(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6824973号
(P6824973)
(45)発行日　令和3年2月3日(2021.2.3)　(24)登録日　令和3年1月15日(2021.1.15)

(51)Int.Cl.　　F I
G16B 30/00 (2019.01)　　G16B 30/00
C12Q 1/68 (2018.01)　　C12Q 1/68

請求項の数 15　(全 32 頁)

(21)出願番号　特願2018-517871(P2018-517871)
(86)(22)出願日　平成28年10月10日(2016.10.10)
(65)公表番号　特表2018-535481(P2018-535481A)
(43)公表日　平成30年11月29日(2018.11.29)
(86)国際出願番号　PCT/US2016/056314
(87)国際公開番号　WO2017/062970
(87)国際公開日　平成29年4月13日(2017.4.13)
審査請求日　令和1年10月1日(2019.10.1)
(31)優先権主張番号　62/239,879
(32)優先日　平成27年10月10日(2015.10.10)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　515059083
ガーダント　ヘルス，　インコーポレイテッド
アメリカ合衆国　カリフォルニア　９４０６３，　レッドウッド　シティ，　ペノブスコット　ドライブ　５０５
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100113413
弁理士　森下　夏樹
(74)代理人　100181674
弁理士　飯田　貴敏
(74)代理人　100181641
弁理士　石川　大輔

最終頁に続く

(54)【発明の名称】無細胞ＤＮＡ分析における遺伝子融合検出の方法および応用

(57)【特許請求の範囲】
【請求項１】
（ａ）ＤＮＡ分子を、ＤＮＡシーケンサーでシーケンシングして、配列のコレクションを生成するステップと；
（ｂ）前記配列のコレクションを、基準ゲノムへとマッピングするステップと；
（ｃ）融合リードを、前記マッピングコレクションから識別するステップであって、融合リードが、部分配列を含有し、第１の部分配列が、第１の遺伝子座へとマッピングされ、第２の部分配列が、第２の別個の遺伝子座へとマッピングされる、ステップと；
（ｄ）前記融合リードについて、前記第１の遺伝子座における第１の切断点と、前記第２の遺伝子座における第２の切断点とを識別するステップであって、切断点が、融合リードの配列が切り詰められた前記基準ゲノム上の点であり、前記第１の切断点と、第２の切断点とが、切断点対を形成する、ステップと；
（ｅ）融合リードのセットを生成するステップであって、各セットが、同じ切断点対を有する融合リードを含む、ステップと；
（ｆ）融合リードのセットをクラスタリングするステップであって、各クラスターを、第１の所定のヌクレオチド距離内の、第１の切断点と、第２の所定のヌクレオチド距離内の、第２の切断点とを有する融合リードのセットから形成し、前記第１の所定の距離と、第２の所定の距離とが各々、２５を超えないヌクレオチドである、ステップと；
（ｇ）１または複数のクラスターについて、遺伝子融合を決定するステップであって、クラスターの遺伝子融合が、第１の融合遺伝子の切断点として、前記クラスター内の、前

記第１の切断点から選択される切断点を有し、かつ、第２の融合遺伝子の切断点として、前記クラスター内の、前記第２の切断点から選択される切断点を有し、前記第１の融合遺伝子切断点と、第２の融合遺伝子切断点とが各々、選択基準に基づき選択され、前記選択基準が、前記クラスター内で、最も多くの融合リードを有する切断点を含む、ステップとを含む方法。

【請求項２】

前記別個の遺伝子座が、異なる染色体上、または同じ染色体の、異なる遺伝子上に位置する、請求項１に記載の方法。

【請求項３】

前記第１の所定の距離と、第２の所定の距離とが各々、５を超えないヌクレオチド、または１０を超えないヌクレオチドである、請求項１または請求項２に記載の方法。

【請求項４】

複数の遺伝子クラスターについて、遺伝子融合を決定するステップを含む、請求項１～３のいずれか一項に記載の方法。

【請求項５】

前記ＤＮＡ分子が、シーケンシングの前に処理するステップに供され、必要に応じて、前記処理するステップが、得るステップ、単離するステップ、断片化するステップ、増幅ステップ、および／またはバーコード処理するステップである、請求項１～４のいずれか一項に記載の方法。

【請求項６】

前記ＤＮＡ分子が、生物から直接得られるか、生物から得られる生物学的試料、例えば、血液、血清、血漿、尿、脳脊髄液、唾液、糞便、リンパ液、滑液、嚢胞液、腹水、胸水、羊水、絨毛膜絨毛試料、着床前胚に由来する流体、胎盤試料、子宮頸部／膣洗浄液および子宮頸部／膣液、間質液、口腔スワブ試料、痰、気管支洗浄液、パップスメア試料、または眼液から得られる、請求項５に記載の方法。

【請求項７】

前記ＤＮＡ分子が、非細胞由来から単離される、請求項５または請求項６に記載の方法。

【請求項８】

バーコード処理するステップは、前記ＤＮＡ分子の一方または両方の末端にバーコード配列を接合させるステップを含む、請求項５～７のいずれか一項に記載の方法。

【請求項９】

前記生物が、がんを有するか、がんを有すると疑われている、請求項６に記載の方法。

【請求項１０】

固有の分子またはリード識別子（リードＩＤ）を各々のリードに割り当てるステップを含む、請求項１～９のいずれか一項に記載の方法。

【請求項１１】

適切な符号を伴う２つの切断点に属する共通のリードＩＤを伴う２つのマッピングリード部分を、潜在的な融合候補として選択するステップを含む、請求項１０に記載の方法。

【請求項１２】

マッピングする前における、元のリード内の、前記潜在的な融合候補の場所が、前記リード部分を、元々互いに隣接して位置するリード部分として示す、請求項１１に記載の方法。

【請求項１３】

リード部分が、１つの鎖にマッピングされる場合、前記切断点の符号の差違について点検するステップを含む、請求項１１または１２に記載の方法。

【請求項１４】

前記基準が、クラスター内に、１つを超える分子を有するか、またはワトソン－クリック鎖の両方を伴う、少なくとも１つの分子を有することを含む、請求項１～１３のいずれか一項に記載の方法。

【請求項１５】
遺伝子情報を解析するシステムであって、
ＤＮＡシーケンサーと；
前記ＤＮＡシーケンサーに結合されたプロセッサーであって、
融合染色体ＤＮＡ分子の、一部のシーケンシングデータを含有する融合リードを決定すること；
前記融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の少なくとも１つの所定の点（切断点）を決定すること；
２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別すること；
切断点対に基づき、１または複数の融合セットを創出し、前記融合セットを、１または複数の融合クラスターへとクラスタリングすること；および
所定の基準を満たす各融合クラスターを、遺伝子融合として識別すること
のための命令を含むコンピュータコードを実行して、試料に由来する遺伝子配列リードデータを処理するプロセッサーと
を含むシステム。

【発明の詳細な説明】
【技術分野】
【０００１】
相互参照
本出願は、２０１５年１０月１０日に出願された米国仮出願番号第６２／２３９，８７９号の利益を主張しており、この仮出願は、その全体が参考として本明細書中に援用される。
【背景技術】
【０００２】
発明の背景
がん性細胞は、一体に融合した染色体を有しうる。このような染色体をシーケンシングすれば、ゲノムの２つの異なるゾーン（または同じ染色体もしくは異なる染色体上）にマッピングされうるリードが生成される。遺伝子融合は、遺伝子アーキテクチャーの進化において役割を果たす。重複、配列分岐、および組換えは、遺伝子進化において作用する主要な寄与因子である。遺伝子融合が非コード配列領域内で起こる場合、今や別の遺伝子のシス調節配列の制御下にある遺伝子の発現の調節異常をもたらしうる。遺伝子融合がコード配列内で起こる場合、新たな遺伝子のアセンブリを引き起こしうることから、マルチドメインタンパク質へと、ペプチドモジュールを追加することにより、新たな機能の出現を可能とする。
【０００３】
染色体バンド形成解析、蛍光ｉｎ　ｓｉｔｕハイブリダイゼーション（ＦＩＳＨ）、および逆転写ポリメラーゼ連鎖反応（ＲＴ－ＰＣＲ）は、診断検査室で利用される、一般的な方法である。これらの方法は全て、がんゲノムの複雑な性格のために、それぞれ別個の欠点を有する。ハイスループットシーケンシングおよびカスタムＤＮＡマイクロアレイなど、近年の発展は、より効率的な方法の導入において有望であるが、いまだに不十分である。ハイスループットゲノムシーケンシング技術は、研究ツールとして使用されており、現在、診療所にも導入されており、オーダーメイド医療の将来では、全ゲノム配列データは、治療的介入を導くのに、重要なツールでありうる。
【発明の概要】
【課題を解決するための手段】
【０００４】
発明の要旨

一態様では、融合染色体ＤＮＡ分子の、少なくとも一部のシーケンシングデータを含有する融合リードを決定し；融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定し；２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別し；切断点対に基づき、１または複数の融合セットを創出し、融合セットを、１または複数の融合クラスターへとクラスタリングし；所定の基準を満たす各融合クラスターを、遺伝子融合として識別することにより、遺伝子融合を決定するための、システムおよび方法が開示される。

【０００５】

一態様では、本開示は、試料に由来する遺伝子配列リードデータを処理するための方法であって、融合染色体ＤＮＡ分子の、少なくとも一部のシーケンシングデータを含有する融合リードを決定するステップと；融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定するステップと；２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別するステップと；切断点対に基づき、１または複数の融合セットを創出し、融合セットを、１または複数の融合クラスターへとクラスタリングするステップと；所定の基準を満たす各融合クラスターを、遺伝子融合として識別するステップとを含む方法を提示する。

【０００６】

一部の実施形態では、方法は、固有の分子またはリード識別子（リードＩＤ）を、各リードへと割り当てるステップを含む。一部の実施形態では、方法は、リードの各マッピング部分を、片側または両側から切り詰めるステップを含む。一部の実施形態では、切断点は、アイデンティティにおいて、リードから独立しており、符号、染色体、および位置により識別される。一部の実施形態では、切断点は、切断点で切り詰められるかまたは分割された多数のリードおよび分子と、切断点を通り越す、多数の野生型のリードおよび分子とを含む統計を保持する。一部の実施形態では、方法は、適切な符号を伴う２つの切断点に属する共通のリードＩＤを伴うあらゆる２つのマッピングリード部分を、潜在的な融合候補として選択するステップを含む。一部の実施形態では、マッピングする前における、元のリード内の、潜在的な融合候補の場所は、リード部分を、元々互いに隣接して位置するリード部分として示す。一部の実施形態では、方法は、リード部分が、１つの鎖にマッピングされる場合、切断点の符号の差違について点検するステップを含む。一部の実施形態では、方法は、融合セット統計を追跡するステップを含む。

【０００７】

一部の実施形態では、融合セット統計は、切断点ＩＤ、セット内に含有されている、分子またはリードの数である。一部の実施形態では、方法は、融合クラスター内の、同様の切断点により、融合セットを群分けするステップを含む。一部の実施形態では、同様の切断点は、５を超えないヌクレオチド、１０を超えないヌクレオチド、または２５を超えないヌクレオチド離れた切断点である。一部の実施形態では、方法は、ゲノム内の２つの領域の間で、融合クラスターを規定するステップを含む。一部の実施形態では、方法は、融合クラスターに関して、各パートナーについて、多数の融合分子を決定するステップを含む。一部の実施形態では、方法は、融合クラスターに関して、各パートナーについて、融合リードの数を決定するステップを含む。一部の実施形態では、方法は、融合クラスターに関して、各パートナーについて、多数の野生型分子を決定するステップを含む。一部の実施形態では、方法は、融合クラスターに関して、各パートナーについて、多数の野生型リードまたは野生型分子を決定するステップを含む。一部の実施形態では、方法は、融合クラスターに関して、各パートナーについて、融合百分率を、各パートナーの、全分子に対する融合分子の比率として決定するステップを含む。一部の実施形態では、全分子は、野生型構成要素と、切り詰められた構成要素とを含む。一部の実施形態では、方法は、融合クラスターに関して、各パートナーについて、遺伝子情報を決定するステップを含む。一部の実施形態では、方法は、融合クラスターの、下流の遺伝子を決定するステップを含む。一部の実施形態では、基準は、クラスター内に、１つを超える分子を有すること、またはワトソン－クリック鎖の両方を伴う、少なくとも１つの分子を有することを含む。

【０００８】
一態様では、本開示は、遺伝子情報を解析するシステムであって、ＤＮＡシーケンサーと；ＤＮＡシーケンサーに結合されたプロセッサーであって、融合染色体ＤＮＡ分子の一部のシーケンシングデータを含有する融合リードを決定し；融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の少なくとも１つの所定の点（切断点）を決定し；２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別し；切断点対に基づき、１または複数の融合セットを創出し、融合セットを、１または複数の融合クラスターへとクラスタリングし；所定の基準を満たす各融合クラスターを、遺伝子融合として識別するための命令を含むコンピュータコードを実行して、試料に由来する遺伝子配列リードデータを処理するプロセッサーとを含むシステムを提示する。

【０００９】
一態様では、本開示は、ＤＮＡ分子を、ＤＮＡシーケンサーでシーケンシングして、配列のコレクションを生成するステップと；配列のコレクションを、基準ゲノムへとマッピングするステップと；融合リードを、マッピングコレクションから識別するステップであって、融合リードが、部分配列を含有し、第１の部分配列が、第１の遺伝子座へとマッピングされ、第２の部分配列が、第２の別個の遺伝子座へとマッピングされる、ステップと；各融合リードについて、第１の遺伝子座における第１の切断点と、第２の遺伝子座における第２の切断点とを識別するステップであって、切断点が、融合リードの配列が切り詰められた基準ゲノム上の点であり、第１の切断点と、第２の切断点とが、切断点対を形成する、ステップと；融合リードのセットを生成するステップであって、各セットが、同じ切断点対を有する融合リードを含む、ステップと；融合リードのセットをクラスタリングするステップであって、各クラスターを、第１の所定のヌクレオチド距離内の、第１の切断点と、第２の所定のヌクレオチド距離内の、第２の切断点とを有する融合リードのセットから形成する、ステップと；１または複数のクラスターについて、遺伝子融合を決定するステップであって、クラスターの遺伝子融合が、第１の融合遺伝子の切断点として、クラスター内の、第１の切断点から選択される切断点を有し、かつ、第２の融合遺伝子の切断点として、クラスター内の、第２の切断点から選択される切断点を有し、第１の融合遺伝子切断点と、第２の融合遺伝子切断点とが各々、選択基準に基づき選択される、ステップとを含む方法を提示する。

【００１０】
一部の実施形態では、別個の遺伝子座は、異なる染色体上、または同じ染色体の、異なる遺伝子上に位置する。一部の実施形態では、第１の所定の距離と、第２の所定の距離とは各々、５を超えないヌクレオチド、１０を超えないヌクレオチド、または２５を超えないヌクレオチドである。一部の実施形態では、選択基準は、クラスター内で、最も多くの融合リードを有する切断点を含む。一部の実施形態では、方法は、複数の遺伝子クラスターについて、遺伝子融合を決定するステップを含む。

