(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6946292号
(P6946292)
(45)発行日　令和3年10月6日(2021.10.6)　(24)登録日　令和3年9月17日(2021.9.17)

(51)Int.Cl.　F I
G16B 30/10 (2019.01)　G16B 30/10

請求項の数 9　(全 50 頁)

(21)出願番号　特願2018-526496(P2018-526496)
(86)(22)出願日　平成28年8月4日(2016.8.4)
(65)公表番号　特表2018-533143(P2018-533143A)
(43)公表日　平成30年11月8日(2018.11.8)
(86)国際出願番号　PCT/US2016/045564
(87)国際公開番号　WO2017/024138
(87)国際公開日　平成29年2月9日(2017.2.9)
審査請求日　令和1年8月2日(2019.8.2)
(31)優先権主張番号　62/201,923
(32)優先日　平成27年8月6日(2015.8.6)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　518043944
エイアールシー　バイオ　リミテッド　ライアビリティ　カンパニー
アメリカ合衆国　マサチューセッツ州　０２１３９　ケンブリッジ　チェリー　ストリート　１６
(74)代理人　100094569
弁理士　田中　伸一郎
(74)代理人　100088694
弁理士　弟子丸　健
(74)代理人　100103610
弁理士　▲吉▼田　和彦
(74)代理人　100067013
弁理士　大塚　文昭

最終頁に続く

(54)【発明の名称】ゲノム分析のためのシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
シーケンシングデバイスによって生成される生の遺伝子配列データをアライメントさせる方法であって、
（ａ）シーケンシングデバイスによって生成される生の遺伝子配列データを得ることと；
（ｂ）１又は２以上の０を含有するマスクを用いて、前記生の遺伝子配列データのリードからｋ－ｍｅｒプロファイルを生成し、前記１又は２以上の０に対応するリード中の塩基が、ｋ－ｍｅｒから除外されるように、前記リードから各ｋ－ｍｅｒを生成することと；
（ｃ）前記ｋ－ｍｅｒプロファイルと、代替パスを有するリファレンス配列から前記マスクを用いて生成されたｋ－ｍｅｒのインデックスとを用いて、前記シーケンシングデバイスによって生成される前記生の遺伝子配列データを、代替パスを含むゲノムバリエーションマップ上のロケーションへマッピングすることと；
（ｄ）前記バリエーションマップ上のそのロケーションに従って、前記シーケンシングデバイスによって生成される前記生の遺伝子配列データをアライメントさせることと、
を含む、方法。
【請求項２】
前記マッピングが、グラフアライメントによって遂行される、請求項１に記載の方法。
【請求項３】
前記グラフアライメントが、少なくとも１つのグラフを使用する、請求項２に記載の方法。

【請求項４】
前記マッピングが、ギャップアライメントを使用して遂行される、請求項１に記載の方法。
【請求項５】
前記マッピングが、セミギャップアライメントを使用して遂行される、請求項１に記載の方法。
【請求項６】
前記代替パスの特定のパスが、前記マッピングの間にマッピングされる回数を蓄積することを更に含む、請求項１に記載の方法。
【請求項７】
前記生の遺伝子配列データが１又は２以上のリードペアを含み、リードペアのサブセットについて可能なアライメントが正確である確率が、（ａ）前記リードペアの個別のリードが正しくアライメントされる確率、および（ｂ）前記ペアにおける前記アライメントされたリードの間の距離および前記ペアにおける両方のリードのアライメント方向性を含む、前記リードペアのアライメントフィーチャの観察についての推定確率の関数として計算される、請求項１に記載の方法。
【請求項８】
前記生の遺伝子配列データが１又は２以上のリードペアを含み、リードペアのサブセットについて可能なアライメントが正確である確率が、（ａ）前記リードペアの個別のリードが正しくアライメントされる確率、（ｂ）前記ペアにおける前記アライメントされたリードの間の距離および前記ペアにおける両方のリードのアライメント方向性を含む、前記ペアのアライメントフィーチャの観察についての推定確率、および（ｃ）前記サブセットにおける１又は２以上の他のリードペアの可能なアライメントフィーチャの観察についての推定確率の関数として計算される、請求項１に記載の方法。
【請求項９】
リードペアのサブセットについて可能なアライメントが正確である確率が、（ａ）前記リードペアの個別のリードが正しくアライメントされる確率、および（ｂ）前記サブセットにおける１又は２以上の他のリードの可能なアライメントフィーチャの観察についての推定確率の関数として計算される、請求項１に記載の方法。
【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、２０１５年８月６日に出願された米国仮特許出願第６２／２０１，９２３号の利益を主張し、その出願の全体は参照として本明細書に援用される。
（技術分野）
本出願は、ゲノム分析のためのシステムおよび方法に関する。
【背景技術】
【０００２】
生物学的シーケンシングは、生体分子（ＤＮＡ、ＲＮＡ、タンパク質および他のポリマー等）内のモノマー（例えばヌクレオチドまたはアミノ酸）の精密な順序を決定するプロセスである。シーケンシング方法および装置の急速な開発は、生物医学的研究を著しく前進させることができる。例えば、次世代核酸シーケンシング技術は、低コストのハイスループットシーケンシングプロセスのための新しいパラダイムを提供することができる。次世代シーケンシング技術は、何千または何百万ものヌクレオチド配列を同時に生ずるために、シーケンシングプロセスを並列処理し、大量の情報をもたらすことができる。また、シーケンシングの精度は、次世代のシーケンシング技術によって著しく促進することができる。研究者は、かかる技術によってより短い時間内で多量の高精度な配列データを収集することができる。全ゲノムＤＮＡ配列およびＲＮＡ配列の決定は、遺伝子検査ならびに疾患の診断および治療のための日常業務になっている。

【０００３】
典型的には、ゲノムデータは、レポジトリ、例えば個人のレポジトリ（例えばゲノムデータを生成する研究室に付随するもの）または公共の配列レポジトリ（それは中央レポジトリ中で様々な研究室から受理したデータを保管する）中で保管することができる。かかる大量のデータの保存は、レポジトリが、巨大なボリュームの保存容量を有する大きな保存ディスクを有することを要求する。さらに、研究の前進により、入って来るゲノムデータの量も増加し、それによって、追加の保存スペースのための維持費および必須要件を増加させる。さらに、ゲノムデータは将来の参照のために利用することができるので、ゲノムデータは情報の損失なしに、解凍および検索（ｒｅｔｒｉｅｖａｌ）を許可するように圧縮形状で保管することができる。

【発明の概要】
【発明が解決しようとする課題】
【０００４】
本明細書では、シーケンシングデバイスによって生成される生の遺伝子配列データをアライメントさせる方法が開示され、本方法は、（ａ）シーケンシングデバイスによって生成される生の遺伝子配列データを検索すること（ｒｅｔｒｉｅｖｉｎｇ）と；（ｂ）シーケンシングデバイスによって生成される該生の遺伝子配列データを、ゲノムバリエーションマップ上のロケーションへアライメントさせることとを含み、ゲノムバリエーションマップが代替パスを含む。いくつかの実施形態において、マッピングはグラフアライメントによって遂行される。いくつかの実施形態において、グラフアライメントは単一のグラフを使用する。いくつかの実施形態において、マッピングはギャップアライメントを使用して遂行される。いくつかの実施形態において、マッピングはセミギャップアライメントを使用して遂行される。いくつかの実施形態において、代替パスからの特定のパスがマッピングステップにおいてマッピングされる回数を蓄積することを更に含む。
【課題を解決するための手段】
【０００５】
本明細書では、１又は２以上のリードペアについて可能なアライメントが正確である確率を定量化する方法が開示され、リードペアのサブセットについて可能なアライメントが正確である確率は、個別のリードが正確にアライメントされる確率、およびペアのアライメントフィーチャ（アライメントされたリードの間の距離および両方のペアのリードのアライメント方向性が含まれるが、これらに限定されない）の観察についての推定確率の関数として計算される。いくつかの実施形態において、１又は２以上のリードペアについて可能なアライメントが正確である確率は、サブセットにおける１又は２以上の他のリードペア（それらは同じバーコードを備えたリードペアとすることができる）のアライメントフィーチャに基づいて、追加でスコアリングすることができる。
【０００６】
本明細書では、１又は２以上のリードについて可能なアライメントが正確である確率を定量化する方法が開示され、リードのサブセットについて可能なアライメントが正確である確率は、個別のリードのアライメントが正確である確率およびサブセットにおける他のリード（それらは同じバーコードを備えたリードとすることができる）のアライメントフィーチャの観察についての確率の関数として計算される。
【０００７】
本明細書では、グラフリファレンスアライメントを、新規バリアントまたは構造バリアント検出と組み合わせることによって改善された、バリアントコーリングのためのシステムおよび方法が開示される。バリアントを検出する方法は、ａ）複数の配列リードを得ることと、ｂ）それらをグラフリファレンスとアライメントさせることと、ｃ）バリエーションまたは構造バリエーションの存在を指示することができるリードを同定し、それらを直接または間接的にバリアントコーラーへ渡す（ｐａｓｓｉｎｇ　ｔｈｅｍ　ｔｏ）ことと（例えば最初にファイルへ書き込まれる）、を含む。いくつかの実施形態において、新規の検出されたバリアントのサブセットは、リファレンスへ自動的に追加され、次いでこ

の更新されたリファレンスは別のアライメントのために使用することができる。いくつかの実施形態において、バリエーションまたは構造バリエーションの存在を指示するリードのサブセットは、アライメントステップが進行中である間に（すなわち、すべての配列リードはまだアライメントされたとは限らない）、同定され、ファイルへ書き込まれるかまたはバリアントコーラーへ渡される。このようにして、バリアントコーリングの前にすべてのアライメントされたリードを通してスキャンすることが必要とされるステップはない。次いで、グラフリファレンスアライメントが進行中である間にまたはそれに後続して、新規バリアント検出は遂行することができる。

【０００８】

本明細書では、直線リファレンスとアライメントされたリードのために使用されるフォーマットとコンパチブルなフォーマットで、グラフリファレンスとの配列リードのアライメントを特徴づけるシステムが開示され、該システムは、ａ）リードのグラフアライメントを受理するように構成され、グラフリファレンス配列が、直線リファレンス配列に対するバリアントパスによって表わされる既知のバリアントを含む、受理モジュールと；ｂ）リファレンス配列の座標に対するそのアライメントの開始と、リードがバリアントパスへアライメントする場合にバリアントの識別番号を表わすリードタグとの報告によって、リードのグラフアライメントを特徴づける報告モジュールと、を含む。いくつかの事例において、リードがバリアントへアライメントする場合、リードフラグは設定することができる。いくつかの事例において、報告モジュールは、バリアントパスの座標に対してアライメントされたリードの開始を指示するリードタグを更に規定する。いくつかの事例において、報告モジュールは、バリアントパスの座標に対してアライメントされたリードの開始および終了を指示するリードタグを更に規定する。いくつかの実施形態において、報告モジュールは、バリアントパスに対するアライメントスコアのセットを含むリードタグを規定する。いくつかの実施形態において、グラフリファレンスによるリードのアライメントは、直線リファレンスの座標へ戻って変換することができる。

【０００９】

本明細書では、シングルフェーズの代替配列パスを生成する方法が開示され、本方法は、リファレンス配列を得ることと；リファレンス配列に代わるリファレンス配列上の相関する遺伝子座を検索することと；相関する遺伝子座を含むシングルフェーズの代替配列パスを生成することと、を含む。いくつかの実施形態において、相関する遺伝子座は２名以上の被験者からのものである。いくつかの実施形態において、相関する遺伝子座は、リファレンス配列上の２つ以上のロケーションへマッピングされる異なる配列のセットである。いくつかの実施形態において、異なる配列のセットのうちの少なくとも２つがフェーズ化される。いくつかの実施形態において、本方法は、異なる配列のセットのうちの少なくとも２つの前記フェーズ化セットをインデックス化することを更に含む。

【００１０】

本明細書では、リファレンス配列を代替パスによりインデックス化する方法が開示され、本方法は、（ａ）リファレンス配列を受理することと；（ｂ）該リファレンス配列中にアンカーされる代替配列を受理することと；（ｃ）リファレンス配列および代替配列の複数のｋ－ｍｅｒを２時間以下で生成することと；（ｄ）ｋ－ｍｅｒを使用して、代替パスによりリファレンス配列をインデックス化することと、を含む。いくつかの実施形態において、リファレンス配列はヒトリファレンスゲノムである。いくつかの実施形態において、リファレンス配列は非ヒトリファレンスゲノムである。いくつかの実施形態において、生成は、直線リファレンス座標系を使用して、ｋ－ｍｅｒを直接インデックス化する。いくつかの実施形態において、生成は、ノードｉｄ、エッジまたはパスをアサインすることを含まない。

【００１１】

本明細書では、代替パスによりリファレンス配列をインデックス化する方法が開示され、本方法は、（ａ）リファレンス配列を受理することと；（ｂ）該リファレンス配列中へとアンカーされる代替配列を受理することと；（ｃ）８０ギガバイト以下のコンピュータ

ー空間中に適合する、リファレンス配列および代替物のインデックス化された複数のｋ－ｍｅｒを生成することと；（ｄ）ｋ－ｍｅｒを使用して、代替パスによりリファレンス配列をインデックス化することと、を含む。いくつかの実施形態において、コンピューター空間は、ディスク、ｒａｍ、またはアドレス空間のうちの１又は２以上から選択される。いくつかの実施形態において、リファレンス配列はヒトリファレンスゲノムである。いくつかの実施形態において、リファレンス配列は非ヒトリファレンスゲノムである。いくつかの実施形態において、生成ステップは、直線リファレンス座標系を直接使用すること、および代替物のｋ－ｍｅｒがその座標系中に現われるように代替物のｋ－ｍｅｒを直接インデックス化することによって遂行される。いくつかの実施形態において、生成は、ノードｉｄ、エッジまたはパスをアサインすることを含まない。

【００１２】

本明細書では、リファレンス配列中のｋ－ｍｅｒのインデックスのクエリを、代替パスにより実行する方法が提供され、本方法は、（ａ）リファレンス配列からの代替パスを含有するリファレンス配列中の複数のｋ－ｍｅｒおよびロケーションを含むインデックスを、代替パスにより検索することと；（ｂ）リファレンス配列中のｋ－ｍｅｒを備えた該インデックスのクエリを、１計算スレッドあたり毎秒６９，０００以上のｋ－ｍｅｒの率で、代替パスにより実行することと、を含む。いくつかの実施形態において、クエリは、１計算スレッドあたり毎秒３４５，０００以上のｋ－ｍｅｒの率で複数の計算スレッド上で遂行される。いくつかの実施形態において、計算スレッドの数は４より大きい。

【００１３】

いくつかの実施形態において、計算スレッドの数は、１より大きい、２より大きい、３より大きい、４より大きい、５より大きい、６より大きい、７より大きい、８より大きい、９より大きい、１０より大きい、２０より大きい、３０より大きい、４０より大きい、５０より大きい、６０より大きい、７０より大きい、８０より大きい、９０より大きい、または１００より大きい。いくつかの実施形態において、クエリは、１計算コアあたり毎秒３４５，０００以上のｋ－ｍｅｒの率で複数の計算コア上で遂行される。

【００１４】

いくつかの実施形態において、プロセッサ作業のうちの９５％を超えるものが、インデックスのクエリ専用である。いくつかの実施形態において、プロセッサ作業のうちの８５％を超えるものが、インデックスのクエリ専用である。いくつかの実施形態において、プロセッサ作業のうちの７５％を超えるものが、インデックスのクエリ専用である。いくつかの実施形態において、プロセッサ作業のうちの６５％を超えるものが、インデックスのクエリ専用である。

【００１５】

いくつかの実施形態において、プロセッサ作業は、インデックスのクエリに単独で専用の作業である。いくつかの実施形態において、プロセッサ作業は、カーネルタスク、メモリスワップ、またはＩ／Ｏからなる群のうちの１又は２以上を含まない。

【００１６】

いくつかの実施形態において、ｋ－ｍｅｒは、少なくとも１０、２０、３０、４０、５０、６０、７０、８０、９０、または１００のヌクレオチド（ｎｔ）の長さである。特定の実施形態において、ｋ－ｍｅｒは少なくとも３２の長さである。

【００１７】

本明細書では、配列を比較する方法が開示され、本方法は、（ａ）シーケンシングアッセイが進行中である間に、シーケンサーからリードを検索することと；（ｂ）シーケンシングアッセイが進行中である間に、リードを配列と比較することと；（ｃ）特異的な遺伝子座がリード中にあるかどうかを決定することと、を含む。いくつかの実施形態において、本方法は、前記特異的な遺伝子座の出現のカウントを蓄積することを更に含む。

【００１８】

本明細書では、代替配列を有するリファレンスゲノム中の遺伝子座をフェーズ化する方法が開示され、本方法は、（ａ）各々が遺伝子座を含む配列中の多重代替パスの出現の数

の蓄積を検索することと；（ｂ）複数の代替パスについての蓄積存在量カウントに基づいて、異なる代替パスからの遺伝子座をグループ化することと；（ｃ）共通のグループ化に基づいて、遺伝子座をフェーズ化することと、を含む。共通のグループ化は、ペアエンド、メイトペア、または遺伝物質の一本鎖から生成されて距離を隔てた他の配列データとして、リンクされるリードを指すことができる。

【００１９】

本明細書では、配列を比較する方法が開示され、本方法は、（ａ）シーケンシングアッセイが進行中である間に、シーケンサーからリードを検索することと；（ｂ）シーケンシングアッセイが進行中である間に、リードを暗号化することと、（ｃ）リードを解読せずにリファレンス配列と比較することと、を含む。いくつかの実施形態において、リファレンス配列はゲノムバリエーションマップである。いくつかの実施形態において、リファレンス配列はリファレンスヒトゲノムである。いくつかの実施形態において、リファレンス配列は非ヒトリファレンスゲノムである。

【００２０】

本明細書では、配列を比較する方法が開示され、本方法は、（ａ）シーケンシングアッセイが進行中である間に、シーケンサーからリードを検索することと；（ｂ）シーケンシングアッセイが進行中である間に、リードを暗号化することと；（ｃ）暗号化されたリードをプロセシングデバイスへ伝送することと、を含む。いくつかの実施形態において、本方法は、前記暗号化されたリードを解読することを更に含む。いくつかの実施形態において、本方法は、シーケンシングアッセイが進行中である間に、前記リードを配列へ比較することを更に含む。いくつかの実施形態において、本方法は、暗号化されたリードに関係する解読鍵を伝送することを更に含む。

【００２１】

本明細書では、配列を比較する方法が開示され、本方法は、（ａ）シーケンシングアッセイが進行中である間に、シーケンサーからリードを検索することと；（ｂ）シーケンシングアッセイが進行中である間に、リードをプロセシングデバイスへ圧縮することと；（ｃ）シーケンシングアッセイが進行中である間に、リードを配列へ解凍することと；（ｄ）シーケンシングアッセイが進行中である間に、リードをリファレンス配列と比較することと、を含む。

【００２２】

本明細書では、既知のバリアントのコールのためのシステムおよび方法が開示され、本方法は、（ａ）リードを検索することと；（ｂ）リードからｋ－ｍｅｒプロファイルを生成することと；（ｃ）、リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに対して、代替パスにより、ｋ－ｍｅｒプロファイルを問い合わせて、バリアントをコールすることと、を含む。

【００２３】

いくつかの実施形態において、既知のバリアントは、１コアあたり毎秒１０のバリアントコール以上の率で、少なくとも単一のコアを使用して同定される。本明細書では、シーケンシングデータの圧縮のためのシステムが開示され、本システムは、（ａ）シーケンシングデータの受理のためのモジュールと；（ｂ）シーケンシングデータの保存のためのメモリユニットと；（ｃ）メモリとその上に保存されたシーケンシングデータへアクセスでき、シーケンシングデータのうちのいくつかまたはすべてをエンコードするように構成されたエンコーディングモジュールと、を含む。いくつかの実施形態において、シーケンシングデータは、受理したシーケンシングデータの８１．５％以上のレベルへ圧縮される。いくつかの実施形態において、シーケンシングデータはフィールドを含み、このフィールドは、配列識別子；塩基コールデータ；アミノ酸コールデータ；コメントのためのライン；および塩基コールデータについてのクオリティ値のうちの１又は２以上を含む。いくつかの実施形態において、エンコーディングモジュールは、シーケンシングデータ中のシーケンシングデータの残りから塩基コールデータを分離する。いくつかの実施形態において、エンコーディングモジュールは、シーケンシングデータ中のシーケンシングデータの残

りからアミノ酸コールデータを分離する。いくつかの実施形態において、塩基コールデータ中のヌクレオチド塩基は、ヌクレオチド塩基：アデニン（Ａ）、チミン（Ｔ）、グアニン（Ｇ）、およびシトシン（Ｃ）に対応する文字と関連し、決定できなかった塩基は（Ｎ）と関連する。いくつかの実施形態において、アミノ酸コールデータはアミノ酸に対応する特徴と関連する。

【００２４】

いくつかの実施形態において、シーケンシングデータ中の塩基を決定することができない場合、エンコーディングモジュールはシーケンシングデータ中のシーケンシングデータの残りから塩基コールデータを分離しないで、その塩基と関連するリードをエンコードする。いくつかの実施形態において、決定されない塩基（ｕｎｄｅｒｍｉｎｅｄ　ｂａｓｅｄ）と関連するリードは、未分離リードとして圧縮される。いくつかの実施形態において、決定できなかった塩基コールを備えたリードで、かかる塩基の位置がセーブされる。いくつかの実施形態において、決定できなかった塩基コールを備えたリードのすべてで、かかる塩基の位置がセーブされる。いくつかの実施形態において、決定できなかった塩基コールを備えたリードのうちの少なくとも９０％で、かかる塩基の位置がセーブされる。いくつかの実施形態において、塩基の長さが２５６塩基未満である場合、塩基のロケーションは１バイト未満でセーブされる。いくつかの実施形態において、長さが６５５３６塩基未満である場合、塩基のロケーションは２バイト未満でセーブされる。いくつかの実施形態において、エンコーディングモジュールは差分エンコーディングを使用して、情報をセーブする。いくつかの実施形態において、エンコーディングモジュールは、ヌクレオチドベースのデータ上で塩基－４エンコーディングを遂行する。いくつかの実施形態において、シーケンシングデータの各々のフィールドは順次セーブされる。

