(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)公表番号
特表2023-534882
(P2023-534882A)
(43)公表日　令和5年8月14日(2023.8.14)

(51)国際特許分類		FI			テーマコード（参考）
C12Q	*1/6837(2018.01)*	C12Q	1/6837	Z ZNA	4B029
C12Q	*1/6869(2018.01)*	C12Q	1/6869	Z	4B063
C12Q	*1/34 (2006.01)*	C12Q	1/34		
C12Q	*1/25 (2006.01)*	C12Q	1/25		
C12M	*1/00 (2006.01)*	C12M	1/00	A	

審査請求　未請求　予備審査請求　未請求　（全69頁）　最終頁に続く

(21)出願番号　特願2023-521274(P2023-521274)
(86)(22)出願日　令和3年6月14日(2021.6.14)
(85)翻訳文提出日　令和5年2月14日(2023.2.14)
(86)国際出願番号　PCT/US2021/037226
(87)国際公開番号　WO2021/257453
(87)国際公開日　令和3年12月23日(2021.12.23)
(31)優先権主張番号　63/039,004
(32)優先日　令和2年6月15日(2020.6.15)
(33)優先権主張国・地域又は機関
米国(US)
(81)指定国・地域　AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,
最終頁に続く

(71)出願人　515236259
ザ・ブロード・インスティテュート・インコーポレイテッド
アメリカ合衆国・マサチューセッツ・０２１４２・ケンブリッジ・メイン・ストリート・４１５
(71)出願人　596060697
マサチューセッツ インスティテュート オブ テクノロジー
アメリカ合衆国マサチューセッツ州０２１３９ケンブリッジ，マサチューセッツ・アヴェニュー・７７
(71)出願人　592017633
ザ ジェネラル ホスピタル コーポレイション
最終頁に続く

(54)【発明の名称】　キメラアンプリコンアレイ配列決定

(57)【要約】
本開示は、核酸配列決定のための組成物及び方法に関し、具体的には、少なくともある態様では、入力配列のキメラアレイを提供することによって、既知のロングレンジ配列決定プラットフォームの有効性、スループット及び／又は収率を増強するための方法及び組成物を提供する。そのようなコンポーネント核酸配列要素のアレイは、バイアスの導入を最小限に抑える方法によって調製することができる。本キメラアンプリコン配列決定プロセスを用いるミトコンドリア系統追跡のための方法と同様に、例えば患者試料からアイソフォーム配列決定情報を得るための現在の方法の適用も具体的に提供される。アレイ核酸配列の処理及び解釈のための方法及びシステムも提供される。
【選択図】図１Ａ