【００１１】
一態様では、本開示は、複数のＤＮＡ分子を、ＤＮＡシーケンサーでシーケンシングするステップと；複数の配列の分子の各々を、識別子でタグ付けするステップと；各タグ付き配列を、基準ゲノムへとマッピングするステップと；切り詰められたリードを、マッピングされたタグ付き配列から識別するステップであって、切り詰められたリードが、マッピング部分と、切り詰められた部分とを含有するタグ付き配列であり、マッピング部分が、遺伝子座へとマッピングされ、切り詰められた部分が、遺伝子座へとマッピングされない、ステップと；各切り詰められたリードの切断点を決定するステップであって、切断点が、切り詰められたリードの配列が切り詰められた基準ゲノム上の点である、ステップと；切断点セットを創出するステップであって、各切断点セットが、同じ切断点を有する切り詰められたリードの識別子を含む、ステップと；切断点セットの対を比較することにより、切断点対のセットを創出するステップであって、切断点対の各セットが、切断点セットの比較される対のいずれのメンバーにおいても存在する識別子を含む、ステップと；切

断点対のセットをクラスタリングするステップであって、各クラスターが、第１の所定の遺伝子距離内にある対の、第１の切断点と、第２の所定の遺伝子距離内にある対の、第２の切断点とを有する、切断点対のセットを含む、ステップと；クラスターのうちの１または複数について、遺伝子融合を決定するステップであって、クラスターの遺伝子融合が、第１の融合遺伝子の切断点として、クラスター内の、第１の切断点から選択される切断点を有し、かつ、第２の融合遺伝子の切断点として、クラスター内の、第２の切断点から選択される切断点を有し、第１の融合遺伝子切断点と、第２の融合遺伝子切断点とが各々、選択基準に基づき選択される、ステップとを含む方法を提示する。一部の実施形態では、選択基準は、クラスター内で、最も多くの融合リードを有する切断点を含む。

【００１２】

一態様では、本開示は、融合遺伝子切断点を識別するための方法であって、融合染色体ＤＮＡ分子の、少なくとも一部のシーケンシングデータを含有する融合リードを決定するステップと；融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定するステップと；２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別するステップと；切断点対に基づき、１または複数の融合セットを創出し、融合セットを、１または複数の融合クラスターへとクラスタリングするステップと；所定の基準を満たす各融合クラスターを、遺伝子融合として識別するステップと；遺伝子融合の切断点を、融合遺伝子切断点として識別するステップとを含む方法を提示する。

【００１３】

一態様では、本開示は、対象における状態を診断するための方法であって、融合染色体ＤＮＡ分子の、少なくとも一部のシーケンシングデータを含有する融合リードを決定するステップと；融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定するステップと；２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別するステップと；切断点対に基づき、１または複数の融合セットを創出し、融合セットを、１または複数の融合クラスターへとクラスタリングするステップと；所定の基準を満たす各融合クラスターを、遺伝子融合として識別するステップとを含み、前記遺伝子融合が、状態を指し示す、方法を提示する。

【００１４】

一部の実施形態では、状態は、がんである。一部の実施形態では、がんは、血液がん、肉腫、および前立腺がんからなる群から選択される。一部の実施形態では、方法は、処置を、対象へと投与するステップをさらに含む。

本発明は、例えば、以下の項目を提供する。

（項目１）

試料に由来する遺伝子配列リードデータを処理するための方法であって、

融合染色体ＤＮＡ分子の少なくとも一部のシーケンシングデータを含有する融合リードを決定するステップと；

前記融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定するステップと；

２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別するステップと；

切断点対に基づき、１または複数の融合セットを創出し、前記融合セットを、１または複数の融合クラスターへとクラスタリングするステップと；

所定の基準を満たす各融合クラスターを、遺伝子融合として識別するステップと

を含む方法。

（項目２）

固有の分子またはリード識別子（リードＩＤ）を、各リードへと割り当てるステップを含む、項目１に記載の方法。

（項目３）

前記リードの各マッピング部分を、片側または両側から切り詰めるステップを含む、項目１に記載の方法。
（項目４）
前記切断点が、アイデンティティにおいて、前記リードから独立しており、符号、染色体、および位置により識別される、項目１に記載の方法。
（項目５）
前記切断点が、前記切断点で切り詰められるかまたは分割された多数のリードおよび分子と、前記切断点を通り越す、多数の野生型のリードおよび分子とを含む統計を保持する、項目４に記載の方法。
（項目６）
適切な符号を伴う２つの切断点に属する共通のリードＩＤを伴うあらゆる２つのマッピングリード部分を、潜在的な融合候補として選択するステップを含む、項目２に記載の方法。
（項目７）
マッピングする前における、元のリード内の、前記潜在的な融合候補の場所が、前記リード部分を、元々互いに隣接して位置するリード部分として示す、項目６に記載の方法。
（項目８）
リード部分が、１つの鎖にマッピングされる場合、前記切断点の符号の差違について点検するステップを含む、項目６に記載の方法。
（項目９）
融合セット統計を追跡するステップを含む、項目１に記載の方法。
（項目１０）
前記融合セット統計が、切断点ＩＤ、前記セット内に含有されている、分子またはリードの数である、項目９に記載の方法。
（項目１１）
融合クラスター内の、同様の切断点により、融合セットを群分けするステップを含む、項目１に記載の方法。
（項目１２）
前記同様の切断点が、５を超えないヌクレオチド、１０を超えないヌクレオチド、または２５を超えないヌクレオチド離れた切断点である、項目１１に記載の方法。
（項目１３）
ゲノム内の２つの領域の間で、融合クラスターを規定するステップを含む、項目１に記載の方法。
（項目１４）
前記融合クラスターに関して、各パートナーについて、多数の融合分子を決定するステップを含む、項目１に記載の方法。
（項目１５）
前記融合クラスターに関して、各パートナーについて、融合リードの数を決定するステップを含む、項目１に記載の方法。
（項目１６）
前記融合クラスターに関して、各パートナーについて、多数の野生型分子を決定するステップを含む、項目１に記載の方法。
（項目１７）
前記融合クラスターに関して、各パートナーについて、多数の野生型リードまたは野生型分子を決定するステップを含む、項目１に記載の方法。
（項目１８）
前記融合クラスターに関して、各パートナーについて、融合百分率を、各パートナーの、全分子に対する融合分子の比率として決定するステップを含む、項目１に記載の方法。
（項目１９）
前記全分子が、野生型構成要素と、切り詰められた構成要素とを含む、項目１に記載の

方法。
（項目２０）
前記融合クラスターに関して、各パートナーについて、遺伝子情報を決定するステップを含む、項目１８に記載の方法。
（項目２１）
前記融合クラスターの、下流の遺伝子を決定するステップを含む、項目１に記載の方法。
（項目２２）
前記基準が、前記クラスター内に、１つを超える分子を有すること、またはワトソン－クリック鎖の両方を伴う、少なくとも１つの分子を有することを含む、項目１に記載の方法。
（項目２３）
遺伝子情報を解析するシステムであって、
ＤＮＡシーケンサーと；
前記ＤＮＡシーケンサーに結合されたプロセッサーであって、
融合染色体ＤＮＡ分子の、一部のシーケンシングデータを含有する融合リードを決定すること；
前記融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の少なくとも１つの所定の点（切断点）を決定すること；
２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別すること；
切断点対に基づき、１または複数の融合セットを創出し、前記融合セットを、１または複数の融合クラスターへとクラスタリングすること；および
所定の基準を満たす各融合クラスターを、遺伝子融合として識別すること
のための命令を含むコンピュータコードを実行して、試料に由来する遺伝子配列リードデータを処理するプロセッサーと
を含むシステム。
（項目２４）
（ａ）ＤＮＡ分子を、ＤＮＡシーケンサーでシーケンシングして、配列のコレクションを生成するステップと；
（ｂ）前記配列のコレクションを、基準ゲノムへとマッピングするステップと；
（ｃ）融合リードを、前記マッピングコレクションから識別するステップであって、融合リードが、部分配列を含有し、第１の部分配列が、第１の遺伝子座へとマッピングされ、第２の部分配列が、第２の別個の遺伝子座へとマッピングされる、ステップと；
（ｄ）各融合リードについて、前記第１の遺伝子座における第１の切断点と、前記第２の遺伝子座における第２の切断点とを識別するステップであって、切断点が、融合リードの配列が切り詰められた前記基準ゲノム上の点であり、前記第１の切断点と、第２の切断点とが、切断点対を形成する、ステップと；
（ｅ）融合リードのセットを生成するステップであって、各セットが、同じ切断点対を有する融合リードを含む、ステップと；
（ｆ）融合リードのセットをクラスタリングするステップであって、各クラスターを、第１の所定のヌクレオチド距離内の、第１の切断点と、第２の所定のヌクレオチド距離内の、第２の切断点とを有する融合リードのセットから形成する、ステップと；
（ｇ）１または複数のクラスターについて、遺伝子融合を決定するステップであって、クラスターの遺伝子融合が、第１の融合遺伝子の切断点として、前記クラスター内の、前記第１の切断点から選択される切断点を有し、かつ、第２の融合遺伝子の切断点として、前記クラスター内の、前記第２の切断点から選択される切断点を有し、前記第１の融合遺伝子切断点と、第２の融合遺伝子切断点とが各々、選択基準に基づき選択される、ステップと
を含む方法。

（項目２５）
前記別個の遺伝子座が、異なる染色体上、または同じ染色体の、異なる遺伝子上に位置する、項目２４に記載の方法。
（項目２６）
前記第１の所定の距離と、第２の所定の距離とが各々、５を超えないヌクレオチド、１０を超えないヌクレオチド、または２５を超えないヌクレオチドである、項目２４に記載の方法。
（項目２７）
前記選択基準が、前記クラスター内で、最も多くの融合リードを有する切断点を含む、項目２４に記載の方法。
（項目２８）
複数の遺伝子クラスターについて、遺伝子融合を決定するステップを含む、項目２４に記載の方法。
（項目２９）
（ａ）複数のＤＮＡ分子を、ＤＮＡシーケンサーでシーケンシングするステップと；
（ｂ）複数の配列の分子の各々を、識別子でタグ付けするステップと；
（ｃ）各タグ付き配列を、基準ゲノムへとマッピングするステップと；
（ｄ）切り詰められたリードを、前記マッピングされたタグ付き配列から識別するステップであって、切り詰められたリードが、マッピング部分と、切り詰められた部分とを含有するタグ付き配列であり、前記マッピング部分が、遺伝子座へとマッピングされ、前記切り詰められた部分が、前記遺伝子座へとマッピングされない、ステップと；
（ｅ）各切り詰められたリードの切断点を決定するステップであって、切断点が、切り詰められたリードの配列が切り詰められた前記基準ゲノム上の点である、ステップと；
（ｆ）切断点セットを創出するステップであって、各切断点セットが、同じ切断点を有する切り詰められたリードの識別子を含む、ステップと；
（ｇ）切断点セットの対を比較することにより、切断点対のセットを創出するステップであって、切断点対の各セットが、切断点セットの比較される対のいずれのメンバーにおいても存在する識別子を含む、ステップと；
（ｈ）切断点対のセットをクラスタリングするステップであって、各クラスターが、第１の所定の遺伝子距離内にある前記対の、第１の切断点と、第２の所定の遺伝子距離内にある前記対の、第２の切断点とを有する、切断点対のセットを含む、ステップと；
（ｉ）前記クラスターのうちの１または複数について、遺伝子融合を決定するステップであって、クラスターの遺伝子融合が、第１の融合遺伝子の切断点として、前記クラスター内の、前記第１の切断点から選択される切断点を有し、かつ、第２の融合遺伝子の切断点として、前記クラスター内の、前記第２の切断点から選択される切断点を有し、前記第１の融合遺伝子切断点と、第２の融合遺伝子切断点とは各々、選択基準に基づき選択される、ステップと
を含む方法。
（項目３０）
前記選択基準が、前記クラスター内で、最も多くの融合リードを有する切断点を含む、項目２９に記載の方法。
（項目３１）
融合遺伝子切断点を識別するための方法であって、
（ａ）融合染色体ＤＮＡ分子の、少なくとも一部のシーケンシングデータを含有する融合リードを決定するステップと；
（ｂ）前記融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定するステップと；
（ｃ）２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別するステップと；
（ｄ）切断点対に基づき、１または複数の融合セットを創出し、前記融合セットを、１

または複数の融合クラスターへとクラスタリングするステップと；
（ｅ）所定の基準を満たす各融合クラスターを、遺伝子融合として識別するステップと；
（ｆ）前記遺伝子融合の切断点を、前記融合遺伝子切断点として識別するステップとを含む方法。
（項目３２）
対象における状態を診断するための方法であって、
（ａ）融合染色体ＤＮＡ分子の、少なくとも一部のシーケンシングデータを含有する融合リードを決定するステップと；
（ｂ）前記融合リードの少なくとも１つのマッピング部分が切り詰められる、ゲノム上の所定の点（切断点）を決定するステップと；
（ｃ）２つの切断点（切断点対）からの２つのマッピングリード部分を、潜在的な融合候補として識別するステップと；
（ｄ）切断点対に基づき、１または複数の融合セットを創出し、前記融合セットを、１または複数の融合クラスターへとクラスタリングするステップと；
（ｅ）所定の基準を満たす各融合クラスターを、遺伝子融合として識別するステップとを含み、
前記遺伝子融合が、前記状態を指し示す、方法。
（項目３３）
前記状態が、がんである、項目３２に記載の方法。
（項目３４）
前記がんが、血液がん、肉腫、および前立腺がんからなる群から選択される、項目３３に記載の方法。
（項目３５）
処置を、前記対象へと投与するステップをさらに含む、項目３４に記載の方法。
【００１５】
参照による組込み
本明細書で言及される全ての刊行物、特許、および特許出願は、各個別の刊行物、特許、または特許出願が、参照により組み込まれることが、具体的、かつ、個別に指し示された場合と同じ程度に、参照により本明細書に組み込まれる。
【００１６】
本発明の新規の特色は、付属の特許請求の範囲において、精緻に明示される。本発明の特色および利点についての、より良好な理解は、本発明の原理が用いられる、例示的な実施形態を明示する、以下の詳細な説明と、添付の図面とを参照することにより得られる。
【図面の簡単な説明】
【００１７】
【図１】図１は、遺伝子融合を検出するための例示的な工程について図示する。
【００１８】
【図２】図２は、２つの他の染色体から、融合染色体を創出する、可能な異なるシナリオについて描示する。
【００１９】
【図３Ａ】図３Ａは、それぞれ、左／右から切り詰められたリード部分を伴う、例示的な＋／－の切断点を示す。
【００２０】
【図３Ｂ】図３Ｂは、遺伝子融合の検出で使用される、例示的なマージング過程を示す。
【００２１】
【図４】図４は、本発明のシステムについて図示する。
【００２２】
【図５】図５は、染色体Ａと染色体Ｂとの間における、例示的な遺伝子融合と、切断点を横切ってマッピングされるＤＮＡ融合リードとを示す。

