、次いで、後の使用のために記憶されることができる。

【００３８】

エラーテーブルフィルタリング統計

【００３９】

エラーテーブル生成プロセスのステップ３（前述）に記載のように、高相違部位は全て、エラーテーブルに対する候補である。候補部位は、一連の統計的試験を通して（ならびにｄｂＳＮＰとの比較を通して）フィルタリング除去されることができる。以下は、２つの例示的試験を含む、候補エラーテーブル部位をフィルタリングするために使用される例示的手順を説明する。

【００４０】

最初に、各部位毎に、Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ試験統計が、算出されることができる。これは、非常にネイティブなジェノタイピングによって行われることができる。例えば、塩基が、リードの２０％未満のサンプル内に認められる場合、ホモ接合型基準（「ＨＯＭ　ＲＥＦ」）と見なされ、リードの２０％～７５％に認められる場合、ヘテロ接合型（「ＨＥＴ」）と見なされ、リードの７５％を上回って認められる場合、ホモ接合型代替（「ＨＯＭ　ＡＬＴ」）と見なされる。次いで、サンプルは、これらの３つのカテゴリ（ＨＯＭ　ＲＥＦ、ＨＥＴ、およびＨＯＭ　ＡＬＴ）にビン化され、Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ試験は、０．５％のアルファに対して標準的カイ二乗統計を使用して行われる。したがって、エラーテーブル内の本部位が実在のＳＮＰから由来し得る可能性が存在する場合、エラーテーブルからの除去が検討される。

【００４１】

しかしながら、これらの部位は、本実施例では、エラーテーブルから直ちに除去されない。エラーテーブルから除去されるためにはまた、Ｂａｙｅｓ　Ｆａｃｔｏｒ試験にも合格しなければならない。Ｂａｙｅｓ　Ｆａｃｔｏｒ試験は、以下のように、２つの異なるモデル、すなわち、ＳＮＰモデルおよび雑音モデルを前提として、データの確率の比率を算出する。

【数６】

$$B = P(data \mid SNP\ model) / P(data \mid noise\ model)$$

【００４２】

Ｂａｙｅｓ　Ｆａｃｔｏｒが高い（例えば、１０を上回る）場合、データは、ＳＮＰモデルに由来するより高い確率を有し、したがって、部位は、エラーテーブルから除去される。

【００４３】

２つのモデルは、リードフラクション分布のモデルである。対立遺伝子の頻度が２０％である場合、対立遺伝子は、雑音であり得、サンプル内の頻度の分布は全て、約２０％となるであろう。すなわち、各サンプル内において、リードの約２０％は、本対立遺伝子を有するであろう。代替として、対立遺伝子は、実在し得、その場合、いくつかのサンプルは、１００％に近い対立遺伝子を有し、いくつかのサンプルは、０％を有し、いくつかのサンプルは、５０％を有するであろう（ＨＯＭ　ＡＬＴ、ＨＯＭ　ＲＥＦ、およびＨＥＴに対応する）。

【００４４】

これらの２つのモデルは、異なる数のパラメータを有する。概して、雑音モデルでは、リード内で雑音を観察する確率（観察される対立遺伝子頻度に対応する）が、必要とされ、ＳＮＰモデルでは、ＨＯＭ　ＡＬＴ、ＨＯＭ　ＲＥＦ、およびＨＥＴサンプルの確率（これらの２つは、１つに総和されなければならないため、２つのみのパラメータ）が、必要とされる。モデルと異なる数のパラメータを比較するために、パラメータは、積分されることができる。したがって、Ｐ（ｄａｔａ｜ｎｏｉｓｅ　ｍｏｄｅｌ）を算出するために、プロセスは、雑音確率の全ての可能な値（０から１）にわたってＰ（ｄａｔａ｜ｎｏｉｓｅ　ｍｏｄｅｌ，ｎｏｉｓｅ　ｐｒｏｂａｂｉｌｉｔｙ）を積分することができる。同様に、Ｐ（ｄａｔａ｜ＳＮＰ　ｍｏｄｅｌ）を算出するために、プロセスは、ｈｏｍ　ｒｅｆおよびｈｅｔ　ｐｒｏｐｏｒｔｉｏｎｓの全ての可能な値にわたってＰ（ｄａｔａ｜ＳＮＰ　ｍｏｄｅｌ，ｈｏｍ　ｒｅｆ　ｐｒｏｐｏｒｔｉｏｎ，ｈｅｔｐｒｏｐｏｒｔｉｏｎ）を積分することができる（ｈｏｍ　ａｌｔ　ｐｒｏｐｏｒｔｉｏｎは、１マイナ

スそれらの２つである）。（積分の面積は、それらの３つの総和がちょうど１であって、それらのいずれも［０，１］範囲外にないように制約される。）本積分は、Ｓｃｉｅｎｔｉｆｉｃ　Ｐｙｔｈｏｎ「ＳｃｉＰｙ」数値積分関数（または均等物）を使用して実装されることができる。