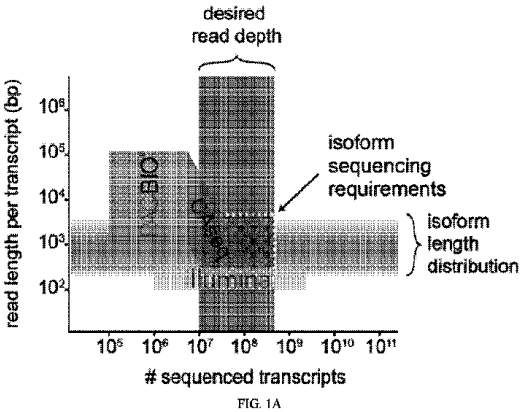


FIG. 1A

【特許請求の範囲】
【請求項１】
アレイ核酸配列を調製するための方法であって、前記方法が、
ｉ）各入力核酸配列が約３０キロベース長以下である複数の入力核酸配列を得ること、
ｉｉ）１つ又は複数のアダプタ配列を前記複数の入力核酸配列に付加し、それによって適合核酸配列の集団を生成すること、
ｉｉｉ）前記適合核酸配列の集団を、前記適合核酸配列の集団内の各適合核酸配列の少なくとも１つの末端に一本鎖末端を生成することができる酵素と接触させ、それにより、一本鎖末端を有する核酸配列の集団を形成すること、及び
ｉｖ）前記一本鎖末端を有する核酸配列の集団をリガーゼと接触させること、
を含み、
それにより、アレイ核酸配列を形成する、方法。
【請求項２】
前記１つ又は複数のアダプタ配列のうち少なくとも１つが、１つの鎖上に内部ｄＵを含む、請求項１に記載の方法。
【請求項３】
前記アレイ核酸配列が、少なくとも２０キロベース、任意選択的に少なくとも５０キロベース、任意選択的に約１００ｋｂ以上の長さを有する、請求項１に記載の方法。
【請求項４】
前記複数の入力核酸配列が、約０．５ｋｂ～２０ｋｂの長さである、請求項１に記載の方法。
【請求項５】
前記複数の入力核酸配列が、１つ又は複数のｃＤＮＡライブラリ、任意選択で１つ又は複数の単一細胞若しくは空間ｃＤＮＡライブラリから得られる、請求項１に記載の方法。
【請求項６】
工程（ｉｉ）が、前記複数の核酸配列を前記対になった増幅プライマーと接触させること、ここで、前記対になった増幅プライマー内の少なくとも１つのプライマーが１つの鎖上に内部ｄＵを含むアダプタ配列を含み、及び、少なくとも１ラウンドの増幅を実施し、それにより、適合核酸配列の集団を生成することを含む、請求項１に記載の方法。
【請求項７】
前記対になった増幅プライマー内の少なくとも１つのプライマーがビオチン化されており、任意選択で、アダプタ配列テールアンプリコンのためのビオチン媒介選択が行われる、請求項６に記載の方法。
【請求項８】
工程（ｉｉｉ）が、前記適合核酸配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する核酸配列の集団を形成することを更に含む、請求項２、６又は７のいずれか一項に記載の方法。
【請求項９】
前記アダプタ配列が５～３０塩基対の長さを含み（標的核酸配列を除く）、任意選択で、前記アダプタ配列が６～２５塩基対の長さであり、任意選択で、前記アダプタ配列が構造５’－Ｎ６－１６＿ｄＵ＿ｔａｒｇｅｔ－ＤＮＡ－３’を有する、請求項１に記載の方法。
【請求項１０】
１つの鎖上に内部ｄＵを含む前記アダプタ配列が、配列番号１～１８からなる群より選択される配列を含む、請求項１に記載の方法。
【請求項１１】
アダプタ配列を有する複数の核酸配列について、各アダプタ配列が、アダプタ配列を有する前記複数の核酸配列のうちの少なくとも１つの他のものと相補的な１つ又は２つの指定された配列を有し、それにより、前記複数のアダプタ配列が相補的なアダプタ配列の集団を形成し、任意選択で、前記相補的なアダプタ配列の集団の各相補的なアダプタ配列が

、前記相補的なアダプタ配列の集団の互いの相補的なアダプタ配列に対して最小の類似性を有し、任意選択で、前記相補的なアダプタ配列の集団の各相補的なアダプタ配列が、前記相補的なアダプタ配列の集団の他の全ての相補的なアダプタ配列から少なくとも１１ハミング距離単位離れている、請求項１に記載の方法。

【請求項１２】

以下の：前記複数の入力核酸配列；前記適合核酸配列の集団；及び／又は前記一本鎖末端を有する核酸配列の集団のうちの１つ又は複数がサイズ選択され、任意選択で前記サイズ選択が電気泳動を介して、任意選択でアガロースゲル上で行われる、請求項１に記載の方法。

【請求項１３】

前記アレイ核酸配列の配列情報が、任意選択でロングリード配列決定プラットフォームを使用して得られる、請求項１に記載の方法。

【請求項１４】

ハプロタイプフェージングの配列情報が前記アレイ核酸配列にわたって得られる、請求項１３に記載の方法。

【請求項１５】

形成される前記アレイ核酸配列が、５つ以上の入力核酸配列、任意選択的に６つ以上、任意選択的に７つ以上、任意選択的に８つ以上、任意選択的に９つ以上、任意選択的に１０以上、任意選択的に１１以上、任意選択的に１２以上、任意選択的に１３以上、任意選択的に１４以上、任意選択的に１５以上、任意選択的に１６以上、任意選択的に１７以上、任意選択的に１８以上、任意選択的に１９以上、任意選択的に２０以上を含む、請求項１に記載の方法。

【請求項１６】

標的化アイソフォーム配列決定情報が、前記複数の入力核酸配列を得る工程（ｉ）中に遺伝子パネルの標的化を介して得られる、請求項１３に記載の方法。

【請求項１７】

前記複数の入力核酸配列が、免疫応答経路のためのｃＤＮＡを含む、請求項１に記載の方法。

【請求項１８】

前記複数の入力核酸配列がミトコンドリアＤＮＡから得られ、任意選択で、前記アレイ核酸配列の配列決定がミトコンドリアＤＮＡ系統追跡に使用される、請求項１に記載の方法。

【請求項１９】

前記適合核酸配列の集団が、ギブソンアセンブリを介して連結される、請求項１に記載の方法。

【請求項２０】

前記アレイ核酸配列が線状アレイである、請求項１に記載の方法。

【請求項２１】

前記アレイ核酸配列が環状アレイである、請求項１に記載の方法。

【請求項２２】

核酸配列の線状アレイのアレイを調製するための方法であって、前記方法が、

ｉ）請求項２０に記載の方法によって入力核酸配列の第１の集団から第１の線状アレイを調製すること、

ｉｉ）請求項２０に記載の方法によって、入力核酸配列の第２の集団から第２の線状アレイを調製すること、ここで、前記第１の線状アレイ及び前記第２の線状アレイは各々、適合する相補的フランキング配列を有し、