【００２３】
【図６】図６は、ＤＮＡ断片の、基準ゲノム内の２つの場所への、例示的なマッピングを示す。
【００２４】
【図７】図７は、マッピング切断点が、異なる場所に位置する、例示的な融合リードを示す。
【００２５】
【図８】図８は、融合遺伝子切断点をコール（ｃａｌｌｉｎｇ）するための、マッピング融合リードの、セットへの例示的な群分けと、セットの、クラスターへの例示的な群分けとを示す。
【００２６】
【図９Ａ】図９Ａ～９Ｃは、遺伝子融合検出工程についての例示的な図解を示す。
【図９Ｂ】図９Ａ～９Ｃは、遺伝子融合検出工程についての例示的な図解を示す。
【図９Ｃ】図９Ａ～９Ｃは、遺伝子融合検出工程についての例示的な図解を示す。
【発明を実施するための形態】
【００２７】
発明の詳細な説明
本発明は、遺伝子融合を検出するためのシステムおよび方法に関する。
【００２８】
融合遺伝子の切断点の正確なマッピングは、難題である。シーケンシングにおけるエラーと、融合遺伝子のアライメントの困難は、切断点をマッピングしようと試みる場合に遭遇する困難のうちの２つに過ぎない。本明細書で記載されるシステムおよび方法は、以下の利点のうちの１または複数をもたらしうる。システムは、非融合遺伝子より活性がはるかに大きい異常なタンパク質をもたらしうるため腫瘍の形成に寄与しうる融合遺伝子を識別しうる。融合遺伝子（これらは、ＢＣＲ－ＡＢＬ、ＴＥＬ－ＡＭＬ１（全て、ｔ（１２；２１）を伴う）、ＡＭＬ１－ＥＴＯ（ｔ（８；２１）を伴う、Ｍ２　ＡＭＬ）、および前立腺がんにおいて生じることが多い、第２１染色体の間質欠失を伴う、ＴＭＰＲＳＳ２－ＥＲＧを含む）は、がんを引き起こすがん遺伝子であるので、システムは、がんの存在を、正確に決定する。ＴＭＰＲＳＳ２－ＥＲＧの場合、アンドロゲン受容体（ＡＲ）によるシグナル伝達を破壊し、発がん性のＥＴＳ転写因子によって、ＡＲの発現を阻害することにより、融合産物は、前立腺がんを調節する。融合遺伝子の大部分は、血液がん、肉腫、および前立腺がんから見出される。発がん性の融合遺伝子は、２つの融合パートナーに由来する、新たな機能または異なる機能を伴う、遺伝子産物をもたらしうる。代替的に、原がん遺伝子が、強力なプロモーターと融合し、これにより、発がん性の機能が、上流の融合パートナーの強力なプロモーターにより引き起こされる上方調節を介して、機能し始める。後者は、リンパ腫において一般的であり、そこでは、がん遺伝子が、免疫グロブリン遺伝子のプロモーターと並置される。発がん性の融合転写物はまた、トランススプライシングまたはリードスルーイベントによっても引き起こされうる。これらの遺伝子融合についての、ゲノム配列およびゲノム構造の視点からの解析は、改善されたがんの診断法および標的化された療法の開発を導くのに、関連するデータをもたらしうるであろう。
【００２９】
図１は、遺伝子融合を決定するための、例示的な工程を示す。一般に、工程は、遺伝子データを、シーケンサーから捕捉し、その挿入サイズが、２つのリード長の合計より小さい、シーケンシングされたペアエンドリードをアライメントし、接続する、マージング法を適用し、マージングの後で、固有のリード識別子（リードＩＤ）を、各リードへと割り当てる（１２）。次に、工程は、全ての切断点を抽出し、融合候補を位置特定する（１６）。次いで、工程は、切断点対に基づき、融合セットを形成し、融合クラスターのための統計を決定する（２０）。次いで、所定の基準に合致させることにより、融合クラスターを、検出された融合として識別する（２４）。
【００３０】

次に、図１の工程についての詳細を論じる。がん性細胞は、一体に融合した染色体を有しうる。このような染色体をシーケンシングすれば、ゲノムの２つの異なるゾーン（または同じ染色体もしくは異なる染色体）へとマッピングされうるリードが生成されるであろう。この挙動を用いて、融合を検出する。

【００３１】

下記で詳述される通り、マッピングの前に、固有のリード識別子（リードＩＤ）を、各リードへと割り当てると、１または複数のＦＡＳＴＱファイル内のリードヘッダー中にコードされるであろう。代替的に、固有のバーコードを含むオリゴヌクレオチドのような、固有の分子を、リードＩＤの代わりに使用することもできる。ＦＡＳＴＱファイルをマッピングしたら、このコードされたリードＩＤを読み出し、どのヒットが、同じ元のリードに由来するのかを、容易に示すことができる。次に、工程は、全ての切断点を抽出する：融合リード（リードは、融合染色体に由来する、部分的なＤＮＡ分子のシーケンシングデータを含有する）を、全体として、ゲノムへとマッピングすることはできず、マッピング装置は、それらの異なる部分を、ゲノム上の異なる場所へとマッピングする。これは、従来型の技法を使用して、切断点マッピングを試みる場合に難題を提示する。このようなリードの各マッピング部分は、片側または両側から切り詰められる（ｃｌｉｐｐｅｄ）。切断点は、融合リードの、少なくとも１つのマッピング部分が切り詰められたゲノム上の点である。切断点は、それらのアイデンティティにおいて、リードから独立しており、それらの符号、染色体、および位置により識別される。＋／－の切断点は、それぞれ、左／右から切り詰められたリード部分を有する。ある位置の同じ側から切り詰められるかまたは分割された全てのリードは、関連する切断点リードリストに列挙される。切断点はまた、切断点で切り詰められるかもしくは分割されたリードおよび分子の数；または切断点を通り越す野生型リードおよび野生型分子の数のような、他の統計も保持しうる。また、切断点位置における遺伝子情報も提供される。クラスタリングを伴うかまたは伴わずに、切断点を割り当てることにより、本明細書で記載される方法およびシステムを使用して、遺伝子融合が生じた遺伝子座を、正確に決定することができる。

【００３２】

次いで、工程は、融合を見出す：適切な符号を伴う２つの切断点に属する共通のリードＩＤを伴うあらゆる２つのマッピングリード部分は、潜在的な融合候補である。それらを真の融合候補とみなすには、それらが、リード部分が元々互いに隣接して位置したことを示す正しい断片の順序（マッピングする前における、元のリード内のそれらの場所）を有することも必要とされる。加えて、結果として得られる融合は、配列鎖に関して、生物学的に可能でなければならない。これは単純に、リード部分が、同じ鎖（いずれも５’鎖、またはいずれも３’鎖）へとマッピングされる場合、切断点の符号は、一致してはならず、逆もまた成り立つことを意味する。この例を、図２に示す。

【００３３】

全ての抽出された融合候補を、切断点対に基づき、融合セットに含める。融合セットはまた、切断点ＩＤ、ならびにセット内に含有される分子およびリードの数などの統計も保持しうる。これらの統計は、追跡することができる。

【００３４】

次いで、クラスタリングを実施する。切断点が十分に近接する全ての融合セットが、融合クラスターに群分けされ得る。結果として、融合クラスターは、ゲノム内の２つの領域の間で規定される。本開示はまた、融合クラスターに関して、各パートナーについて、多数の融合分子を決定するか、各パートナーについて、融合リードの数を決定するか、各パートナーについて、多数の野生型分子を決定するか、各パートナーについて、多数の野生型リードまたは野生型分子を決定するか、または各パートナーについて、融合百分率を、各パートナーの、全分子に対する融合分子の比率として決定することも提示する。

【００３５】

図５は、染色体Ａと染色体Ｂとの間の、仮説的な遺伝子融合を示す。交差の結果として、遺伝子融合は、各染色体の一部を含有する。交差点を、切断点と称する。無細胞ＤＮＡ

では、融合リード１、融合リード２、および融合リード３など、ＤＮＡ断片は、切断点を横切ってマッピングされうる。

【００３６】

シーケンシングは、ＤＮＡ断片の配列をもたらす。ソフトウェアは、識別タグにより、各配列をマークする。ソフトウェアまた、これらの結果として得られる配列も、基準ゲノムへとマッピングする。図６は、融合リード１の、基準ゲノムへの仮説的なマッピングを示す。マッピングソフトウェアは、融合リードの配列を、基準ゲノム内のいずれの場所であれ、十分な相同性が見出される場所へとマッピングする。あいまいな配列は、基準ゲノム内の複数の場所にマッピングされうる。

【００３７】

融合遺伝子の切断点を横切ってマッピングされる融合リードの場合、ソフトウェアは典型的には、融合リードの配列を、各染色体へと１回ずつの２回にわたりマッピングする。しかし、各場合に、マッピングソフトウェアは、配列の一部（部分配列）を、基準ゲノムへと、適正にマッピングしえない。したがって、マッピング配列は、基準ゲノムへとマッピングされる部分配列、および相同性が小さい結果として、同じ遺伝子座へとマッピングされない部分配列の両方を含む。このような部分配列を、「切り詰められた（ｃｌｉｐｐｅｄ）」配列と称する。リードが切り詰められた基準ゲノム上の点が、切断点である。

【００３８】

各配列は、識別タグを保有するため、２つの異なる場所へとマッピングされる配列は、同一なタグに起因して、同じ元の配列に由来するものとして識別することができる。こうして、例えば、十分な相同性を有する配列の部分配列は、染色体Ａへとマッピングされ、不十分な相同性を有する配列の部分配列は、切り詰められる。同様に、マッピングソフトウェアは、配列を、染色体Ｂへとマッピングし、相同性が不十分な場合、配列を切り詰める。

【００３９】

しかし、シーケンシングにおけるエラー、およびマッピングアルゴリズムの特徴を含むいくつかの因子の結果として、融合遺伝子の切断点を含むＤＮＡ断片が、基準染色体の各々上の切断点遺伝子座へと、正確にマッピングされないことがある。例えば、マッピングソフトウェアは、配列の切断点を、実際の切断点のやや上流に識別する場合もあり、やや下流に識別する場合もある。

【００４０】

各配列は、識別タグを保有するため、２つの異なる場所へとマッピングされる配列は、同一なタグに起因して、同じ元の配列に由来するものとして識別することができる。こうして、例えば、十分な相同性を有する配列の部分配列は、染色体Ａへとマッピングされ、不十分な相同性を有する配列の部分配列は、切り詰められる。同様に、マッピングソフトウェアは、配列を、染色体Ｂへとマッピングし、相同性が不十分である場合、配列を切り詰める。

【００４１】

しかし、シーケンシングにおけるエラー、およびマッピングアルゴリズムの特徴を含む、いくつかの因子の結果として、融合遺伝子の切断点を含むＤＮＡ断片が、基準染色体の各々における切断点遺伝子座へと、正確にマッピングされないことがある。例えば、マッピングソフトウェアは、配列の切断点を、実際の切断点のやや上流に識別する場合もあり、やや下流に識別する場合もある。これらのエラーは、例えば、正確な遺伝子融合情報に依存する、がんの診断法の精度に影響を及ぼしうる。

【００４２】

いくつかの仮説的なマッピングエラーを、図７に示す。融合リード１は、適正にマッピングされ、切断点は、基準染色体内で、切断点Ａ１および切断点Ｂ１（第１の切断点および第２の切断点）として表示される。この融合リードは、切断点対Ａ１－Ｂ１を有する。融合リード２は、不適正にマッピングされ、染色体Ａについての切断点は上流、切断点Ａ２（第１の切断点）にあると決定されている。しかし、染色体Ｂ内の切断点は、切断点Ｂ

１（第２の切断点）に正しくマッピングされている。この融合リードは、切断点対Ａ２－Ｂ１を有する。融合リード３もまた、不適正にマッピングされ、染色体Ａについての切断点は、切断点Ａ１（第１の切断点）に正しくマッピングされているが、染色体Ｂについての切断点は下流、切断点Ｂ２（第２の切断点）にあると決定されている。この融合リードは、切断点対Ａ１－Ｂ２を有する。このような状況下では、ソフトウェアは、融合遺伝子について、いくつかの切断点を識別している。

【００４３】

本開示の方法に従い、融合遺伝子内の切断点をコールするために、マッピング配列を、共通の切断点対に基づき、セットへと群分けし、次いで、基準ゲノム内で、所定の塩基距離内にある切断点に基づき、クラスターへと群分けする。

【００４４】

このような方法について、図８に記載する。融合リード１、２、３、４、５および６の配列は、基準ゲノム染色体ＡおよびＢへとマッピングされる。融合の両側における切断点が、配列は、セットへと群分けされる。例では、融合リード１と、融合リード４とは、切断点対Ａ１およびＢ１を共有し、セットＩへと群分けされる。融合リード２と、融合リード５とは、切断点対Ａ２およびＢ１を共有し、セットＩＩへと群分けされる。融合リード３と、融合リード６とは、切断点対Ａ１およびＢ２を共有し、セットＩＩＩへと群分けされる。

【００４５】

この例では、切断点Ａ１と、切断点Ａ２とは、所定の遺伝子距離Ａ（例えば、１０塩基）内にあり、切断点Ｂ１と、切断点Ｂ２とは、所定の遺伝子距離Ｂ内にある。したがって、セットＩ、ＩＩ、およびＩＩＩは、クラスターへと群分けされる。

【００４６】

融合遺伝子切断点は、使用者により選択される選択基準を使用してコールされる。一部の実施形態では、基準は、クラスター内に、１つを超える分子を有すること、および／またはワトソン－クリック鎖の両方を伴う、少なくとも１つの分子を有することを含む。一方法では、全ての切断点の間で、最も多くの関連する融合リード（the most associated fused reads）を有する切断点を、融合遺伝子切断点としてコールする投票法により、各染色体内の切断点を決定する。他の方法では、品質アルゴリズムを使用して、異なる配列の切断点に重みづけすることができる。図８の例では、染色体Ａでは、切断点Ａ１が、４つの融合リードと関連するのに対し、切断点Ａ２は、２つの融合リードと関連する。したがって、第１の遺伝子融合切断点は、Ａ１にあるとコールされる。染色体Ｂでは、切断点Ｂ１が、４つの融合リードと関連するのに対し、切断点Ｂ２は、２つの融合リードと関連する。したがって、第２の遺伝子融合切断点は、Ｂ１にあるとコールされる。

【００４７】

別の例示的な方法を、図９Ａ～９Ｃに示す。ハイスループットシーケンサーなどのＤＮＡシーケンシングシステムを使用して、ＤＮＡ分子をシーケンシングする。配列を解析して、コレクション内の、元の分子コンセンサス配列を生成することができる。生成された配列のコレクションに、固有の識別子でタグ付けする（この場合、１～７）。配列は、基準ゲノムへとマッピングされる。この例では、配列は各々、基準ゲノム内の、２つの異なる場所へとマッピングされる。マッピングされた部分を、バーとして描示するのに対し、切り詰められた部分は、破線として描示する。全ての配列の切断点を識別する。この例では、染色体Ａ上の切断点は、Ａ１、Ａ２、およびＡ３である。染色体Ｂ上の切断点は、Ｂ１、Ｂ２、およびＢ３である。マッピングされたリードを、共通の切断点に基づき、セットへと組織化する。切断点対を、各染色体上で、同じ識別子および同じ切断点を有する配列の対として決定する。染色体上の、クラスター切断点への所定の距離を決定する。この例では、クラスターは、切断点Ａ１およびＡ２、ならびにＢ１およびＢ２を含む。切断点Ａ３およびＢ３は、所定の距離外にあり、したがって、クラスター内に含まれない。選択された基準に基づき、元の分子内の切断点対がコールされる。この例では、基準は、投票に基づく。したがって、切断点Ａ１およびＢ１は、クラスター内の大部分の分子を有する