【００２５】

いくつかの実施形態において、本明細書において開示されるシステムは、分離したファイル中でセーブされる少なくとも２つのフィールドを有する。いくつかの実施形態において、少なくとも２つの異なるファイルが、各々のフィールド上で遂行される異なる圧縮アルゴリズムを使用してセーブされる。いくつかの実施形態において、配列識別子フィールド中のデータは、差分エンコーディングを使用して圧縮される。いくつかの実施形態において、塩基コールデータはＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換を使用して処理される。いくつかの実施形態において、塩基コールデータのプロセシングは、ランレングスエンコーディングを実行すること、およびＨｕｆｆｍａｎエンコーディングを使用して圧縮することを更に含む。いくつかの実施形態において、コメントのためのラインは差分エンコーディングを使用して圧縮される。いくつかの実施形態において、フィールドが空の場合に、コメントのための追加のラインは無視される。いくつかの実施形態において、クオリティ値データはＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換を使用して処理される。いくつかの実施形態において、プロセシングは、ランレングスエンコーディングを実行すること、およびＨｕｆｆｍａｎエンコーディングを使用して圧縮することを更に含む。

【００２６】

本明細書では、配列アライメントマップ（ＳＡＭ）データの圧縮のためのシステムが開示され、本システムは、（ａ）ＳＡＭデータをその上に保存したメモリと；（ｂ）メモリとその上に保存されたＳＡＭへのアクセスを有し、ＳＡＭデータを８０％以上のレベルへ圧縮するように構成された、エンコーディングモジュールと、を含む。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のクエリテンプレート名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のリファレンス配列名を圧縮する。エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中の左端のマッピング位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のメイトリードのリファレンス名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のメイトリードの位置を

圧縮する。いくつかの実施形態において、エンコーディングモジュールは、Ｈｕｆｆｍａｎコーディング方法を使用して、データ形式をｃｉｇａｒ文字列に圧縮する。いくつかの実施形態において、エンコーディングモジュールは、辞書ベースの方法を使用して、データ形式をｃｉｇａｒ文字列に圧縮する。いくつかの実施形態において、エンコーディングモジュールは、塩基４エンコーディングを使用して、ＳＡＭデータからの塩基コールデータを圧縮する。いくつかの実施形態において、エンコーディングモジュールは、ＳＡＭデータからのクオリティデータを圧縮する。いくつかの実施形態において、（ａ）クエリテンプレート名、リファレンス配列名、左端のマッピング位置、メイトリードのリファレンス名、およびメイトリードの位置のうちの１又は２以上を含むＳＡＭデータからのデータの各々は、差分エンコーディングを使用して圧縮することができる；（ｂ）ｃｉｇａｒ文字列を含むＳＡＭデータからのデータは、Ｈｕｆｆｍａｎコーディングまたは辞書ベースの方法を使用して圧縮することができる；（ｃ）塩基コールデータを含むＳＡＭデータからのデータは、塩基４エンコーディングを使用して圧縮することができる；ならびに（ｄ）クオリティデータを含むＳＡＭデータからのデータは圧縮することができる。いくつかの実施形態において、ＳＡＭデータは順次順序付けられる。

【００２７】

本明細書では、バリアントコールフォーマット（ＶＣＦ）データの圧縮のためのシステムが開示され、本システムは、ＶＣＦデータをその上に保存したメモリと；メモリとその上に保存されたＶＣＦへのアクセスを有し、ゲノムデータをＶＣＦデータの９５％以上のレベルへエンコードするように構成された、エンコーディングモジュールと、を含む。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のクエリテンプレート名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のリファレンス配列名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中の左端のマッピング位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のメイトリードのリファレンス名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のメイトリードの位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、Ｈｕｆｆｍａｎコーディング方法を使用して、データ形式をｃｉｇａｒ文字列に圧縮する。いくつかの実施形態において、エンコーディングモジュールは、辞書ベースの方法を使用して、データ形式をｃｉｇａｒ文字列に圧縮する。いくつかの実施形態において、エンコーディングモジュールは、塩基４エンコーディングを使用して、ＶＣＦデータからの塩基コールデータを圧縮する。いくつかの実施形態において、エンコーディングモジュールはＶＣＦデータからのクオリティデータを圧縮する。

【００２８】

本開示の態様は、シーケンシングデバイスによって生成される生の遺伝子配列データをアライメントさせる方法を開示し、本方法は、（ａ）シーケンシングデバイスによって生成される生の遺伝子配列データを得ることと；（ｂ）シーケンシングデバイスによって生成される生の遺伝子配列データを、代替パスを含むゲノムバリエーションマップ上のロケーションへマッピングすることと；（ｃ）バリエーションマップ上のそのロケーションに従って、シーケンシングデバイスによって生成される生の遺伝子配列データをアライメントさせることと、を含む。

【００２９】

いくつかの実施形態において、マッピングはグラフアライメントによって遂行される。いくつかの実施形態において、グラフアライメントは少なくとも１つのグラフを使用する。いくつかの実施形態において、マッピングはギャップアライメントを使用して遂行される。いくつかの実施形態において、マッピングはセミギャップアライメントを使用して遂行される。いくつかの実施形態において、本方法は、代替パスの特定のパスが、マッピングの間にマッピングされる回数を蓄積することを更に含む。いくつかの実施形態において

、生の遺伝子配列データは１又は２以上のリードペアを含み、そこで、リードペアのサブセットについて可能なアライメントが正確である確率は、（ａ）リードペアの個別のリードが正しくアライメントされる確率、および（ｂ）リードペアのアライメントフィーチャ（ペアにおけるアライメントされたリードの間の距離およびペアにおける両方のリードのアライメント方向性が含まれる）の観察についての推定確率の関数として計算される。いくつかの実施形態において、生の遺伝子配列データは１又は２以上のリードペアを含み、そこで、リードペアのサブセットについて可能なアライメントが正確である確率は、（ａ）リードペアの個別のリードが正しくアライメントされる確率、（ｂ）ペアのアライメントフィーチャ（ペアにおけるアライメントされたリードの間の距離およびペアにおける両方のリードのアライメント方向性が含まれる）の観察についての推定確率、および（ｃ）サブセットにおける１又は２以上の他のリードペアの可能なアライメントフィーチャの観察についての推定確率の関数として計算される。いくつかの実施形態において、リードのサブセットについて可能なアライメントが正確である確率は、（ａ）リードペアの個別のリードが正しくアライメントされる確率、および（ｂ）サブセットにおける１又は２以上の他のリードの可能なアライメントフィーチャの観察についての推定確率の関数として計算される。

【００３０】

本開示の態様は、新規バリアントを同定する方法を提供し、本方法は、（ａ）複数の配列リードを得ることと；（ｂ）代替パスによって表わされる既知のバリアントを含むグラフリファレンスに対して複数の配列リードをアライメントさせることと；（ｃ）１又は２以上の代替パスに対して変則的にアライメントする複数の配列リードのサブセットを使用して、新規バリアントを同定することと、を含む。

【００３１】

いくつかの実施形態において、新規バリアントは構造バリアントを含む。いくつかの実施形態において、新規バリアントの同定に使用される複数の配列リードのサブセットは、グラフリファレンス中のすべての代替パスへ変則的にアライメントする。いくつかの実施形態において、配列リードはリードペアを含み、そこで変則的なアライメントは、大多数のアライメントされたリードペアのものとは異なる、アライメントされたリードペア方向性を含む。いくつかの実施形態において、配列リードはリードペアを含み、そこで変則的なアライメントは、大多数のアライメントされたリードペアよりも有意に小さいかまたは大きい、アライメントされたリードペアインサート長を含む。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の中央値よりも１０％を超えて大きいかまたは小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の中央値よりも５０％を超えて大きいかまたは小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の中央値よりも１００％を超えて大きいかまたは小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の中央値よりも２００％を超えて大きいかまたは小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の中央値よりも３００％を超えて大きいかまたは小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の９９番目のパーセンタイル値よりも大きいか、または１番目のパーセンタイル値よりも小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の９８番目のパーセンタイル値よりも大きいか、または２番目のパーセンタイル値よりも小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の９７番目のパーセンタイル値よりも大きいか、または３番目のパーセンタイル値よりも小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサート長の９５番目のパーセンタイル値よりも大きいか、または５番目のパーセンタイル値よりも小さい。いくつかの実施形態において、インサート長は、アライメントされたリードのサブセットのインサー

ト長の９０番目のパーセンタイル値よりも大きいか、または１０番目のパーセンタイル値よりも小さい。いくつかの実施形態において、インサート長は、いくつかのユーザー指定値よりも大きいかまたは小さい。いくつかの実施形態において、配列リードはリードペアを含み、そこで、変則的なアライメントは、１つのリードがアライメントされ、１つのリードがアライメントされないリードペアを含む。いくつかの実施形態において、変則的なアライメントは、リードの一部がクリップされたリードを含む。いくつかの実施形態において、クリップされたリードの部分は１０％よりも大きい。いくつかの実施形態において、クリップされたリードの部分は５％よりも大きい。いくつかの実施形態において、クリップされたリードの部分は２０％よりも大きい。いくつかの実施形態において、クリップされたリードの部分は３０％よりも大きい。いくつかの実施形態において、同定された新規のバリアントは標的適用について以前に文書化されないバリアントである。いくつかの実施形態において、同定された新規バリアントはグラフリファレンス中に存在しないバリアントである。いくつかの実施形態において、同定された新規バリアントのサブセットはグラフリファレンスへ自動的に追加されて更新されたグラフリファレンスを生じ、そこで、更新されたグラフリファレンスは別のアライメントのために使用される。いくつかの実施形態において、本方法は、グラフリファレンス中の代替パスへアライメントするリード数をカウントすること、およびグラフリファレンス中の代替パスへアライメントするリードの数を使用して既知のバリアントを同定することを更に含む。いくつかの実施形態において、同定された新規バリアントは構造バリアントを含む。いくつかの実施形態において、既知のバリアントは標的適用のために以前に文書化されている。いくつかの実施形態において、新規バリアントは標的適用のために以前に文書化されていない。いくつかの実施形態において、既知のバリアントはグラフリファレンス中に存在するバリアントである。いくつかの実施形態において、新規バリアントはグラフリファレンス中に存在しないバリアントである。いくつかの実施形態において、変則的なアライメントは、ａ）大多数のアライメントされたリードペアのものとは異なる、アライメントされたリードペア方向性；ｂ）大多数のアライメントされたリードペアよりも有意に小さいかまたは大きい、アライメントされたリードペアインサート長；ｃ）１つのリードがアライメントされ、１つのリードがアライメントされない、リードペア；ｄ）リードの一部がクリップされたリード；ｅ）インサート長が、アライメントされたリードペアのサブセットのインサート長の９９番目、９８番目、９７番目、９５番目、もしくは９０番目のパーセンタイル値よりも大きいか、または１番目、２番目、３番目、５番目、もしくは１０番目のパーセンタイル値よりも小さい、リードペア；およびｆ）リードが異なるリファレンス配列へアライメントする、リードペア、のうちの１又は２以上を含む。いくつかの実施形態において、異なるリファレンス配列は、異なる染色体からのものである。いくつかの実施形態において、本方法は、あらかじめ定義されたクオリティ尺度または検出確実性尺度を満たす同定された新規バリアントのサブセットを同定すること、およびサブセットをグラフリファレンスへ追加することを更に含む。いくつかの実施形態において、本方法は、あらかじめ定義されたサイズ範囲内である同定された新規バリアントのサブセットを同定すること、およびサブセットをグラフリファレンスへ追加することを更に含む。いくつかの実施形態において、本方法は、ゲノムのあらかじめ定義された領域内に位置する同定された新規バリアントのサブセットを同定すること、およびサブセットをグラフリファレンスへ追加することを更に含む。いくつかの実施形態において、本方法は、あらかじめ定義された相対値または絶対値を超える頻度を備えた配列リードのうちの１又は２以上において検出される同定された新規バリアントのサブセットを同定すること、およびサブセットをグラフリファレンスへ追加することを更に含む。いくつかの実施形態において、更新されたグラフリファレンスは、後続のアライメントおよびバリアント検出のために使用される。いくつかの実施形態において、グラフリファレンスは、２つ以上のアライメントおよびバリアント検出において使用され漸進的に更新される。いくつかの実施形態において、グラフリファレンスは、同じコンピューター上で２つ以上のアライメントおよびバリアント検出において使用され漸進的に更新される。いくつかの実施形態において、グラフリファレンスは、１又は２

以上のコンピューターの中で共有および更新される。いくつかの実施形態において、グラフリファレンスは中央レポジトリ中で保存および更新され、１又は２以上のコンピューターの中で共有される。いくつかの実施形態において、既知のバリアントまたは新規バリアントは種内バリアントを含む。いくつかの実施形態において、既知のバリアントまたは新規バリアントは種間バリアントを含む。

【００３２】

本開示の態様は、配列バリアントを検出する方法を開示し、本方法は、ａ）複数の配列リードを得ることと；ｂ）代替パスによって表わされた既知のバリアントを含むグラフリファレンスに対して複数の配列リードのサブセットをアライメントさせることを含むプロセスによってアライメントされたリードのバッチを生成することと；ｃ）アライメントされたリードのバッチ内の１又は２以上の変則的にアライメントされたリードを同定することと；ｄ）１又は２以上の変則的にアライメントされたリードを使用して新規構造バリアントを同定することと、を含む。

【００３３】

いくつかの実施形態において、本方法は、グラフリファレンス中の代替パスへアライメントする、アライメントされたリードのバッチ中のリードの数をカウントすること、およびリードの数を使用して既知のバリアントを同定することを更に含む。いくつかの実施形態において、本方法は、少なくとも１つの追加のバッチのためにステップａ）～ｄ）を遂行することを更に含む。いくつかの実施形態において、本方法は、少なくとも１つの追加のバッチのためにステップａ）～ｄ）を遂行することを更に含む。いくつかの実施形態において、既知のバリアントは標的適用のために以前に文書化されている。いくつかの実施形態において、新規構造バリアントは標的適用のために以前に文書化されていない。いくつかの実施形態において、既知のバリアントはグラフリファレンス中に存在するバリアントである。いくつかの実施形態において、新規構造バリアントはグラフリファレンス中に存在しないバリアントである。いくつかの実施形態において、バッチからの変則的にアライメントされたリードのサブセットはファイルへ書き込まれ、続いて、新規構造バリアントの同定に使用される。いくつかの実施形態において、バッチからの変則的にアライメントされたリードのサブセットをコンピュータープログラムへ渡して、リードのサブセットをファイルへ書き込むことなしに、新規構造バリアントを同定する。いくつかの実施形態において、変則的なアライメントは、ａ）大多数のアライメントされたリードペアのものとは異なる、アライメントされたリードペア方向性；ｂ）大多数のアライメントされたリードペアよりも有意に小さいかまたは大きい、アライメントされたリードペアインサート長；ｃ）１つのリードがアライメントされ、１つのリードがアライメントされない、リードペア；ｄ）リードの一部がクリップされたリード；ｅ）インサート長が、アライメントされたリードペアのサブセットのインサート長の９９番目、９８番目、９７番目、９５番目、もしくは９０番目のパーセンタイル値よりも大きいか、または１番目、２番目、３番目、５番目、もしくは１０番目のパーセンタイル値よりも小さい、リードペア；およびｆ）リードが異なるリファレンス配列へアライメントする、リードペアのうちの１又は２以上を含む。いくつかの実施形態において、異なるリファレンス配列は、異なる染色体からのものである。いくつかの実施形態において、本方法は、追加のフィーチャを使用してグラフリファレンス中の代替パスへアライメントするリードの追加のフィーチャをトラッキングして、既知のバリアントを同定することを更に含む。いくつかの実施形態において、複数の配列リードの１％未満は、ファイルの２回以上からのリードである。いくつかの実施形態において、複数の配列リードの５％未満は、ファイルの２回以上からのリードである。いくつかの実施形態において、複数の配列リードの１０％未満は、ファイルの２回以上からのリードである。いくつかの実施形態において、複数の配列リードの１５％未満は、ファイルの２回以上からのリードである。いくつかの実施形態において、既知のバリアントまたは新規構造バリアントは種内バリアントを含む。いくつかの実施形態において、既知のバリアントまたは新規構造バリアントは種間バリアントを含む。

【００３４】

本開示の態様は、直線リファレンスがアライメントされたリードのために使用されたフォーマットとコンパチブルなフォーマットで、配列リードのグラフリファレンスアライメントを簡潔に特徴づけるシステムを提供し、本システムは、ａ）リードのグラフリファレンス配列へのグラフアライメントを受理するように構成され、そこで、グラフリファレンス配列が、リファレンス配列と比べたバリアントパスによって表わされる既知のバリアントを含む、受理モジュールと；ｂ）リファレンス配列の座標に対するそのアライメントの開始と、リードがバリアントパスへアライメントする場合のバリアントパスの識別番号を表わすリードタグとの報告によって、リードのグラフアライメントを特徴づける、報告モジュールと、を含む。

【００３５】

いくつかの実施形態において、報告モジュールは、リードがバリアントパスへアライメントする場合に、設定されるリードフラグを更に報告する。いくつかの実施形態において、ｂ）のリードタグが提供される場合、報告モジュールは、バリアントパスの座標に対してアライメントされたリードの開始を指示する第２のリードタグを更に出力する。いくつかの実施形態において、ｂ）のリードタグが提供される場合、報告モジュールは、バリアントパスの座標に対してアライメントされたリードの開始および終了を指示する第２のリードタグを更に出力する。いくつかの実施形態において、ｂ）のリードタグが提供される場合、報告モジュールは、バリアントパスに対する文字列アライメントスコアを含む第２のリードタグを更に出力する。いくつかの実施形態において、アライメントスコアには、マッチ、挿入または欠失の数が含まれる。いくつかの実施形態において、ｂ）のリードタグが提供される場合、報告モジュールは、どれだけのリードがバリアントパスへマッピングされるかを含む第２のリードタグを更に出力する。いくつかの実施形態において、報告モジュールは、どれだけのリードがリファレンス配列へマッピングされるかを含む第２のリードタグを更に出力する。いくつかの実施形態において、報告モジュールは、リファレンス配列へマッピングされるリードを含む第２のリードタグを更に出力する。いくつかの実施形態において、報告モジュールは、バリアントパスへ最初にマッピングされるリードを指示する第２のリードタグを更に出力する。いくつかの実施形態において、アライメントの開始は、リファレンス配列の上への射影を指示する。

【００３６】

本開示の態様は、配列リードペアの変則的なグラフアライメントを決定するシステムを提示し、本システムは、ａ）直線リファレンスパスを含むグラフリファレンスへアライメントされたリードペアを受理し、そこで、リードペアのうちの少なくとも１つのリードが、そのアライメントのうちのいくつかまたはすべてを代替パス上に有する、受理モジュールと；ｂ）少なくとも１つのリードを直線リファレンス座標系へ変換し、変換操作情報をメタデータとして保存する、変換モジュールと；ｃ）リードペアにおける、変換されたリード、メタデータ、および第２のリードを、インプットとして採用し、リードペアへ特異的な特性を計算する、計算モジュールと；ｄ）特性を採用し、グラフリファレンスへ変則的にアライメントされるかまたは否かとしてペアを分類する、判断モジュールと、を含む。

【００３７】

いくつかの実施形態において、特性は、直線リファレンスパスと比べたインサート長を含む。いくつかの実施形態において、特性は、直線リファレンスパスと比べたＣＩＧＡＲスコアを含む。いくつかの実施形態において、特性は、直線リファレンスパスに対するアライメント位置を含む。いくつかの実施形態において、アライメント報告は下流の分析ツールとコンパチブルである。いくつかの実施形態において、コンパチビリティーは、コンパチブルなファイルフォーマットであることを含む。いくつかの実施形態において、コンパチブルなファイルフォーマットはＳＡＭである。いくつかの実施形態において、コンパチブルなファイルフォーマットはＢＡＭである。いくつかの実施形態において、コンパチブルなファイルフォーマットはＶＣＦである。

【００３８】

本開示の態様は、少なくとも１つのフェーズ化された代替配列パスを生成する方法を提供し、本方法は、ａ）リファレンス配列を得ることと；ｂ）代替リファレンス配列上の相関する遺伝子座を検索することと；ｃ）相関する遺伝子座を含む、少なくとも１つのフェーズ化された代替配列パスを生成することと、を含む。

【００３９】

いくつかの実施形態において、相関する遺伝子座は２つ以上の別個の起源からのものである。いくつかの実施形態において、相関する遺伝子座は、リファレンス配列上の２つ以上のロケーションへマッピングされる、異なる配列のセットを含む。いくつかの実施形態において、異なる配列のセットのうちの少なくとも２つがフェーズ化される。いくつかの実施形態において、本方法は、異なる配列の前記フェーズ化されたセットをインデックス化することを更に含む。

【００４０】

本開示の態様は、代替パスによりリファレンス配列をインデックス化する方法を開示し、本方法は、（ａ）リファレンス配列を受理することと；（ｂ）リファレンス配列へマッピングされる代替配列を受理することと；（ｃ）リファレンス配列および代替配列のｋ－ｍｅｒを２時間以下で生成することと；（ｄ）ｋ－ｍｅｒを使用して、代替パスによりリファレンス配列をインデックス化することと、を含む。

【００４１】

いくつかの実施形態において、リファレンス配列は核酸配列である。いくつかの実施形態において、核酸配列はゲノム配列である。いくつかの実施形態において、核酸配列は、二本鎖ＤＮＡ、一本鎖ＤＮＡ、ＤＮＡ／ＲＮＡハイブリッド、一本鎖ＲＮＡ、二本鎖ＲＮＡ、または相補的ＤＮＡ（ｃＤＮＡ）を含む。いくつかの実施形態において、核酸配列は合成配列である。いくつかの実施形態において、ゲノム配列はヒトゲノムからのものである。いくつかの実施形態において、ゲノム配列は非ヒトゲノムからのものである。いくつかの実施形態において、非ヒトゲノムは、細菌ゲノム、ウイルスゲノム、真菌ゲノム、原生動物ゲノム、および植物ゲノムからなる群から選択される。いくつかの実施形態において、リファレンス配列はアミノ酸配列である。いくつかの実施形態において、アミノ酸配列は既知の配列である。いくつかの実施形態において、アミノ酸配列は機能性配列である。いくつかの実施形態において、アミノ酸配列は合成配列である。いくつかの実施形態において、アミノ酸配列はヒトである。いくつかの実施形態において、アミノ酸配列は非ヒトである。いくつかの実施形態において、非ヒトアミノ酸配列は、細菌配列、ウイルス配列、真菌配列、原生動物配列、および植物（ｆｌｏｒａｌ）（植物（ｐｌａｎｔ））配列からなる群から選択される。いくつかの実施形態において、代替パスは未知のアミノ酸配列を含む。いくつかの実施形態において、生成は、直線リファレンス座標系を使用して、ｋ－ｍｅｒを直接インデックス化する。いくつかの実施形態において、生成は、ノードＩＤ、エッジ、またはパスをアサインすることを含まない。