【００４５】

これらモデル（雑音およびＳＮＰモデル）は両方とも、リードがある種類のＢｅｒｎｏｕｌｌｉ分布から求められている、すなわち、プロセスは、ある確率ｐを用いて、問題の対立遺伝子を認めるかどうかという仮定に基づく。雑音モデルに関して、ｐは、パラメータ（雑音確率）であって、プロセスは、そのｐにわたって積分する。確率Ｐ（ｄａｔａ｜ｎｏｉｓｅ　ｍｏｄｅｌ，ｐ）は、二項分布確率質量関数を使用することによって算出されることができ、ｐは、プロセスが当該対立遺伝子を認める確率である。ＰＭＦに対するｘおよびｎパラメータは、単に、対立遺伝子が認められた回数およびサンプル内の合計リード数である。これは、所与のサンプルの確率を算出することを可能にし、データセット内の全サンプルにわたって全それらの確率をともに乗算することは、パラメータｐを前提としたモデルの全体的確率を提供する。（注記：例示的計算におけるアンダーフローを回避するために、プロセスは、各確率を１０で乗算してもよい。したがって、算出される確率は、１０＾Ｎでスケーリングされ、Ｎは、データセット内のサンプルの数である。）

【００４６】

ＳＮＰモデルに関して、例示的プロセスは、サンプルがＨＯＭ　ＲＥＦである可能性に関するものと、ＨＥＴに関するものと、ＨＯＭ　ＡＬＴに関するものとの３つの二項分布を含む。しかしながら、各場合において、プロセスは、サンプルがＨＯＭ　ＲＥＦまたはＨＯＭ　ＡＬＴである場合でも、汚染が、依然として、ある基準をもたらし得るため、確率ｐを把握しない。同様に、ＨＥＴの場合も、汚染および他の影響（マッピング品質等）が、正確に５０％ではないｐをもたらし得る。これに対処するために、プロセスは、ｐをベータ分布を伴う無作為変数であるようにし得る。すなわち、ｐの全ての可能な値にわたって積分することは、ベータ二項分布を与え、これは、ＳＮＰモデル内のこれらの３つの場合における単純二項の代わりに使用され得る。事前確率情報（ＨＯＭ　ＲＥＦ、ＨＥＴ、またはＨＯＭ　ＡＬＴである）をモデル化するために、プロセスは、ベータ事前確率のために、アルファおよびベータパラメータを使用して、分布を適切に我々の歪ませることができる。ＨＯＭ　ＲＥＦおよびＨＯＭ　ＡＬＴの場合、プロセスは、アルファ＝２０およびベータ＝１（またはその逆）を使用して、図３Ａに示されるもののようなプロットをもたらすことができる。ＨＥＴの場合、プロセスは、アルファ＝２０およびベータ＝２０を使用して、図３Ｂに示されるもののようなプロットをもたらすことができる。

【００４７】

Ｂａｙｅｓ　Ｆａｃｔｏｒ試験に不合格の任意の部位は、Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ比例計算において生じた雑音であると仮定され、したがって、エラーテーブル内に保たれる。

【００４８】

Ｂａｙｅｓ　Ｆａｃｔｏｒ試験に加え、一実施例では、部位がエラーテーブルから外されるためには、Ｓｔｒａｎｄ　Ｂｉａｓ試験に合格しなければならない。Ｓｔｒａｎｄ　Ｂｉａｓ試験は、非常に単純である。すなわち、基準に関するリードおよび対立遺伝子に関するリードが、どのストランドのトラックがカウントされているのかを維持しながら、全サンプルにわたって集められる。全体的対立遺伝子頻度ｐもまた、算出される。次いで、フォワードリードの確率を算出し（それらが確率ｐを用いた二項分布に由来すると仮定する）、バックワードリードに関する同一確率を算出する。それらの確率の比が、非常に高いか、または非常に低い場合、対立遺伝子の分布が一方のストランドまたは他方のストランドに向かって非常にバイアスされていることを示す。したがって、その比の対数がある閾値を上回る（例えば、１０を上回る）大きさを有する場合、部位は、ストランドバイアスされていると見なされ、エラーテーブル内に含まれる。

【００４９】

故に、一実施例では、部位が、Ｈａｒｄｙ－Ｗｅｉｎｂｅｒｇ試験、Ｂａｙｅｓ　Ｆａｃｔｏｒ試験、およびＳｔｒａｎｄ　Ｂｉａｓ試験に合格する場合、エラーテーブル候補部位から除去される。

【００５０】

種々の他の試験または試験の組み合わせも、エラーテーブルを生成（またはフィルタリング）するために採用されてもよいことを認識されたい。さらに、他の変数または閾値も、本明細書に説明される実施例とともに採用され、シーケンサエラーと実在バリアントとの間の差異を判定してもよい。