ことに基づき、切断点対としてコールされる。

【００４８】

システムは、一般にスプライシングにおいて見られる短い配列についての、複数のアライメントを取り扱いうる。それらのアライメントを確認するために、融合点の周りでは、より長い配列を得ることができる。偽陽性率を低減するために、リード数フィルター、配列類似性フィルター、リード位置分布フィルターを含む、一連のフィルターを使用して、特異性の大きな結果をもたらすことができる。加えて、キメラ転写物の存在度を推定するのに、発現推定ツールである、ＲＳＥＭ（ＲＮＡ－Ｓｅｑ　ｂｙ　Ｅｘｐｅｃｔａｔｉｏｎ　Ｍａｘｉｍｉｚａｔｉｏｎ）を使用してＥＭ（Ｅｘｐｅｃｔａｔｉｏｎ　Ｍａｘｉｍｉｚａｔｉｏｎ）アルゴリズムを、まばら度最適化と共に適用することができる。さらに、この存在度の定量化により、識別の精度を増大させることもできる。まとめると、これらの特色は、遺伝子融合イベントについて、より完全に検討するシステムを可能とする。

【００４９】

システムは、複数の配列リードを、１または複数の試料から、任意の適するファイルフォーマットで得ることと、重複する配列リードのセットを識別することと、重複する配列リードの各セットについて、１つのリードだけを保存することとを包含しうる。適切なファイルフォーマットは、ＦＡＳＴＡファイルフォーマットおよびＦＡＳＴＱファイルフォーマットを含む。ＦＡＳＴＡと、ＦＡＳＴＱとは、ハイスループットシーケンシングからの生配列リードを保存するのに使用される、共通のファイルフォーマットである。ＦＡＳＴＱファイルは、各配列リードのための識別子、配列、および各リードの品質スコアストリングを保存する。ＦＡＳＴＡファイルは、識別子および配列だけを保存する。これらの２つのファイルフォーマットは、多くの共通のシーケンシングアライメントアルゴリズムおよびアセンブリアルゴリズムへの入力である。本発明は、ＦＡＳＴＱファイルおよびＦＡＳＴＡファイルのためのリード配列情報が、試料内および試料間で、高度に冗長であるかまたは重複する傾向にあることを認識する。このことは、配列リードの多くが、同じ配列からなることを意味する。本発明の方法は、この冗長性を利用して、ファイルサイズの、何分の１もの低減を達成し、保存されたデータの読出しは無損失である。例えば、本発明は、試料と関連するＦＡＳＴＡ／ＦＡＳＴＱファイルを読み取り、マスターのリード配列ファイル内の固有のリード配列だけを保存するのに使用されうる。

【００５０】

システムは、識別された固有の配列と同じ配列を有する各リードについてのリード識別子などのメタ情報を収集することもさらに包含する。次いで、このメタ情報を、メタ情報を元のＦＡＳＴＡ／ＦＡＳＴＱファイル内で識別された固有の配列リードと相関させる、試料についてのファイルへと書き込み、この時点で、マスターリード配列ファイル内に保存することができる。この新たなファイルは、元のファイル内で見出される重複情報を含有しないため、元のファイルより小さく、転送が容易である。さらに、圧縮ファイルは、任意の実際の配列データを含有する必要がまったくない。ある特定の態様では、圧縮ファイルは、マスターファイル内に保存された固有の配列へとインデックスづけされた、配列リードについての識別子だけを含有しうる。

【００５１】

配列データは、複数の配列リードを得ること（不揮発性メモリに結合されたプロセッサーを含むコンピュータシステムを使用して）により、圧縮することができる。各配列リードは、配列ストリングのほか、メタ情報も含みうる。配列リードは、例えば、記載行（「＞」記号を先行させた）と、任意選択で、ＦＡＳＴＱの場合には、品質スコアとを含むメタ情報を伴う、１または複数のＦＡＳＴＡファイルまたはＦＡＳＴＱファイルのフォーマットで提示することができる。配列ストリングは、例えば、ＩＵＰＡＣヌクレオチドコードを使用する、ヌクレオチド配列データを表すことが好ましい。固有のエントリーだけを含有する配列ストリングのサブセットが識別される。次いで、本発明のシステムおよび方法を使用して、識別されたサブセットと、その配列リードについての（複数の配列リードの各々についての）メタ情報とを、その配列リードを表すサブセット内の固有のエントリ

ーについての指標と共に含む出力を書き込む。

【００５２】

一部の実施形態では、サブセット（すなわち、固有の配列リードだけを含有する）を、テキストファイルでありうる、マスターリードファイルへと書き込む。ファイルが、人間に読取り可能であり、さらなる処理（例えば、ＰｅｒｌまたはＰｙｔｈｏｎなどのスクリプト言語を使用して）を、容易に実施しうるように、ＩＵＰＡＣヌクレオチドコードを使用して、固有の配列リードを、マスターリードファイル内に表示することが好ましい。メタ情報は、入力のＦＡＳＴＡファイルまたはＦＡＳＴＱファイルに対応する、圧縮出力ファイルへと書き込むことができる。

【００５３】

方法は、元の入力を、出力だけから再構成するステップを含むことが可能であり、ある特定の実施形態では、読出しは、無損失であり、なお、完全に無損失でもある。すなわち、複数の配列リードを含む、新たなＦＡＳＴＡファイルまたはＦＡＳＴＱファイルを創出するように、出力を処理することができる。読出しが無損失である場合、新たなＦＡＳＴＡファイルまたはＦＡＳＴＱファイルは、ＦＡＳＴＡファイルまたはＦＡＳＴＱファイルと同じ情報を含有する。

【００５４】

本発明は、任意の適切な種類のデータファイルに適合する。前述のＦＡＳＴＡファイルおよびＦＡＳＴＱファイルに加えて、配列リードはまた、ＶＣＦ（Ｖａｒｉａｎｔ　Ｃａｌｌ　Ｆｏｒｍａｔ）ファイルでも捕捉することができる。ハイスループットシーケンシングの進歩により、複数のシーケンシング施設が、ヒトゲノム内の変異体を検出し、それらを、これらのＶＣＦファイルを介して報告することが一般的である。本発明は、異なる情報源に由来する変異体情報を、研究者が、複雑な、対立遺伝子レベル、試料レベル、および集団レベルの検索を、施設を越えて実施することを可能とする形で、ＶＣＦファイル内に保存する、統一データベースの開発を容易としうる。統一データベースは、１つの汎用の対立遺伝子表上に、あらゆる固有の対立遺伝子（例えば、固有の配列リード）を保存し、これらの固有の対立遺伝子の、関連する試料および試料レベルのメタデータへの照会を保存することにより、変異体情報を、異なる試料に由来するＶＣＦファイルに統合しうる。

【００５５】

システムの実装は、配列データを圧縮するための方法を含みうる。方法は、複数の配列リードを得るステップ（不揮発性メモリに結合されたプロセッサーを含むコンピュータシステムを使用して）であって、各配列リードが、配列ストリングおよびメタ情報を含む、ステップと；固有のエントリーだけを含有する配列ストリングのサブセットを識別するステップと；サブセットと、その配列リードについての（複数の配列リードの各々についての）メタ情報とを、その配列リードを表すサブセット内の固有のエントリーについての指標と共に含む出力を書き込むステップとを含む。出力は、ＩＵＰＡＣヌクレオチドコードを使用して、サブセットを保存する、１または複数のテキストファイルを含むことが好ましい。出力は、プレーンテキストとして保存することが好ましい（例えば。これはテキストエディタープログラムを使用して開くことができ、人間がスクリーン上で読み取ることができる）。好ましい実施形態では、配列リードデータは、損失を伴わずに保存される。方法は、複数の配列リードを含む、新たなＦＡＳＴＡファイルまたはＦＡＳＴＱファイルを創出するように、出力を処理するステップを含みうる。複数の配列リードは、ＦＡＳＴＡファイルまたはＦＡＳＴＱファイルとして得ることができ、新たなＦＡＳＴＡファイルまたはＦＡＳＴＱファイルは、ＦＡＳＴＡファイルまたはＦＡＳＴＱファイルと同じ情報を含有しうる。一部の実施形態では、出力が占有するディスクスペースは、得られた複数の配列リードを保存するのに要求されるディスクスペースの％未満である。

【００５６】

ここで、本発明を実施するために有用な、試料を得るための方法、シーケンシングリードを生成するための方法、および多様な種類のシーケンシングについて記載する。これら

の例示的な方法は、限定的なものではなく、当業者の必要に応じて改変しうることを理解されたい。

【００５７】

複数の配列リードを得るステップは、試料に由来する核酸をシーケンシングして、配列リードを生成することを含みうる。下記で詳細に説明する通り、複数の配列リードを得るステップはまた、シーケンシングデータを、シーケンサーから受信することも含みうる。試料中の核酸は、例えば、組織試料中のゲノムＤＮＡ、検査室試料中の特定の標的から増幅されたｃＤＮＡ、または複数の生物から混合されたＤＮＡを含む、任意の核酸でありうる。一部の実施形態では、試料は、一倍体生物または二倍体生物に由来する、ホモ接合性のＤＮＡを含む。例えば、試料は、希少な劣性対立遺伝子についてホモ接合性の患者に由来するゲノムＤＮＡを含みうる。他の実施形態では、試料は、２つの類縁の核酸が、５０または１００％以外の対立遺伝子頻度、すなわち、２０％、５％、１％、０．１％、または他の任意の対立遺伝子頻度で存在するように、体細胞突然変異を伴う、二倍体生物または倍数性生物に由来する、ヘテロ接合性の遺伝子素材を含む。

【００５８】

一実施形態では、核酸鋳型分子（例えば、ＤＮＡまたはＲＮＡ）を、タンパク質、脂質、および非鋳型核酸など、他の様々な構成要素を含有する生物学的試料から単離する。核酸鋳型分子は、動物、植物、細菌、真菌、または他の任意の細胞性生物から得られる、任意の細胞素材から得ることができる。本発明における使用のための生物学的試料はまた、ウイルス粒子またはウイルス調製物も含む。核酸鋳型分子は、生物から直接得ることもでき、生物から得られる生物学的試料、例えば、血液、血清、血漿、尿、脳脊髄液、唾液、糞便、リンパ液、滑液、嚢胞液、腹水、胸水、羊水、絨毛膜絨毛試料、着床前胚に由来する流体、胎盤試料、子宮頸部／膣洗浄液および子宮頸部／膣液、間質液、口腔スワブ試料、痰、気管支洗浄液、パップスメア試料、または眼液から得ることもできる。任意の組織検体または体液検体（例えば、ヒト組織検体またはヒト体液検体）を、本発明において使用するための核酸の供給源として使用することができる。核酸鋳型分子はまた、初代細胞培養物または細胞株などの培養細胞からも単離することができる。鋳型核酸が得られる細胞または組織には、ウイルスまたは他の細胞内病原体を感染させることができる。試料はまた、生物学的検体から抽出された全ＲＮＡ、ｃＤＮＡライブラリー、ウイルスＤＮＡ、またはゲノムＤＮＡでもありうる。試料はまた、非細胞由来のＤＮＡからも単離することができる。

【００５９】

生物学的試料から得られた核酸は、解析に適する断片を作製するように断片化することができる。鋳型核酸は、様々な、機械的方法、化学的方法、および／または酵素的方法を使用して、所望の長さへと、断片化またはせん断処理することができる。ＤＮＡは、例えば、Ｃｏｖａｒｉｓ（Ｗｏｂｕｒｎ、Ｍａｓｓ．）により販売されている超音波処理機を使用する超音波処理を介してランダムにせん断処理することもでき、短時間にわたる、ＤＮアーゼへの曝露により断片化することもでき、１もしくは複数の制限酵素の混合物、またはトランスポザーゼもしくはニッキング酵素を使用して断片化することもできる。ＲＮＡは、短時間にわたる、ＲＮアーゼへの曝露により断片化することもでき、熱に加えた磁気により断片化することもでき、せん断処理によりにより断片化することもできる。ＲＮＡは、ｃＤＮＡへと転換することができる。断片化を利用する場合、断片化の前または後で、ＲＮＡを、ｃＤＮＡへと転換することができる。一実施形態では、核酸を、超音波処理により断片化する。別の実施形態では、核酸を、流体せん断装置により断片化する。一般に、個々の核酸鋳型分子は、約２ｋｂ～約４０ｋｂでありうる。特定の実施形態では、核酸は、約６ｋｂ～１０ｋｂの断片である。核酸分子は、一本鎖の場合もあり、二本鎖の場合もあり、一本鎖領域を伴う二本鎖（例えば、ステムループ構造）の場合もある。

【００６０】

生物学的試料は、必要に応じて、界面活性剤または表面活性剤の存在下で、溶解させることもでき、ホモジナイズすることもでき、分画することもできる。適切な界面活性剤は

、イオン性界面活性剤（例えば、ドデシル硫酸ナトリウムまたはＮ－ラウロイルサルコシン）を含む場合もあり、非イオン性界面活性剤を含む場合もある。核酸は、試料から抽出または単離したら、増幅することができる。

【００６１】

増幅とは、核酸配列のさらなるコピーの作製を指し、一般に、ポリメラーゼ連鎖反応（ＰＣＲ）または当技術分野で公知の他の技術を使用して実行する。増幅反応は、ＰＣＲなど、核酸分子を増幅する、当技術分野で公知の、任意の増幅反応でありうる。他の増幅反応は、ネステッドＰＣＲ、ＰＣＲ－一本鎖コンフォメーション多型、リガーゼ連鎖反応、鎖置換増幅、および制限断片長多型、転写ベースの増幅システム、ローリングサークル増幅、および超分枝型ローリングサークル増幅、定量的ＰＣＲ、定量的蛍光ＰＣＲ（ＱＦ－ＰＣＲ）、マルチプレックス蛍光ＰＣＲ（ＭＦ－ＰＣＲ）、リアルタイムＰＣＲ（ＲＴＰＣＲ）、制限断片長多型ＰＣＲ（ＰＣＲ－ＲＦＬＰ）、ｉｎ　ｓｉｔｕローリングサークル増幅（ＲＣＡ）、ブリッジＰＣＲ、ピコ滴定ＰＣＲ、エマルジョンＰＣＲ、転写増幅、自己持続配列複製、コンセンサス配列プライミングＰＣＲ、任意プライミングＰＣＲ、オリゴヌクレオチドプライミングＰＣＲ、および核酸ベースの配列増幅（ＮＡＢＳＡ）を含む。使用しうる増幅法は、米国特許第５，２４２，７９４号；同第５，４９４，８１０号；同第４，９８８，６１７号；および同第６，５８２，９３８号において記載されている増幅法を含む。ある特定の実施形態では、増幅反応は、例えば、参照により本明細書に組み込まれる、米国特許第４，６８３，１９５号；および米国特許第４，６８３，２０２号において記載されているＰＣＲである。ＰＣＲ、シーケンシング、および他の方法のためのプライマーは、クローニング、直接化学合成、および当技術分野で公知の他の方法により調製することができる。プライマーはまた、Ｅｕｒｏｆｉｎｓ　ＭＷＧ　Ｏｐｅｒｏｎ（Ｈｕｎｔｓｖｉｌｌｅ、Ａｌａ．）またはＬｉｆｅ　Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｃａｒｌｓｂａｄ、Ｃａｌｉｆ．）などの販売元から得ることもできる。