【００４２】

本開示の態様は、代替パスによりリファレンス配列をインデックス化する方法を開示し、本方法は、（ａ）リファレンス配列を受理することと；（ｂ）リファレンス配列へマッピングされる代替配列を受理することと；（ｃ）リファレンス配列および代替配列のインデックス化された複数のｋ－ｍｅｒを生成し、そこで、インデックス化された複数のｋ－ｍｅｒが８０ギガバイト以下のサイズであることと；（ｄ）ｋ－ｍｅｒを使用して、代替パスによりリファレンス配列をインデックス化することと、を含む。

【００４３】

いくつかの実施形態において、コンピューター空間は、ディスク、ｒａｍ、またはアドレス空間のうちの１又は２以上から選択される。いくつかの実施形態において、リファレンス配列は核酸配列である。いくつかの実施形態において、核酸配列はゲノム配列である。いくつかの実施形態において、核酸配列は、二本鎖ＤＮＡ、一本鎖ＤＮＡ、ＤＮＡ／ＲＮＡハイブリッド、一本鎖ＲＮＡ、二本鎖ＲＮＡ、または相補的ＤＮＡ（ｃＤＮＡ）を含む。いくつかの実施形態において、核酸配列は合成配列である。いくつかの実施形態にお

いて、ゲノム配列はヒトゲノムからのものである。いくつかの実施形態において、ゲノム配列は非ヒトゲノムからのものである。いくつかの実施形態において、非ヒトゲノムは、細菌ゲノム、ウイルスゲノム、真菌ゲノム、原生動物ゲノム、および植物ゲノムからなる群から選択される。いくつかの実施形態において、リファレンス配列はアミノ酸配列である。いくつかの実施形態において、アミノ酸配列は既知の配列である。いくつかの実施形態において、アミノ酸配列は機能性配列である。いくつかの実施形態において、アミノ酸配列は合成配列である。いくつかの実施形態において、アミノ酸配列はヒトである。いくつかの実施形態において、アミノ酸配列は非ヒトである。いくつかの実施形態において、非ヒトアミノ酸配列は、細菌配列、ウイルス配列、真菌配列、原生動物配列、および植物（ｆｌｏｒａｌ）（植物（ｐｌａｎｔ））配列からなる群から選択される。いくつかの実施形態において、生成ステップは、直線リファレンス座標系を直接使用すること、および代替配列のｋ－ｍｅｒが直線座標系中に現われるように代替配列のｋ－ｍｅｒを直接インデックス化することによって遂行される。いくつかの実施形態において、生成は、ノードＩＤ、エッジ、またはパスをアサインすることを含まない。

【００４４】

本開示の態様は、リファレンス配列中のｋ－ｍｅｒのインデックスのクエリを代替パスにより実行する方法を開示し、本方法は、（ａ）リファレンス配列からの複数のｋ－ｍｅｒを含むインデックスを、代替パスにより検索することと；（ｂ）ｋ－ｍｅｒを備えたインデックスのクエリを、１計算スレッドあたり毎秒６９，０００以上のｋ－ｍｅｒの率で、実行することと、を含む。

【００４５】

いくつかの実施形態において、クエリは、１計算スレッドあたり毎秒３４５，０００以上のｋ－ｍｅｒの率で複数の計算スレッド上で遂行される。いくつかの実施形態において、計算スレッドの数は４より大きい。いくつかの実施形態において、クエリは、１計算コアあたり毎秒３５５，０００以上のｋ－ｍｅｒの率で複数の計算コア上で遂行される。いくつかの実施形態において、プロセッサ作業のうちの９５％を超えるものが、インデックスのクエリ専用である。いくつかの実施形態において、プロセッサ作業は、インデックスのクエリに単独で専用の作業である。いくつかの実施形態において、プロセッサ作業は、カーネルタスク、メモリスワップ、またはＩ／Ｏからなる群から選択される１又は２以上のタスクを含まない。いくつかの実施形態において、ｋ－ｍｅｒは少なくとも２０の長さである。いくつかの実施形態において、ｋ－ｍｅｒは少なくとも３２の長さである。

【００４６】

本開示の態様は、配列を比較する方法を開示し、本方法は、（ａ）シーケンサーがシーケンシングアッセイを遂行している間に、シーケンサーからのリードを検索することと；（ｂ）シーケンサーがシーケンシングアッセイを遂行している間に、リードを配列へ比較することと；（ｃ）特異的な遺伝子座がリード中にあるかどうかを決定することと、を含む。

【００４７】

いくつかの実施形態において、本方法は、特異的な遺伝子座の出現のカウントを蓄積することを更に含む。

【００４８】

本開示の態様は、代替配列を有するリファレンス配列中の遺伝子座をフェーズ化する方法を提供し、本方法は、（ａ）各々が遺伝子座を含む配列中の複数の代替パスの出現の数を検索することと；（ｂ）複数の代替パスについての出現の数に基づいて、異なる代替パスからの遺伝子座を共通のグループへとグループ化することと；（ｃ）共通のグループ化に基づいて、遺伝子座をフェーズ化することと、を含む。

【００４９】

本開示の態様は、配列を比較する方法を提供し、本方法は、（ａ）シーケンサーがシーケンシングアッセイを遂行している間に、シーケンサーからのリードを検索することと；（ｂ）シーケンサーがシーケンシングアッセイを遂行している間に、リードを暗号化する

こととと；（ｃ）リードの解読なしに、リードをリファレンス配列と比較することと、を含む。

【００５０】

いくつかの実施形態において、リファレンス配列は配列バリエーションマップである。いくつかの実施形態において、リファレンス配列はリファレンス核酸配列である。いくつかの実施形態において、核酸配列はゲノム配列である。いくつかの実施形態において、核酸配列は、二本鎖ＤＮＡ、一本鎖ＤＮＡ、ＤＮＡ／ＲＮＡハイブリッド、一本鎖ＲＮＡ、二本鎖ＲＮＡ、または相補的ＤＮＡ（ｃＤＮＡ）を含む。いくつかの実施形態において、核酸配列は合成配列である。いくつかの実施形態において、ゲノム配列はヒトゲノムからのものである。いくつかの実施形態において、ゲノム配列は非ヒトゲノムからのものである。いくつかの実施形態において、非ヒトゲノムは、細菌ゲノム、ウイルスゲノム、真菌ゲノム、原生動物ゲノム、および植物ゲノムからなる群から選択される。いくつかの実施形態において、リファレンス配列はアミノ酸配列である。いくつかの実施形態において、アミノ酸配列は既知の配列である。いくつかの実施形態において、アミノ酸配列は機能性配列である。いくつかの実施形態において、アミノ酸配列は合成配列である。いくつかの実施形態において、アミノ酸配列はヒトである。いくつかの実施形態において、アミノ酸配列は非ヒトである。いくつかの実施形態において、非ヒトアミノ酸配列は、細菌配列、ウイルス配列、真菌配列、原生動物配列、および植物配列からなる群から選択される。

【００５１】

本開示の態様は、配列を比較する方法を提供し、本方法は、（ａ）シーケンサーがシーケンシングアッセイを遂行している間に、シーケンサーからのリードを検索することと；（ｂ）シーケンサーがシーケンシングアッセイを遂行している間に、リードを暗号化することと；（ｃ）暗号化されたリードをプロセシングデバイスへ伝送することと、を含む。

【００５２】

いくつかの実施形態において、本方法は、暗号化されたリードを解読することを更に含む。いくつかの実施形態において、本方法は、シーケンシングアッセイが進行中である間に、リードをリファレンス配列と比較することを更に含む。いくつかの実施形態において、本方法は、暗号化されたリードに関係する解読鍵を伝送することを更に含む。

【００５３】

本開示の態様は、配列を比較する方法を提供し、本方法は、（ａ）シーケンサーがシーケンシングアッセイを遂行している間に、シーケンサーからのリードを検索することと；（ｂ）シーケンサーがシーケンシングアッセイを遂行している間に、プロセシングデバイス上のリードを圧縮することと；（ｃ）シーケンサーがシーケンシングアッセイを遂行している間に、リードを解凍することと；（ｄ）シーケンサーがシーケンシングアッセイを遂行している間に、リードをリファレンス配列と比較することと、を含む。

【００５４】

本開示の態様は、既知のバリアントをコールする方法を提供し、本方法は、（ａ）リードを検索することと；（ｂ）リードからｋ－ｍｅｒプロファイルを生成することと；（ｃ）リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに対して、代替パスにより、ｋ－ｍｅｒプロファイルを問い合わせて、既知のバリアントをコールすることと、を含む。

【００５５】

いくつかの実施形態において、既知のバリアントは、１コアあたり毎秒１０のバリアントコール以上の率で、少なくとも単一のコアを使用してコールされる。

【００５６】

本開示の態様は、シーケンシングデータの圧縮のためのシステムを提供し、本システムは、（ａ）シーケンシングデータの受理のための、受理モジュールと；（ｂ）シーケンシングデータの保存のための、メモリユニットと；（ｃ）メモリおよびその上に保存されたシーケンシングデータへのアクセスを有し、シーケンシングデータのうちのいくつかまたはすべてをセーブするように構成された、エンコーディングモジュールと、を含む。

【００５７】
いくつかの実施形態において、シーケンシングデータは、受理したシーケンシングデータの９０％以上のレベルへ圧縮される。いくつかの実施形態において、シーケンシングデータは、配列識別子、塩基コールデータ、コメントライン、および塩基コールデータについてのクオリティ値のうちの１又は２以上から選択されるフィールドを含む。いくつかの実施形態において、シーケンシングデータはアミノ酸コールデータを含む。いくつかの実施形態において、エンコーディングモジュールは、シーケンシングデータ中の塩基コールデータを、シーケンシングデータの残りから分離する。いくつかの実施形態において、エンコーディングモジュールは、シーケンシングデータ中のアミノ酸コールデータを、シーケンシングデータの残りから分離する。いくつかの実施形態において、塩基コールデータ中のヌクレオチド塩基は、アデニン、チミン、グアニン、シトシン、および決定されなかった塩基からなる群から選択されるヌクレオチド塩基と関連する。いくつかの実施形態において、アミノ酸コールデータ中のアミノ酸は、アラニン（ａｌａ、Ａ）、アルギニン（ａｒｇ、Ｒ）、アスパラギン（ａｓｎ、Ｎ）、アスパラギン酸（ａｓｐ、Ｄ）、システイン（ｃｙｓ、Ｃ）、グルタミン（ｇｌｎ、Ｑ）、グルタミン酸（ｇｌｕ、Ｅ）、グリシン（ｇｌｙ、Ｇ）、ヒスチジン（ｈｉｓ、Ｈ）、イソロイシン（ｉｌｅ、Ｉ）、ロイシン（ｌｅｕ、Ｌ）、リジン（ｌｙｓ、Ｋ）、メチオニン（ｍｅｔ、Ｍ）、フェニルアラニン（ｐｈｅ、Ｆ）、プロリン（ｐｒｏ、Ｐ）、セリン（ｓｅｒ、Ｓ）、スレオニン（ｔｈｒ、Ｔ）、トリプトファン（ｔｒｐ、Ｗ）、チロシン（ｔｙｒ、Ｙ）、バリン（ｖａｌ、Ｖ）、および決定されなかったアミノ酸からなる群から選択されるアミノ酸と関連する。いくつかの実施形態において、決定されなかった塩基について、エンコーディングモジュールは、塩基コールデータをシーケンシングデータ中のシーケンシングデータの残りから分離せず、決定されなかった塩基と関連するリードをエンコードする。いくつかの実施形態において、決定されなかった塩基と関連するリードは、未分離リードとして圧縮される。いくつかの実施形態において、決定されなかった塩基を備えたリードのロケーションがセーブされる。いくつかの実施形態において、決定されなかった塩基を備えたすべてのリードがセーブされる。いくつかの実施形態において、決定されなかった塩基と関連するリードは長さで２５６塩基未満であり、決定されなかった塩基のロケーションは１バイト未満でセーブされる。いくつかの実施形態において、決定されなかった塩基と関連するリードは長さで６５５３６塩基未満であり、決定されなかった塩基のロケーションは２バイト未満でセーブされる。いくつかの実施形態において、エンコーディングモジュールは差分エンコーディングを使用して情報をセーブする。いくつかの実施形態において、エンコーディングモジュールは、ヌクレオチドベースのデータ上で塩基－４エンコーディングを遂行する。いくつかの実施形態において、エンコーディングモジュールは、化学的特性によるアミノ酸に関する短縮したアルファベットを使用してアミノ酸ベースのデータ上でエンコードすることを遂行する。いくつかの実施形態において、シーケンシングデータの各々のフィールドは順次セーブされる。いくつかの実施形態において、少なくとも２つのフィールドが分離したファイル中でセーブされる。いくつかの実施形態において、少なくとも２つの異なるファイルが、各々のフィールド上で遂行される異なる圧縮アルゴリズムを使用してセーブされる。いくつかの実施形態において、配列識別子フィールド中のデータは、差分エンコーディングを使用して圧縮される。いくつかの実施形態において、塩基コールデータはＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換を使用してプロセシングされる。いくつかの実施形態において、塩基コールデータの処理は、ランレングスエンコーディングを実行すること、およびＨｕｆｆｍａｎエンコーディングを使用して圧縮することを更に含む。いくつかの実施形態において、コメントのためのラインは差分エンコーディングを使用して圧縮される。いくつかの実施形態において、フィールドが空の場合に、コメントのための追加のラインは無視される。いくつかの実施形態において、クオリティ値データはＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換を使用して処理される。いくつかの実施形態において、処理は、ランレングスエンコーディングを実行すること、およびＨｕｆｆｍａｎエンコーディングを使用して圧縮することを更に含む。

【００５８】
配列アライメントマップ（ＳＡＭ）データの圧縮のためのシステムがであって、本システムは、（ａ）ＳＡＭデータをその上に保存したメモリと；（ｂ）メモリおよびその上に保存されたＳＡＭデータへのアクセスを有し、ＳＡＭデータを８０％以上のレベルへ圧縮するように構成された、エンコーディングモジュールと、を含む。
【００５９】
いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のクエリテンプレート名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のリファレンス配列名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中の左端のマッピング位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のメイトリードのリファレンス名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＳＡＭデータ中のメイトリードの位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、Ｈｕｆｆｍａｎコーディング方法を使用して、ＣＩＧＡＲ文字列からのデータを圧縮する。いくつかの実施形態において、エンコーディングモジュールは、辞書ベースの方法を使用して、ＣＩＧＡＲ文字列からのデータを圧縮する。いくつかの実施形態において、エンコーディングモジュールは、塩基４エンコーディングを使用して、ＳＡＭデータからの塩基コールデータを圧縮する。いくつかの実施形態において、エンコーディングモジュールは、ＳＡＭデータからのクオリティデータを圧縮する。いくつかの実施形態において、（ａ）エンコーディングモジュールは、差分エンコーディングを使用して、クエリテンプレート名、リファレンス配列名、左端のマッピング位置、メイトリードのリファレンス名、およびメイトリードの位置のうちの１又は２以上を含むＳＡＭデータからのデータを圧縮するように構成され；（ｂ）エンコーディングモジュールは、Ｈｕｆｆｍａｎコーディングまたは辞書ベースの方法を使用して、ＣＩＧＡＲ文字列を含むＳＡＭデータからのデータを圧縮するように構成され；（ｃ）エンコーディングモジュールは、塩基－４エンコーディングを使用して、塩基コールデータを含むＳＡＭデータからのデータを圧縮するように構成され；（ｄ）エンコーディングモジュールは、クオリティデータを含むＳＡＭデータからのデータを圧縮するように構成される。いくつかの実施形態において、ＳＡＭデータは順次順序付けられる。
【００６０】
本開示の態様は、ＶＣＦデータの圧縮のためのシステムを提供し、本システムは、ＶＣＦデータをその上に保存したメモリと；メモリおよびその上に保存されたＶＣＦデータへアクセスでき、ゲノムデータをＶＣＦデータの９５％以上のレベルへエンコードするように構成された、エンコーディングモジュールと、を含む。
【００６１】
いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のクエリテンプレート名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のリファレンス配列名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中の左端のマッピング位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のメイトリードのリファレンス名を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、差分エンコーディングを使用して、ＶＣＦデータ中のメイトリードの位置を圧縮する。いくつかの実施形態において、エンコーディングモジュールは、Ｈｕｆｆｍａｎコーディング方法を使用して、データ形式をｃｉｇａｒ文字列に圧縮する。いくつかの実施形態において、エンコーディングモジュールは、辞書ベースの方法を使用して、データ形式をｃｉｇａｒ文字列に圧縮する。いくつかの実施形態において、エンコーディングモジュールは、塩基４エンコーディン

グを使用して、ＶＣＦデータからの塩基コールデータを圧縮する。いくつかの実施形態において、エンコーディングモジュールはＶＣＦデータからのクオリティデータを圧縮する。いくつかの実施形態において、決定されなかった塩基コールのリードについて、決定されなかった塩基コールの位置がセーブされる。いくつかの実施形態において、決定されなかった塩基コールを備えたすべてのリードについて、決定されなかった塩基コールの位置がセーブされる。

【００６２】

本開示の態様は、生のプロテオーム配列データをアライメントさせる方法を提供し、本方法は、（ａ）生のプロテオーム配列データを検索することと；（ｂ）該生のプロテオーム配列データを、バリエーションマップ上のロケーションへマッピングし、そこで、バリエーションマップが代替パスを含むことと；（ｃ）バリエーションマップ上のそのロケーションに従って、生のプロテオーム配列データをアライメントさせることと、を含む。

【００６３】

いくつかの実施形態において、マッピングはグラフアライメントによって遂行される。いくつかの実施形態において、グラフアライメントは少なくとも１つのグラフを使用する。いくつかの実施形態において、マッピングはギャップアライメントを使用して遂行される。いくつかの実施形態において、マッピングはセミギャップアライメントを使用して遂行される。いくつかの実施形態において、本方法は、代替パスからの特定のパスがマッピングステップにおいてマッピングされる回数を蓄積することを更に含む。

【００６４】

本開示の態様は、少なくとも代替配列パスを生成する方法を提供し、本方法は、ａ）リファレンス配列を得ることと；ｂ）リファレンス配列に代わるリファレンス配列上の相関する遺伝子座を検索することと、ｃ）相関する遺伝子座を含む、少なくとも１つの代替配列パスを生成することと、を含む。

【００６５】

いくつかの実施形態において、相関する遺伝子座は２つ以上の別個の起源からのものである。いくつかの実施形態において、相関する遺伝子座は、リファレンス配列上の２つ以上のロケーションへマッピングされる異なる配列のセットである。

【００６６】

本開示の態様は、アミノ酸配列を比較する方法を提供し、本方法は、ａ）アミノ酸配列を検索することと；ｂ）アミノ酸配列からｋ－ｍｅｒプロファイルを生成することと；ｃ）データベース内の複数の配列からのｋ－ｍｅｒプロファイルのインデックスに対して、ｋ－ｍｅｒプロファイルを問い合わせることと、を含む。

【００６７】

本開示の態様は、アミノ酸配列に対して既知のバリアントをコールするためのシステムを提供し、本システムは、（ａ）アミノ酸配列を検索することと；（ｂ）アミノ酸配列からｋ－ｍｅｒプロファイルを生成することと；（ｃ）既知のアミノ酸およびポリマー配列のデータセットからのｋ－ｍｅｒプロファイルのインデックスに対して、代替パスにより、ｋ－ｍｅｒプロファイルを問い合わせて、バリアントをコールすることと、を含む。

【００６８】

いくつかの実施形態において、既知のバリアントは、１コアあたり毎秒１０のバリアントコール以上の率で、少なくとも単一のコアを使用して同定される。

【００６９】

本開示の態様は、アミノ酸配列データの圧縮のためのシステムを提供し、本システムは、（ａ）アミノ酸配列データを受理するためのモジュールと；（ｂ）アミノ酸配列データを保存するためのメモリユニットと；（ｃ）メモリおよびその上に保存されたアミノ酸配列データへアクセスでき、アミノ酸配列データのうちのいくつかまたはすべてをエンコードするように構成された、エンコーディングモジュールと、を含む。

【００７０】

本開示の態様は、サンプル中の種および／または株を同定する方法を提供し、本方法は

、ａ）リードを検索することと；ｂ）リードからｋ－ｍｅｒプロファイルを生成することと；ｃ）リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに対して、代替パスにより、ｋ－ｍｅｒプロファイルを問い合わせて、バリアントをコールすることと；ｄ）コールされたバリアントに基づいて、サンプル中に存在する種または株を決定することと、を含む。

【００７１】

いくつかの実施形態において、ｋ－ｍｅｒプロファイルは、ギャップを導入したｋ－ｍｅｒを含む。いくつかの実施形態において、ｋ－ｍｅｒプロファイルは、１，０００，０００の塩基中最大１の頻度で異なる配列を圧縮する。いくつかの実施形態において、代替パスのインデックスはフェーズ化された情報を含む。いくつかの実施形態において、株間の差と直接関連するｋ－ｍｅｒのみが使用される。いくつかの実施形態において、ｋ－ｍｅｒインデックスのサイズは、リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに比較して、少なくとも９９％低減される。いくつかの実施形態において、ｋ－ｍｅｒインデックスのサイズは、リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに比較して、少なくとも９９．９％低減される。いくつかの実施形態において、株間の差と直接関連するｋ－ｍｅｒ決定のみがバリアント決定のために使用される。いくつかの実施形態において、ｋ－ｍｅｒインデックスのサイズは、リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに比較して、少なくとも９９％低減される。いくつかの実施形態において、そこで、ｋ－ｍｅｒインデックスのサイズは、リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに比較して、少なくとも９９．９％低減される。