ｉｉｉ）前記第１の線状アレイ及び前記第２の線状アレイを溶液中で組み合わせること、及び

ｉｖ）溶液中の前記第１の線状アレイ及び前記第２の線状アレイをリガーゼと接触させること、

を含み、
それにより、核酸配列の線状アレイのアレイを形成する、方法。
【請求項２３】
前記第１の線状アレイ若しくは前記第２の線状アレイ、又はその両方が線状アレイのアレイを含む、請求項２２に記載の方法。
【請求項２４】
ｖ）請求項２０に記載の方法によって入力核酸配列の第３の集団から第３の線状アレイを調製すること、ここで、前記線状アレイのアレイ及び第３の線状アレイは各々、適合する相補的フランキング配列を有し、
ｖｉ）前記線状アレイのアレイ及び前記第３の線状アレイを溶液中で組み合わせること、
ｖｉｉ）溶液中の前記線状アレイのアレイ及び前記第３の線状アレイをリガーゼと接触させ、それにより、核酸配列の線状アレイのより大きなアレイを形成すること、
を更に含み、任意選択で、工程（ｖ）～（ｖｉｉ）が繰り返されて、第４の線状アレイ、第５の線状アレイ、及び／又はより多くの線状アレイが、線状アレイのより大きなアレイに組み込まれる、請求項２２又は請求項２３に記載の方法。
【請求項２５】
入力ｃＤＮＡ配列の集団からアイソフォーム配列決定情報を得るための方法であって、前記方法が、
ｉ）複数の入力ｃＤＮＡ配列を得ること、
ｉｉ）前記複数の入力ｃＤＮＡ配列を対になった増幅プライマーと接触させること、ここで、前記対になった増幅プライマー内の少なくとも１つのプライマーは、１つの鎖上に内部ｄＵを含むアダプタ配列を含み、少なくとも１ラウンドの増幅を行い、それにより、適合ｃＤＮＡ配列の集団を生成させ、
ｉｉｉ）前記適合ｃＤＮＡ配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより、一本鎖末端を有する適合ｃＤＮＡ配列の集団を形成すること、
ｉｖ）前記一本鎖末端を有する適合ｃＤＮＡ配列の集団をリガーゼと接触させ、それによって線状アレイ核酸配列を形成すること、
ｖ）前記線状アレイ核酸配列から、任意選択でロングリード配列決定によって配列情報を得ること、及び
ｖｉ）前記線状アレイ核酸配列から得られた前記配列情報を分析して、アイソフォーム配列決定情報を得ること、
を含み、
それにより、前記入力ｃＤＮＡ配列の集団からアイソフォーム配列決定情報を得る、方法。
【請求項２６】
入力ミトコンドリアｃＤＮＡ配列の集団からミトコンドリア系統追跡を行うための方法であって、前記方法が、
ｉ）複数の入力ミトコンドリアｃＤＮＡ配列を得ること、
ｉｉ）前記複数の入力ミトコンドリアｃＤＮＡ配列を対になった増幅プライマーと接触させること、ここで、前記対になった増幅プライマー内の少なくとも１つのプライマーが、１つの鎖上に内部ｄＵを含むアダプタ配列を含み、少なくとも１ラウンドの増幅を行い、それにより、適合ミトコンドリアｃＤＮＡ配列の集団を生成し、
ｉｉｉ）前記適合ミトコンドリアｃＤＮＡ配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより、一本鎖末端を有する適合ミトコンドリアｃＤＮＡ配列の集団を形成すること、
ｉｖ）前記一本鎖末端を有する適合ミトコンドリアｃＤＮＡ配列の集団をリガーゼと接触させ、それによりアレイ核酸配列を形成すること、
ｖ）配列情報を前記アレイ核酸配列から、任意選択でロングリード配列決定によって得

ること、及び
ｖｉ）前記アレイ核酸配列から得られた前記配列情報を分析してミトコンドリア系統を追跡すること、
を含み、
それにより、前記入力ミトコンドリアｃＤＮＡ配列の集団からミトコンドリア系統追跡を行う、方法。
【請求項２７】
アレイ核酸配列を調製するための方法であって、前記方法が、
ｉ）複数の入力核酸配列を得ること、ここで、前記複数の入力配列内の各入力核酸配列が約３００キロベース以下の長さであり、
ｉｉ）前記複数の入力核酸配列を対になった増幅プライマーと接触させること、ここで、前記対になった増幅プライマー内の少なくとも１つのプライマーが１つの鎖上に内部ｄＵを含むアダプタ配列を含み、少なくとも１ラウンドの増幅を実施し、それにより、適合核酸配列の集団を生成し、
ｉｉｉ）前記適合核酸配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する適合核酸配列の集団を形成すること、及び
ｉｖ）前記一本鎖末端を有する適合核酸配列の集団をリガーゼと接触させること、
を含み、
それにより、アレイ核酸配列を形成する、方法。
【請求項２８】
アレイ核酸配列を調製するための方法であって、前記方法が、
ｉ）複数の入力核酸配列を得ること、ここで、前記複数の入力配列内の各入力核酸配列が約３００キロベース以下の長さであり、
ｉｉ）前記複数の入力核酸配列を、一本の鎖上の内部ｄＵを含むアダプタ配列及びリガーゼと接触させ、それにより、適合核酸配列の集団を生成すること、
ｉｉｉ）前記適合核酸配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する適合核酸配列の集団を形成すること、及び
ｉｖ）前記一本鎖末端を有する適合核酸配列の集団をリガーゼと接触させること、
を含み、
それにより、線状アレイ核酸配列を形成する、方法。
【請求項２９】
前記複数の入力配列内の各入力核酸配列は、約３０キロベース以下の長さである、請求項２７又は２８に記載の方法。
【請求項３０】
複数の核酸配列を含み、前記複数の核酸配列の少なくとも２つが、配列番号１～１８からなる群から選択されるアダプタ配列を含む、組成物。
【請求項３１】
配列番号１～１８からなる群から選択される複数のアダプタ配列及びその使用説明書を含むキット。
【請求項３２】
核酸配列リードの集団の個々の核酸配列リード内の別個の配列要素を同定するための方法であって、前記個々の核酸配列リードは、配列要素の線状アレイを有し、
前記配列要素の線状アレイの各々は、高複雑度のライブラリから引き出された２つ以上の核酸配列要素を含み、高複雑度のライブラリから引き出された各核酸配列要素は、低複雑度のライブラリから引き出された１つ若しくは複数の予想される核酸配列、又は低複雑度のライブラリから引き出された１つ若しくは複数の予想される核酸配列及び配列リード終端に隣接し、前記方法は、
（ａ）前記核酸配列リードの集団の配列データに１つ又は複数の統計的アノテーション