【００６２】

増幅アダプターを、断片化核酸へと接合させることができる。アダプターは、Ｉｎｔｅｇｒａｔｅｄ　ＤＮＡ　Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｃｏｒａｌｖｉｌｌｅ、Ｉｏｗａ）などから、市販品を購入することができる。ある特定の実施形態では、アダプター配列は、酵素により、鋳型核酸分子へと接合させる。酵素は、リガーゼまたはポリメラーゼでありうる。リガーゼは、オリゴヌクレオチド（ＲＮＡまたはＤＮＡ）を、鋳型核酸分子へとライゲーションすることが可能な、任意の酵素でありうる。適切なリガーゼは、Ｔ４　ＤＮＡリガーゼおよびＴ４　ＲＮＡリガーゼを含み、Ｎｅｗ　Ｅｎｇｌａｎｄ　Ｂｉｏｌａｂｓ（Ｉｐｓｗｉｃｈ、Ｍａｓｓ．）から市販されている。当技術分野では、リガーゼを使用するための方法が周知である。ポリメラーゼは、ヌクレオチドを、鋳型核酸分子の３’末端および５’末端へと付加することが可能な、任意の酵素でありうる。

【００６３】

ライゲーションは、平滑末端ライゲーションの場合もあり、相補的な突出末端を用いるライゲーションの場合もある。ある特定の実施形態では、断片の末端は、平滑末端を形成するように、断片化に続き、修復することもでき、トリミングすることもでき（例えば、エクソヌクレアーゼを使用して）、充填することもできる（例えば、ポリメラーゼおよびｄＮＴＰを使用して）。一部の実施形態では、Ｅｐｉｃｅｎｔｒｅ　Ｂｉｏｔｅｃｈｎｏｌｏｇｉｅｓ（Ｍａｄｉｓｏｎ、Ｗｉｓ．）から市販されているキットなど、市販のキットを使用して、平滑末端の５’リン酸化核酸末端を作出するように、末端修復を実施する。平滑末端を作出したら、断片の３’末端および５’末端への、鋳型非依存的な付加を形成し、これにより、単一のＡ突出を作製するように、末端を、ポリメラーゼおよびｄＡＴＰで処理することができる。Ｔ－Ａクローニングと称する方法では、この単一のＡを使用して、断片の、５’末端からの単一のＴ突出とのライゲーションを導く。代替的に、制限消化の後で、制限酵素により残される、突出の可能な組合せは公知であるため、末端は、そのまま放置する、すなわち、粘着末端とすることもできる。ある特定の実施形態では、相補的な突出末端を伴う、二本鎖オリゴヌクレオチドを使用する。

【００６４】
本発明の実施形態は、バーコード配列を、鋳型核酸へと接合させることを伴う。ある特定の実施形態では、バーコードを、各断片へと接合させる。他の実施形態では、複数のバーコード、例えば、２つのバーコードを、各断片へと接合させる。バーコード配列は一般に、配列を、シーケンシング反応において有用とする、ある特定の特色を含む。例えば、バーコード配列は、バーコード配列内に、最小限のホモポリマー領域（すなわち、ＡＡまたはＣＣＣなど、連続で２つまたはこれを超える同じ塩基）を有するか、またはホモポリマー領域を有さないようにデザインされる。バーコード配列はまた、１塩基ずつのシーケンシングを実施する場合に、それらが、塩基付加順序（ｂａｓｅ　ａｄｄｉｔｉｏｎ　ｏｒｄｅｒ）から、少なくとも１編集距離隔たっているようにデザインして、最初の塩基および最後の塩基が、予測される配列の塩基とマッチしないことを確実にする。
【００６５】
バーコード配列は、各配列が核酸の特定の部分と相関するようにデザインされ、配列リードが、それらが由来した部分へと戻る形でそれと相関させることが可能となる。ある特定の実施形態では、バーコード配列は、約５ヌクレオチド～約１５ヌクレオチドの範囲である。特定の実施形態では、バーコード配列は、約４ヌクレオチド～約７ヌクレオチドの範囲である。バーコード配列は、鋳型核酸に沿ってシーケンシングされるので、接合させた鋳型核酸に由来する最長のリードを許容するように、オリゴヌクレオチドの長さは、最小限の長さであるものとする。例えば、複数のＤＮＡバーコードは、多様な数のヌクレオチド配列を含みうる。ある特定の実施形態では、バーコード配列は、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、またはこれを超えるヌクレオチドを含む。ポリヌクレオチドの一方だけの末端へと接合させる場合、複数のＤＮＡバーコードは、２つ、３つ、４つ、５つ、６つ、７つ、８つ、９つ、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、２１、２２、２３、２４、２５、２６、２７、２８、２９、３０、またはこれを超える異なる識別子をもたらしうる。代替的に、ポリヌクレオチドの両方の末端へと接合させる場合、複数のＤＮＡバーコードは、４つ、９つ、１６、２５、３６、４９、６４、８１、１００、１２１、１４４、１６９、１９６、２２５、２５６、２８９、３２４、３６１、４００、またはこれを超える異なる識別子（ＤＮＡバーコードを、ポリヌクレオチドの一方だけの末端へと接合させる場合の２乗個の識別子である）をもたらしうる。
【００６６】
一般に、バーコード配列は、鋳型核酸分子から、少なくとも１塩基隔てる（ホモポリマー的な組合せを最小化する）。ある特定の実施形態では、バーコード配列を、鋳型核酸分子へと、例えば、酵素により接合させる。酵素は、下記で論じられる通り、リガーゼまたはポリメラーゼでありうる。
【００６７】
増幅アダプターもしくはシーケンシングアダプター、またはバーコード、あるいはこれらの組合せを、断片化核酸へと接合させることができる。このような分子は、Ｉｎｔｅｇｒａｔｅｄ　ＤＮＡ　Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｃｏｒａｌｖｉｌｌｅ、Ｉｏｗａ）などから、市販品を購入することができる。ある特定の実施形態では、このような配列を、リガーゼなどの酵素により、鋳型核酸分子へと接合させる。適切なリガーゼは、Ｎｅｗ　Ｅｎｇｌａｎｄ　Ｂｉｏｌａｂｓ（Ｉｐｓｗｉｃｈ、Ｍａｓｓ．）から市販されている、Ｔ４　ＤＮＡリガーゼおよびＴ４　ＲＮＡリガーゼを含む。ライゲーションは、平滑末端ライゲーションの場合もあり、相補的な突出末端の使用を介するイゲーションの場合もある。ある特定の実施形態では、断片の末端は、平滑末端を形成するように、断片化に続き、修復することもでき、トリミングすることもでき（例えば、エクソヌクレアーゼを使用して）、充填することもできる（例えば、ポリメラーゼおよびｄＮＴＰを使用して）。一部の実施形態では、Ｅｐｉｃｅｎｔｒｅ　Ｂｉｏｔｅｃｈｎｏｌｏｇｉｅｓ（Ｍａｄｉｓｏｎ、Ｗｉｓ．）から市販されているキットなど、市販のキットを使用して、平滑末端の

５’リン酸化核酸末端を作出するように、末端修復を実施する。平滑末端を作出したら、断片の３’末端および５’末端への、鋳型非依存的な付加を形成し、これにより、単一のＡ突出を作製するように、末端を、ポリメラーゼおよびｄＡＴＰで処理することができる。Ｔ－Ａクローニングと称する方法では、この単一のＡにより、断片の、５’末端からの単一のＴ突出とのライゲーションを導くことができる。代替的に、制限消化の後で、制限酵素により残される、突出の可能な組合せは公知であるため、末端は、そのまま放置する、すなわち、粘着末端とすることもできる。ある特定の実施形態では、相補的な突出末端を伴う、二本鎖オリゴヌクレオチドを使用する。

【００６８】

任意の処理するステップ（例えば、得るステップ、単離するステップ、断片化するステップ、増幅ステップ、またはバーコード処理するステップ）の後、核酸をシーケンシングすることができる。

【００６９】

シーケンシングは、当技術分野で公知の任意の方法によるシーケンシングでありうる。ＤＮＡシーケンシング技法は、標識ターミネーターまたはプライマーと、スラブまたはキャピラリーによるゲル分離とを使用する、古典的なジデオキシシーケンシング反応（サンガー法）、可逆的に終結させた標識ヌクレオチドを使用する合成によるシーケンシング、パイロシーケンシング、４５４シーケンシング、Ｉｌｌｕｍｉｎａ／Ｓｏｌｅｘａシーケンシング、標識オリゴヌクレオチドプローブのライブラリーへの、対立遺伝子特異的ハイブリダイゼーション、ライゲーションに続く、標識クローンのライブラリーへの、対立遺伝子特異的ハイブリダイゼーションを使用する合成によるシーケンシング、重合化ステップにおける、標識ヌクレオチドの組込みの、リアルタイムのモニタリング、ポロニーシーケンシング、ＳＯＬｉＤシーケンシング、標的化されたシーケンシング、一分子リアルタイムシーケンシング、エクソンシーケンシング、電子顕微鏡ベースのシーケンシング、パネルシーケンシング、トランジスター媒介型シーケンシング、直接シーケンシング、ランダムショットガンシーケンシング、全ゲノムシーケンシング、ハイブリダイゼーションによるシーケンシング、キャピラリー電気泳動、ゲル電気泳動、デュプレックスシーケンシング、サイクルシーケンシング、一塩基伸長シーケンシング、固相シーケンシング、ハイスループットシーケンシング、超並列署名シーケンシング、エマルジョンＰＣＲ、低変性温度における共増幅ＰＣＲ（ＣＯＬＤ－ＰＣＲ）、マルチプレックスＰＣＲ、可逆的色素ターミネーターによるシーケンシング、ペアドエンドシーケンシング、短期シーケンシング、エクソヌクレアーゼシーケンシング、ライゲーションによるシーケンシング、短リードシーケンシング、一分子シーケンシング、リアルタイムシーケンシング、リバースターミネーターシーケンシング、ナノ小孔シーケンシング、ＭＳ－ＰＥＴシーケンシング、およびこれらの組合せを含む。一部の実施形態では、シーケンシング法は、超並列シーケンシング、すなわち、少なくとも１００、１０００、１０，０００、１００，０００、１００万、１０００万、１億、または１０億のポリヌクレオチド分子のうちのいずれかを、同時に（または間断なく）シーケンシングすることである。一部の実施形態では、シーケンシングは、例えば、ＩｌｌｕｍｉｎａまたはＡｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓから市販されている遺伝子解析器などの遺伝子解析器により実施することができる。より近年では、ポリメラーゼまたはリガーゼを使用する逐次的伸長反応または単一の伸長反応のほか、プローブライブラリーとの、単一の示差的ハイブリダイゼーションまたは逐次的な示差的ハイブリダイゼーションによっても、個別分子のシーケンシングが裏付けられている。シーケンシングは、ＤＮＡシーケンサー（例えば、シーケンシング反応を実施するようにデザインされたマシン）により実施することができる。

【００７０】

使用しうるシーケンシング技法は、例えば、合成システムによるシーケンシングの使用を含む。第１のステップでは、ＤＮＡを、約３００～８００塩基対の断片へとせん断処理し、断片を平滑末端処理する。次いで、オリゴヌクレオチドアダプターを、断片の末端へとライゲーションする。アダプターは、断片の増幅およびシーケンシングのためのプライ

マーとして用いられる。断片を、ＤＮＡ捕捉ビーズ、例えば、５’ビオチンタグを含有する、アダプターＢを使用する、例えば、ストレプトアビジンコーティングビーズへと接合させることができる。ビーズへと接合させた断片を、油－水エマルジョンの液滴内でＰＣＲ増幅する。結果は、各ビーズ上でクローン増幅されたＤＮＡ断片の複数のコピーである。第２のステップでは、ビーズを、ウェル（ピコリットルサイズの）内に捕捉する。各ＤＮＡ断片上で並行して、パイロシーケンシングを実施する。１または複数のヌクレオチドの付加は、光シグナルを発生させ、これは、シーケンシング装置内のＣＣＤカメラにより記録される。シグナルの強度は、組み込まれるヌクレオチドの数に比例する。パイロシーケンシングは、ヌクレオチド付加時に放出されるピロリン酸（ＰＰｉ）を使用する。ＰＰｉは、アデノシン５’ホスホ硫酸の存在下で、ＡＴＰスルフリラーゼにより、ＡＴＰへと転換される。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と転換し、この反応は、光を発生させ、これを検出および解析する。

【００７１】

使用しうるＤＮＡシーケンシング技法の別の例は、Ｌｉｆｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ Ｃｏｒｐｏｒａｔｉｏｎ（Ｃａｒｌｓｂａｄ、Ｃａｌｉｆ．）からのＡｐｐｌｉｅｄ ＢｉｏｓｙｓｔｅｍｓによるＳＯＬｉＤ技術である。ＳＯＬｉＤシーケンシングでは、ゲノムＤＮＡを、断片へとせん断処理し、アダプターを、断片の５’末端および３’末端へと接合させて、断片ライブラリーを生成する。代替的に、アダプターを、断片の５’末端および３’末端へとライゲーションし、断片を環状化させ、環状化させた断片を消化して、内部アダプターを作出し、アダプターを、結果として得られる断片の５’末端および３’末端へと接合させて、メイトペアドライブラリーを生成することにより、内部アダプターを導入することもできる。次に、クローンビーズ集団を、ビーズ、プライマー、鋳型、およびＰＣＲ構成要素を含有するマイクロリアクター内で調製する。ＰＣＲに続き、鋳型を変性させ、ビーズを濃縮して、鋳型を伸長させたビーズを分離する。選択されたビーズ上の鋳型を、３’修飾にかけ、これにより、スライドガラスへの結合を可能とする。配列は、部分的にランダムなオリゴヌクレオチドの、決定される中心塩基（または塩基対）であって、特異的フルオロフォアにより識別される塩基との、逐次的なハイブリダイゼーションおよびライゲーションにより決定することができる。色を記録した後で、ライゲーションされたオリゴヌクレオチドを除去し、次いで、処理を反復する。