【図面の簡単な説明】

【００７２】

【図１】ゲノム分析のための例示的な連続的モデルを示す。

【図２】ゲノム分析のための例示的なストリーミングモデルを示す。

【図３】ゲノム分析のための例示的な自己更新ストリーミングモデルを示す。

【図４】２つの配列のｋ－ｍｅｒプロファイルの例を示す。

【図５】ＩＤと共に、リファレンスおよび代替パスの例を示す。

【図６Ａ】例示的な候補アライメントロケーション（ＣＡＬ）生成およびリードグラフアライメントワークフローを示す。

【図６Ｂ】例示的な候補アライメントロケーション（ＣＡＬ）生成およびリードグラフアライメントワークフローを示す。

【図６Ｃ】例示的な候補アライメントロケーション（ＣＡＬ）生成およびリードグラフアライメントワークフローを示す。

【図６Ｄ】例示的な候補アライメントロケーション（ＣＡＬ）生成およびリードグラフアライメントワークフローを示す。

【図６Ｅ】例示的な候補アライメントロケーション（ＣＡＬ）生成およびリードグラフアライメントワークフローを示す。

【図６Ｆ】例示的な候補アライメントロケーション（ＣＡＬ）生成およびリードグラフアライメントワークフローを示す。

【図７Ａ】ｋ－ｍｅｒのコンパチビリティーまたはインコンパチビリティーについての例示的な定義を示す。

【図７Ｂ】ｋ－ｍｅｒのコンパチビリティーまたはインコンパチビリティーについての例示的な定義を示す。

【図７Ｃ】ｋ－ｍｅｒのコンパチビリティーまたはインコンパチビリティーについての例示的な定義を示す。

【図７Ｄ】ｋ－ｍｅｒのコンパチビリティーまたはインコンパチビリティーについての例示的な定義を示す。

【図７Ｅ】ｋ－ｍｅｒのコンパチビリティーまたはインコンパチビリティーについての例示的な定義を示す。

【図８Ａ】ＣＡＬの数を低減するオフセット正規化を例証する例示的な概略図を示す。
【図８Ｂ】ＣＡＬの数を低減するオフセット正規化を例証する例示的な概略図を示す。
【図９】配列グラフへのダイナミックプログラミングまたはアライメントの開始に使用するシードを決定するための例示的なプロセスを示す。
【発明を実施するための形態】
【００７３】
（定義）
当業者によるこれらの用語の理解に加えて、以下の用語を以下で議論してこの明細書において使用されるような用語の意味を例証する。本明細書および請求項において使用される場合、単数形「１つの（ａ）」、「１つの（ａｎ）」および「その（ｔｈｅ）」には、文脈が明確に指示しない限り複数形の相応が含まれ得る。例えば、「細胞」という用語には複数の細胞が含まれ、それらにはその混合物が含まれる。
【００７４】
本明細書において使用される場合、「アライメント」という用語は、シーケンサーによって生成された全ての配列文字列をリファレンス文字列へマッチさせる任意のコンピュータープロセスとすることができる。アライメントは、例えばＳｍｉｔｈ　Ｗａｔｅｒｍａｎのローカルアライメント、ギャップアライメント、またはセミギャップアライメントとすることができる。
【００７５】
ゲノム中の変動は「代替パス」として表わすことができる。例えば、第１のゲノムは、ＤＮＡ塩基（文字Ａ、Ｃ、ＴおよびＧによって表わされる）の直線配列とすることができる。第２のゲノムは、第１の被験体と第２の被験体との間の生物学的多様性を表わすＤＮＡ塩基の異なる配列を有することができる。
【００７６】
「グラフリファレンス」は、１又は２以上の配列の圧縮表現を表わし、すべての配列によって共有される配列インターバルは１つの配列パスへ折り畳まれ、異なる配列インターバルは代替パスとして維持される。
【００７７】
「直線リファレンス」は、２つ以上のオプションが各々の要素の同一性について規定されない配列の表示とすることができる。いくつかの規格において、配列は核酸であり、他のものにおいて、それらはタンパク質である。
【００７８】
「関連する遺伝子座」は、一般的には同じゲノム領域を表わす、２つのゲノム、または被験体ゲノムおよびリファレンスゲノムからの配列を意味することができる。それは、１つのゲノムであるが２又は３以上の異なる領域からの配列も意味することができる。一般的には、相関する遺伝子座は、同じ種内にあるであろう。それらは、一般的には同じ被験体内にもあるであろう。相関する遺伝子座は、連鎖不平衡、ハプロイド上の保存された領域、先験的データ（１０００ゲノム等）または同種のものを介して相関させることができる。
【００７９】
ゲノム情報は「フェーズ化」することができる。フェーズ化された配列は、染色体コピーにわたって異なり得る突然変異を含むユニークな染色体含有物を捕捉する。フェーズ化されたシーケンシングは、いくつかの事例において、母方および父方から遺伝する対立遺伝子を識別することができる。
【００８０】
「ｋ－ｍｅｒ」という用語は、配列中に含有される長さｋのすべての可能なサブ配列を指すことができる。
【００８１】
マップ構造の中へ入る個別の被験体ゲノムが、一次配列とマッチするポイントでリファレンスゲノムへとマージされ、ゲノムに沿った追加の代替パスとして現われるバリエーシ

ョンがある場合に、「ゲノムバリエーションマップ」は構築することができる。結果として生じるマップは、ゲノムバリエーションの複数の形状を含むことになる。ゲノムバリエーションマップはグラフとして表わすことができる。

【００８２】

「アセンブリー」という用語は、すべての配列文字列のセットが由来するもとの配列文字列を再構築する目的で、シーケンサーによって生成された配列文字列が互いの間でマージされる任意のコンピュータープロセスとすることができる。

【００８３】

「リモートアライメント」という用語は、アライメントが独立したサブタスクの特定のあらかじめ定義された数へと分割され、サブタスクが、配列文字列を受理すること、配列文字列をアライメントさせること、およびすべてのサブタスクの最終的な全体の完全なアライメントを提供する適切な計算デバイスへ配列文字列を伝送することができる独立したコンピューターデバイスによって遂行することができる、任意のコンピュータープロセスとすることができる。

【００８４】

「インデックス」という用語は、データのアクセスの最適化に使用される任意のデータベースとすることができる。データベースは鍵で構成できる。これらの鍵は、もとのデータベース上のサーチが基づくであろう属性とすることができる。ゲノムバリエーションマップまたは配列グラフのインデックスは、配列グラフ中の短配列のオフセットおよび配列グラフ中で短配列が属する代替パスと一緒に配列グラフ中に含有される、短配列（ｋ－ｍｅｒ）のデータベースを含有することができる。ゲノムバリエーションマップまたは配列グラフのインデックスは、配列グラフ上でＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換（ＢＷＴ）からなるデータベースとすることができ、それは位置マーカーを使用して、変換された配列内の代替パスのロケーションを注釈することができる。この後者のインデックスは当業者に公知のウェーブレット木を使用して保存することができる。他の事例において、インデックスは、エンコードのためにウェーブレット木を使用する位置マーカーによるＢＷＴを含まない。

【００８５】

「ハッシュテーブル」という用語は、インデックス内の高速化されたサーチを可能にすることができる方法または構造を表すことができる。

【００８６】

「リファレンス配列」という用語は、手元にある分子の定義に要求される情報から構成される配列文字列を指すことができる。例えば、全体のヒトゲノムは、ヒトゲノムの定義に準拠する約３０億の塩基からなるヌクレオチドの配列文字列とすることができる。リファレンスゲノム（あるいはリファレンスアセンブリー）はリファレンス配列とすることができる。リファレンスゲノムは、関連する核酸のセットの代表的な例としてアセンブルされたデジタル核酸配列データベースとすることができる。リファレンスゲノムは例えば特定の種のゲノムの例とすることができる。いくつかの実例において、リファレンスゲノムは代替パスを含むことができる。

【００８７】

「メタデータ」という用語は、整合性があり得る順序づけられた様式で追加された異なるタイプの構造の構成物を表す。

【００８８】

「生の遺伝子配列データ」はシーケンシング反応から得られるデータである。生の遺伝子配列データはテキストベースとすることができ、例えばそれはＦＡＳＴＡフォーマットを有することができる。ＦＡＳＴＡフォーマットは、ヌクレオチド配列またはペプチド配列のいずれかを表わすためのテキストベースのフォーマットであり、ヌクレオチドまたはアミノ酸は１文字コードを使用して表わされる。生の遺伝子配列データは、生物学的配列（例えば塩基コールデータまたはアミノ酸コールデータ）およびその対応するクオリティスコアの両方ならびに他の関連データまたはメタデータの保存のための、テキストベース

のフォーマットとすることができる。例えば、それはＦＡＳＴＱフォーマットを有することができる。ＦＡＳＴＱフォーマットは、生物学的配列およびその対応するクオリティスコアの両方の保存のためのテキストベースのフォーマットである。いくつかの実例において、配列文字およびクオリティスコアは各々簡潔性のための単一のＡＳＣＩＩ文字によりエンコードされる。いくつかの実例において、生の遺伝子配列データは、フォーマットコンバータを使用して、１つのフォーマットから別のフォーマットに転換することができる。いくつかの実例において、生の遺伝子配列データは「リード」と呼ばれる。

【００８９】

「シーケンシングデバイス」はシーケンシング反応を遂行するデバイスである。シーケンシングデバイスを使用して、生の遺伝子配列データを生成することができる。いくつかの実例において、シーケンシングデバイスがシーケンシング反応を遂行している間に、本明細書において記載される方法を遂行することができる。例えば、配列データがシーケンシングデバイスによって生成される時に、それらのデータは暗号化され、暗号化される間にアライメントすることができる。いくつかの実例において、シーケンシングデバイスはＳＡＭデータを出力することができる。

【００９０】

「リードペア」は、少なくとも２つの領域がシーケンスされた、接続された核酸配列を起源とする、シーケンスされたリードのペアを意味することができる。シーケンスされたリードの間のヌクレオチド文字列の配列は、いくつかの事例において、既知でない。いくつかのリードペアの生成技法において、全ヌクレオチド文字列の長さはあまり変動しない。シーケンスされたリードがサンプルに類似するリファレンス配列へアライメントされる場合に、全体のヌクレオチド文字列の長さ（すなわちインサート長）についての分布は、いくつかの事例において、推測することができる。この情報は、構造バリアントコーリング（この分布において非常に低い確率を有するインサート長とアライメントするリードペア）のために使用することができ、サンプル中に存在する構造バリアントがあることを示す。加えて、いくつかのリードペア生成技法について、ペアの２つのシーケンスされたリードは、特定の配向によりリファレンス配列へアライメントする可能性が最も高く、例えば、左端のリードはリファレンス配列へそのままアライメントし（「フォワード」配向性）、その一方で、右端のリードはリファレンス配列の相補物へアライメントする（「リバース」配向性）。アライメントされたリードペアにおける最も可能性の高い配向からの逸脱は構造バリエーションについての指標とすることができる。

【００９１】

リードペアにおけるインサート長の長さは、利用される特定のシーケンシング技術に依存して変動することができる。ＮＧＳプラットフォームは、インサート長が何百から何千または何万もの塩基対のサイズで変動することができるリードペアを提供することができる。

【００９２】

いくつかの規格において、インサート長の分布は特異的なモデルに従う。いくつかの規格において、「標的適用」は、グラフリファレンスがバリエーションを表わすヒト集団を指すことができる。他の事例において、「標的適用」は、対象となる１又は２以上の集団（種、特異的患者集団、地理的集団、植物集団、真菌、細菌もしくはウイルスの株もしくは株のセットまたはその組み合わせ等）を指すことができる。標的適用は、１又は２以上の個体または種の２倍体または多倍数体の特徴を包含することもできる。

【００９３】

ＳＡＭフォーマット（または「ＳＡＭデータ」）は、タブ区切りの一連のＡＳＣＩＩカラム中で配列データを保存するためのテキスト形式である。ＳＡＭデータは、その姉妹ＢＡＭフォーマット（「ＢＡＭデータ」）（それは圧縮されインデックス化されたバイナリ形式で同じデータを保存する）の人間可読バージョンとして生成することができる。ＳＡＭフォーマットデータはアライナーから出力することができ、アライナーは、ＦＡＳＴＱファイルを読み取り、既知のリファレンスゲノムに関する位置へ配列をアサインする。Ｓ

ＡＭを使用して、シーケンシング機械から直接生成された非アライメント配列データを保管することもできる。いくつかの実例において、ＳＡＭデータはＣＩＧＡＲ文字列を含む。ＣＩＧＡＲ文字列は塩基長の配列さおよび関連操作である。それらを使用して、特性（例えばその塩基はリファレンスとアライメントする（マッチ／ミスマッチのいずれか））が、リファレンスから欠失されるか、またはリファレンス中にない挿入であることを示す。

【００９４】

バリアントコールフォーマット（ＶＣＦ）は、バイオインフォマティクスの学際的分野における配列バリエーションの保存のために使用されるテキストファイルのフォーマットを規定する。「ＶＣＦデータ」はＶＣＦフォーマットで保存されたデータである。バリアントコールフォーマットは、リファレンス配列と一緒に保存されるのに必要とされるバリエーションのみを保存する。

【００９５】

ジェネラルフィーチャフォーマット（ＧＦＦ）はすべての遺伝子データを保存し、それはゲノムにわたって共有されるので、そのほとんどは冗長である。「ＧＦＦデータ」はＧＦＦフォーマットで保存されたデータである。

【００９６】

「グラフアライメント」は、グラフおよびグラフ表現を使用するゲノムデータの分析を包含することができる。例えば、ゲノムバリエーションマップグラフを使用して、グラフアライメントによって生の配列データを分析することができる。

【００９７】

「被験体」という用語は、本明細書において使用される場合、一般的には発現される遺伝物質を含有する生物学的存在を指す。生物学的存在は、植物、動物、または微生物（例えば細菌、ウイルス、真菌、原生動物が含まれる）とすることができる。被験体は、インビボで得られるかまたはインビトロで培養される、生物学的存在の組織、細胞およびそれらの子孫とすることができる。被験体は哺乳類とすることができる。哺乳類はヒトとすることができる。

【００９８】

「サンプル」または「核酸サンプル」は、核酸を含有するかまたは含有すると推定される任意の物質を指すことができる。サンプルは、被験体から得られる生物学的サンプルとすることができる。核酸は、ＲＮＡ、ＤＮＡ（例えばゲノムＤＮＡ、ミトコンドリアＤＮＡ、ウイルスＤＮＡ、合成ＤＮＡ、またはＲＮＡから逆転写されたｃＤＮＡ）とすることができる。核酸サンプル中の核酸は、一般的にはハイブリダイズされたプライマーの伸長のためのテンプレートとして供される。いくつかの実施形態において、生物学的サンプルは液体サンプルである。液体サンプルは、全血、血漿、喀痰、滑液、血清、腹水、脳脊髄液、汗、尿、涙液、唾液、頬腔サンプル、腔すすぎ液、または器官すすぎ液とすることができる。液体サンプルは、本質的に無細胞の液体サンプル（例えば血漿、血清、汗、尿、涙液）とすることができる。他の実施形態において、生物学的サンプルは、固体の生物学的サンプル、例えば糞便、組織生検（例えば腫瘍生検）である。サンプルは、インビトロの細胞培養構成物（細胞培養培地中の細胞の増殖から生じる馴化培地、組換え細胞および細胞構成要素が含まれるがこれらに限定されない）も含むことができる。

【００９９】

「ヌクレオチド」は核酸を形成できる生体分子とすることができる。ヌクレオチドは、公知のプリン塩基およびピリミジン塩基だけでなく修飾された他のヘテロ環塩基も含有する部分を有することができる。かかる修飾は、メチル化されたプリンもしくはピリミジン、アシル化されたプリンもしくはピリミジン、アルキル化されたリボース、または他のヘテロ環を包含する。加えて、「ヌクレオチド」という用語は、ハプテン、ビオチン、または蛍光標識を含有し、従来のリボース糖およびデオキシリボース糖だけでなく他の糖も同様に含有することができるこれらの部分を包含する。修飾されたヌクレオシドまたはヌクレオチドは、糖部分上の修飾も包含し、例えば、そこで、ヒドロキシル基のうちの１又は

２以上は、ハロゲン原子もしくは脂肪族基と置き換えられるか、エーテル、アミンとして官能化されるか、または同種のものである。

【０１００】

「ヌクレオチド」には、ロック核酸（ＬＮＡ）または架橋核酸（ＢＮＡ）を含むこともできる。ＢＮＡおよびＬＮＡは、一般的にはリボース部分が２’酸素と４’炭素を接続する架橋により修飾される、修飾リボヌクレオチドを指す。一般的に、架橋は、多くの場合Ａ型二重鎖において見出される３’－エンド（Ｎｏｒｔｈ）立体配座でリボースを「ロックする」。「ロック核酸」（ＬＮＡ）という用語は、一般的にはＢＮＡのクラスを指し、リボース環は、２’－Ｏ原子を４’－Ｃ原子と接続するメチレン架橋により「ロックされる」。ＤＮＡおよびＲＮＡ中に現われる６つの一般的な核酸塩基（Ｔ、Ｃ、Ｇ、Ａ、Ｕおよびｍ Ｃ）を含有するＬＮＡヌクレオシドは、標準的なＷａｔｓｏｎ－Ｃｒｉｃｋ塩基対合ルールに従って、それらの相補的なヌクレオシドと塩基対を形成することができる。したがって、所望される場合は常に、ＢＮＡおよびＬＮＡのヌクレオチドは、オリゴヌクレオチド中のＤＮＡ塩基またはＲＮＡ塩基と混合することができる。ロックされたリボース立体配座は、塩基のスタッキングおよび骨格の前組織化を促進する。塩基のスタッキングおよび骨格の前組織化は、増加した二重鎖の熱安定性（例えば増加したＴｍ）および識別力を生じることができる。ＬＮＡは、他の核酸で可能でない条件下で単一の塩基ミスマッチを識別することができる。

【０１０１】

「ポリヌクレオチド」、「核酸」、「ヌクレオチド」、「配列」および「オリゴヌクレオチド」という用語は、互換的に使用することができる。それらは、任意の長さのヌクレオチドのポリマー形状（デオキシリボヌクレオチドまたはリボヌクレオチド、またはその類似体のいずれか）を指すことができる。ポリヌクレオチドは任意の三次元構造も有し、既知または未知の任意の機能を遂行することができる。以下は、ポリヌクレオチドの非限定的な例である。遺伝子または遺伝子断片のコーディング領域または非コーディング領域、連鎖分析から定義される遺伝子座（複数可）、エクソン、イントロン、メッセンジャーＲＮＡ（ｍＲＮＡ）、転移ＲＮＡ、リボソームＲＮＡ、リボザイム、ｃＤＮＡ、組換えポリヌクレオチド、分岐ポリヌクレオチド、プラスミド、ベクター、任意の配列の単離されたＤＮＡ、任意の配列の単離されたＲＮＡ、核酸プローブ、およびプライマー。ポリヌクレオチドは、修飾ヌクレオチド（メチル化ヌクレオチドおよびヌクレオチド類似体）を含むことができる。存在する場合、ヌクレオチド構造への修飾は、ポリマーのアセンブリーの前または後に与えられ得る。ヌクレオチドの配列は非ヌクレオチド構成要素によって中断することができる。ポリヌクレオチドは、標識構成要素とのコンジュゲーション等によって重合の後に更に修飾することができる。

【０１０２】

「バリアント」は、核酸配列またはアミノ酸配列（例えば遺伝子または遺伝子産物）の正常な配列中の変更とすることができる。いくつかの実例において、遺伝子型および対応する表現型はバリアントと関連する。他の実例において、バリアントに公知の機能はない。バリアントは、リファレンス配列と比べた配列差も意味することができる。バリアントはＳＮＰとすることができる。バリアントはＳＮＶとすることができる。バリアントは複数のヌクレオチドの挿入とすることができる。バリアントは複数のヌクレオチドの欠失とすることができる。バリアントは突然変異とすることができる。バリアントはコピー数変動とすることができる。バリアントは構造バリアントとすることができる。バリアントは同義突然変異へと変換されるヌクレオチドの挿入または欠失とすることができる。バリアントは非同義突然変異へと変換されるヌクレオチドの挿入または欠失とすることができる。

【０１０３】

「一塩基多型（ＳＮＰ）」は、１つの塩基長であるバリアントを意味することができる。

【０１０４】

「インデル」は、２以上の塩基長の小さなバリアントを意味することができる。インデルは挿入または欠失とすることができる。いくつかの実例において、インデルは小さな構造バリアントとすることができる。

【０１０５】

「既知のバリアント」は、以前に報告されたバリアントを意味することができる。既知のバリアントはグラフリファレンスの中へ含まれるバリアントとすることができる。いくつかの規格において、既知のバリアントは、外部媒体（データベース、ジャーナル、医療記録等）において報告される。いくつかの規格において、報告は内部であると判断することができる。

【０１０６】

「新規のバリアント」は、グラフリファレンス中に含まれないサンプル中のバリアントとすることができる。いくつかの規格において、新規バリアントは、以前に報告されたが含まれていなかったバリアントとすることができる。他の規格において、新規バリアントは、これまでに未知のバリアントとすることができる。

【０１０７】

「構造バリアント」は、通常５０又はそれ以上の塩基と判断されるより長いバリアントを意味することができる。

【０１０８】

「リードサイクル」は、リードのセットのバルクを通してスキャンするプロセスとすることができる。少数のリードは、リードサイクルにおいて廃棄されるかまたは二回以上含むことができる。他のリードサイクルにおいて、リード上での異なる操作に取り掛かることができる。例えば、これらには、リードクオリティの再較正、再アライメント、フィルタリング、他の統計的操作を含むことができるが、これらに限定されない。

【０１０９】

「バリアントコーリング」は、バリアントが配列中に存在するかどうかを決定するプロセスとして定義することができる。バリアントには、ＳＮＰ、インデル、構造バリアント、および同義または非同義の誘発突然変異を含むことができるが、これらに限定されない。

【０１１０】

「標的ポリヌクレオチド」という用語は、本明細書において使用される場合、一般的には研究下で対象となるポリヌクレオチドを指す。ある特定の実施形態において、標的ポリヌクレオチドは、研究下で対象となる１又は２以上の配列を含有する。標的ポリヌクレオチドは、例えばゲノム配列を含むことができる。標的ポリヌクレオチドは標的配列を含み、その存在、量および／もしくはヌクレオチド配列、またはこれらにおける変化は決定されることが所望される。いくつかの実例において、標的ポリヌクレオチドは代替パスへアライメントされる。

【０１１１】

「ゲノム配列」という用語は、本明細書において使用される場合、ゲノム中に出現する配列を指すことができる。ＲＮＡがゲノムから転写されるので、この用語は、生物体の核ゲノム中に存在する配列に加えて、かかるゲノムから転写されたＲＮＡ（例えばｍＲＮＡ）のｃＤＮＡコピー中に存在する配列を包含する。「ゲノム配列」は、細胞質上でまたはミトコンドリア中に出現する配列とすることもできる。