【００５１】

コマンドラインインターフェース：

【００５２】

以下の節は、例示的バリアントコーラーと、それとともに提供され得るツールの実践的インストールおよび使用を説明する。本明細書に説明される例示的バリアントコーラーは、標準的Ｐｙｔｈｏｎパッケージとして実装されることができる（一実施例では、唯一の依存物は、配列アライメントのためのＣ＋＋ライブラリｓｅｑａｎである）。当然ながら、当業者は、他のプログラミング言語、データフォーマット、および同等物も、可能性として考えられ、検討されることを認識するであろう。

【００５３】

一実施例では、例示的バリアントコーラーは、エラー補正のために、事前に構築されたエラーテーブル（例えば、本明細書に説明されるように）に依拠する。エラーテーブルを生成するために、プロセスは、コーリングのための領域に関するデータを伴う複数のサンプル（例えば、数百サンプルまたはそれを上回る）を収集する。エラーテーブルが、次いで、以下の例示的コマンドを介して、具体的領域に関して生成されることができる（ｃｈｒ１：１００－２００等）。

【数７】

```
python -m kcall gen-table
    --reference /path/to/hg19.fa
    --output my_error-table.err
     --from /directory/with/bam/files
    --threads $NTHREADS
    --region chr1:100-200
--dbsnp dbsnp.vcf
```

【００５４】

代替として、プロセスは、以下の＊．ｂｅｄファイルを提供することができる。

【数８】

```
python -m kcall gen-table
    --reference /path/to/hg19.fa
    --output my_error-table.err
    --from /directory/with/bam/files
    --threads $NTHREADS
    --bed /path/to/my/bedfile.bed
--dbsnp dbsnp.vcf
```

【００５５】

最後に、ディレクトリの代わりに、＊．ｂａｍファイルのリストを用いて、プロセスは、－－ｆｒｏｍの代わりに、そのリストを提供することができる。

【数９】

```
python -m kcall gen-table
    --reference /path/to/hg19.fa
    --output my_error-table.err
    --from /path/to/list-of-bam-files.txt
    --threads $NTHREADS
    --bed /path/to/my/bedfile.bed
--dbsnp dbsnp.vcf
```

【００５６】

ユーザが、クラスタ内のいくつかのノードにわたってエラーテーブル生成を並列処理することを所望する場合、プロセスは、＊．ｂｅｄファイル内の領域毎に別個のジョブを引き起こすことができる。プロセスは、次いで、生成された断片の全てを単一テーブルに組み合わせることができる。エラーテーブルが単純ｊｓｏｎフォーマットであるため、プロセスは、ｊｑツールを使用して、これを行うことができる。
＃全エラーテーブル断片は、ｐｉｅｃｅｓ／ａｓ　ｊｓｏｎ　ｆｉｌｅｓ．ｃａｔ　ｐｉｅｃｅｓ／＊．ｊｓｏｎ｜ｊｑ－ｓａｄｄ＞ｃｏｍｂｉｎｅｄ＿ｔａｂｌｅ．ｊｓｏｎ”内に記憶されると仮定する。

【００５７】

エラーテーブルが生成されると、プロセスは、以下のコマンドを用いて、Ｋｃａｌｌバリアントコーラーを起動することができる。

【数１０】

```
python -m kcall call
    --reference /path/to/hg19.fa
    --errors my_error-table.jso
    --bam /path/to/sample.bam
    --threads $NTHREADS
    --bed /path/to/bed/file.bed
    --output-json output.json
    --output-vcf-full full.vcf
 --output-vcf-snp snp.vcf
```

【００５８】

例示的バリアントコーラーは、前述に示される対応するフラグの下で、少なくとも３つのフォーマット、例えば、ｊｓｏｎ、ｖｃｆ－ｓｎｐ、およびｖｃｆ－ｆｕｌｌにおいて出力を提供することができる。プロセスは、これらのフラグの任意のサブセットを有してもよい。すなわち、いずれも提供されない場合、プロセスは、ｖｃｆ－ｓｎｐフォーマットを出力し、標準化する。ｊｓｏｎフォーマットは、概して、最も単純であって、単に、ディクショナリを伴うＪＳＯＮファイルをもたらし、各キーは、領域を記述するストリング（「ｃｈｒ１：１００－２００」等）であって、値は、無コール理由を記述するストリング（領域がコーリングされない場合）、またはディプロタイプおよび領域に関する配列を提供する信頼キーを伴うディクショナリのいずれかである。ｖｃｆ－ｆｕｌｌフォーマットは、ＶＣＦと同一情報を出力し、各領域は、ちょうど１行に対応する。無コールについての情報は、ＶＣＦから利用可能である（ジェノタイプＧＴフィールドが．／．となるであろうため）が、無コール理由は、ＪＳＯＮ出力フォーマットから利用可能であることに留意されたい。最後に、ｖｃｆ－ｓｎｐフォーマットは、個々のハプロタイプコールを介して、出力ＶＣＦを分割し、それらが数塩基の分離より近い場合、ＳＮＰＳをともに結合する。これは、ＧＡＴＫおよびＦｒｅｅｂａｙｅｓに類似するコールを生成する。