【００７２】

使用しうるＤＮＡシーケンシング技法の別の例は、例えば、Ｌｉｆｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｓｏｕｔｈ Ｓａｎ Ｆｒａｎｃｉｓｃｏ、Ｃａｌｉｆ．）下のＩｏｎ Ｔｏｒｒｅｎｔにより、ＩＯＮ ＴＯＲＲＥＮＴの商品名で販売されているシステムを使用する、イオン半導体シーケンシングである。イオン半導体シーケンシングは、例えば、それらの各々の内容が、参照によりそれらの全体において組み込まれる、Rothbergら、An integrated semiconductor device enabling non-optical genome sequencing、Nature、４７５巻：３４８～３５２頁（２０１１年）；米国公開第２０１０／０３０４９８２号；米国公開第２０１０／０３０１３９８号；米国公開第２０１０／０３００８９５号；米国公開第２０１０／０３００５５９号；および米国公開第２００９／００２６０８２号において記載されている。

【００７３】

使用しうるシーケンシング技術の別の例は、Ｉｌｌｕｍｉｎａシーケンシングである。Ｉｌｌｕｍｉｎａシーケンシングは、フォールドバックＰＣＲと、アンカリングプライマーとを使用する、固体表面上のＤＮＡの増幅に基づく。ゲノムＤＮＡを断片化し、アダプターを、断片の５’末端および３’末端へと付加する。フローセルチャネルの表面へと接合させたＤＮＡ断片を伸長させ、ブリッジ増幅する。断片は二本鎖となり、二本鎖分子を変性させる。複数サイクルの固相増幅に続き、変性は、フローセルの各チャネル内の同じ鋳型の一本鎖ＤＮＡ分子の約１，０００コピーの何百万ものクラスターを創出しうる。プライマー、ＤＮＡポリメラーゼ、および４つのフルオロフォアで標識された可逆的終結ヌクレオチドを使用して、逐次的シーケンシングを実施する。ヌクレオチド組込みの後、レ

ーザーを使用して、フルオロフォアを励起し、画像を捕捉し、第１の塩基が何かを記録する。組み込まれた各塩基から、３’ターミネーターおよびフルオロフォアを除去し、組込みステップ、検出ステップ、および識別ステップを反復する。この技術に従いシーケンシングについては、それらの各々が、参照によりそれらの全体において組み込まれる、米国特許第７，９６０，１２０号；米国特許第７，８３５，８７１号；米国特許第７，２３２，６５６号；米国特許第７，５９８，０３５号；米国特許第６，９１１，３４５号；米国特許第６，８３３，２４６号；米国特許第６，８２８，１００号；米国特許第６，３０６，５９７号；米国特許第６，２１０，８９１号；米国公開第２０１１／０００９２７８号；米国公開第２００７／０１１４３６２号；米国公開第２００６／０２９２６１１号；および米国公開第２００６／００２４６８１号において記載されている。

【００７４】

使用しうるシーケンシング技術の別の例は、Ｐａｃｉｆｉｃ　Ｂｉｏｓｃｉｅｎｃｅｓ（Ｍｅｎｌｏ　Ｐａｒｋ、Ｃａｌｉｆ．）の一分子リアルタイム（ＳＭＲＴ）技術を含む。ＳＭＲＴでは、４つのＤＮＡ塩基の各々を、４つの異なる蛍光色素のうちの１つへと接合させる。これらの色素は、リン酸基に連結されている。単一のＤＮＡポリメラーゼを、鋳型である一本鎖ＤＮＡの単一の分子と共に、ゼロモードウェーブガイド（ＺＭＷ）の底部に固定化させる。ヌクレオチドを、成長する鎖へと組み込むには、数ミリ秒を要する。この時間中に、蛍光標識が励起され、蛍光シグナルをもたらし、蛍光タグが切り離される。色素の対応する蛍光の検出は、どの塩基が組み込まれたのかを指し示す。処理を反復する。

【００７５】

使用しうるシーケンシング技法の別の例は、ナノ小孔シーケンシング（SoniおよびMeller、２００７年、Progress toward ultrafast DNA sequence using solid-state nanopores、Clin Chem、５３巻（１１号）：１９９６～２００１頁）である。ナノ小孔とは、直径を１ナノメートルのオーダーとする小孔である。ナノ小孔を、導電性流体中に浸漬し、これにわたって電位をかける結果として、ナノ小孔を介するイオンの導電に起因して、わずかな電流がもたらされる。流れる電流の量は、ナノ小孔のサイズを感知する。ＤＮＡ分子がナノ小孔を通過するとき、ＤＮＡ分子上の各ヌクレオチドは、ナノ小孔を、異なる程度に閉塞させる。したがって、ＤＮＡ分子が、ナノ小孔を通過するときの、ナノ小孔を通過する電流の変化は、ＤＮＡ配列の読取りを表す。

【００７６】

使用しうるシーケンシング技法の別の例は、化学感受性電界効果トランジスター（ｃｈｅｍＦＥＴ）アレイを使用して、ＤＮＡをシーケンシングすること（例えば、米国公開第２００９／００２６０８２号において記載されている通りに）を伴う。技法の１つの例では、ＤＮＡ分子を、反応チャンバーに入れ、鋳型分子を、ポリメラーゼに結合させたシーケンシングプライマーとハイブリダイズさせることができる。１または複数の三リン酸の、シーケンシングプライマーの３’末端における、新たな核酸鎖への組込みは、ｃｈｅｍＦＥＴにより、電流の変化を介して検出することができる。アレイは、複数のｃｈｅｍＦＥＴセンサーを有しうる。別の例では、単一の核酸を、ビーズへと接合させ、核酸を、ビーズ上で増幅し、個々のビーズを、ｃｈｅｍＦＥＴアレイ上の、個々の反応チャンバーであって、各チャンバーがｃｈｅｍＦＥＴセンサーを有する、反応チャンバーへと移し、核酸をシーケンシングすることができる。

【００７７】

使用しうるシーケンシング技法の別の例は、例えば、Moudrianakis, E. N.およびBeer M.、Base sequence determination in nucleic acids with the electron microscope、III. Chemistry and microscopy of guanine-labeled DNA、PNAS、５３巻：５６４～７１頁（１９６５年）により記載されている通りに、電子顕微鏡を使用することを伴う。技法の１つの例では、個々のＤＮＡ分子を、電子顕微鏡を使用して識別可能な、金属標識を使用して標識する。次いで、これらの分子を、平面上に展開し、電子顕微鏡を使用してイメージングして、配列を定める。

【００７８】
本発明の実施形態に従うシーケンシングは、複数のリードを生成する。本発明に従うリードは一般に、約１５０塩基未満の長さ、または約９０塩基未満の長さのヌクレオチド配列のデータを含む。ある特定の実施形態では、リードは、約８０～約９０塩基の間、例えば、約８５塩基の長さである。一部の実施形態では、本発明の方法を、極めて短いリード、すなわち、約５０塩基または約３０塩基未満の長さのリードへと適用する。配列リードデータは、配列データのほか、メタ情報も含みうる。配列リードデータは、任意の適切なファイルフォーマットであって、例えば、当業者に公知の、ＶＣＦファイル、ＦＡＳＴＡファイル、またはＦＡＳＴＱファイルを含むフォーマットで保存することができる。

【００７９】
ＦＡＳＴＡとは元来、配列データベースを検索するためのコンピュータプログラムであり、ＦＡＳＴＡという名称はまた、標準的なファイルフォーマットも指す。PearsonおよびLipman、１９８８年、Improved tools for biological sequence comparison、PNAS、８５巻：２４４４～２４４８頁を参照されたい。ＦＡＳＴＡフォーマットの配列は、１行の記載で始まり、配列データ行が続く。記載行は、第１列の大なり（「＞」）記号により、配列データと区別される。「＞」記号に続く語は、配列の識別子であり、行の残りは、記載である（いずれも、任意選択である）。「＞」と、識別子の第１の文字との間には、スペースを置かないものとする。テキストの全ての行は、８０文字より短いことが推奨される。「＞」で始まる別の行が現れたら、配列は終了し、これは、別の配列の開始を指し示す。

【００８０】
ＦＡＳＴＱフォーマットは、生物学的配列（通例、ヌクレオチド配列）、およびその対応する品質スコアの両方を保存するための、テキストベースのフォーマットである。ＦＡＳＴＱフォーマットは、ＦＡＳＴＡフォーマットと類似するが、配列データに品質スコアを後続させている。配列文字および品質スコアのいずれも、簡略のために、単一のＡＳＣＩＩ記号でコードされている。ＦＡＳＴＱフォーマットは、Ｉｌｌｕｍｉｎａ Ｇｅｎｏｍｅ Ａｎａｌｙｚｅｒなど、ハイスループットシーケンシング装置の出力を保存するための、事実上の標準フォーマットである（Cock ら、２００９年、The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants、Nucleic Acids Res、３８巻（６号）：１７６７～１７７１頁）。

【００８１】
ＦＡＳＴＡファイルおよびＦＡＳＴＱファイルでは、メタ情報は、記載行は含むが、配列データ行は含まない。一部の実施形態では、ＦＡＳＴＱファイルでは、メタ情報は、品質スコアを含む。ＦＡＳＴＡファイルおよびＦＡＳＴＱファイルでは、配列データは、記載行の後で始まり、典型的に、任意選択で、「－」を伴うＩＵＰＡＣの両義性コードの何らかのサブセットを使用して提示される。好ましい実施形態では、配列データは、必要に応じて、任意選択で、「－」またはＵ（例えば、ギャップまたはウラシルを表す）を含む、Ａ、Ｔ、Ｃ、Ｇ、およびＮの記号を使用する。

【００８２】
一部の実施形態では、少なくとも１つのマスター配列リードファイルおよび出力ファイルを、プレーンテキストファイル（例えば、ＡＳＣＩＩ；ＩＳＯ／ＩＥＣ６４６；ＥＢＣＤＩＣ；ＵＴＦ－８；またはＵＴＦ－１６などのコード法を使用する）として保存する。本発明により提供されるコンピュータシステムは、プレーンテキストファイルを開くことが可能な、テキストエディタープログラムを含みうる。テキストエディタープログラムとは、テキストファイル（プレーンテキストファイルなど）のコンテンツを、コンピュータスクリーン上に提示し、人間が、テキストを編集する（例えば、モニター、キーボード、およびマウスを使用して）ことが可能なコンピュータプログラムを指す場合がある。例示的なテキストエディターは、限定なしに述べると、Ｍｉｃｒｏｓｏｆｔ Ｗｏｒｄ、ｅｍａｃｓ、ｐｉｃｏ、ｖｉ、ＢＢＥｄｉｔ、およびＴｅｘｔＷｒａｎｇｌｅｒを含む。好ましくは、テキストエディタープログラムは、プレーンテキストファイルを、コンピュータ

スクリーン上に表示し、メタ情報および配列リードを、人間に読取り可能なフォーマット（例えば、バイナリーコード化せず、印字による人間の書記において使用される英数字記号を使用する）で示すことが可能である。

【００８３】

ＦＡＳＴＡファイルまたはＦＡＳＴＱファイルを参照しながら、方法について論じてきたが、本発明の方法およびシステムは、例えば、ＶＣＦ（Ｖａｒｉａｎｔ　Ｃａｌｌ　Ｆｏｒｍａｔ）フォーマットによるファイルを含む、任意の適切な配列ファイルフォーマットを圧縮するのに使用することができる。典型的なＶＣＦファイルは、ヘッダーセクションおよびデータセクションを含むであろう。ヘッダーは、各々が「＃＃」記号で始まり、タブで区切られたフィールド定義行が、単一の「＃」記号で始まる、任意の数のメタ情報行を含有する。フィールド定義行は、８つの必須の列に名付け、本文セクションは、フィールド定義行により定義される列に投入される、データの行を含有する。ＶＣＦフォーマットについては、Danecekら、２０１１年、The variant call format and VCF tools、Bioinformatics、２７巻（１５号）：２１５６～２１５８頁において記載されている。ヘッダーセクションは、圧縮ファイルへと書き込むメタ情報として取り扱うことができ、データセクションは、固有である場合に限り、それらの各々がマスターファイル内に保存される、行として取り扱うことができる。

【００８４】

本発明のある特定の実施形態は、配列リードのアセンブリを提供する。アライメントによるアセンブリでは、例えば、リードを、互いに照らして、または基準に照らしてアライメントする。転じて、各リードを、基準ゲノムに照らしてアライメントすることにより、リードの全てを、互いとの関係で位置づけて、アセンブリを創出する。加えて、配列リードを、基準配列に照らしてアライメントするか、または基準配列へとマッピングすることはまた、配列リード内の変異体配列を識別するのにも使用することができる。変異体配列の識別を、本明細書で記載される方法およびシステムと組み合わせて使用して、さらに、疾患もしくは状態の診断もしくは予後診断の一助とするか、または処置の決定を導くこともできる。

【００８５】

一部の実施形態では、本発明のステップのうちの、いずれかまたは全てを自動化する。代替的に、本発明の方法は、全体的または部分的に、１または複数の専用プログラムであって、例えば、各々が、任意選択で、Ｃ＋＋などのコンパイラー言語で書き込まれ、次いで、バイナリー言語としてコンパイルおよび分散される、専用プログラムにより具体化することができる。本発明の方法は、全体的または部分的に、既存の配列解析プラットフォーム内のモジュールとして実装することもでき、既存の配列解析プラットフォーム内で機能を呼び出すことにより実装することもできる。ある特定の実施形態では、本発明の方法は、単一の開始キュー（例えば、人間の活動、別のコンピュータプログラム、またはマシンから供給される誘起イベントの１つまたは組合せ）に自動的に応答して呼び出される全てのステップである、多数のステップを含む。したがって、本発明は、ステップのうちのいずれか、またはステップの任意の組合せが、キューに自動的に応答して起動しうる方法を提供する。自動的にとは一般に、人間による入力、影響、または相互作用を介在しないこと（すなわち、元の人間活動またはキュー以前の人間活動だけに応答すること）を意味する。

【００８６】

システムはまた、対象核酸についての、正確かつ高感度の解釈を含む、多様な形態の出力も包含する。読出しの出力は、コンピュータファイルのフォーマットで提供することができる。ある特定の実施形態では、出力は、ＦＡＳＴＡファイル、ＦＡＳＴＱファイル、またはＶＣＦファイルである。出力は、基準ゲノムの配列に照らしてアライメントされた核酸配列などの配列データを含有する、テキストファイルまたはＸＭＬファイルを作製するように処理することができる。他の実施形態では、処理は、対象核酸における、基準ゲノムと比べた、１または複数の突然変異について記載する、座標またはストリングを含有

する出力をもたらす。当技術分野で公知のアライメントストリングは、ＳＵＧＡＲ（Ｓｉｍｐｌｅ ＵｎＧａｐｐｅｄ Ａｌｉｇｎｍｅｎｔ Ｒｅｐｏｒｔ）、ＶＵＬＧＡＲ（Ｖｅｒｂｏｓｅ Ｕｓｅｆｕｌ Ｌａｂｅｌｅｄ Ｇａｐｐｅｄ Ａｌｉｇｎｍｅｎｔ Ｒｅｐｏｒｔ）、およびＣＩＧＡＲ（Ｃｏｍｐａｃｔ Ｉｄｉｏｓｙｎｃｒａｔｉｃ Ｇａｐｐｅｄ Ａｌｉｇｎｍｅｎｔ Ｒｅｐｏｒｔ）（Ning, Z.ら、Genome Research、１１巻（１０号）：１７２５～９頁（２００１年））を含む。これらのストリングは、例えば、Ｅｕｒｏｐｅａｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｈｉｎｘｔｏｎ、ＵＫ）製の、Ｅｘｏｎｅｒａｔｅ配列アライメントソフトウェアに実装されている。