【０１１２】

「決定すること」、「測定すること」、「評価すること」、「査定すること」、「アッセイすること」、および「分析すること」という用語は、測定の任意の形状を指すように本明細書において互換的に使用することができ、要素が存在するかどうか決定することを包含する。これらの用語は、定量的決定および／または定性的決定の両方を含むことができる。評価は、相対的または絶対的とすることができる。「～の存在を査定すること」は、それが存在するか存在しないかどうかを決定することに加えて、存在するものの量を決定することを含むことができる。

【０１１３】
「ゲノム断片」という用語は、本明細書において使用される場合、ゲノム、例えば動物または植物のゲノム（ヒト、サル、ラット、魚類もしくは昆虫または植物のゲノム等）の領域を指すことができる。ゲノム断片は、アダプターライゲーションされるかまたはされなくてもよい。ゲノム断片は、アダプターライゲーションされるか（その事例において、それは断片の１つのまたは両方の末端へ（少なくとも分子の５’末端へ）ライゲーションされたアダプターを有する）、またはアダプターライゲーションされなくてもよい。
【０１１４】
「シーケンシング」という用語は、本明細書において使用される場合、ポリヌクレオチドのうちの少なくとも１０の連続するヌクレオチドの同一性（例えば少なくとも２０、少なくとも５０、少なくとも１００、少なくとも２００、または少なくとも５００以上の連続するヌクレオチドの同一性）が得られる方法を指すことができる。
【０１１５】
「バーコード配列」という用語は、本明細書において使用される場合、一般的にはアッセイに関する情報をエンコードできるヌクレオチドのユニーク配列を指す。バーコード配列は、問い合わせられる対立遺伝子の同一性、標的ポリヌクレオチドもしくはゲノム遺伝子座の同一性、サンプルの同一性、被験体、またはその任意の組み合わせに関する情報をエンコードできる。バーコード配列は、プライマー、レポータープローブ、または両方の部分とすることができる。バーコード配列は、オリゴヌクレオチドの５’末端もしくは３’末端にあること、またはオリゴヌクレオチドの任意の領域中に所在することができる。
【０１１６】
「突然変異」という用語は、本明細書において使用される場合、一般的にはゲノムまたは機能遺伝子のヌクレオチド配列の変化を指す。突然変異は、ＤＮＡの大きなセクション（例えばコピー数変動）を包含することができる。突然変異は染色体全体（例えば異数性）を包含することができる。突然変異はＤＮＡの小さなセクションを包含することができる。ＤＮＡの小さなセクションを包含する突然変異の例には、例えば点突然変異または一塩基変異多型、多塩基多型、挿入（例えば遺伝子座での１又は２以上のヌクレオチドの挿入）、複数のヌクレオチド変化、欠失（例えば遺伝子座での１又は２以上のヌクレオチドの欠失）、および逆位（例えば１又は２以上のヌクレオチドの配列の反転）が含まれる。
【０１１７】
「遺伝子座」という用語は、本明細書において使用される場合、染色体上の遺伝子、ヌクレオチド、またはヌクレオチド配列のロケーションを指すことができる。遺伝子座の「対立遺伝子」は、本明細書において使用される場合、遺伝子座でのヌクレオチドまたは配列の代替の型を指すことができる。「野生型対立遺伝子」は、一般的には、被験体の集団中で最も高い頻度を有する対立遺伝子を指す。「野生型」対立遺伝子は、一般的には疾患と関連しない。「突然変異対立遺伝子」は、一般的には、「野生型対立遺伝子」よりも低い頻度を有し、疾患と関連することができる対立遺伝子を指す。「突然変異対立遺伝子」は、必ずしも疾患と関連しない。「問い合わせられる対立遺伝子」という用語は、一般的には、検出するためのアッセイがデザインされている対立遺伝子を指す。
【０１１８】
「一塩基多型」または「ＳＮＰ」という用語は、一般的には、本明細書において使用される場合、配列内の単一ヌクレオチド置換から結果として生じるタイプのゲノム配列バリエーションを指す。「ＳＮＰ対立遺伝子」または「ＳＮＰの対立遺伝子」は、一般的には特定の遺伝子座でのＳＮＰの代替型を指す。「問い合わせられるＳＮＰ対立遺伝子」という用語は、一般的には、検出するためのアッセイがデザインされているＳＮＰ対立遺伝子を指す。
【０１１９】
配列アライメント
多くの次世代シーケンシング技法は短いリード配列を生成し、次いでそれはアライメントされ、より長い配列情報へとアセンブルすることができる。短いリード配列は、複数の

良好な候補アライメントロケーションがある場合に、正確にアライメントさせるのが難しくなる可能性がある。短いリード配列はサンプル中に存在するバリエーションがある場合に、正確にアライメントさせるのが難しくなる可能性がある。本明細書において、これらの問題に対処する方法が提供される。これらの方法において、リードのペアのための最も良好なアライメントは、ペアにおける個別のリードのアライメントクオリティおよびペアのアライメントのフィーチャ（ペアにおけるアライメントされたリードの間の距離およびペアにおけるアライメントされたリードの相対的方向性等）を考慮することで見出すことができる。いくつかの実施形態において、リードのペアにおけるこれらのアライメントフィーチャを観察する確率は、シーケンシング技術およびサンプルの性質の知識に基づいて推測することができ、ペアのアライメントのスコアリングに使用することができる。

【０１２０】

例えば、典型的なペアエンドシーケンシングライブラリーおよびサンプル配列に類似するリファレンス配列を使用する場合に、大多数のリードペアは、ペアを作るリードにおいて同じ相対的方向性でアライメントし、それは「正常な方向性」と呼ぶことができる。この正常な方向性からの逸脱は、実験誤差またはサンプル中のバリエーションの存在に起因する可能性がある。正常な方向性とは異なる任意のタイプの方向性でアライメントするリードペアの確率は、サンプル中のバリエーション（異なるリードペア方向性と関連する異なるタイプのバリエーションが含まれる）の予想される割合および予想される実験誤差率に基づいて、推測することができる。リードペアにおけるリードの各々の相対的方向性について、インサート長の分布を推測することもできる。リードペア方向性確率とインサート長確率の積は、可能なリードペアアライメントが正確である可能性がどのくらいかを指示するのに使用することができる。この因数を使用して、リードのペアのための可能なアライメントのクオリティに加えてペアにおける個別のリードのアライメントクオリティをスコアリングすることができる。

【０１２１】

リードまたはリードのペアについてのアライメントクオリティは、他のリードまたはリードのペア（例えば同じバーコードを備えた他のリードまたはリードペア）のアライメントフィーチャに依存することもできる。このようにして、同じバーコードを備えたリードまたはリードペアの起源に関する予備的知識を使用して、正確である可能性が最も高いアライメントを同定することができる。

【０１２２】

リードのサブセットにおけるリードのアライメントクオリティは、個別のリードのアライメントクオリティ、およびサブセットにおける他のリードのアライメントフィーチャを観察する推定確率に基づくことができる。例えば、サブセットにおけるリードは同じバーコードを備えたリードとすることができる。

【０１２３】

バリアント（例えばゲノムバリアント）は、類似したリファレンス配列と比べた配列（例えば核酸配列）における差である。構造バリアント（ＳＶ）は、通常の短いリード配列長と比べて大きいバリアントであり（例えば核酸配列について、構造バリアントは通常５０ｂｐよりも大きなバリアントであると判断される）、したがって、短いリード技術により検出することが難しい可能性がある。本明細書において開示されるグラフリファレンスアライメント方法は、より良好な感受性、特異性およびスピードでこれらのバリアントを検出することができるように、アライメントさせる場合に予備的知識としてバリアントを含むことを可能にすることができる。構造バリアントもグラフリファレンスの中へ含むことができ、それらへリードをアライメントさせることによって検出することができる。

【０１２４】

グラフリファレンス中に含まれるバリアントの数は、実際には、メモリ制約または効率の考慮によって制限される可能性がある。そのような理由で、サンプル中に存在するバリアントは、グラフリファレンス中に含まれない場合もあり、それらを「新規」バリアントとして検出する必要がある。新規構造バリアントは、典型的な短いリード長と比べてそれ

らの大きなサイズに起因して検出するのが特に困難になる可能性がある。グラフリファレンスパラダイム内の新規バリアントを検出する方法が、本明細書において開示される。本方法は、（ａ）複数の配列を得ることと、（ｂ）グラフリファレンスに対して複数の配列をアライメントさせることと、（ｃ）変則的にアライメントする複数の配列を使用して新規バリアントを同定することと、を含むことができる。

【０１２５】

新規構造バリアントを検出するために、１）配列リードを得ることと；２）それらをアライメントさせ、それらをファイルへ書き込むことと；３）構造バリエーションを指示すリードについて、アライメントされたリードのファイルをスキャンすることと；４）これらのリードに基づいた構造バリアントを検出することと、ができる。この手順は、例えば図１中で描写される。この例において、配列は、シーケンシング１００から得られた、ＦＡＳＴＱファイル１０１中に含まれるリードデータである。これらのリードのアライメント１１０は、リファレンス配列１１１に関して行なわれる。第１のリードサイクル１１２はアライメントによってリードをプロセシングすることを包含する。生成物はアライメントさせたリード１１３を備えたＳＡＭまたはＢＡＭのファイルである。ファイルは、第１のリードサイクルにおいて分析されたものと同じ数または異なる数のリードを含むことができる。変則的にアライメントされたリードの検出１２０は、アライメントされたリードを備えたＳＡＭまたはＢＡＭファイルを通してスキャンして、変則的にアライメントされたリードを検出することを包含する、第２のリードサイクル１２１において遂行される。変則的にアライメントされたリードは、別個のＳＡＭまたはＢＡＭファイル１２２中でセーブすることができる。次いで、バリアントおよび／または構造バリアントはこれらの変則的にアライメントされたリードに基づいて検出１３０することができる。

【０１２６】

構造バリエーションの存在を指示するリードの割合は、５％未満または１％でさえある。すべてのアライメントされたリードを通してスキャンしてこの割合のリードを収集することは非能率的であり、このステップはしばしば構造バリエーション検出において最も時間のかかるステップである可能性がある。グラフリファレンスパラダイムにおいて新規バリアントを効率的に検出する方法が、本明細書において開示される。本方法は、ａ）複数の配列リードを得ることと；ｂ）グラフリファレンスに対して複数の配列リードのサブセットをアライメントさせることを含むプロセスによって、アライメントされたリードのバッチを生成し、そこで、グラフリファレンスが、代替パスによって表わされた既知のバリアントを含むことと；ｃ）アライメントされたリードのバッチ内の１又は２以上の変則的にアライメントされたリードを同定し、１又は２以上の変則的にアライメントされたリードを使用して未知の構造バリアントを同定することと、を含むことができる。いくつかの実施形態において、本方法は、ｎのリードのバッチについて遂行される。

【０１２７】

図２は、グラフアライメントパラダイムにおいて新規バリアントを効率的に検出する手順の例を描写する。シーケンシング２００は、配列リードデータ（例えばＦＡＳＴＱファイルで）２０１を生ずる。リードは、１又は２以上のリードのバッチ２１２、２１３中でグラフリファレンス２１１を使用して、アライメント２１０される。サンプル中のシーケンシングリードの総数に比較して、バッチ中のリードの数は少ない可能性がある。バッチ中のリードがアライメントされる際に、変則的にアライメントされたものは同定および分離される。それらはＳＡＭまたはＢＡＭファイル２１４へ書き込むことができる。このようにして、アライメントおよびバリアント検出はすべてのアライメントされたリードを通してスキャンする必要なしに遂行されて、変則的にアライメントされたリードを同定することができる。リードはアライメント時に変則的にアライメントされたとしてマークすることができるので、多数のリードにわたって１つのリードサイクル２１５のみが存在し、第２のリードサイクルを必要としない。変則的にアライメントされたリードは、バリアント（構造バリアントが含まれる）をコールする２２０ために使用することができる。

【０１２８】

いくつかの事例において、異なるバッチからの変則的にアライメントされたリードを使用して、新規構造バリアントを同定する。いくつかの事例において、同じバッチからの変則的にアライメントされたリードを使用して、新規構造バリアントを同定する。いくつかの事例において、１又は２以上のバッチからの変則的にアライメントされたリードのサブセットを使用して、例えばファイルへ書き込まれた後に新規構造バリアントを同定する。いくつかの事例において、１又は２以上のバッチからの変則的にアライメントされたリードのサブセットをコンピュータープログラムへ渡して、リードのサブセットを最初にファイルへ書き込むことなしに、新規構造バリアントを同定する。

【０１２９】

いくつかの事例において、既知のバリアントは以前に文書化されたバリアントである。いくつかの事例において、新規バリアントは以前に標的適用について文書化されないバリアントである。いくつかの事例において、既知のバリアントはグラフリファレンス中に含まれているバリアントである。いくつかの事例において、新規バリアントはグラフリファレンス中に含まれていないバリアントである。

【０１３０】

いくつかの事例において、アライメントされたリードのバッチ中で、グラフリファレンス中の代替パスへアライメントされたリードの数をカウントし使用して、既知のバリアントを同定する。

【０１３１】

いくつかの事例において、グラフリファレンス中の代替グラフによって表わされたバリアントは、構造バリアントとすることができる。いくつかの事例において、同定された新規バリアントは構造バリアントである。

【０１３２】

いくつかの事例において、アライメントはギャップアライメントを使用して遂行される。いくつかの事例において、アライメントはセミギャップアライメントを使用して遂行される。

【０１３３】

いくつかの事例において、新規構造バリアントの同定に使用される複数の配列リードのサブセットは、グラフリファレンス中のすべての代替パスへ変則的にアライメントされたリードを含む。

【０１３４】

変則的なアライメントは、大多数のアライメントされたリードペアとは異なるペア方向性（すなわちペア中の２つのアライメントされたリードの方向性）でアライメントすることを含むことができる。変則的なアライメントは、アライメントされたリードペアの平均インサート長または中央インサート長よりも有意に小さいかまたは大きいインサート長を備えたリードペアを含むことができる。変則的なインサート長は、９９番目、９０番目、９５番目、９７番目、９８番目、９７番目、９６番目、９５番目、９４番目、９３番目、９２番目、９１番目、または９０番目のパーセンタイル値よりも大きくすることができる。変則的なインサート長は、アライメントされたリードのサブセットのインサート長の１番目、２番目、３番目、４番目、５番目、６番目、７番目、８番目、９番目、または１０番目のパーセンタイル値よりも小さくすることができる。いくつかのユーザー指定値よりも大きいかまたは小さいような変則的なインサート長を設定することができる。いくつかの事例において、変則的なアライメントは、アライメントされる１つのリードを包含し、その一方で、他のものはアライメントされない。いくつかの事例において、変則的なアライメントは、クリップされた配列の割合を包含する。クリップされた割合は、配列のこの割合がアライメントされないことを意味する。いくつかの実施形態において、クリップされたリードの部分は、リードの少なくとも５％、１０％、１５％、２０％、２５％、３０％、３５％、４０％、４５％、または５０％である。

【０１３５】

例えば図３中で描写されるように、いくつかの事例において、次いで、見出されたバリ

アントのサブセットをグラフへ自動的に追加することができ、その結果、グラフは自己更新するようになる。この例において、配列は、シーケンシング３００から得られたリードデータ３０１（例えばＦＡＳＴＱファイル中に含まれた）である。配列リードは、１又は２以上のリードのバッチ３１２、３１３中でグラフリファレンス３１１を使用して、アライメント３１０される。バッチは、ＦＡＳＴＱファイル中のリードの総数に比べて、少数のリードを含むことができる。バッチ中のリードがアライメントされる際に、変則的にアライメントされたものは同定および分離することができる。それらはＳＡＭまたはＢＡＭファイル３１４へ書き込むことができる。このようにして、変則的にアライメントされたリードのアライメントおよび同定は、ただ１つのリードサイクルにおいて遂行される。次いで、新規バリアントおよび／または構造バリアントはこれらの変則的にアライメントされたリードに基づいて検出３２０することができる。次いで、見出されたバリアントのサブセットを自己更新様式３３０でグラフへ追加することができる。この手順を介して、グラフアライメントおよびバリアントの検出方法は、後続の分析に際して自己改善されるようになる。

【０１３６】

いくつかの実施形態において、バリアントは、グラフへ追加されるために特定の条件を満たす必要がある（サンプルセット中の頻度、長さ、タイプまたはクオリティ条件等）。かかる条件を課すことは、グラフリファレンスが適用のために簡潔で適切なままであることを保証することができる。

【０１３７】

いくつかの事例において、グラフリファレンスは、２以上のアライメントおよびバリアント検出において使用され漸進的に更新される。それは同じコンピューターまたは複数のコンピューター上で使用および更新することができる。いくつかの事例において、グラフリファレンスは中央レポジトリ中で保存および更新され、１又は２以上のコンピューターの中で共有される。

【０１３８】

グラフリファレンスを使用してアライメントされるリードは、直線リファレンスに対してアライメントされたリードのためのＳＡＭフォーマットとコンパチブルなＳＡＭフォーマット中に書き込むことができる。１又は２以上の異なるビットフラグまたはリードタグは、追加情報を伝えるように含むことができる。例えば、グラフリファレンスを使用してアライメントされるリードを出力するフォーマットは、リードアライメントがバリアントにオーバーラップする場合に設定される随意のビットフラグ、リファレンスおよび／またはバリアントパスに関するアライメントのロケーションを特徴づけるリードタグ、ならびにリードがどのバリアントへアライメントするかを指示するリードタグを含むことができる。いくつかの事例において、代替パスによるオーバーラップにアライメントするリードのアライメントは、直線リファレンス座標へ戻って変換される。バリアントパスの座標に対してアライメントされた配列の開始を示す追加のリードタグを、使用することができる。バリアントパスの座標と比べてアライメントされたリードの開始および終了の両方を指示する追加のリードタグを、使用することができる。バリアントパスに関係するマッチ、ミスマッチ、挿入、欠失および開始位置の数が含まれるが、これらに限定されないアライメントスコアを含む追加のリードタグを、使用することができる。かかるリードタグには、マッピングに依存するリファレンスパスに関するアライメントスコアも含むことができる。いくつかの事例において、アライメントの開始は、直線リファレンスパスへの射影を指示する。追加のリードタグは、リードが代替パスを通過することができたか、その代りにそれがリファレンスパスへマッピングされたどうかを詳述するのに、使用することができる。追加のリードタグは、リードがどれだけの代替パスを通過したかを詳述するのに使用することができる。追加のリードタグは、リードがどれだけの代替パスを通過しなかったか、その代りにリファレンスパスへマッピングされたかを詳述するのに、使用することができる。追加のリードタグは、リードがバリアントパスへマッピングされるかどうかを詳述するのに使用することができる。

【０１３９】
次世代シーケンシングプラットフォーム
アライメントされるか、アセンブルされるか、またはそうでなければ本開示の技法を使用してプロセシングされる、シーケンシング情報は、次世代シーケンシング（ＮＧＳ）プラットフォームからのものとすることができる。本開示の技法は、異なるソースプラットフォーム、異なるファイルフォーマット、異なるリード長、異なる正確性、異なるクオリティスコア、異なる誤差率、および異なる優勢タイプまたは誤差源のシーケンシング情報により、使用することができる。
【０１４０】
ＮＧＳプラットフォームは商業的に入手可能なプラットフォームとすることができる。商業的に入手可能なプラットフォームには、逐次合成によるシーケンシング（ｓｅｑｕｅｎｃｉｎｇ－ｂｙ－ｓｙｎｔｈｅｓｉｓ）、イオン半導体シーケンシング、ピロシーケンシング、可逆的ダイターミネーターシーケンシング、ライゲーションによるシーケンシング（ｓｅｑｕｅｎｃｉｎｇ ｂｙ ｌｉｇａｔｉｏｎ）、単一分子シーケンシング、ハイブリダイゼーションによるシーケンシング、およびナノポアシーケンシングのためのプラットフォームを含むことができるが、これらに限定されない。合成によるシーケンシングのためのプラットフォームは、例えばＩｌｌｕｍｉｎａ、４５４ Ｌｉｆｅ Ｓｃｉｅｎｃｅｓ、Ｈｅｌｉｃｏｓ Ｂｉｏｓｃｉｅｎｃｅｓ、およびＱｉａｇｅｎから入手可能である。Ｉｌｌｕｍｉｎａプラットフォームには、例えばＩｌｌｕｍｉｎａのＳｏｌｅｘａプラットフォーム、Ｉｌｌｕｍｉｎａ のＧｅｎｏｍｅ Ａｎａｌｙｚｅｒが含まれ得る。例示的なＩｌｌｕｍｉｎａプラットフォームは、Ｇｕｄｍｕｎｄｓｓｏｎ ｅｔ ａｌ（Ｎａｔ．Ｇｅｎｅｔ．２００９ ４１：１１２２－６）、Ｏｕｔ ｅｔ ａｌ（Ｈｕｍ．Ｍｕｔａｔ．２００９ ３０：１７０３－１２）およびＴｕｒｎｅｒ（Ｎａｔ．Ｍｅｔｈｏｄｓ ２００９ ６：３１５－６）、米国特許公開第２００８０１６０５８０号および第２００８０２８６７９５号、米国特許第６，３０６，５９７号、第７，１１５，４００号および第７，２３２，６５６号に記載される。４５４ Ｌｉｆｅ Ｓｃｉｅｎｃｅプラットフォームには、例えばＧＳ ＦｌｅｘおよびＧＳ Ｊｕｎｉｏｒが含まれ得る。例示的な４５４ Ｌｉｆｅ Ｓｃｉｅｎｃｅプラットフォームは、米国特許第７，３２３，３０５号に記載される。Ｈｅｌｉｃｏｓ Ｂｉｏｓｃｉｅｎｃｅｓからのプラットフォームには、Ｔｒｕｅ Ｓｉｎｇｌｅ Ｍｏｌｅｃｕｌｅ Ｓｅｑｕｅｎｃｉｎｇプラットフォームが含まれる。イオン半導体シーケンシングのためのプラットフォームには、Ｉｏｎ Ｔｏｒｒｅｎｔ Ｐｅｒｓｏｎａｌ Ｇｅｎｏｍｅ Ｍａｃｈｉｎｅ（ＰＧＭ）が含まれ、例えば米国特許第７，９４８，０１５号に記載される。ピロシーケンシングのためのプラットフォームには、ＧＳ Ｆｌｅｘ ４５４システムが含まれ、例えば米国特許第７，２１１，３９０号、第７，２４４，５５９号および第７，２６４，９２９号に記載される。ライゲーションによるシーケンシングのためのプラットフォームおよび方法には、ＳＯＬｉＤシーケンシングプラットフォームが含まれ、例えば米国特許第５，７５０，３４１号に記載される。単一分子シーケンシングのためのプラットフォームには、Ｐａｃｉｆｉｃ ＢｉｏｓｃｉｅｎｃｅからのＳＭＲＴシステムおよびＨｅｌｉｃｏｓ Ｔｒｕｅ Ｓｉｎｇｌｅ Ｍｏｌｅｃｕｌｅ Ｓｅｑｕｅｎｃｉｎｇプラットフォームが含まれる。
【０１４１】
自動化サンガー法は「第一世代」技術として判断することができるが、自動化サンガーシーケンシングが含まれるサンガーシーケンシングも、本開示の方法によって用いることができる。この技術は、シーケンシングの相対的容易性および正確性を与える約１０００塩基対までであるが、これらに限定されないＤＮＡの短いセグメントを包含することができる。原子間力顕微鏡（ＡＦＭ）または透過電子顕微鏡法（ＴＥＭ）が含まれるがこれらに限定されない核酸画像化技術の開発の使用を含む、追加のシーケンシング方法も、本開示の方法によって包含される。例示的なシーケンシング技術は更に後述される。
【０１４２】
次世代シーケンシング技術はＩｏｎ Ｔｏｒｒｅｎｔシーケンシングプラットフォーム