【００５９】

いったん例示的バリアントコーラーがコールを生成すると、プロセスは、それらを別の

コールのセットと比較することができる。例えば、バリアントコーラーは、本目的のために、基準ゲノム内のそれらの場所によってインデックス化された塩基毎の差異を見出す、積分比較ツールを含んでもよい。これは、プロセスが、ＶＣＦと異なる出力フォーマットを比較することを可能にし、したがって、コールセットは、容易に、Ｆｒｅｅｂａｙｅｓ、ＧＡＴＫ１、またはＧＡＴＫ２コールセットと比較されることができる。２つのＶＣＦを比較するために、以下のコマンドが、使用されることができる。

【数１１】

```
python -m kcall compare first_vcf.vcf second_vcf.vcf
    --reference /path/to/hg19.fa
    --output output.diff
    --stats output.stats
    --name $SAMPLE_NAME
    --bed /path/to/bed/file.bed
```

【００６０】

生成された出力は、前述の２つのタブ分離テーブル（ｏｕｔｐｕｔ．ｄｉｆｆおよびｏｕｔｐｕｔ．ｓｔａｔｓ）内に含有される。これらの２つのＴＳＶファイルは、それぞれ、２つのコールセット間の差異および差異の頻度についてのいくつかの統計を含有する。

【００６１】

例示的アーキテクチャおよび処理環境：

【００６２】

本明細書に説明されるシステムおよびプロセスのある側面および実施例が動作し得る、例示的環境およびシステムが、ここで説明される。図４に示されるように、いくつかの実施例では、本システムは、クライアント－サーバモデルに従って実装されることができる。本システムは、ユーザデバイス１０２上で実行されるクライアント側部分と、サーバシステム１１０上で実行されるサーバ側部分とを含むことができる。ユーザデバイス１０２は、デスクトップコンピュータ、ラップトップコンピュータ、タブレットコンピュータ、ＰＤＡ、携帯電話（例えば、スマートフォン）、または同等物等の任意の電子デバイスを含むことができる。

【００６３】

ユーザデバイス１０２は、インターネット、イントラネット、または任意の他の有線もしくは無線公共もしくはプライベートネットワークを含み得る、１つまたはそれより多くのネットワーク１０８を通して、サーバシステム１１０と通信することができる。ユーザデバイス１０２上の例示的システムのクライアント側部分は、ユーザ対応入力および出力処理ならびにサーバシステム１１０との通信等、クライアント側機能性を提供することができる。サーバシステム１１０は、個別のユーザデバイス１０２上に常駐する任意の数のクライアントのために、サーバ側機能性を提供することができる。さらに、サーバシステム１１０は、クライアント対応Ｉ／Ｏインターフェース１２２と、１つまたはそれより多くの処理モジュール１１８と、データおよびモデル記憶１２０と、外部サービスとのＩ／Ｏインターフェース１１６とを含み得る、１つまたはそれより多くのコーラーサーバ１１４を含むことができる。クライアント対応Ｉ／Ｏインターフェース１２２は、コーラーサーバ１１４のためのクライアント対応入力および出力処理を促進することができる。１つまたはそれより多くの処理モジュール１１８は、本明細書に説明されるように、種々の問題および候補スコアリングモデルを含むことができる。いくつかの実施例では、コーラーサーバ１１４は、タスク完了または情報取得のために、ネットワーク１０８を通して、テキストデータベース、サブスクリプションサービス、政府記録サービス、および同等物等の外部サービス１２４と通信することができる。外部サービスとのＩ／Ｏインターフェース１１６は、そのような通信を促進することができる。

【００６４】

サーバシステム１１０は、コンピュータの１つまたはそれより多くの独立型データ処理

デバイスまたは分散型ネットワーク上に実装されることができる。いくつかの実施例では、サーバシステム１１０は、第三者サービスプロバイダ（例えば、第三者クラウドサービスプロバイダ）の種々の仮想デバイスおよび／またはサービスを採用し、サーバシステム１１０の下層コンピューティングリソースおよび／またはインフラストラクチャリソースを提供することができる。

【００６５】

コーラーサーバ１１４の機能性は、クライアント側部分およびサーバ側部分の両方を含むように図４に示されるが、いくつかの実施例では、本明細書に説明されるある機能（例えば、ユーザインターフェース特徴およびグラフィカル要素に関して）は、ユーザデバイス上にインストールされた独立型アプリケーションとして実装されることができる。加えて、システムのクライアント部分とサーバ部分との間における機能性の分割は、異なる実施例において変動することができる。例えば、いくつかの実施例では、クライアント上で実行されるユーザデバイス１０２は、ユーザ対応入力および出力処理機能のみを提供し、システムの全ての他の機能性をバックエンドサーバに委譲する、シンクライアントであることができる。