【００８７】

一部の実施形態では、ＣＩＧＡＲストリングを含む配列アライメント（例えば、ＳＡＭ（ｓｅｑｕｅｎｃｅ ａｌｉｇｎｍｅｎｔ ｍａｐ）ファイルまたはＢＡＭ（ｂｉｎａｒｙ ａｌｉｇｎｍｅｎｔ ｍａｐ）ファイルなど）を作製する（ＳＡＭフォーマットについては、例えば、Liら、The Sequence Alignment/Map format and SAM tools、Bioinformatics、２００９年、２５巻（１６号）：２０７８～９頁において記載されている）。一部の実施形態では、ＣＩＧＡＲは、１行当たり１つずつのギャップ処理アライメントを、表示するかまたは含む。ＣＩＧＡＲとは、ＣＩＧＡＲストリングとして報告される、対応のある、圧縮アライメントフォーマットである。ＣＩＧＡＲストリングは、対応のある、長い（例えば、ゲノム）アライメントを表すのに有用である。ＣＩＧＡＲストリングは、リードのアライメントを、基準ゲノム配列に照らして表すように、ＳＡＭフォーマットで使用される。

【００８８】

ＣＩＧＡＲストリングは、確立されたモチーフに従う。各記号は数を先行させ、イベントの塩基カウントをもたらす。使用される文字は、Ｍ、Ｉ、Ｄ、Ｎ、およびＳ（Ｍ＝マッチ；Ｉ＝挿入；Ｄ＝欠失；Ｎ＝ギャップ；Ｓ＝置換）を含みうる。ＣＩＧＡＲストリングは、マッチ／ミスマッチおよび欠失（またはギャップ）の配列を規定する。例えば、ＣＩＧＡＲストリングである、２ＭＤ３Ｍ２Ｄ２Ｍは、アライメントが、２つのマッチ、１つの欠失（数字１は、幾分のスペースを節約するために省略する）、３つのマッチ、２つの欠失、および２つのマッチを含有することを意味するであろう。

【００８９】

本発明により想定される通り、上記で記載した機能は、ソフトウェア、ハードウェア、ファームウェア、配線、またはこれらの任意の組合せを含む、本発明のシステムを使用して実装することができる。機能を実装するフィーチャはまた、機能のうちの一部を、異なる物理的な場所で実装するように、分散させることを含め、物理的に多様な位置に配置することもできる。

【００９０】

当業者が、本発明の方法を実施するのに必要であるか、または最も適すると認識する通り、本発明のコンピュータシステムまたはマシンは、バスを介して互いと連絡する、１または複数のプロセッサー（例えば、中央処理装置（ＣＰＵ）、グラフィックスプロセシングユニット（ＧＰＵ）、またはこれらの両方）、メインメモリ、およびスタティックメモリを含む。

【００９１】

図４は、本発明の方法を実施するのに適するシステム７０１について図示する。図７に示される通り、システム７０１は、サーバーコンピュータ７０５、ターミナル７１５、シーケンサー７１５、シーケンサーコンピュータ７２１、コンピュータ７４９、またはこれらの任意の組合せのうちの１または複数を含みうる。このような各コンピュータデバイスは、ネットワーク７０９を介して連絡しうる。シーケンサー７２５は、任意選択で、それ自身の、例えば、専用のシーケンサーコンピュータ７２１（任意の入力／出力機構（Ｉ／Ｏ）、プロセッサー、およびメモリを含む）を含むか、またはこれと作動可能に結合することができる。加えて、または代替的に、シーケンサー７２５は、ネットワーク７０９を

介して、サーバー７０５またはコンピュータ７４９（例えば、ラップトップ、デスクトップ、またはタブレット）と作動可能に結合することができる。コンピュータ７４９は、１または複数のプロセッサー、メモリ、およびＩ／Ｏを含む。本発明の方法が、クライアント／サーバー型アーキテクチャーを利用する場合、本発明の方法の任意のステップは、プロセッサー、メモリ、およびＩ／Ｏのうちの１または複数を含み、データ、命令などを得るか、またはインターフェースモジュールを介して、結果を提示するか、またはファイルとして結果を提示することが可能なサーバー７０５を使用して実施することができる。サーバー７０５は、コンピュータ７４９またはターミナル７１５を介して、ネットワーク７０９にわたり関与する場合もあり、またはサーバー７０５は、ターミナル７１５へと、直接接続することもできる。ターミナル５１５は、コンピュータデバイスであることが好ましい。本発明に従うコンピュータは、Ｉ／Ｏ機構およびメモリに結合された１または複数のプロセッサーを含むことが好ましい。

【００９２】

プロセッサーは、例えば、単一のコアまたはマルチコアプロセッサーのうちの１または複数を含む、１または複数のプロセッサーにより用意することができる。Ｉ／Ｏ機構は、ビデオディスプレイユニット（例えば、液晶ディスプレイ（ＬＣＤ）またはブラウン管（ＣＲＴ））、英数字入力デバイス（例えば、キーボード）、カーソル制御デバイス（例えば、マウス）、ディスクドライブユニット、信号発生デバイス（例えば、スピーカー）、加速度計、マイクロフォン、セル型ラジオ周波数アンテナ、およびネットワークインターフェースデバイス（例えば、ネットワークインターフェースカード（ＮＩＣ）、Ｗｉ－Ｆｉカード、セル型モデム、データジャック、イーサーネットポート、モデムジャック、ＨＤＭＩ（登録商標）ポート、ミニＨＤＭＩ（登録商標）ポート、ＵＳＢポート）、タッチスクリーン（例えば、ＣＲＴ、ＬＣＤ、ＬＥＤ、ＡＭＯＬＥＤ、Ｓｕｐｅｒ　ＡＭＯＬＥＤ）、ポインティングデバイス、トラックパッド、ライト（例えば、ＬＥＤ）、ライト／画像投影デバイス、またはこれらの組合せを含みうる。本発明に従うメモリは、本明細書で記載される方法または機能のうちの、任意の１または複数を果たす命令の、１または複数のセット（例えば、ソフトウェア）を保存する、１または複数のマシン読取り型メディアを含むことが好ましい、１または複数の有形デバイスにより用意される、不揮発性メモリを指す。ソフトウェアはまた、システム７０１内のコンピュータ、それらもまた、マシン読取り型メディアを構成する、メインメモリ、およびプロセッサーによるその実行時に、メインメモリ内、プロセッサー内、またはこれらの両方の中にも、完全に、または少なくとも部分的に存在しうる。ソフトウェアはさらに、ネットワークインターフェースデバイスを介して、ネットワークにわたり、送信または受信することもできる。

【００９３】

例示的な実施形態では、マシン読取り型メディアは、単一のメディアでありうるが、「マシン読取り型メディア」という用語は、１または複数の命令のセットを保存する、単一のメディアまたは複数のメディア（例えば、集中型もしくは分散型データベース、ならびに／または関連するキャッシュおよびサーバー）を含むように理解されるものとする。「マシン読取り型メディア」という用語はまた、マシンによる実行のための命令のセットを保存するか、コードするか、または保有することが可能であり、マシンに、本発明の方法のうちの、任意の１または複数を実施させる任意のメディアも含むように理解されるものとする。メモリは、例えば、ハードディスクドライブ、ソリッドステートドライブ（ＳＳＤ）、光ディスク、フラッシュメモリ、ジップディスク、テープドライブ、「クラウド」保存ロケーションのうちの１もしくは複数、またはこれらの組合せでありうる。ある特定の実施形態では、本発明のデバイスは、メモリのための、有形、不揮発性のコンピュータ読取り型メディアを含む。メモリとしての使用のための例示的なデバイスは、半導体メモリデバイス（例えば、ＥＰＲＯＭデバイス、ＥＥＰＲＯＭデバイス、ソリッドステートドライブ（ＳＳＤ）デバイス、およびフラッシュメモリデバイス、例えば、ＳＤカード、マイクロＳＤカード、ＳＤＸＣカード、ＳＤＩＯカード、ＳＤＨＣカード）；磁気ディスク（例えば、内部ハードディスクまたはリムーバブルディスク）；および光ディスク（例え

ば、ＣＤディスクおよびＤＶＤディスク）を含む。

【００９４】

一部の実施形態では、本開示の方法およびシステムを使用して、疾患または状態、例えば、がんを診断することができる。本明細書で使用される「診断」という用語は、患者が、所与の疾患または状態を患っているのか否かを、当業者が推定および／または決定する方法を指す。一部の実施形態では、本発明の方法を、疾患または状態、例えば、がんの予後診断において使用することができる。本明細書で使用される「予後診断」という用語は、疾患または状態の再発を含む、疾患または状態の進行の可能性を指す。一部の実施形態では、本発明の方法を使用して、疾患または状態、例えば、がんを発症する危険性を評価することができる。例えば、本明細書で記載される方法およびシステムを使用して、疾患または状態の発症についての、特定の診断、予後診断、または疾患もしくは状態を発症する危険性と関連する、切断点または遺伝子融合を識別することができる。さらに、本明細書で記載される方法およびシステムを使用して、予測される治療転帰と関連する、切断点または遺伝子融合を識別することができる。したがって、方法およびシステムを使用して、疾患または状態の処置を導く（例えば、化合物または薬剤を、対象へと投与することにより）こともでき、疾患または状態を処置するための医薬の調製を導くこともできる。

【００９５】

本明細書で使用される、疾患または状態を「処置すること」とは、臨床結果を含む、有益なまたは所望の結果を得るように、方策を講じることを指す。有益な臨床結果または所望の臨床結果は、疾患または状態と関連する、１または複数の症状の緩和または軽快を含むがこれらに限定されない。本明細書で使用される、化合物もしくは薬剤を、対象へと「投与すること」、または化合物もしくは薬剤の、対象への「投与」は、当業者に公知の様々な方法のうちの１つを使用して実行することができる。例えば、化合物または薬剤は、静脈内投与、動脈内投与、皮内投与、筋内投与、腹腔内投与、静脈内投与、皮下投与、眼内投与、舌下投与、経口（服用による）投与、鼻腔内（吸入による）投与、脊髄内投与、脳内投与、および経皮（例えば、皮膚導管を介する吸収による）投与することができる。化合物または薬剤はまた、充電式デバイスもしくは生体分解性ポリマーデバイス、または他のデバイス、例えば、化合物または薬剤の、持続放出、徐放、または制御放出をもたらす、パッチおよびポンプ、または製剤によっても適切に導入することができる。投与することはまた、例えば、単回で実施することもでき、複数回で実施することもでき、および／または１もしくは複数の期間にわたり実施することもできる。一部の態様では、投与は、自己投与を含む直接投与、および薬物を処方する行為を含む間接的投与の両方を含む。例えば、本明細書で使用される通り、患者が薬物を自己投与するか、もしくは別の医師が薬物を投与するように指示する医師、および／または患者に薬物についての処方を施す医師は、患者へと、薬物を投与している。一部の実施形態では、化合物または薬剤を、例えば、対象へと、服用により経口投与するか、または例えば、対象へと、注射により静脈内投与する。一部の実施形態では、経口投与される化合物または薬剤は、持続放出製剤もしくは徐放製剤であるか、またはこのような徐放もしくは持続放出のためのデバイスを使用して投与される。

【００９６】

本明細書で使用される「がん」という用語は、それらの大半が、周囲の組織に浸潤する可能性があり、異なる部位へと転移しうる、多様な種類の悪性新生物を含むがこれらに限定されない（例えば、全ての目的で、参照によりその全体において本明細書に組み込まれる、PDR Medical Dictionary、１版（１９９５年）を参照されたい）。「新生物」および「腫瘍」という用語は、細胞の増殖により、正常組織より迅速に増殖し、増殖を誘発した刺激が解除された後でも増殖し続ける異常組織を指す。このような異常組織は、構造的な組織化、および正常組織との機能的な協調の、部分的または完全な欠如であって、良性（良性腫瘍など）または悪性（悪性腫瘍など）でありうる欠如を示す。がんの一般的な類型の例は、癌腫（例えば、乳がん、前立腺がん、肺がん、および結腸がんの共通の形態など、上皮細胞に由来する悪性腫瘍）、肉腫（結合組織または間葉細胞に由来する悪性腫瘍

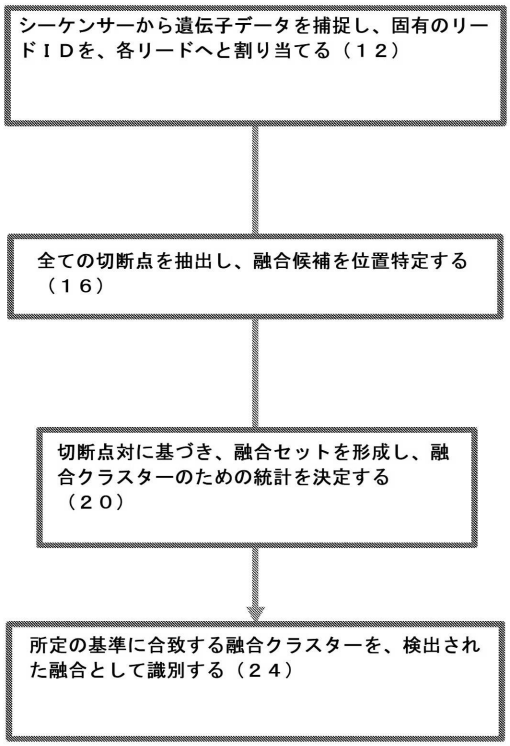
）、リンパ腫（造血細胞に由来する悪性腫瘍）、白血病（造血細胞に由来する悪性腫瘍）、および生殖細胞腫瘍（全能性細胞に由来する腫瘍であって、成人では、精巣内または卵巣内に見出されることが最も多く；胎児、乳児、および若齢小児では、体内の正中線上、特に、鼻骨の先端に見出されることが最も多い腫瘍）、芽球性腫瘍（典型的に、未成熟組織または胚性組織に酷似する悪性腫瘍）などを含むがこれらに限定されない。本発明により包含されることが意図される新生物の種類の例は、神経組織、造血組織、乳腺、皮膚、骨、前立腺、卵巣、子宮、子宮頸部、肝臓、肺、脳、喉頭、胆嚢、膵臓、直腸、副甲状腺、甲状腺、副腎、免疫系、頭頸部、結腸、胃、気管支、および／または腎臓のがんと関連する新生物を含むがこれらに限定されない。特定の実施形態では、検出されうるがんの種類および数は、血液がん、脳がん、肺がん、皮膚がん、鼻腔がん、咽頭がん、肝臓がん、骨がん、リンパ腫、膵臓がん、皮膚がん、腸がん、直腸がん、甲状腺がん、膀胱がん、腎臓がん、口腔がん、胃がん、充実性腫瘍、異種性腫瘍、同種性腫瘍などを含むがこれらに限定されない。特定の実施形態では、がんは、血液がん、肉腫、または前立腺がんである。