モデルを適用して、前記核酸配列リードの集団内で、高複雑度のライブラリから引き出された個々の核酸配列要素の領域及び低複雑度のライブラリから引き出された核酸配列の領域を予測すること、ここで、前記１つ又は複数の統計的アノテーションモデルが、
ｉ）核酸配列リード全体に散在する１つ又は複数の予想される核酸配列を認識するための生成統計的アライメントモデル、
ｉｉ）既知ではない、又は高複雑度の配列の辞書から引き出された配列を認識するためのランダム統計アライメントモデル、を含み、
予測された転位部位は、各モデルの末端に配置され、前記生成統計的アライメントモデルの内部位置内では許容されず、
（ｂ）複数の核酸配列リードに対して工程（ａ）を繰り返し、それにより、前記１つ又は複数の統計的モデルを前記複数の核酸配列リードの各核酸配列リードに順相補性配向及び逆相補性配向の両方で適用し、最大対数尤度値を有するモデルを同定することによって、選択された最大事後状態経路の最終的リード当たりのモデル（ｍａｘｉｍｕｍ ａ ｐｏｓｔｅｒｉｏｒｉ ｓｔａｔｅ ｐａｔｈ Ｆｉｎａｌ ｐｅｒ－ｒｅａｄ ｍｏｄｅｌ）選択を決定すること、及び
（ｃ）前記複数の核酸配列リードの各核酸配列リードを、工程（ｂ）の最大事後状態経路の最終的リード当たりのモデル選択によって同定される転位部位によって区画された別個の配列要素にセグメント化すること、
を含み、
それにより、前記核酸配列リードの集団内の別個の配列要素を同定する、方法。
【請求項３２】
前記高複雑度のライブラリが、１，０００を超える異なる要素、任意選択で１０，０００を超える異なる要素を含むか、又は含む可能性がある、請求項３２に記載の方法。
【請求項３３】
前記高複雑度のライブラリ及び／又は知られていない若しくは高複雑度の配列の辞書から引き出された配列が、ｃＤＮＡ転写物配列、バーコード配列及び固有の分子識別子からなる群から選択される要素を含む、請求項３２に記載の方法。
【請求項３４】
前記低複雑度のライブラリが、１００個以下の異なる配列、任意選択的に５０個以下の異なる配列、任意選択的に２５個以下の異なる配列、任意選択的に１５個以下の異なる配列を含む、請求項３２に記載の方法。
【請求項３５】
前記低複雑度のライブラリがアダプタ及び／又はリンカー配列を含む、請求項３５に記載の方法。
【請求項３６】
前記先験的に予想される核酸配列が、アダプタ及び／又はリンカー配列を含む、請求項３２に記載の方法。
【請求項３７】
先験的に知られていない配列、又は高複雑度の配列の辞書から引き出された配列が、ｃＤＮＡ配列、バーコード配列及び固有の分子識別子配列からなる群から選択される１つ又は複数の配列を含み、任意選択で前記バーコード配列が単一細胞バーコード配列を含む、請求項３２に記載の方法。
【請求項３８】
複数の核酸配列リードの個々の配列リード内の別個の配列要素を同定し、配列要素データを保存するためのシステムであって、前記システムは、
ネットワークと通信するための１つ又は複数のネットワークインターフェース；
前記ネットワークインターフェースに結合される、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び
前記プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリを含み、