【００６６】

サーバシステム１１０およびクライアント１０２はさらに、例えば、処理ユニット、メモリ（本明細書に説明される機能の一部または全部を実施するための論理またはソフトウェアを含み得る）、および通信インターフェース、ならびに他の従来のコンピュータ構成要素（例えば、キーボード／タッチスクリーン等の入力デバイスおよびディスプレイ等の出力デバイス）を有する、種々のタイプのコンピュータデバイスの任意の１つを含んでもよいことに留意されたい。さらに、サーバシステム１１０およびクライアント１０２の一方または両方は、概して、論理（例えば、ｈｔｔｐウェブサーバ論理）を含む、またはローカルもしくは遠隔データベースもしくは他のデータおよびコンテンツのソースからアクセスされる、データをフォーマットするようにプログラムされる。この目的を達成するために、サーバシステム１１０は、共通ゲートウェイインターフェース（ＣＧＩ）プロトコルおよび関連付けられたアプリケーション（または「スクリプト」）、Ｊａｖａ（登録商標）「ｓｅｒｖｌｅｔｓ」、すなわち、サーバシステム１１０上で起動するＪａｖａ（登録商標）アプリケーション、または情報を提示し、クライアント１０２からの入力を受信するための同等物等の種々のウェブデータインターフェース技法を利用してもよい。サーバシステム１１０は、単数形で本明細書に説明されるが、実際には、複数のコンピュータ、デバイス、データベース、関連付けられたバックエンドデバイス、および同等物を備え、（有線および／または無線で）通信し、本明細書に説明される機能の一部または全部を行うように協働してもよい。サーバシステム１１０はさらに、アカウントサーバ（例えば、電子メールサーバ）、モバイルサーバ、メディアサーバ、および同等物を含むか、またはそれと通信してもよい。

【００６７】

さらに、本明細書に説明される例示的方法およびシステムは、種々の機能を行うために別個のサーバおよびデータベースシステムの使用を説明するが、他の実施形態も、説明される機能性が行われる限り、設計選択肢上、説明される機能を単一デバイスまたは複数のデバイスの任意の組み合わせ上で生じさせるように動作する、ソフトウェアまたはプログラミングを記憶することによって実装され得ることに留意されたい。同様に、説明されるデータベースシステムも、単一データベース、分散型データベース、分散型データベースの集合、冗長オンラインもしくはオフラインバックアップもしくは他の冗長性を伴うデータベース、または同等物として実装されることができる、分散型データベースまたは記憶ネットワークおよび関連付けられた処理インテリジェンスを含むことができる。図に描写されないが、サーバシステム１１０（ならびに本明細書に説明される他のサーバおよびサービス）は、概して、限定ではないが、プロセッサ、ＲＡＭ、ＲＯＭ、クロック、ハードウェアドライバ、関連付けられた記憶、および同等物を含む、サーバシステム内で通常見出されるような当該技術で認識される構成要素を含む（例えば、以下に論じられる図５参

照）。さらに、説明される機能および論理は、ソフトウェア、ハードウェア、ファームウェア、またはそれらの組み合わせ内に含まれてもよい。

【００６８】

図５は、種々のコーリングおよびスコアリングモデルを含む前述の説明されたプロセスのうちの任意の１つを実施するために構成された例示的コンピューティングシステム１４００を図示する。本文脈では、コンピューティングシステム１４００は、例えば、プロセッサ、メモリ、記憶装置、および入力／出力デバイス（例えば、モニタ、キーボード、ディスクドライブ、インターネット接続等）を含んでもよい。しかしながら、コンピューティングシステム１４００は、プロセスのいくつかまたは全ての側面を実行するための電気回路または他の特殊ハードウェアを含んでもよい。いくつかの動作設定において、コンピューティングシステム１４００は、１つまたはそれを上回るユニットを含むシステムとして構成されてよく、それぞれは、ソフトウェア、ハードウェア、またはそのいくつかの組み合わせにおいて、プロセスのいくつかの側面を実行するために構成される。

【００６９】

図５は、前述の説明されたプロセスを実施するために使用され得る、いくつかの構成要素を有するコンピューティングシステム１４００を図示する。メインシステム１４０２は、入力／出力（「Ｉ／Ｏ」）セクション１４０６と、１つまたはそれを上回る中央処理装置（「ＣＰＵ」）１４０８と、メモリセクション１４１０とを有する主回路基板１４０４を含み、それに関連されるフラッシュメモリカード１４１２を有してよい。Ｉ／Ｏセクション１４０６は、ディスプレイ１４２４、キーボード１４１４、ディスク記憶装置１４１６、およびメディアドライブ装置１４１８に接続される。メディアドライブ装置１４１８は、コンピュータ可読媒体１４２０の読み取り／書き込みが可能で、プログラム１４２２および／またはデータを含有することができる。