【００９７】

本明細書では、本発明の好ましい実施形態について示し、記載してきたが、当業者には、このような実施形態は、例だけを目的として提示されることが明らかであろう。本発明から逸脱することなく、当業者は今や、多数の変形形態、変更形態、および代用に想到するであろう。本明細書で記載される、本発明の実施形態に対する、多様な代替物を、本発明の実施において利用しうることを理解されたい。以下の特許請求の範囲は、本発明の範囲を規定するものであり、これらの特許請求の範囲の範囲内にある方法および構造、ならびにそれらの均等物は、その対象となることが意図される。

【図１】

FIG. 1



【図２】

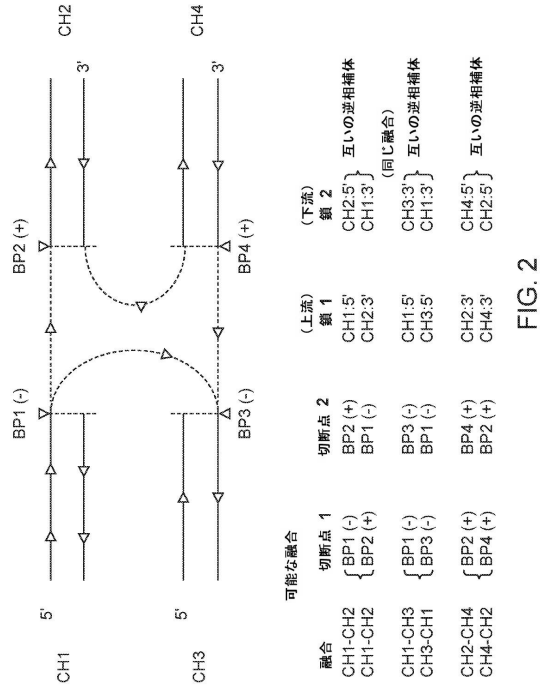


FIG. 2

【図３Ａ】

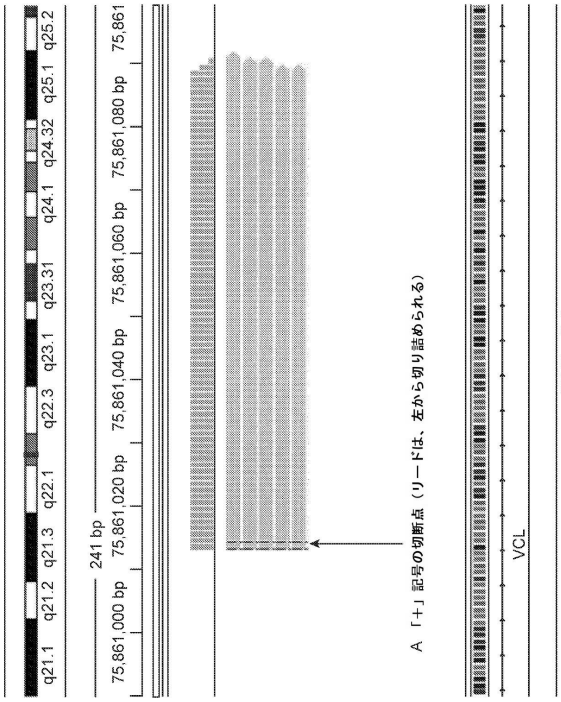


FIG. 3A

【図３Ｂ】

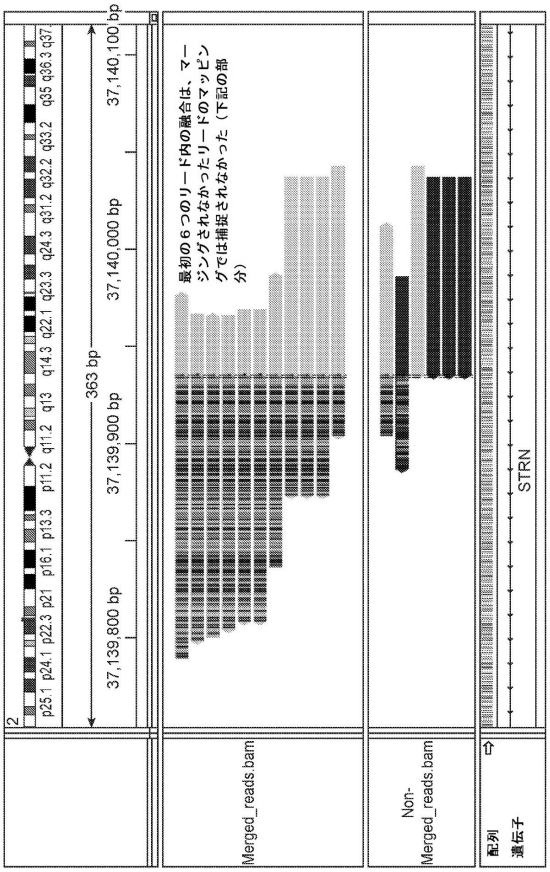
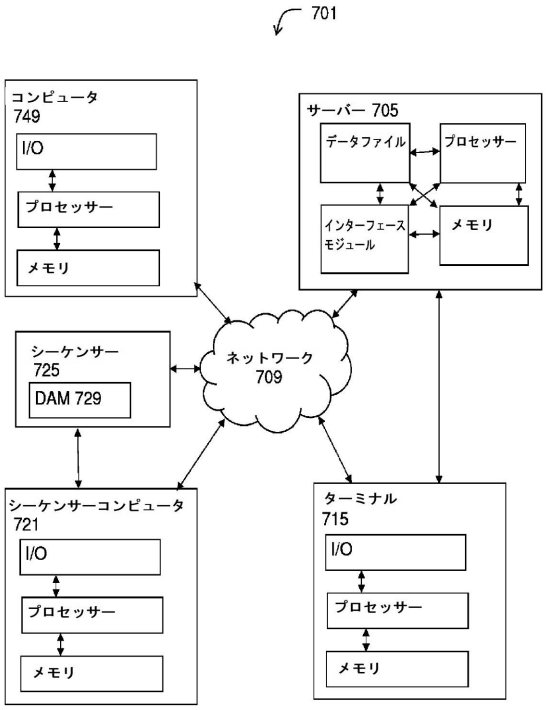


FIG. 3B

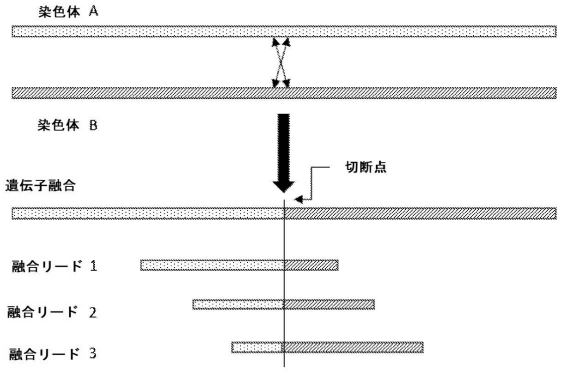
【図４】

FIG. 4



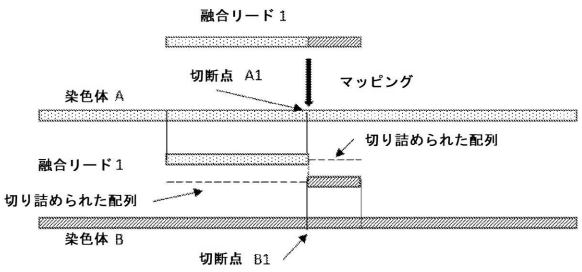
【図５】

FIG. 5



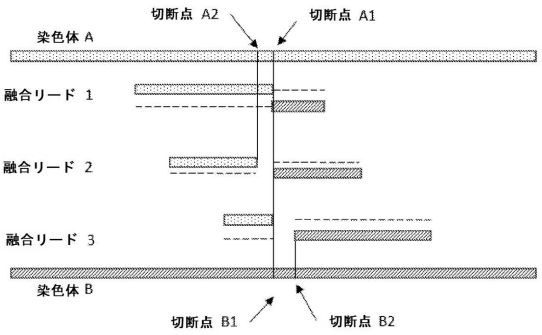
【図６】

FIG. 6



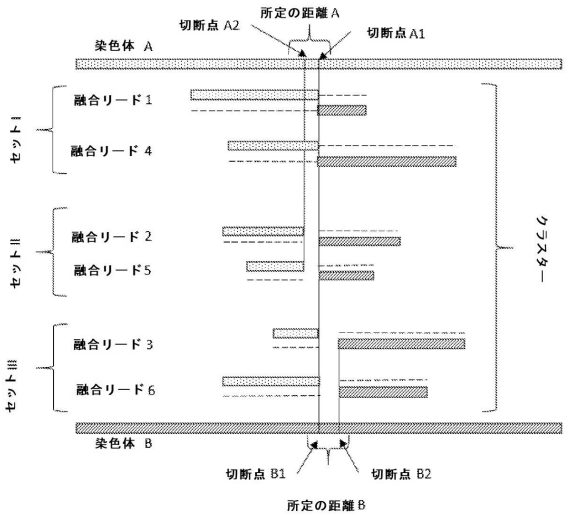
【図７】

FIG. 7



【図８】

FIG. 8



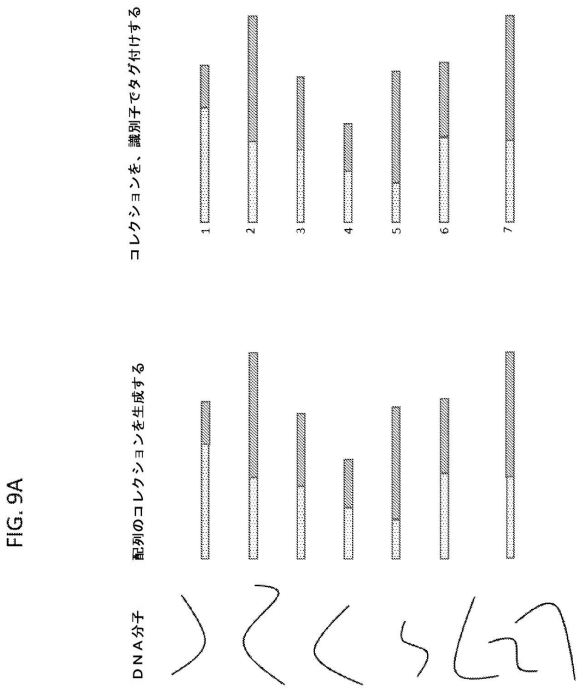
染色体 A:
切断点Ａ１：融合リード１、４、３、６＝４つ
切断点Ａ２：融合リード２、５＝２つ
第１の遺伝子融合切断点の判定：Ａ１

染色体 B:
切断点Ｂ１：融合リード１、４、２、５＝４つ
切断点Ｂ２：融合リード３、６＝２つ
第２の遺伝子融合切断点の判定：Ｂ１

遺伝子融合切断点対：Ａ１－Ｂ１

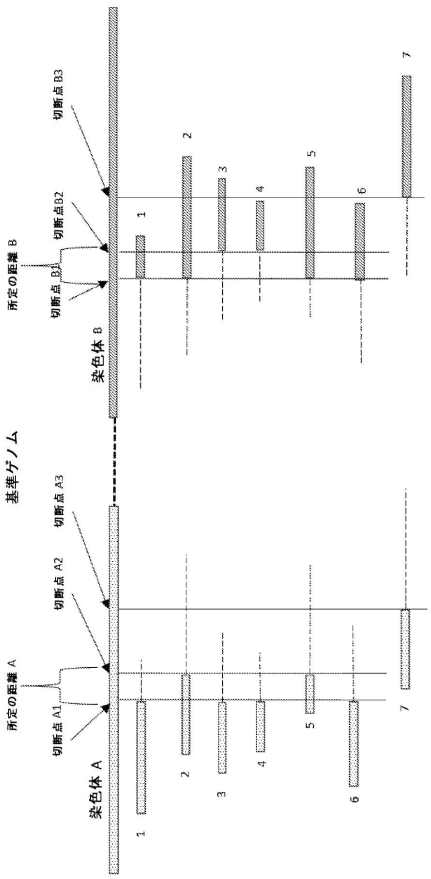
【図９Ａ】

FIG. 9A



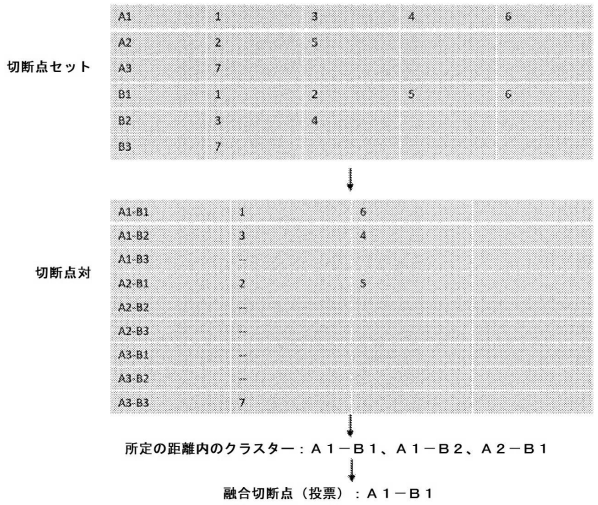
【図９Ｂ】

FIG. 9B



【図９Ｃ】

FIG. 9C


切断点セット
A1 1 3 4 6
A2 2 5
A3 7
B1 1 2 5 6
B2 3 4
B3 7
切断点対
A1-B1 1 6
A1-B2 3 4
A1-B3 --
A2-B1 2 5
A2-B2 --
A2-B3 --
A3-B1 --
A3-B2 --
A3-B3 7
所定の距離内のクラスター：Ａ１－Ｂ１、Ａ１－Ｂ２、Ａ２－Ｂ１
融合切断点（投票）：Ａ１－Ｂ１

フロントページの続き

(74)代理人 230113332
弁護士 山本 健策
(72)発明者 モクタリ, モハマド アール.
アメリカ合衆国 カリフォルニア ９４４０４, フォスター シティ, カタマラン ストリート ６４９, アパートメント ３
(72)発明者 ケルマニ, バハラーム ガッファザデー
アメリカ合衆国 カリフォルニア ９４０２２, ロス アルトス, サード ストリート ７３, アパートメント １２

審査官 関 博文

(56)参考文献 米国特許出願公開第２０１５／０１４２３２８（ＵＳ，Ａ１）
米国特許出願公開第２００９／０２０２９９９（ＵＳ，Ａ１）
中国特許出願公開第１０４８９４２７１（ＣＮ，Ａ）
SCHRODER, Jan , Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads, BIOINFORMATICS, 英国, ２０１４年 1月 2日, Vol.30, no.8, pp.1064-1072

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ１６Ｂ １０／００－９９／００
Ｃ１２Ｑ １／６８