前記プロセスは、実行されると、
（ａ）配列要素の線状アレイを有する個々の核酸配列リードを含む複数の核酸配列リードを取得する、
ここで、配列要素の線状アレイを有する各リードが高複雑度のライブラリから引き出された２つ以上の個々の核酸配列要素を含み、高複雑度のライブラリから引き出された各核酸配列要素が低複雑度の１つ若しくは複数の予想される核酸配列、又は低複雑度の１つ若しくは複数の予想される核酸配列及び配列リード終端のいずれかに隣接している、
（ｂ）核酸配列リード内で、高複雑度のライブラリから引き出された個々の核酸配列要素の前記複数の領域及び低複雑度のライブラリから引き出された核酸配列の領域を予測するために、１つ又は複数の統計的アノテーションモデルを前記複数の核酸配列リードの配列データに適用する、
ここで、前記１つ又は複数の統計的アノテーションモデルは、
ｉ）核酸配列リード全体に散在する１つ又は複数の予想される核酸配列を認識するための生成統計的アライメントモデル、及び
ｉｉ）既知ではない配列、又は高複雑度の配列の辞書から引き出された配列を認識するためのランダム統計アライメントモデルを含み、
前記生成統計的アライメントモデルでは、予測された転位部位は各モデルの末端に配置され、かつ内部位置内では許容されない、
（ｃ）複数の核酸配列リードに対して工程（ａ）を繰り返し、それにより、前記１つ又は複数の統計的モデルを順相補配向及び逆相補配向の両方で前記複数の核酸配列リードの各核酸配列リードに適用し、最大対数尤度値を有するモデルを同定することによって選択された最終リード当たりのモデル選択により、各モデルの最大事後状態経路を決定し、それにより、前記核酸配列リード内の既知のセグメントを標識する、
（ｄ）前記複数の核酸配列リードの各核酸配列リードを、工程（ｃ）の最大事後状態経路の最終的リード当たりのモデルによって同定される転位部位によって区画された標識された既知のセグメントの個別の配列要素にセグメント化し、
それにより、前記複数の核酸配列リード内の別個の配列要素を同定する、及び
（ｅ）前記複数の核酸配列リード内で同定された前記別個の配列要素を配列要素データファイルに保存する、
ように構成される、システム。

【請求項３９】
前記高複雑度のライブラリが、１，０００を超える異なる要素、任意選択で１０，０００を超える異なる要素を含むか、又は含む可能性がある、請求項３９に記載のシステム。

【請求項４０】
前記高複雑度のライブラリ及び／又は先験的に知られていない配列若しくは高複雑度の配列の辞書から引き出された配列が、ｃＤＮＡ転写物配列、バーコード配列及び固有の分子識別子からなる群から選択される要素を含む、請求項３９に記載のシステム。

【請求項４１】
前記低複雑度のライブラリが、１００個以下の異なる配列、任意選択的に５０個以下の異なる配列、任意選択的に２５個以下の異なる配列、任意選択的に１５個以下の異なる配列を含む、請求項３９に記載のシステム。

【請求項４２】
前記低複雑度のライブラリがアダプタ及び／又はリンカー配列を含む、請求項４２に記載のシステム。

【請求項４３】
前記先験的に予想される核酸配列が、アダプタ及び／又はリンカー配列を含む、請求項３９に記載のシステム。

【請求項４４】
先験的に知られていない配列、又は高複雑度の配列の辞書から引き出された配列が、ｃＤＮＡ配列、バーコード配列及び固有の分子識別子配列からなる群から選択される１つ又

は複数の配列を含み、任意選択で前記バーコード配列が単一細胞バーコード配列を含む、請求項３９に記載のシステム。

【請求項４５】

複数の核酸配列リードの個々の配列リードを、低品質として同定し、除去し、配列データを保存するためのシステムであって、前記システムは、

ネットワークと通信するための１つ又は複数のネットワークインターフェース；

前記ネットワークインターフェースに結合される、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び

前記プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリを含み、

前記プロセスは、実行されると、

ｉ）複数の核酸配列リードの個々の配列リードに対して請求項３９に記載の工程（ａ）～（ｅ）を実施する、

ｉｉ）ライブラリ調製より予想される順序で起こらない別個の配列要素を含む任意のリードを低品質として同定し、除去する、

ここで、最初の別個の配列要素の後で開始するが、残りの別個の配列要素が順番であるリード、及び最後の別個の配列要素の前で終わるが、前のセクションが全て順番であるリード、並びにこれらの場合の組合わせは除去されず、

ｉｉｉ）低品質リードが除去された前記複数の核酸配列リードを配列データファイルに保存する、ように構成される、システム。

【請求項４６】

サーキュラーコンセンサスシーケンシングソフトウェアによって高品質であると同定された１つ又は複数の核酸配列リードが低品質であると同定され、除去される、請求項４６に記載のシステム。

【請求項４７】

更なる分析のために十分に高品質の個々の配列リードを同定し、複数の核酸配列リードの個々の配列リードを配列データに付加し、配列データを保存するためのシステムであって、前記システムは、