【００７０】

前述の説明されたプロセスの結果に基づく少なくともいくつかの値は、その後の使用のために保存されることができる。加えて、非一過性コンピュータ可読媒体は、コンピュータを用いて、前述の説明されたプロセスのうちの任意の１つを実施するための１つまたはそれを上回るコンピュータプログラムを記憶する（例えば、有形に具現化する）ために使用されることができる。コンピュータプログラムは、例えば、汎用プログラミング言語（例えば、Ｐａｓｃａｌ、Ｃ、Ｃ＋＋、Ｐｙｔｈｏｎ、Ｊａｖａ（登録商標））またはある特殊用途専用言語で書き込まれてもよい。

【００７１】

種々の例示的実施形態が、本明細書に記載される。これらの実施例は、非限定的意味で参照される。それらは、公開された本技術のより広く適用できる側面を例証するために提供される。種々の実施形態の厳密な精神および範囲から逸脱することなく、種々の変更がなされ、また、均等物が代用されてよい。加えて、多くの修正が、特定の状況、材料、組成物、プロセス、プロセス行為、またはステップを、種々の実施形態の目的、精神、または範囲に適合させるためになされてよい。さらに、当業者によって理解されるであろうように、本明細書に記載および例証される個々の変形例はそれぞれ、種々の実施形態の範囲または精神から逸脱することなく、任意の他のいくつかの実施形態の特徴から容易に分離されてよい、またはそれらと併用されてよい個別の構成要素および特徴を有する。全てのそのような修正は、本開示と関連付けられる請求項の範囲内であることが意図される。