を利用することができ、それはシーケンシングケミストリーと半導体技術をペアにして、半導体チップ上で化学的にコードされた情報（Ａ、Ｃ、Ｇ、Ｔ）をデジタル情報（０、１）の中へ直接変換するものである。理論により束縛されることは意図しないが、ヌクレオチドがポリメラーゼによってＤＮＡ鎖へと取り込まれる場合に、水素イオンが副産物として放出される。Ｉｏｎ　Ｔｏｒｒｅｎｔプラットフォームは、ｐＨの変化として水素原子の放出を検出する。ｐＨの検出された変化を使用して、ヌクレオチド取り込みを指示することができる。Ｉｏｎ　Ｔｏｒｒｅｎｔプラットフォームは微細機械加工されたウェルの高密度アレイを含み、超並列手法においてこの生化学的プロセスを遂行する。各々のウェルは異なるライブラリーメンバーを保持し、それはクローン的に増幅することができる。ウェルの下はイオン感受性層であり、その下はイオンセンサーである。プラットフォームは、次々とヌクレオチドによりアレイを順次あふれさせる。ヌクレオチド（例えばＣ）がＤＮＡテンプレートへ追加され、次いでＤＮＡ鎖の中へ取り込まれる場合に、水素イオンが放出されることになる。そのイオンからの電荷は溶液のｐＨを変化させ、それはＩｏｎ　Ｔｏｒｒｅｎｔのイオンセンサーによって同定することができる。ヌクレオチドが取り込まれなければ、電圧変化は記録されず、塩基はコールされないことになる。ＤＮＡ鎖上に２つの同一の塩基があれば、電圧は２倍であり、チップはコールされた２つの同一の塩基を記録することになる。直接的な同定は、数秒でのヌクレオチド取り込みの記録を可能にする。Ｉｏｎ　Ｔｏｒｒｅｎｔプラットフォームのためのライブラリー調製は、一般的にはＤＮＡ断片の両方の末端での２つの別個のアダプターのライゲーションを包含する。これらのライブラリーは、任意のシーケンシングの前に配列を増幅する分離したエマルションＰＣＲを包含することができ、それはプロセスを複雑にして減速させる場合がある。加えて、この２ステップのプロセスは、他の次世代シーケンシングプラットフォームに比較して、特にホモポリマーについて、より高い誤差率（例えば０．５～２．５％；１００塩基対あたり１．５のインデル誤差率）をもたらす場合がある。更に、ＡＴリッチセグメントおよびＧＣリッチセグメントからなる複雑な領域は、低いカバレッジをもたらす場合がある。例えば、Ｉｏｎ　３１８（商標）Ｃｈｉｐ　ｖ２を備えたＩｏｎ　Ｔｏｒｒｅｎｔ　ＰＧＭシーケンサーのための調製およびシーケンシングの時間は、それぞれ８時間および４～７時間までである。従来のセットアップに基づいて、システムは、２００塩基対または４００塩基対の単一ヌクレオチド配列からなる、ランあたり６００メガバイト～２ギガバイトの範囲のデータを出力することができる。各々のリードについてのｐｈｒｅｄクオリティスコア（Ｑ）は１０～３０の範囲とすることができ、それはそれぞれ９０％～９９．９％のシーケンシング正確性へ換算される。

【０１４３】

次世代シーケンシング技術はＩｌｌｕｍｉｎａシーケンシングプラットフォームを利用することができ、それは一般的にはフローセルの上のライブラリーメンバーのクラスター増幅および逐次合成によるシーケンシングアプローチを用いる。クラスター増幅されたライブラリーメンバーに、ポリメラーゼ依存性の単一塩基伸長の反復サイクルを行う。単一塩基の伸長は、可逆的ターミネーターのｄＮＴＰ（異なる除去可能なフルオロフォアにより標識された各々のｄＮＴＰ）の取り込みを含むことができる。可逆的ターミネーターのｄＮＴＰは一般的には３’修飾されて、ポリメラーゼによる更なる伸長を防止する。取り込みの後に、取り込まれたヌクレオチドは、蛍光イメージングによって同定することができる。蛍光イメージングに続いて、フルオロフォアを除去することができ、３’修飾を除去することができ、３’ヒドロキシル基を結果として生じ、それによって単一塩基伸長の別のサイクルを可能にする。Ｉｌｌｕｍｉｎａプラットフォームのためのライブラリー調製は、一般的にはＤＮＡ断片の両方の末端での２つの別個のアダプターのライゲーションを含む。これらのライゲーションされたＤＮＡ断片は、所望される出力リードサイズに依存して、３００塩基対までであるが、これらに限定されない長さで変動することができる（一般的には短いリードと称される）。最近のライブラリー調製（ＴｒｕＳｅｑ長いリード技術等）は、１０キロベースまでの合成されたリードを可能にすることができるが、これらはＨｉＳｅｑプラットフォームバージョンへ限定することができる。ライブラリー調

製は、単一末端からのリードまたはペア末端からのリードを包含することができる。ペア末端の調製のいくつかの例は、２×３００塩基対、２×２５０塩基対、または２×１５０塩基対のヌクレオチド配列である。平均準備時間はおよそ８時間を含む。いくつかの一般的で商業的に入手可能なシステムには、ＭｉＳｅｑ、ＮｅｘｔＳｅｑ　５００およびＨｉＳｅｑ　２５００を含み、変動するデータ出力サイズおよびシーケンシング時間を有する。ＭｉＳｅｑシーケンシングランは最大６０時間かかり、１ランあたりおよそ１３～１６ギガバイトを出力することができる一方で、ＮｅｘｔＳｅｑ　５００およびＨｉＳｅｑ　２５００は最大３０時間および６０時間かかり、それぞれ１ランあたり１００～１２０ギガバイトおよび２５０～３００ギガバイトを出力することができる。顕著なことに、すべてのシステムについてのシーケンシング誤差率は約０．１％であり、９９．９％の高精度とすることができる（３０のｐｈｒｅｄクオリティスコア（Ｑ））。

【０１４４】

次世代シーケンシング技術はＨｅｌｉｃｏｓ　Ｔｒｕｅ　Ｓｉｎｇｌｅ　Ｍｏｌｅｃｕｌｅ　Ｓｅｑｕｅｎｃｉｎｇ（ｔＳＭＳ）とすることができ、それは逐次合成によるシーケンシング技術を用いることができる。ｔＳＭＳ技法において、ポリＡアダプターはＤＮＡ断片の３’末端へライゲーションすることができる。適合させた断片は、ｔＳＭＳフローセル上で固定化されたポリＴオリゴヌクレオチドへハイブリダイズさせることができる。ライブラリーメンバーは、約１億テンプレート／cm^2の密度でフローセルの上へ固定化することができる。次いでフローセルは装置（例えばＨＥＬＩＳＣＯＰＥ（商標）シーケンサー）の中へロードされ、レーザーはフローセルの表面を照らし、各々のテンプレートの位置を明らかにすることができる。ＣＣＤカメラは、フローセル表面上のテンプレートの位置をマッピングすることができる。ライブラリーメンバーに、ポリメラーゼ依存性の単一塩基伸長の反復サイクルを行うことができる。シーケンシング反応はＤＮＡポリメラーゼおよび蛍光標識されたヌクレオチドの導入によって開始することができる。ポリメラーゼは、標識されたヌクレオチドをプライマーへテンプレート指向様式で取り込むことができる。ポリメラーゼおよび取り込まれないヌクレオチドを除去することができる。蛍光標識されたヌクレオチドの指示された取り込みを有するテンプレートは、フローセル表面の画像化によって識別することができる。画像化の後に、切断ステップは蛍光標識を除去することができ、所望されるリード長が達成されるまで、プロセスは他の蛍光標識されたヌクレオチドにより反復することができる。配列情報は、各々のヌクレオチド追加ステップにより収集することができる。

【０１４５】

次世代シーケンシング技術は、例えばＭａｒｇｕｌｉｅｓ，Ｍ．ｅｔ　ａｌ．Ｎａｔｕｒｅ　４３７：３７６－３８０［２００５］中で記載されるような、４５４（Ｒｏｃｈｅ）シーケンシングプラットフォームを利用することができる。４５４シーケンシングは一般的には２つのステップを含む。第１のステップにおいて、ＤＮＡは断片へ切断することができる。断片は平滑末端とすることができる。オリゴヌクレオチドアダプターは、断片の末端へライゲーションすることができる。アダプターは、一般的には断片の増幅およびシーケンシングのためのプライマーとして機能する。少なくとも１つのアダプターは捕捉試薬（例えばビオチン）を含むことができる。断片はＤＮＡ捕捉ビーズ（例えばストレプトアビジンコートビーズ）へ添付することができる。ビーズへ添付された断片は油－水エマルションの液滴内でＰＣＲ増幅され得、各々のビーズ上でクローン的に増幅されたＤＮＡ断片の複数のコピーをもたらす。第２のステップにおいて、ビーズはウェル中で捕捉され、それはピコリットルサイズとすることができる。ピロシーケンシングは各々のＤＮＡ断片上で並列して遂行することができる。ピロシーケンシングは、一般的にはヌクレオチド取り込みに際してのピロリン酸（ＰＰｉ）の放出を検出する。ＰＰｉは、アデノシン５’－ホスホ硫酸の存在下においてＡＴＰスルフリラーゼによってＡＴＰへ変換することができる。ルシフェラーゼはＡＴＰを使用してルシフェリンをオキシルシフェリンへ変換することができ、それによって検出されるシグナルを生成する。検出された光シグナルを使用して、取り込まれたヌクレオチドを同定することができる。Ｉｏｎ　Ｔｏｒｒｅｎｔと

同様に、４５４のシステムは、任意のシーケンシングの前に、分離したエマルションＰＣＲによって増幅されたライブラリーを要求し、それはシーケンシングプロセスを複雑にして減速させる可能性がある。このシステムは、同様に高い誤差率（例えば０．５～１％；１００塩基対あたり０．４のインデル誤差率）ももたらす可能性がある。例えば、ＧＳ Ｊｕｎｉｏｒ Ｐｌｕｓ構成を備えたＲｏｃｈｅ ４５４ ＧＳシーケンサーのための調製およびシーケンシングの時間は、それぞれ８時間および１８時間までである。このセットアップは、７００塩基対の単一ヌクレオチド配列からなる、ランあたり５０～７０メガバイトの範囲のデータを出力することが期待することができる。Ｔｉｔａｎｉｕｍ ＸＬ＋構成を使用する類似したセットアップは同じ準備時間を有することができるが、より長い最大３０時間のシーケンシングランである。このセットアップは、７００塩基対の単一ヌクレオチドリードからなる、ランあたり１００～１２０ギガバイトの範囲のデータを出力することが期待できる。全体的には、これらのシステムにおけるリードについてのｐｈｒｅｄクオリティスコア（Ｑ）は２０～３０の範囲であり、それはそれぞれ９９％～９９．９％のシーケンシング正確性へ換算される。

【０１４６】

次世代シーケンシング技術は、ＳＯＬｉＤ（商標）技術（Ａｐｐｌｉｅｄ Ｂｉｏｓｙｓｔｅｍｓ）を利用することができる。ＳＯＬｉＤプラットフォームは、一般的にはライゲーションによるシーケンシングのアプローチを利用する。ＳＯＬｉＤプラットフォームによる使用のためのライブラリー調製は、一般的には断片ライブラリーを生成するために断片の５’末端および３’末端へ添付されるアダプターのライゲーションを含む。あるいは、内部アダプターは、断片の５’末端および３’末端へアダプターをライゲーションすること、断片を環状化すること、環状化された断片を消化して内部アダプターを生成すること、ならびにメイトペアにされたライブラリーを生成するためにもたらされた断片の５’末端および３’末端へアダプターを添付することによって導入することができる。次に、クローンのビーズ集団は、ビーズ、プライマー、テンプレートおよびＰＣＲ構成要素を含有するマイクロリアクター中で調製することができる。ＰＣＲに続いて、テンプレートは変性することができる。ビーズは、延長されたテンプレートを備えたビーズについて濃縮することができる。選択されたビーズ上のテンプレートに、スライドグラスへ結合することを可能にする３’修飾を行うことができる。配列は、特異的なフルオロフォアによって同定される中央の決定された塩基（または塩基対）を備えた部分的にランダムなオリゴヌクレオチドの順次のハイブリダイゼーションおよびライゲーションによって決定することができる。色が記録された後に、ライゲーションされたオリゴヌクレオチドを除去することができ、次いでプロセスを反復することができる。

【０１４７】

次世代シーケンシング技術は、単一分子リアルタイム（ＳＭＲＴ（商標）シーケンシングプラットフォーム（Ｐａｃｉｆｉｃ Ｂｉｏｓｃｉｅｎｃｅｓ））を利用することができる。ＳＭＲＴシーケンシングにおいて、色素標識ヌクレオチドの継続的な取り込みはＤＮＡ合成の間に画像化することができる。単一ＤＮＡポリメラーゼ分子は個別のゼロモード波長識別子（ＺＭＷ識別子）の底部の表面へ添付することができ、この識別子は、リン酸結合したヌクレオチドが成長しているプライマー鎖の中へ取り込まれている間に、配列情報を得る。ＺＭＷは、一般的には、マイクロ秒スケールでＺＭＷの内外で急速に拡散する蛍光ヌクレオチドのバックグラウンドに対して、ＤＮＡポリメラーゼによる単一ヌクレオチドの取り込みの観察を可能にする閉じ込め構造を指す。これとは対照的に、ヌクレオチドの取り込みは、一般的には、ミリ秒の時間スケールで出現する。この時間の間に、蛍光標識を励起して蛍光シグナルを産生することができ、それが検出される。蛍光シグナルの検出を使用して、配列情報を生成することができる。次いでフルオロフォアは除去され、プロセスは反復することができる。ＳＭＲＴプラットフォームのためのライブラリー調製は、一般的には、ＤＮＡ断片の末端へのヘアピン型アダプターのライゲーションを包含する。これらのライゲーションされたＤＮＡ断片は、所望される出力リードサイズに依存して、４０，０００塩基対までであるが、これらに限定されない長さで変動することがで

きる（一般的には長いリードと称される）。平均調製時間はおよそ８時間を含み、ＤＮＡポリメラーゼ合成のためにまたはその間のＤＮＡ配列の変更を要求せず、したがって反復ゲノム領域および可能性のあるＤＮＡの修飾（例えばＤＮＡメチル化）を解決する能力がある。強力な技法ではあるが、それは、次世代シーケンシング技術の中で最も高い誤差率のうちの１つ（１４パーセント）を生じ得る。例えば、ＲＳ　ＩＩ構成によるＳＭＲＴ（商標）プラットフォームを使用して、ランは最大４時間であり、上記の長いリードからなる、ランあたり０．５～１ギガバイトのデータを生じ得る。全体的には、このシステムにおけるリードについてのｐｈｒｅｄクオリティスコア（Ｑ）は、大抵は３０であり、それは９９．９％のシーケンシング正確性へ換算される。

【０１４８】

次世代シーケンシング技術は、ナノポアシーケンシングを利用することができる（例えばＳｏｎｉ　ＧＶ　ａｎｄ　Ｍｅｌｌｅｒ　Ａ．Ｃｌｉｎ　Ｃｈｅｍ　５３：１９９６－２００１［２００７］中で記載されているように）。ナノポアシーケンシングＤＮＡ分析技法は、Ｏｘｆｏｒｄ　Ｎａｎｏｐｏｒｅ　Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｏｘｆｏｒｄ、イギリス）を含むがこれらに限定されない、多くの会社によって産業用に開発されている。ナノポアシーケンシングは、それがナノポアを介して通過するかまたはそれに近接するにつれて、それによってＤＮＡの単一分子が直接シーケンスされる単一分子シーケンシング技術である。ナノポアは小さな穴（直径が１ナノメートルのオーダー）とすることができる。導電性流体中でナノポアが浸されることおよび横切る電位（電圧）を適用することは、ナノポアを介するイオンの伝導に起因するわずかな電流をもたらし得る。流れる電流の量は、ナノポアのサイズおよび形、ならびに例えばＤＮＡ分子による閉塞への感受性がある。ＤＮＡ分子がナノポアを介して通過するにつれて、ＤＮＡ分子上の各々ヌクレオチドは異なる程度にナノポアを閉塞させ、異なる程度でナノポアを介する電流の大きさを変化させる。したがって、ＤＮＡ分子がナノポアを介して通過する際の電流のこの変化は、ＤＮＡ配列のリーディングを表わす。他のナノポアベースの検出様式も用いることができる。ナノポアシーケンシングは、現在のところ任意のシーケンシング技術について最も高い誤差率を示すことができ、限定されるものではないが、最大２５～３０％に達する。それにもかかわらず、最近の開発は、例えば複数の次元を介するシーケンシングによってこの誤差率を低減させることに注目している。最初に１－Ｄでシーケンスしていたのであるが、Ｏｘｆｏｒｄ　Ｎａｎｏｐｏｒｅ　Ｔｅｃｈｎｏｌｏｇｉｅｓは、より低い誤差率をもたらし、正確性を増加させることができる２－Ｄシーケンシングへ発展した。この技術はユーザーが要求するまでＤＮＡ分子をリーディングし続けることができるので、シーケンシングランタイムはユーザーに依存することができ、フローセルは磨耗するか、またはより多くの試薬もしくはサンプルが必要である。

【０１４９】

次世代シーケンシング技術は、化学感応性電界効果トランジスタ（ｃｈｅｍＦＥＴ）アレイ（例えば米国特許公開第２００９００２６０８２号に記載されるような）を利用することができる。技法の一例において、ＤＮＡ分子を反応チャンバーの中へ置くことができ、テンプレート分子はポリメラーゼへ結合されたシーケンシングプライマーへハイブリダイズすることができる。シーケンシングプライマーの３’末端での新しい核酸鎖の中への１又は２以上の三リン酸塩の取り込みは、ｃｈｅｍＦＥＴによる電流中の変化によって識別することができる。アレイは複数のｃｈｅｍＦＥＴセンサーを有することができる。別の例において、単一核酸をビーズへ添付することができ、核酸はビーズ上で増幅することができ、個別のビーズは、ｃｈｅｍＦＥＴセンサーを有する各々のチャンバーを備えたｃｈｅｍＦＥＴアレイ上の個別の反応チャンバーへ移すことができ、核酸をシーケンスすることができる。

【０１５０】

次世代シーケンシング技術は、透過電子顕微鏡法（ＴＥＭ）を利用することができる。この方法（Ｉｎｄｉｖｉｄｕａｌ　Ｍｏｌｅｃｕｌｅ　Ｐｌａｃｅｍｅｎｔ　Ｒａｐｉｄ　Ｎａｎｏ　Ｔｒａｎｓｆｅｒ（ＩＭＰＲＮＴ）と称される）は、一般的には重原子マー

カーにより選択的に標識された高分子量（１５０ｋｂ以上）ＤＮＡを単一原子分解能の透過電子顕微鏡で画像化すること、および一貫した塩基間の間隔により超高密度（３ｎｍ鎖間）の並列アレイ中の超薄フィルム上にこれらの分子をアレンジすることを含む。電子顕微鏡を使用してフィルム上の分子を画像化して、重原子マーカーの位置を決定し、ＤＮＡから塩基配列情報を抽出する。この方法は、国際特許公開第２００９／０４６４４５号に更に記載される。この方法は、１０分間未満で完全なヒトゲノムをシーケンスすることを可能にする。

【０１５１】

本方法は、ハイブリダイゼーション（ＳＢＨ）によるシーケンスを利用することができる。ＳＢＨは、一般的には複数のポリヌクレオチドプローブと複数のポリヌクレオチド配列を接触させることを含み、そこで各々の複数のポリヌクレオチドプローブは、随意で基板へつなぐことができる。基板は、既知のヌクレオチド配列のアレイを含む平らな表面とすることができる。アレイへのハイブリダイゼーションのパターンを使用して、サンプル中に存在するポリヌクレオチド配列を決定することができる。他の実施形態において、各々のプローブはビーズ（例えば磁性ビーズまたは同種のもの）につながれる。ビーズへのハイブリダイゼーションを同定および使用して、サンプル内の複数のポリヌクレオチド配列を同定することができる。

【０１５２】

配列リードの長さは利用された特定のシーケンシング技術に依存して変動することができる。ＮＧＳプラットフォームは、何十～何百、もしくは何千塩基対、または何万もしくは何十万塩基対のサイズで変動する配列リードでさえも提供することができる。本明細書において記載される方法のうちのいくつかの実施形態において、配列リードは、約２０塩基長、約２５塩基長、約３０塩基長、約３５塩基長、約４０塩基長、約４５塩基長、約５０塩基長、約５５塩基長、約６０塩基長、約６５塩基長、約７０塩基長、約７５塩基長、約８０塩基長、約８５塩基長、約９０塩基長、約９５塩基長、約１００塩基長、約１２０塩基長、約１３０塩基長、約１４０塩基長、約１５０塩基長、約２００塩基長、約２５０塩基長、約３００塩基長、約３５０塩基長、約４００塩基長、約４５０塩基長、約５００塩基長、約６００塩基長、約７００塩基長、約８００塩基長、約９００塩基長、約１０００塩基長、または１０００塩基長を超える。