ネットワークと通信するための１つ又は複数のネットワークインターフェース；

前記ネットワークインターフェースに結合される、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び

前記プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリ、を含み、

前記プロセスは、実行される場合、

ｉ）複数の核酸配列リードの各々における各ヌクレオチドである複数の核酸配列リードの個々の配列リードに対して請求項３９に記載の工程（ａ）～（ｅ）を実施する、

ｉｉ）更なる分析のための十分に高い品質に関して、最初の予想されるセグメントの後で開始するが残りのセクションが順番であるリード、及び最後の予想されるセグメントの前で終わるが前のセクションが順番であるリード、並びにこれらの場合の任意の組合わせを含む、ライブラリ調製により出現すると予想される順序で標識されたセクションを含む任意のリードを同定する、及び

ｉｉｉ）更なる分析のために十分に高品質であると同定された前記核酸配列リードを配列データファイルに保存する、ように構成される、システム。

【請求項４８】

サーキュラーコンセンサスシーケンシングソフトウェアによって低品質であると同定された１つ又は複数の核酸配列リードが、更なる分析に対して十分に高品質であると同定される、請求項４８に記載のシステム。

【請求項４９】

請求項４６に記載の低品質として同定されたリード、又は請求項４８に記載の高品質として同定されたリードの品質を概算し、推定品質スコアをデータに付加し、データを保存

するためのシステムであって、
前記システムは、
ネットワークと通信するための１つ又は複数のネットワークインターフェース；
前記ネットワークインターフェースに結合される、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び
前記プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリ、を含み、
前記プロセスは、実行されると、
（ｉ）請求項４６に記載の低品質として同定された各リード又は請求項４８に記載の高品質として同定された各リードでの各別個の配列要素について、別個の配列要素内のヌクレオチドと個別の配列要素に対する予想される配列との間の観察されたアライメントスコアを計算し、個別の配列要素内のヌクレオチドと個別の配列要素に対する予想される配列との間の最良の可能なアライメントスコアを計算する；
（ｉｉ）各セクションの品質スコアを取得するために、任意選択的に、工程（ｉ）で計算された前記アライメントスコアを前記最良の可能なアライメントスコアで除算する、
（ｉｉｉ）全体的な観察されたアライメントスコアを得るために工程（ｉ）で計算された全ての観察されたアライメントスコアを合計し、全体的な最良のアライメントスコアを得るために工程（ｉ）で計算された全ての最良の可能なアライメントスコアを合計し、及び、前記全体的な観察されたアライメントスコアと前記全体的な可能な最良のアライメントスコアとの比を得ることによって、前記核酸配列リードの推定品質スコアを計算する；及び
（ｉｖ）前記核酸配列リードについての前記推定された品質スコアをデータファイルに保存する、ように構成される、システム。
【請求項５０】
前記観察されたアライメントスコアが、工程（ｉ）において、動的プログラミングアルゴリズムを直接的に使用して、又は直接的に前記別個の配列要素と前記予想される配列との間のレーベンシュタイン距離を計算し、前記予想される配列の長さからその距離を減算することによって計算され、任意選択で、前記動的プログラミングアルゴリズムが、Ｓｍｉｔｈ－Ｗａｔｅｒｍａｎアルゴリズム、Ｎｅｅｄｌｅｍａｎ－Ｗｕｎｓｃｈアルゴリズム、及びペア隠れマルコフモデルアルゴリズムからなる群から選択される、請求項５０に記載のシステム。
【請求項５１】
前記最良の可能なアライメントスコアが、前記予想される配列とそれ自体との間のアライメントスコアを計算することによって得られる、請求項５０に記載のシステム。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願の相互参照
本出願は、「キメラアンプリコンアレイ配列決定（Ｃｈｉｍｅｒｉｃ　Ａｍｐｌｉｃｏｎ　Ａｒｒａｙ　Ｓｅｑｕｅｎｃｉｎｇ）」と題する、２０２０年6月１５日に出願された米国仮特許出願第６３／０３９，００４号の利益を主張する。上記出願の全内容は、参照により本明細書に組み込まれる。
【０００２】
連邦政府による資金提供を受けた研究に関する記載
本発明は、国立衛生研究所によって授与された助成金番号Ｕ１９ＡＩ０８２６３０の下で政府の支援を受けてなされた。政府は、本発明に一定の権利を有する。
【０００３】
本発明は、一般に、核酸配列決定のための方法及び組成物、特に配列決定のための核酸集団の調製に関する。
【背景技術】