【図１】

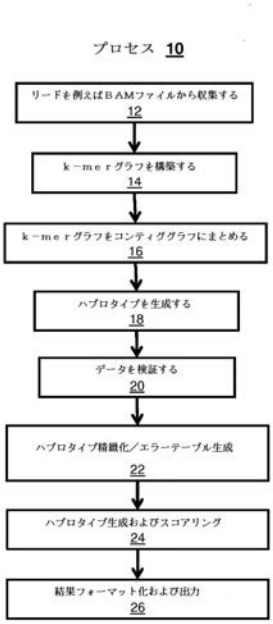

FIG. 1

【図２Ａ】

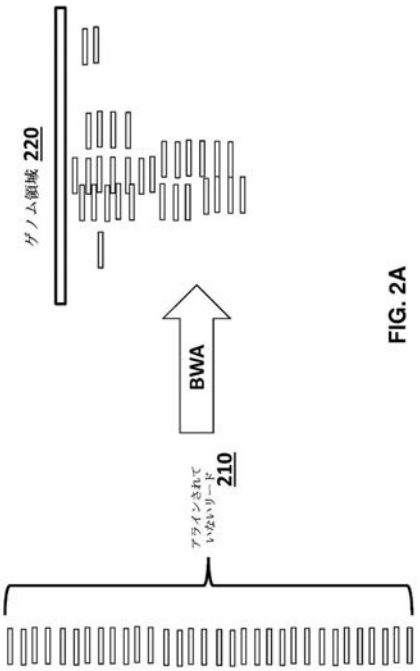

FIG. 2A

【図２Ｂ】

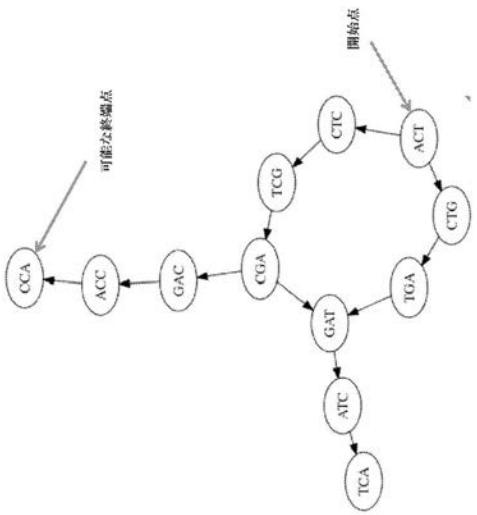

FIG. 2B

【図２Ｃ】

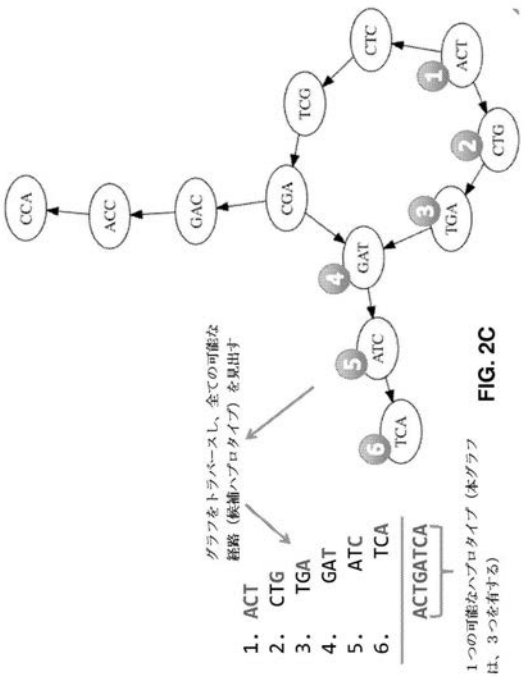

FIG. 2C

【図３Ａ】

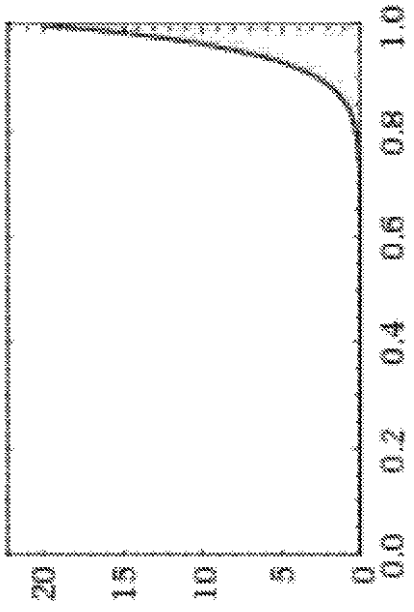

FIG. 3A

【図３Ｂ】

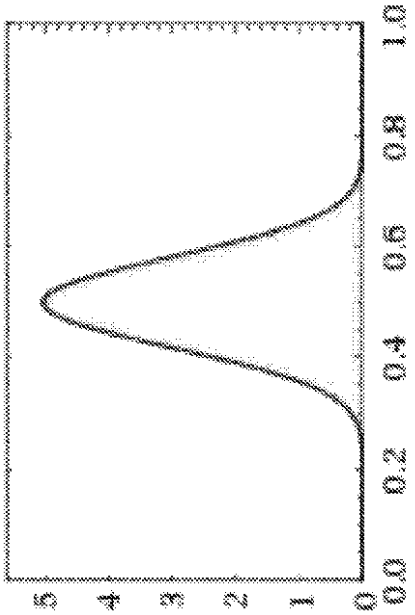

FIG. 3B

【図４】

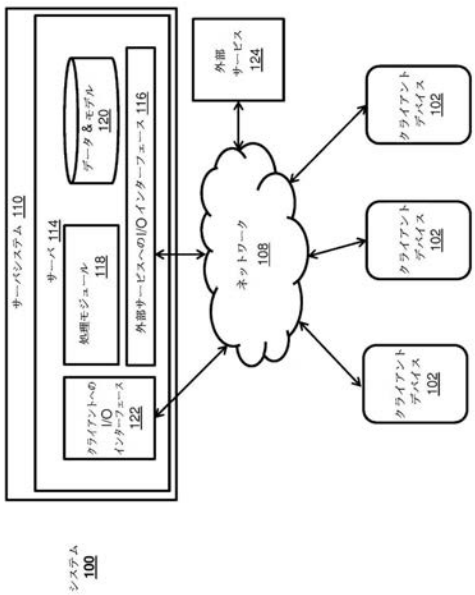

FIG. 4

【図５】

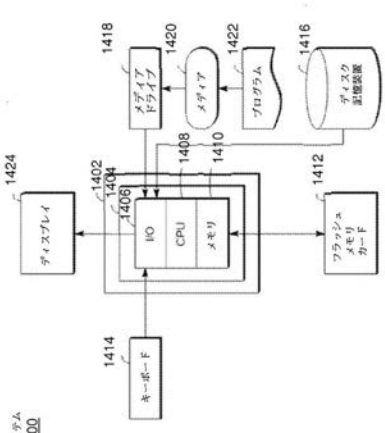

FIG. 5

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 15/55807

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G01N 33/48, C40B 20/00, G06F 19/18 (2016.01)
CPC - G06F19/18, C07H 21/00, G06F19/24
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)
IPC(8): G01N 33/48, C40B 20/00, G06F 19/18 (2016.01)
CPC: G06F19/18, C07H 21/00, G06F19/24

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 702/20, 506/2, 702/20

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatBase, Google Patents, Google Scholar, Google Web, search terms: computer-implemented, genome sample, variants, reference genomic sequence, reads, error table, diplotypes, haplotypes, posterior probability, k-mer graph, Hardy-Weinberg test, Bayes Factor test, Strand Bias Test, sequencer error, output variants, scoring