【０１５３】

サンプル中に存在するＤＮＡ断片の部分的シーケンシングを遂行することができる。

【０１５４】

暗号化

本明細書において開示される方法およびシステムは、暗号化を用いることもできる。暗号化は、ワンタイムパッド暗号を暗号化のために使用して遂行することができる。暗号化方法の追加の非限定的な例は、暗号論的擬似乱数生成器、情報理論的に安全なアルゴリズム、整数因数分解アルゴリズム、素数判定、アドバンスドアクセスコンテンツシステム、対称鍵アルゴリズム、破壊暗号化アルゴリズム（ｂｒｏｋｅｎ　ｃｒｙｐｔｏｇｒａｐｈｙ　ａｌｇｏｒｉｔｈｍｓ）、暗号解読アルゴリズム、および暗号学的ハッシュ関数を含むことができる。さらに、暗号化方法は、公開鍵、秘密鍵および／またはパスフレーズ（安全な電子メール転送において使用されるものに類似する）を利用する鍵ペア概念を利用することができる。例えば、暗号化分析デバイスは意図される受理者デバイスの公開鍵を有することができる。同様に、意図される受理者デバイスは暗号化分析デバイスの公開鍵を有する必要がある。また、鍵付ハッシュメッセージ認証コード（ＨＭＡＣ）を使用して、秘密の暗号化鍵と組み合わせて暗号学的ハッシュ関数を用いてメッセージ認証コードを生成することができる。このメッセージ認証コードを使用して、データ保全性を検証することに加えて、伝送された配列またはデータを認証することの両方ができる。暗号鍵が配列データの送信および受理のために使用される場合に、鍵は、例えば無作為に生成することができ、十分なエントロピーを有することができる。エントロピーは予測不能のコンピューターオペレーションに由来することができる。例えば、ディスクドライブヘッドの動

きである。
【０１５５】
暗号化された情報（配列情報等）は解読せずに比較することができる。
【０１５６】
代替の暗号化方法は単独でまたは組み合わせて用いることができる。例えば、デジタル署名は鍵ペアの秘密鍵を使用して生成することができる。デジタル署名は、送信されている生物学的配列が送信者によって署名されたことを裏付けることができる。
【０１５７】
シーケンサーがシーケンシングアッセイを遂行している間に、暗号化を遂行することができる。本開示の技法は、分析、暗号化、およびシーケンシングアッセイの時間スケールでの他のプロセシング（リアルタイム分析が含まれる）を可能にする迅速な計算を可能にすることができる。
【０１５８】
コンピューターシステム
コンピューターシステムは、非一時的コンピューター可読媒体中に含有される命令を使用して、本明細書において開示される方法を実行することができる。非一時的コンピューター可読媒体は、いくつかの事例において、一時的な伝搬シグナル以外のコンピューター可読媒体をすべて含むことができる。
【０１５９】
いくつかの実施形態において、プロセッサは、１又は２以上のコントローラー、計算ユニット、および／もしくはコンピューターシステムの他のユニットと関連するか、またはファームウェア中に埋め込まれる。いくつかの実施形態において、方法のうちの１又は２以上のステップはハードウェアにおいて実装される。いくつかの実施形態において、方法のうちの１又は２以上のステップはソフトウェアにおいて実装される。ソフトウェアルーチンは、任意のコンピューター可読メモリユニット（フラッシュメモリ、ＲＡＭ、ＲＯＭ、磁気ディスク、レーザーディスク（登録商標）、または本明細書において記載されるようなもしくは当技術分野において公知の他のストレージ媒体等）中に保存することができる。ソフトウェアは、任意の公知の通信方法、例えば、電話線、インターネット、無線接続などの通信チャンネル上で、またはコンピューター可読ディスク、フラッシュドライブなどの可搬型媒体によって計算デバイスへ通信することができる。本明細書において記載される方法のうちの１又は２以上のステップは、様々なオペレーション、ツール、ブロック、モジュールおよび技法として実装することができ、同様にして、ファームウェア、ハードウェア、ソフトウェア、またはファームウェア、ハードウェアおよびソフトウェアの任意の組み合わせで実装することができる。ハードウェアにおいて実装された場合に、ブロック、オペレーション、技法などのうちのいくつかまたはすべては、例えば特定用途向け集積回路（ＡＳＩＣ）、カスタム集積回路（ＩＣ）、フィールドプログラマブル論理アレイ（ＦＰＧＡ）またはプログラマブルロジックアレイ（ＰＬＡ）において実装することができる。
【０１６０】
システムは、例示的な本明細書において記載される方法を実装するようにプログラムされる中央コンピューターサーバーを包含することができる。サーバーは、シングルコアプロセッサ、マルチコアプロセッサ、または並列処理のための複数のプロセッサとすることができる、中央処理装置（ＣＰＵ、または「プロセッサ」）を含むことができる。いくつかの実例において、システムは、１０、９、８、７、６、５、４、３未満、または２未満のプロセッサを含む。
【０１６１】
１つの計算スレッドはプロセッサ上の最小の実現可能な命令ユニットである。複数のスレッドは、同時に（他のものが終了する前に１つが開始する）実行して同じプロセス内に存在することができ、リソース（メモリ等）を共有する。しながら、計算スレッドはプロセッサそれ自体の定義に使用される場合もある。例えば、１つのプロセッサが１つの物理

コアである場合、しかしそれは４つのスレッド、または論理コアを有することができる。したがって、本明細書において使用される場合、「計算スレッド」はプロセッサまたはスレッドとすることができる。

【０１６２】

いくつかの実例において、本明細書において記載されるシステムは、マルチスレッディングを使用することができる。いくつかの実例において、システムはマルチタスクオペレーティングシステムを含む。マルチスレッディングは、複数のスレッドが単一プロセスのコンテキスト内に存在することを可能にする、広く普及したプログラミングおよび実行モデルである。これらのスレッドはプロセスのリソースを共有するが、独立して実行することができる。マルチスレッディングは、マルチプロセッシングシステム上の並列実行を可能にするように単一プロセスへ適用することもできる。

【０１６３】

サーバーは、メモリ（例えばランダムアクセスメモリ、リードオンリメモリ、フラッシュメモリ）；電子記憶装置（例えばハードディスク）；１又は２以上の他のシステムと通信するための通信用インターフェース（例えばネットワークアダプタ）；ならびにキャッシュ、他のメモリ、データストレージ、および／または電子ディスプレイアダプターが含まれ得る周辺デバイスを含むことができる。メモリ、記憶装置、インターフェース、および周辺デバイスは、通信バス（マザーボード等）を介してプロセッサと通信することができる。記憶装置は、データの保存のためのデータ記憶装置とすることができる。サーバーは、通信用インターフェースの補助によりコンピューターネットワーク（「ネットワーク」）へ作動可能に結合することができる。ネットワークは、インターネット、イントラネットならびに／または、エクストラネット、インターネットと通信するイントラネットおよび／もしくはエクストラネット、テレコミュニケーションまたはデータネットワークとすることができる。いくつかの事例におけるネットワークは、サーバーの補助により、ピアツーピアネットワークを実装することができ、それにより、サーバーへ接続されたデバイスは、クライアントまたはサーバーとして動作することができる。

【０１６４】

記憶装置は、ファイル（被験体レポートおよび／または介護者との通信等）、シーケンシングデータ、個体に関するデータ、または本開示と関連するデータの任意の態様を保存することができる。

【０１６５】

サーバーはネットワークを介して１又は２以上の遠隔コンピューターシステムと通信することができる。１又は２以上の遠隔コンピューターシステムは、例えばパソコン、ラップトップ、タブレット、電話、スマートフォン、または個人用デジタル情報処理端末とすることができる。

【０１６６】

いくつかの状況において、システムは単一サーバーを含む。他の状況において、システムは、イントラネット、エクストラネットおよび／またはインターネットを介して互いに通信する複数のサーバーを含む。

【０１６７】

サーバーは、シーケンシング情報、クライアントもしくは患者の情報（例えば生のシーケンシングデータ、圧縮した配列データ、配列データを含有するグラフ、リファレンスゲノム、代替パスを含むリファレンスゲノム、多型、突然変異、患者歴および人口層データ等）、および／または可能性のある関連の他の情報を保存するように適合させることができる。かかる情報は記憶装置またはサーバー上で保存することができ、かかるデータはネットワークを介して伝送することができる。

【０１６８】

本明細書において記載されるような方法は、サーバーの電子ストレージロケーション上で（例えばメモリまたは電子記憶装置上で）保存された、機械（またはコンピュータープロセッサ）実行可能コード（またはソフトウェア）によって実装することができる。使用

時、コードはプロセッサによって実行することができる。いくつかの事例において、コードは記憶装置から検索され、プロセッサによる即時アクセスのためにメモリ上で保存することができる。いくつかの状況において、電子記憶装置は除外することができ、機械実行可能命令はメモリ上で保存される。いくつかの状況において、コードは第２のコンピューターシステム上で実行することができる。

【０１６９】

本明細書において提供されるシステムおよび方法の態様は、プログラミングにおいて実施することができる。技術の様々な態様は、典型的には、一種の機械可読媒体で送られるかまたはその中で具現化される機械（またはプロセッサ）実行可能コードおよび／または関連データの形態の「製品」または「製造品」として考えることができる。機械実行可能コードは、電子記憶装置、かかるメモリ（例えばリードオンリメモリ、ランダムアクセスメモリ、フラッシュメモリ）またはハードディスク上で保存することができる。「ストレージ」タイプの媒体には、コンピューター、プロセッサもしくは同種のものの有形メモリのうちの任意もしくはすべて、またはその関連するモジュール（様々な半導体メモリ、テープドライブ、ディスクドライブおよび同種のもの等）が含まれ、それらはソフトウェアプログラミングのための任意の時間における非一時的なストレージを提供することができる。ソフトウェアのすべてまたは一部は、場合によってはインターネットまたは様々な他の通信ネットワークを介して通信することができる。かかる通信は、例えば１つのコンピューターまたはプロセッサから別のものの中へ（例えば管理サーバーまたはホストコンピューターからアプリケーションサーバーのコンピュータープラットフォームの中へ）のソフトウェアのローディングを可能にすることができる。したがって、ソフトウェアエレメントを記録することができる別のタイプの媒体としては、有線および光学地上通信ネットワークを介して様々なエアリンク上で、ローカルデバイスの間の物理インターフェースを横切って使用されるような、光波、電波および電磁波を挙げることができる。有線または無線様のもの、または光リンクなどといった、かかる波を送る物理的エレメントは、ソフトウェアを記録す媒体として考えることができる。

【０１７０】

不揮発性ストレージ媒体は、例えば、何らかのコンピューター（複数可）内の何らかのストレージデバイスなどの、光学ディスクまたは磁気ディスクを含むことができ、かかるものを使用してシステムを実装することができる。有形伝送媒体は、同軸ケーブル、銅線、およびファイバーオプティクス（コンピューターシステム内のバスを備えるワイヤーが含まれる）を含むことができる。搬送波伝送媒体は、電気シグナルもしくは電磁シグナルまたは音波もしくは光波（高周波（ＲＦ）データ通信および赤外線（ＩＲ）データ通信の間に生成されるもの）の形態とすることができる。したがって、コンピューター可読媒体の一般的形式は、例えばフロッピーディスク、フレキシブルディスク、ハードディスク、磁気テープ、任意の他の磁気メディア、ＣＤ－ＲＯＭ、ＤＶＤ、ＤＶＤ－ＲＯＭ、任意の他の光媒体、パンチカード、ペーパーテイム（ｐａｐｅｒ　ｔａｍｅ）、ホールパターンを備えた任意の他の物理ストレージ媒体、ＲＡＭ、ＲＯＭ、ＰＲＯＭおよびＥＰＲＯＭ、ＦＬＡＳＨ－ＥＰＲＯＭ、任意の他のメモリチップもしくはカートリッジ、データもしくは命令を輸送する搬送波、かかる搬送波を輸送するケーブルもしくはリンク、またはコンピューターがプログラミングコードおよび／もしくはデータを読み取ることができる任意の他の媒体を含む。コンピューター可読媒体のこれらの形式の多くは、実行のためのプロセッサへの１又は２以上の命令の１又は２以上のシーケンスを保有することを伴うことができる。

【０１７１】

コンピューターシステムは、１又は２以上のステップのために使用することができ、それらのステップには、例えばサンプル採取、サンプルプロセシング、シーケンシング、リファレンスゲノムへの配列比較、配列アライメント、グラフィカルインターフェースへの出力、レポートの生成、および受理者への結果の報告が含まれる。

【０１７２】

クライアントサーバーアーキテクチャおよび／またはリレーショナルデータベースアーキテクチャは、本開示の技法において使用することができる。一般に、クライアントサーバーアーキテクチャは、ネットワーク上の各々のコンピューターまたはプロセスがクライアントまたはサーバーのいずれかであるネットワークアーキテクチャである。サーバーコンピューターは、ディスクドライブ（ファイルサーバー）、プリンター（プリントサーバー）、またはネットワークトラフィック（ネットワークサーバー）の管理専用の強力なコンピューターとすることができる。クライアントコンピューターは、ユーザーがアプリケーションを実行するＰＣ（パーソナルコンピューター）またはワークステーションに加えて、本明細書において開示されるような例示の出力デバイスを含むことができる。クライアントコンピューターは、ファイル、デバイス、そしてプロセッシング能力などのリソースのためのサーバーコンピューターに依存することができる。サーバーコンピューターはすべてのデータベース機能性に対処する。クライアントコンピューターは、フロントエンドデータ管理に対処するソフトウェアを有し、ユーザーからデータ入力を受け取ることができる。

【０１７３】

計算を遂行した後に、プロセッサは、例えば元の入力デバイスもしくは記憶装置へ、同じもしくは異なるコンピューターシステムの別の記憶装置へ、または出力デバイスへ、出力(計算等からの)を供給することができる。プロセッサからの出力は、データディスプレイ、例えば表示画面（例えばデジタルデバイス上のモニターまたは画面）、プリントアウト、データ信号（例えばパケット）、グラフィカルユーザーインターフェース（例えばウェブページ）、アラーム（例えばフラッシュライトまたはサウンド）、または上記のもののうちのいずれかの組み合わせによって表示することができる。一実施形態において、出力はネットワーク（例えば無線ネットワーク）を通して出力デバイスへ伝送される。ユーザーは出力デバイスを使用して、データ処理コンピューターシステムからの出力を受け取ることができる。ユーザーが出力を受け取った後、ユーザーは行動計画を決定することができるか、またはユーザーが医療関係者である場合に行動計画（医学的治療等）を遂行することができる。いくつかの実施形態において、出力デバイスは入力デバイスと同じデバイスである。例示の出力デバイスには、電話、無線電話、携帯電話、ＰＤＡ、フラッシュメモリドライブ、光源、サウンドジェネレーター、ファックス、コンピューター、コンピューターモニター、プリンター、ｉＰｏｄ（登録商標）、およびウェブページが含まれるが、これらに限定されない。ユーザーステーションはプリンターまたは表示モニターと通信して、サーバーによって処理された情報を出力することができる。かかるディスプレイ、出力デバイスおよびユーザーステーションを使用して、被験者またはその介護者へ警報を出すことができる。

【０１７４】

本開示に関するデータは、受理者が受け取るおよび／またはレビューするためにネットワークまたは接続を通して伝送することができる。受理者は、レポートが関連する被験者；またはその介護者（例えば医療提供者、マネージャー、他の医療従事者、または他の保護者）；遺伝子型決定分析を遂行したおよび／もしくは命じた人もしくは事業者；遺伝カウンセラーとすることができるがこれらに限定されない。受理者は、さらにかかるレポートの保存のためのローカルシステムまたはリモートシステム（例えばサーバーまたは「クラウドコンピューティング」アーキテクチャの他のシステム）とすることができる。一実施形態において、コンピューター可読媒体には、生物学的サンプルの分析の結果の伝達に好適な媒体が含まれる。

【０１７５】

ヒトゲノムのための明示的な手法で保存された配列グラフは、４０ＧＢのストレージを必要とする可能性がある。長さＮおよびＭの代替パスの配列による明示的な配列グラフストレージのための最小のデータ構造は、１０００ゲノムのフェーズＩＩＩコールセットからのヒトリファレンスゲノムおよびバリアントを保持することができ、以下の通りである。

【数１】

$$N*13\text{バイト}+\sum_{k=0}^{\text{バリアントの数}}(\text{バリアント}k\text{の長さ})*13\text{バイト}+16\text{バイト},$$

式中、リファレンス上の各々のヌクレオチドは最小サイズの関連するノードＩＤおよびエッジを有し、各々の代替パスはリファレンスノードＩＤおよびエッジへ対応するポインターを有する。本開示の技法において、配列グラフのストレージは、以下の通りとすることができる。

【数２】

$$N+\sum_{k=0}^{\text{バリアントの数}}(\text{バリアント}k\text{の長さ})+8\text{バイト}$$

式中、リファレンス上の各々のヌクレオチドは、シングルバイトで最小で保存され、代替パス上の各々のヌクレオチドはシングルバイトで保存される。さらに、各々の代替パスは開始位置および終了位置を含むことができる。このデータ構造により、ヒト直線リファレンスおよび１０００ゲノムのフェーズＩＩＩコールセットは、３．５ＧＢ未満のストレージ（現在の業界基準未満の大きさのオーダー）を必要とする可能性がある。

【０１７６】

本開示の技法は他の配列（細菌等）へ適用することができる。例えば、Ｍｙｃｏｂａｃｔｅｒｉｕｍ　ｔｕｂｅｒｃｕｌｏｓｉｓの事例において、リファレンスとしてＨ３７Ｒｖおよび代替パスについてＭｙｃｏｂａｃｔｅｒｉｕｍ　ｃａｎｅｔｔｉｉを使用して、現在の産業実践では少なくとも５５メガバイトのグラフが生成されるだろうが、本発明は約４．５メガバイトのサイズのグラフを生成することができる。

【０１７７】

前述の業界基準グラフ（最小のサイズで４０ＧＢ）上に構築されるｋ－ｍｅｒインデックスが、最終的にあまりにも大きいと（＞５００ＧＢ）考えられたので、使用し続けることができず、圧縮技術と組み合わせたＢｕｒｒｏｗｓ　Ｗｈｅｅｌｅｒ変換を用いる様々な代替物が業界で探索された。これらの技法は、変換を介して同じ情報を抽出する代替手法に相当したが、本開示の技法は、例えば効率的な配列グラフストレージがあるのでこの大きなインデックスを回避することができる。本発明の配列グラフおよび前記グラフのｋ－ｍｅｒインデックス（あるいはマスクされたｋ－ｍｅｒ）により、１０００ゲノムのフェーズＩＩＩコールセットを備えた全体のヒトリファレンスゲノムのｋ－ｍｅｒインデックスは、７２ＧＢ未満のコンピューターストレージ（例えば３３のｋ－ｍｅｒマスクサイズを使用して）に適合することができる。

【０１７８】

オフセット、ならびに代替パスの配列、開始オフセットおよび終了オフセットを含有するデータ構造をポイントする各々の代替パスのためのユニークなポインターと共に、塩基４において各々のｋ－ｍｅｒを保存することによって、インデックスは以下の通り増大する。

【数３】

$$\left(N-k+k*\sum_{j=0}^{\text{バリアントの数}}(k+(\text{バリアント}j\text{の長さ})-1)\right)*16\text{バイト},$$

式中、Ｎはリファレンス配列の長さであり、ｋはｋ－ｍｅｒの生成に使用されるマスクの長さである。マスクは「１」の文字列であり、リファレンス配列グラフへの完全なマッチを表わすことができるか、またはマスクは「０」を含有することができ、それはｋ－ｍｅｒからマスクする塩基を除外する。代替パスとして１０００ゲノムを備えたヒトリファレンス染色体１配列について、および各々の代替物は単一のＳＮＰであると想定して、マスクされたｋ－ｍｅｒインデックスは３．９８ＧＢのサイズである。したがって、一実施例

において、本開示の技法は、リファレンス中の１塩基あたり１６バイトおよび１バリアントあたり５２８バイトの割合で増大するマスクされたｋ－ｍｅｒインデックスを生成することができる。

【０１７９】

次いでこのインデックスを使用して、配列グラフ上に戻してアライメントさせる配列の候補アライメントロケーションを見出すことができる。アライメントさせる配列のｋ－ｍｅｒは生成され（例えば４０，６４２配列／秒／計算スレッドの率で）、インデックス中でサーチすることができる。いくつかのｋ－ｍｅｒはリファレンス配列を指示することができ；いくつかは代替パスを指示することができる。

【０１８０】

ＢＦＡＳＴは候補アライメントロケーション（ＣＡＬ）としてすべてのｋ－ｍｅｒを処理することができ、それはいくつかの等価ＣＡＬおよび疑似ＣＡＬをもたらす。これを避けるために、等価ＣＡＬが単一ＣＡＬへ統合されるかまたはと同期されるように、ｋ－ｍｅｒ位置を正規化することができ、アライメントモジュールにおいて試験される必要のあるＣＡＬがより少なくなる。ｋ－ｍｅｒの各々、リファレンス配列に対して相対的なオフセットを有することができ；これはリファレンス配列におけるオフセットまで減算されて、正規化オフセットを得ることができる。

【０１８１】

図６は例示的な候補アライメントロケーション生成およびリードグラフアライメントワークフローを示す。図６Ａは、アライメントされる配列６０１を持つ、配列受理モジュール６００を示す。図６Ｂは、マスクを適用して配列のｋ－ｍｅｒ分解６１１を得る、ｋ－ｍｅｒ化モジュール６１０を示す。図６Ｃは、リファレンスグラフ６２２においてｋ－ｍｅｒ６２１を見出す、グラフインデックスクエリモジュール６２０を示す。図６Ｄは、コンパチブルなｋ－ｍｅｒを単一の候補アライメントロケーション６３１へと同期させる、ｋ－ｍｅｒグラフ同期化モジュール６３０を示す。図６Ｅは、配列から最長のカバーされたセクションをとることによってシード６４１を生成する、グラフシーディングモジュール６４０を示す。図６Ｆは、ダイナミックプログラミングアルゴリズムを使用してシードをグラフ６５１へと延長することによってグラフアライメントを遂行する、グラフアライメントモジュール６５０を示す。