【０００４】
次世代ＤＮＡ配列決定の出現は生物学的研究に革命をもたらしたが、現在の配列決定プラットフォームによって解決が依然として不十分である多数の重要な遺伝的特徴が存在する。例えば、ｍＲＮＡ成熟中にエクソンの差次的スプライシングを介して遺伝子機能の深く本質的な多様化を可能にするコア生物学的プロセスである選択的スプライシングは、公知の単一細胞配列決定法によって十分に捕捉されていない。腫瘍のクローン進化研究のために、単一細胞のマーカ対立遺伝子からクローン関係を導き出す能力は、頑強な配列決定カバレッジを必要とし、単一細胞遺伝子発現ワークフローでもこれまで達成出来ていない試みを必要とする。更に、潜在する遺伝的障害に起因する疾患には、診断及び病因の解明の両方のためにゲノム組成を忠実に再構築する能力が必要とされる。特に、接合後の変異の結果であり、重度の神経障害に寄与することが知られている体細胞モザイク現象を特徴付けることは、多数の個々の細胞のサンプリングを必要とし、これは現在の方法では扱いにくい作業である。以前に記載されたアプローチではこれらの重要な特徴を解決することができないことは、複雑な生物学的系を忠実に特徴付ける当技術分野の能力が著しく不足していることを強調している。これらの制限は、既知のアプローチが現在の配列決定技術でロングレンジＤＮＡ情報を効率的に捕捉することができないことから生じる。したがって、現在のロングリード配列決定プラットフォームでのロングレンジＤＮＡ情報の捕捉を最適化することができるアプローチが必要とされている。

【発明の概要】

【課題を解決するための手段】

【０００５】
本開示は、少なくとも部分的には、特に、ロングリード配列決定プラットフォームを使用してキメラ核酸に対して核酸配列決定を行うための組成物及び方法に関する。ある態様において、本開示は、ハイスループット構築のための方法及び組成物、並びにロングリード配列決定プラットフォームへの適用のための、（本明細書において「キメラアレイ配列決定」又は「ＣＡｓｅｑ」と呼ばれるプロセスを介した）核酸のキメラアレイの使用を提供する。そのようなキメラアレイは、以前は不明瞭であった遺伝的特徴の解明、例えば選択的スプライシングの検出；腫瘍クローン進化等のクローン進化の改善された検出；例えば、疾患診断及び疾患病因の解明のための、ゲノム組成の忠実な再構成；体細胞モザイク現象の特徴付け；及びより一般的には改良されたゲノムハプロタイプ評価を可能にする。

【０００６】
本開示は、そのロングリードプラットフォームの固有の特徴を利用して、複数の共通配列決定ライブラリの出力を増強するための一般化可能なワークフローを提供する。ロングリードシーケンサは、非常に大きな配列決定出力を有するが（例えば、ＰａｃＢｉｏ（登録商標）Ｓｅｑｕｅｌ　ＩＩは約３００ＧＢである）、ラン当たりのリードの総数は限られている（例えば、ＰａｃＢｉｏ（登録商標）Ｓｅｑｕｅｌ　ＩＩは約４Ｍである）。出力を最大化するために、より小さい断片のライブラリをアレイにアセンブルし、ロングリードシーケンサで効率的に配列決定し、配列決定されたライブラリメンバーの数をアレイ中の断片の数に対して線形に増加させることができる。したがって、本開示のある態様は、単一細胞の遺伝子発現試料からのハイスループット完全転写物配列決定を可能にするという本開示の主な利点を有する、高効率ロングリード配列決定のためのアレイのアセンブリのための合理化され、一般化可能な方法を詳述する。

【０００７】
一態様では、本開示は、アレイ核酸配列を調製する方法を提供し、方法は、（ｉ）それぞれが約３００キロベース長以下（任意選択で３０キロベース長以下）である、複数の入力核酸配列を取得すること、（ｉｉ）１つ又は複数のアダプタ配列を複数の核酸配列に付着させ、それにより、適合（ａｄａｐｔｅｄ）核酸配列の集団を作製すること、（ｉｉｉ）適合核酸配列の集団を、適合核酸配列の集団内の各二本鎖適合核酸配列の少なくとも１つの末端に一本鎖末端を生成することができる酵素と接触させ、それにより一本鎖末端を有する核酸配列の集団を形成すること、及び（ｉｖ）一本鎖末端を有する核酸配列の集団

をリガーゼと接触させること、を含み、それによりアレイ核酸配列を形成する。

【０００８】

いくつかの実施形態において、アダプタ配列の少なくとも１つは、１つの鎖上に内部ｄＵを含む。

【０００９】

実施形態では、アレイ核酸配列は、少なくとも２０キロベースの長さを有する。任意選択で、アレイ核酸配列は、少なくとも５０キロベースの長さを有する。関連する実施形態では、アレイ核酸配列は、約１００キロベース以上の長さを有する。

【００１０】

一実施形態では、複数の入力核酸配列は、約０．５ｋｂ～２０ｋｂの長さである。

【００１１】

ある実施形態において、複数の入力核酸配列は、１つ又は複数のｃＤＮＡライブラリから得られる。任意選択で、複数の入力核酸配列は、１つ又は複数の単一細胞又は空間ｃＤＮＡライブラリから得られる。

【００１２】

実施形態では、工程（ｉｉ）は、複数の核酸配列を対になった増幅プライマーと接触させること、この際、対になった増幅プライマーの少なくとも１つが１つの鎖上の内部ｄＵを含むアダプタ配列を含む、及び、少なくとも１ラウンドの増幅を実行すること、を含み、それにより適合核酸配列の集団を生成する。