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2012/0053845 A1 (BRUSTLE et al.) 01 March 2012 (01.03.2012) para [0009], [0010], [0015], [0019], [0024], [0025], [0026], [0028], Fig. 1	1-6, 9-20
Y	US 2013/0054508 A1 (KERMANI et al.) 28 February 2013 (28.02.2013) para [0004], [0048], [0060], [0069], [0217]	1-6, 9-20
Y	ZUO et al. Use of diplotypes-matched haplotype pairs from homologous chromosomes-in gene-disease association studies. Shanghai Arch Psychiatry. (June 2014) vol 26, no 3, pp 165-170, abstract, pg 166, col 1, para 2 3, pg 168 col 1 para 2	1-6, 9-20
Y	US 2004/0265816 A1 (TANAKA et al.) 30 December 2004 (30.12.2004) para [0012], [0049], [0071]	3, 11, 17
Y	US 2005/0214811 A1 (MARGULIES et al.) 29 September 2005 (29.09.2016) abstract, para [0073]	6, 14, 20

☐ Further documents are listed in the continuation of Box C. ☐

* 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
27 January 2016	16 FEB 2016
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 15/55807

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

********** See Supplemental Sheet to continue **********

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-6, 9-20

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 15/55807

Continuation of Box No. III, Observations where unity of invention is lacking:

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I: Claims 1-6, 9-20 drawn to a computer-implemented method or system for determining variants from a genome sample relative to a reference genomic sequence

Group II: Claims 7-8, drawn to a computer-implemented method for generating an error table of sequence data.

The inventions listed as Groups I and II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

Special Technical Features:

Group I has the special technical feature of ACCESSING AN ERROR TABLE of sequence data from previously sequenced samples, not required by Group II.

Group I has the special technical feature of determining a set of possible haplotypes from a set of collected reads, where the set of possible haplotypes is FILTERED BY THE ERROR TABLE, not required by Group II.

Group II has the special technical feature of aligning a set of collected reads to a reference sample, determining mismatches with the reference table, and ADDING SITES THAT HAVE A MISMATCH TO AN ERROR TABLE, not required by Group I.

Common Technical Feature:

Groups I and II have the common technical feature of an error table, where Group I accesses and uses a preconstructed error table, and Group II constructs and generates an error table for future use.

However, said common technical feature does not represent a contribution over the prior art, and is anticipated by US 2013/0332081 A1 to REESE et al. (hereinafter "Reese").

As to the common technical feature, Reese teaches (para [0116-0117]; "Variant Masking. In one aspect, disclosed herein are methods of identifying and/or prioritizing phenotype causing variants by comparing a target cohort to a background cohort, wherein variants within certain regions of the reference sequence are excluded or masked from a CLRT [i.e. composite likelihood ratio test]. Variant calling can be error-prone in repetitive and/or homologous sequences. Variant calling errors in repetitive or homologous sequences can be caused by short sequence reads aligning to multiple sites within the reference sequence. Variant masking can decrease the effect of sequencing platform bias (FIG. 3B). A CLRT with variant masking can be performed with AAS information and/or variant frequency estimation"; para [0153]; "The variant masking option allows the user to exclude a list of nucleotide sites from the likelihood calculations based on information obtained prior to the genome analysis. The masking files used in these analyses excludes sites where short reads would map to more than one position in the reference genome. This procedure mitigates the effects introduced by cross-platform biases by excluding sites that are likely to produce spurious variant calls due to improper alignment of short reads to the reference sequence. The three masking schemes employed are a) 60-bp single-end reads, b) 35-bp single-end reads, and c) 35-bp paired-end reads separated by 400-bp. These three masking files are included with the VAAST distribution")

As the common technical feature was known in the art at the time of the invention, this cannot be considered a common special technical feature that would otherwise unify the groups. The inventions lack unity with one another.

Therefore, Groups I and II lack unity of invention under PCT Rule 13 because they do not share a same or corresponding special technical feature

NOTE, Claims 16-18 depends from "the system of claim 9", as drafted, is objected to, because claim 9 does not teach "a system". For this International Search and Opinion, claims 16-18 are re-construed as dependent claims of claim 15.

フロントページの続き

(81)指定国 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,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,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,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,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

（特許庁注：以下のものは登録商標）
１．ＪＡＶＡＳＣＲＩＰＴ

(72)発明者 ギビアンスキー, アンドリュー レオニドヴィッチ
アメリカ合衆国 カリフォルニア ９４０８０, サウス サンフランシスコ, キンボール ウェイ １８０, カウンシル, インコーポレイテッド 気付
(72)発明者 ハケ, イムラン サイーダル
アメリカ合衆国 カリフォルニア ９４０８０, サウス サンフランシスコ, キンボール ウェイ １８０, カウンシル, インコーポレイテッド 気付
(72)発明者 マグワイア, ジャレッド ロバート
アメリカ合衆国 カリフォルニア ９４０８０, サウス サンフランシスコ, キンボール ウェイ １８０, カウンシル, インコーポレイテッド 気付
(72)発明者 ロバートソン, アレクサンダー デ ヨング
アメリカ合衆国 カリフォルニア ９４０８０, サウス サンフランシスコ, キンボール ウェイ １８０, カウンシル, インコーポレイテッド 気付