【０１８２】

例えば図７中で示されるように、いくつかのｋ－ｍｅｒは直接的にコンパチブルであり、いくつかのｋ－ｍｅｒは間接的にコンパチブルである。両方がリファレンス配列へ属するか、または両方が同じ代替パスへ属する場合、同じ正規化されたオフセット（点線の矢印）を備えた２つのｋ－ｍｅｒは直接的にコンパチブルであり；１つがリファレンス配列へ、および他のものが代替パスへ属する場合に、それらはインコンパチブルである。図７Ａは、両方のｋ－ｍｅｒが同じ正規化されたオフセットを有し、リファレンス配列へ属する例を示し、これは直接的にコンパチブルである。図７Ｂは、両方のｋ－ｍｅｒが同じ正規化されたオフセットを有し、同じ代替物パスへ属する例を示し、これは直接的にコンパチブルである。図７Ｃは、両方のｋ－ｍｅｒがリファレンス配列へ属するが、異なる正規化されたオフセットを有する例を示し、これは直接的にインコンパチブルである。図７Ｄは、１つのｋ－ｍｅｒがリファレンス配列へ、他のものが代替パスへ属し、両方が同じ正規化されたオフセットを有する例を示し、これは間接的にコンパチブルである。図７Ｅは、両方のｋ－ｍｅｒが同じ正規化されたオフセットを有するが、異なる代替パスへ属する例を示し、これは間接的にインコンパチブルである。直接的にコンパチブルなｋ－ｍｅｒは同じリファレンスパスまたは代替パスへ属し、同じ正規化されたオフセットを有する（例えば図７Ａを参照）。間接的にコンパチブルなｋ－ｍｅｒはアライメントされる配列からのｋ－ｍｅｒであり、そこで、いくつかのｋ－ｍｅｒはリファレンス配列を指示し、いくつかのｋ－ｍｅｒは代替パスを指示するが、ｋ－ｍｅｒのセットは同じ正規化されたオフセットを有する（例えば図７Ｃを参照）正規化オフセットは、コンパチブルなｋ－ｍｅｒによって支援されたリードの候補アライメントロケーションである（例えば図６Ｄを参

照）。図８は、候補アライメントロケーション（ＣＡＬ）の数を低減するオフセット正規化を例証する例示的な概略図を示す。図８Ａにおいて、アライメントされるべき配列８０１は、各々が配列に関するその相対的なオフセットを備えたより小さなｋ－ｍｅｒ　８０２へと分割される。図８Ｂにおいて、ｋ－ｍｅｒ　８０３の第２の群は単一の候補アライメントロケーションを形成する。

【０１８３】

図９は、配列グラフ９００へのダイナミックプログラミングまたはアライメントの開始に使用するシードを決定するための例示的なプロセスを示す。配列から生成できる多量の候補アライメントロケーションに起因して、グラフアライメントモジュールを介してそれらをすべて実行することは、所望されないかまたはできないだろう。この理由のために、最も良好なシードは、全カバレッジを使用してそれらをランキングすることによって実践的に選択することができる。例えば、シード１　９０１は５０％のカバレッジを有し、シード２　９０２は８０％のカバレッジを有し、シード３　９０３は２０％のカバレッジを有し；シード２を最も良好なものとして選択することができる。

【０１８４】

配列グラフを表わす本ストレージスキームおよびデータ構造内で、インデックス中のｋ－ｍｅｒのクエリを、１計算スレッドあたり１，０００、２，０００、３，０００、４，０００、５０００、６，０００、７，０００、８，０００、９，０００、１０，０００、１５，０００、２０，０００、２５，０００、３０，０００、３５，０００、４０，０００、４５，０００、５０，０００、５５，０００、６０，０００、６５，０００、６６，０００、６７，０００、６８，０００、６９，０００、７０，０００、７５，０００、８０，０００、８５，０００、９０，０００、９５，０００、１００，０００、１０５，０００、１１０，０００、１１５，０００、１２０，０００、１２５，０００、１３０，０００、１３５，０００、１４０，０００、１４５，０００、１５０，０００、１５５，０００、１６０，０００、１６５，０００、１７０，０００、１７５，０００、１８０，０００、１８５，０００、１９０，０００、１９５，０００、２００，０００、２０５，０００、２１０，０００、２１５，０００、２２０，０００、２２５，０００、２３０，０００、２３５，０００、２４０，０００、２４５，０００、２５０，０００、２５５，０００、２６０，０００、２６５，０００、２７０，０００、２７５，０００、２８０，０００、２８５，０００、２９０，０００、２９５，０００、３００，０００、３０５，０００、３１０，０００、３１５，０００、３２０，０００、３２５，０００、３３０，０００、３３５，０００、３４０，０００、３４５，０００、３５０，０００、または３５５，０００　ｋ－ｍｅｒ／秒以上の率で実行することができる。いくつかの実施形態において、これらのデータ構造を使用して、アライメントさせる配列からのｋ－ｍｅｒのクエリを、３５５，０００　ｋ－ｍｅｒ／秒／計算スレッド以上の率で、配列グラフインデックスにおいて実行することができる。業界基準の明示的なグラフインデックスに基づいて構築された代替のｋ－ｍｅｒインデックスのクエリは、７０～１０００　ｋ－ｍｅｒ／秒／計算スレッドの率で実行することができる。

【０１８５】

ｋ－ｍｅｒがリファレンス配列グラフにおいて設置された後に、正規化されたオフセットは各々について計算され、配列は候補アライメントロケーションを有し、リファレンス配列グラフに沿った塩基の最長のカバレッジ（ｋ－ｍｅｒによってカバーされた塩基の合計によって決定されるように）を使用して、グラフ中の配列をシーディングすることができる（例えば図６Ｅを参照）。一実施例において、これらのデータ構造を使用することによって、配列は８７０４配列／秒／計算スレッドの率（ｋ－ｍｅｒ同期化のための時間を含む）でシーディングすることができる。配列は、約１００、２００、３００、４００、５００、６００、７００、８００、９００、１０００、１１００、１２００、１３００、１４００、１５００、１６００、１７００、１８００、１９００、２０００、２１００、２２００、２３００、２４００、２５００、２６００、２７００、２８００、２９００、３０００、３１００、３２００、３３００、３４００、３５００、３６００、３７００、

３８００、３９００、４０００、４１００、４２００、４３００、４４００、４５００、４６００、４７００、４８００、４９００、５０００、５１００、５２００、５３００、５４００、５５００、５６００、５７００、５８００、５９００、６０００、６１００、６２００、６３００、６４００、６５００、６６００、６７００、６８００、６９００、７０００、７１００、７２００、７３００、７４００、７５００、７６００、７７００、７８００、７９００、８０００、８１００、８２００、８３００、８４００、８５００、８６００、８７００、８８００、８９００、９０００、９１００、９２００、９３００、９４００、９５００、９６００、９７００、９８００、９９００、または１００００配列／秒／計算スレッド以上の率（ｋ－ｍｅｒ同期化のための時間を含む）でシーディングすることができる。

【０１８６】

最も高いカバレッジを備えたものがグラフアライメントモジュールへ渡されるように、候補アライメントロケーションをランキングすることができる（例えば図９を参照）。一実施例において、１配列あたり単一のシードをアライメントさせることによって、配列は１３，７５４リード／秒／計算スレッドの率でアライメントされる。一実施例において、１配列あたり多くとも５つのシードをアライメントさせることによって、配列は４，６０７リード／秒／計算スレッドの率でアライメントされる。一実施例において、１配列あたり多くとも３２のシードをアライメントさせることによって、配列は９７８リード／秒／計算スレッドの率でアライメントされる。

【０１８７】

表１は、従来技術の直線アライナーと比較した、本明細書において開示される本グラフベースの方法の感受性および真の発見率（すなわち１－偽発見率）を示す。これらの結果は、３０×カバレッジでＶａｒＳｉｍを使用して、染色体１のシミュレーションから生成された。結果は、真の発見率について０．９％の改善、およびＢＷＡに関する感受性について０．４％の差を示す。

【表１】

感受性および真の発見率

方法	感受性(%)	真の発見率(%)
BWA	98.3	96.1
グラフベースの方法	**97.9**	**97.0**

【０１８８】

ｋ－ｍｅｒプロファイルは、そのｋ－要素またはｋ－ｍｅｒへと細分化された配列を表わすことができる。いくつかの事例において、ｋ－ｍｅｒのセットを表わすプロファイルは、他の要素配列毎のｋ－ｍｅｒへの配列の細分化を表わすことができる。いくつかの事例において、ｋ－ｍｅｒプロファイルは、有することができる最も少数の要素へのｋ－ｍｅｒの分割を含むことができる。例えば、図４は、配列のｋ－ｍｅｒプロファイルを有する、２つの例示的な手法を示す。左側では、配列４００は、６つのｋ－ｍｅｒ　４０２　４０３　４０４　４０５　４０６　４０７（各々５のサイズ）へと細分化され４０１、順次加えられる。右側では、同じ配列４１０は、２つのオーバーラップしないｋ－ｍｅｒ　４１１　４１２へと細分化される。

【０１８９】

本開示の技法を使用して、リードを検索すること、このリードからｋ－ｍｅｒプロファイルを生成すること、およびｋ－ｍｅｒプロファイルのクエリを、リファレンス配列からのｋ－ｍｅｒプロファイルのインデックスに対して、代替パスにより実行して、バリアントをコールすることができる。いくつかの事例において、クエリの実行を使用して、配列の特異的な断片を検出することができる。いくつかの事例において、その断片を使用して、バリアントの存在についてクエリを実行することができる。

【０１９０】

図５は、それらのＩＤ　５０３と共に、リファレンス５０１および代替パス５０２の例を示す。いくつかの事例において、代替パスはバブルと称される。様々な追加のリードタグ５０４が使用することができる。リードがクロスされた代替パスを言及する「ＶＬ」タグが使用することができる。「ＶＮ」タグも、リードがどれだけのバリアントを通過するかを詳述するのに使用することができる。「ＮＬ」タグは、バブルへアライメントするリファレンスパスを詳述するのに使用することができる（例えばそれは代替パスを通過しなかった。）。「ＶＶ」リードタグは、バリアントにマッピングされたリードを詳述するのに使用することができる。「ＧＤ」タグは、バリアントパスに関係するマッチ、ミスマッチ、挿入、欠失および開始位置の数が含まれるが、これらに限定されないアライメントスコアを含有することができる。「ＧＲ」タグは、バリアントパスの座標と比べてアライメントされたリードの開始および終了の両方を指示することができる。

【０１９１】

ｋ－ｍｅｒプロファイルが代替パスを備えたリファレンス配列のインデックスを通過するので、システムは、ｋ－ｍｅｒが代替パス中にあるか否かを問い合わせことができる。いくつかの事例において、これはバリアントをコールするのに十分な根拠である。他の事例において、質の高いスコアを備えたｋ－ｍｅｒのみがバリアントへ帰着することができる。他の事例において、バリアントは統計モデルを使用して選択することができる。

【０１９２】

いくつかの事例において、ｋ－ｍｅｒプロファイル形成は、パスのインデックスの形成に加えてギャップを導入したｋ－ｍｅｒを含むことができる。

【０１９３】

いくつかの事例において、ｋ－ｍｅｒインデックスは、１，０００塩基毎に１までの率で配列を圧縮することができ；他では、それは１，０００，０００の塩基において１で圧縮することができる。他では、それは１０，０００，０００塩基以上において１で圧縮することができる。

【０１９４】

ｋ－ｍｅｒインデックスはフェーズ化情報を含み、簡潔な代替パスを生成することができる。

【０１９５】

いくつかの事例において、代替パスおよびその対応するリファレンス（すなわちバブル）に関係するインデックスのｋ－ｍｅｒが使用される一方で、インデックスの残りは廃棄される。いくつかの事例において、これは、ｋ－ｍｅｒインデックスのサイズを９９％を超えて低減させる。いくつかの事例において、それは９９．９％を超える。ｋ－ｍｅｒ空間をこのサイズに低減することは、バリアント、亜種および異なる配列の間の差を強調することができる。いくつかの事例において、これは、クエリを実行するプロセスを、毎秒１．１×、１．２×、１．３×、１．４×、１．５×、１．６×、１．７×、１．８×、１．９×、２×、３×、４×、５×、６×、７×、８×、９×、１０×、２０×、３０×、４０×、５０×、６０×、７０×、８０×、９０×、１００×、２００×、３００×、４００×、５００×、６００×、７００×、８００×、９００×、または１０００×ｋ－ｍｅｒを超えて加速することができる。

【０１９６】

いくつかの事例において、バリアントのコールはインデックスを使用して遂行され、そこでは、代替パスおよびその対応するリファレンス（すなわちバブル）のみが使用される一方で、インデックスの残りは廃棄することができる。これは、バリアントコーリングプロセスを、毎秒１．１×、１．２×、１．３×、１．４×、１．５×、１．６×、１．７×、１．８×、１．９×、２×、３×、４×、５×、６×、７×、８×、９×、１０×、２０×、３０×、４０×、５０×、６０×、７０×、８０×、９０×、１００×、２００×、３００×、４００×、５００×、６００×、７００×、８００×、９００×、または１０００×ｋ－ｍｅｒを超えて加速することができる。

【０１９７】

ｋ－ｍｅｒは、２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、２４、２５、２６、２７、２８、２９、３０、３１、３２、３３、３４、３５、３６、３７、３８、３９、４０、４１、４２、４３、４４、４５、４６、４７、４８、４９、５０、５５、６０、６５、７０、７５、８０、８５、９０、９５、１００以上のサイズとすることができる。

【０１９８】

本発明は１又は２以上の好ましい実施形態に関して記載され、明示的に述べられたもの以外の多くの均等物、代替物、変形物、および変更物が可能であり、本発明の範囲内であることが認識されるべきである。

【図１】

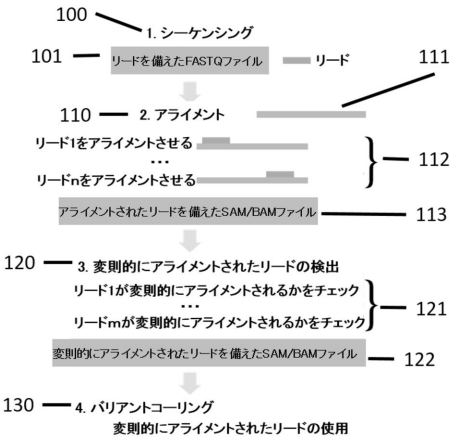

FIG. 1

【図２】

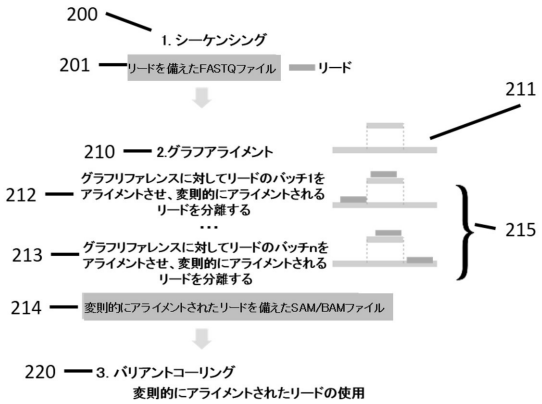

FIG. 2

【図３】

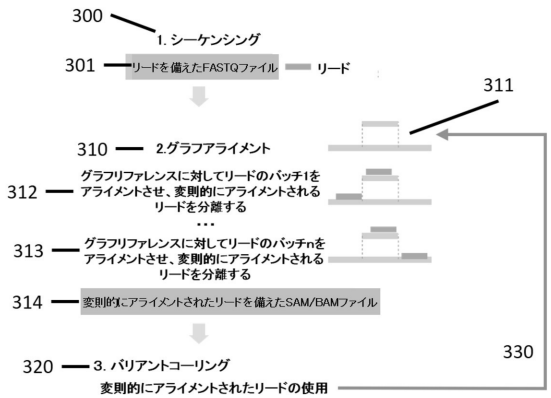

FIG. 3

【図４】

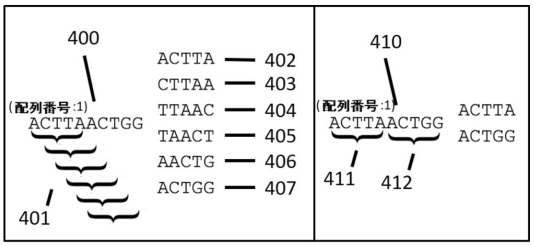

FIG. 4

【図５】

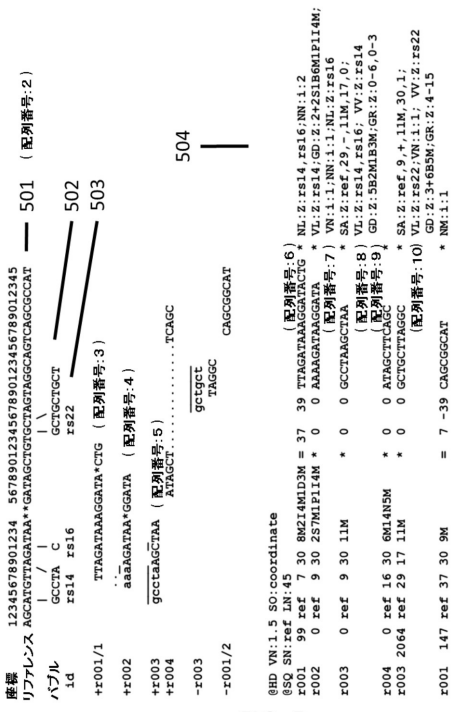

FIG. 5

【図６Ａ】

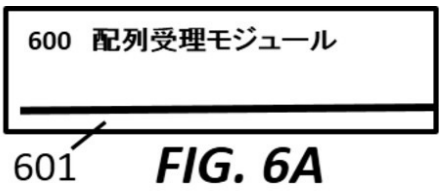

FIG. 6A

【図６Ｂ】

FIG. 6B

【図６Ｃ】

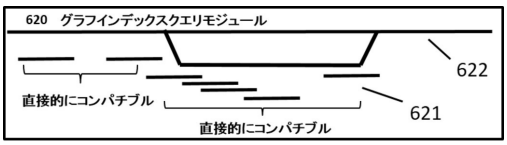

FIG. 6C

【図６Ｄ】

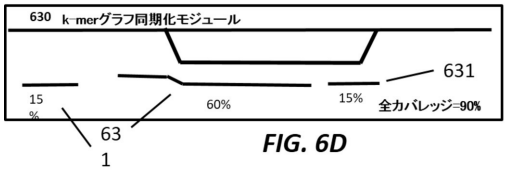

FIG. 6D

【図６Ｅ】

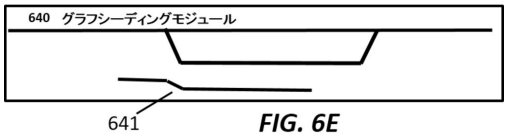

FIG. 6E

【図６Ｆ】

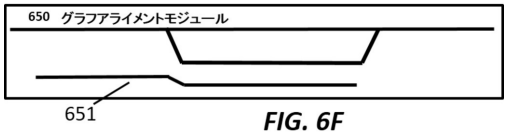

FIG. 6F

【図７Ａ】

FIG. 7A

【図７Ｂ】

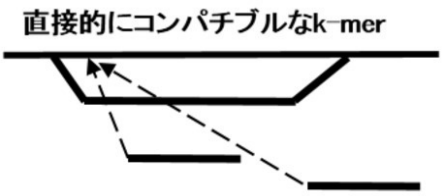

FIG. 7B

【図７Ｃ】

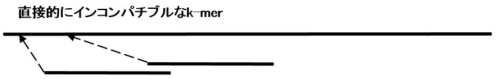

FIG. 7C

【図７Ｄ】

FIG. 7D

【図７Ｅ】

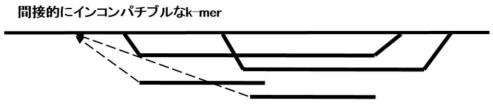

FIG. 7E

【図８Ａ】

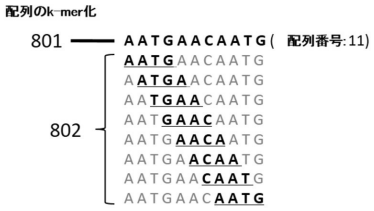

FIG. 8A

【図８Ｂ】

FIG. 8B

【図９】

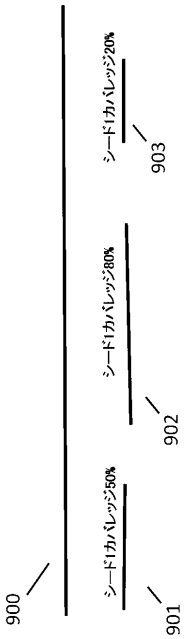

FIG. 9

【配列表】
0006946292000001.app

フロントページの続き

(74)代理人 100086771
弁理士 西島 孝喜
(74)代理人 100109070
弁理士 須田 洋之
(74)代理人 100109335
弁理士 上杉 浩
(74)代理人 100120525
弁理士 近藤 直樹
(74)代理人 100139712
弁理士 那須 威夫
(74)代理人 100176418
弁理士 工藤 嘉晃
(72)発明者 キロス サラテ アレハンドロ
アメリカ合衆国 マサチューセッツ州 ０２１３８ ケンブリッジ ヒューロン アベニュー ７００ ２０１
(72)発明者 オリヴァレス－アマージャ ロベルト
アメリカ合衆国 マサチューセッツ州 ０２１４４ サマーヴィル カレッジ アベニュー ２０１ ユニット ２
(72)発明者 ワトソン トマス ジェイムズ
アメリカ合衆国 マサチューセッツ州 ０２４６６ オーバーンデール オークランド アベニュー ２８
(72)発明者 ファン アヘレン ヘレン セシル
アメリカ合衆国 マサチューセッツ州 ０２１４４ サマーヴィル カレッジ アベニュー ２０１ ユニット ２
(72)発明者 コロナド スロカ エドゥアルド
アメリカ合衆国 マサチューセッツ州 ０２２１５ ボストン クイーンズベリー ストリート ９８ アパートメント ５
(72)発明者 アングロ セルメノ カルロス アントニオ
メキシコ サン ルイス ポトシ ３７２７０ サン ルイス ポトシ フアン デル ハロ ４６０ インテリオール １６
(72)発明者 フィンブレス フラド フェルナンド
メキシコ グアナフアト ３７５３０ レオン ハルディン デ ロス ラウレレス １０１ア
(72)発明者 ソリス ガルシア－インダ アブラハム
メキシコ グアナフアト ３６６６０ イラプアト オラス ８０３
(72)発明者 フォントベ エルレラ フェルナンド
メキシコ グアナフアト ３６６４３ イラプアト ビージャ アルタ ２８６
(72)発明者 コステ パブロ ジー
アメリカ合衆国 マサチューセッツ州 ０２４５９ ニュートン スピアーズ ロード ３５

審査官 山内 裕史

(56)参考文献 米国特許出願公開第２０１５／０１９９４７５（ＵＳ，Ａ１）
特開２０１２－２３９４３０（ＪＰ，Ａ）
特表２０１４－５０５９３５（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ ５／００ － ９９／００