【００１３】

いくつかの実施形態では、増幅プライマーの各対の少なくとも１つがビオチン化されている。任意選択で、アダプタ配列テールアンプリコンのためのビオチン媒介選択が行われる。

【００１４】

実施形態では、工程（ｉｉｉ）は、適合核酸配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する核酸配列の集団を形成することを含む。

【００１５】

いくつかの実施形態では、アダプタ配列は、５～３０塩基対の長さを含む（標的核酸配列を除く）。任意選択で、アダプタ配列は６～２５塩基対の長さである。任意選択で、アダプタ配列は、構造５’－Ｎ６－１６＿ｄＵ＿ｔａｒｇｅｔ－ＤＮＡ－３’を有する。

【００１６】

実施形態では、一方の鎖に内部ｄＵを有するアダプタ配列は配列番号：１～１８の配列を含む。

【００１７】

いくつかの実施形態では、アダプタ配列を有する複数の核酸配列について、各アダプタ配列は、アダプタ配列を有する複数の核酸配列のうちの少なくとも１つの他のものと相補的な１つ又は２つの指定配列を有し、それにより、複数のアダプタ配列は相補的なアダプタ配列の集団を形成する。任意選択で、相補的なアダプタ配列の集団の各相補的なアダプタ配列は、相補的なアダプタ配列の集団の互いに相補的なアダプタ配列に対して最小の類似性を有する。関連する実施形態では、相補的なアダプタ配列の集団の各相補的なアダプタ配列は、相補的なアダプタ配列の集団の他の全ての相補的なアダプタ配列から少なくとも１１ハミング距離単位離れている。

【００１８】

ある実施形態では、以下の１つ又は複数がサイズ選択される：複数の入力核酸配列；適合核酸配列の集団；及び／又は一本鎖末端を有する核酸配列の集団。任意選択で、サイズ選択は電気泳動を介して行われる。関連する実施形態では、サイズ選択は、アガロースゲルを使用して行われる。

【００１９】

一定の実施形態では、アレイ核酸配列の配列情報が得られる。任意選択で、アレイ核酸

配列の配列情報は、ロングリード配列決定プラットフォームを使用して得られる。

【００２０】

関連する実施形態では、ハプロタイプフェージングの配列情報がアレイ核酸配列にわたって得られる。

【００２１】

別の実施形態では、形成されるアレイ核酸配列は、５つ以上の入力核酸配列を含む。任意選択で、形成されるアレイ核酸配列は、６個以上、７個以上、８個以上、９個以上、１０個以上、１１個以上、１２個以上、１３個以上、１４個以上、１５個以上、１６個以上、１７個以上、１８個以上、１９個以上、又は２０個以上の入力核酸配列を含む。

【００２２】

ある実施形態において、標的化アイソフォーム配列決定情報は、複数の入力核酸配列を得る工程（ｉ）の間に遺伝子パネルの標的化を介して得られる。

【００２３】

実施形態では、複数の入力核酸配列は、免疫応答経路のためのｃＤＮＡを含む。

【００２４】

いくつかの実施形態では、複数の入力核酸配列は、ミトコンドリアＤＮＡから得られる。任意選択で、アレイ核酸配列の配列決定は、ミトコンドリアＤＮＡ系統追跡に使用される。

【００２５】

ある実施形態では、適合核酸配列の集団は、ギブソンアセンブリを介して結合される。

【００２６】

いくつかの実施形態では、アレイ核酸配列は線状アレイである。

【００２７】

ある実施形態において、アレイ核酸配列は、環状アレイである。

【００２８】

本開示の更なる態様は、入力ｃＤＮＡ配列の集団からアイソフォーム配列決定情報を得るための方法を提供し、方法は、（ｉ）複数の入力ｃＤＮＡ配列を得ること、（ｉｉ）複数のｃＤＮＡ配列を対になった増幅プライマーと接触させ、それにより、適合ｃＤＮＡ配列の集団を生成させること、この際、対になった増幅プライマーのうちの少なくとも１つは１つの鎖上に内部ｄＵを含むアダプタ配列を提示し、少なくとも１回の増幅を行い、（ｉｉｉ）適合ｃＤＮＡ配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する適合ｃＤＮＡ配列の集団を形成すること、（ｉｖ）一本鎖末端を有する適合ｃＤＮＡ配列の集団をリガーゼと接触させ、それにより線状アレイ核酸配列を形成すること、（ｖ）線状アレイ核酸配列から配列情報を得ること（任意選択で、配列は、ロングリード配列決定によって得られる）、及び（ｖｉ）線状アレイ核酸配列から得られた配列情報を分析して、アイソフォーム配列決定情報を得ること、を含み、それにより、入力ｃＤＮＡ配列の集団からアイソフォーム配列決定情報を得る。

【００２９】

本開示の別の態様は、入力ミトコンドリアｃＤＮＡ配列の集団からミトコンドリア系統追跡を行うための方法を提供し、方法は、（ｉ）複数の入力ミトコンドリアｃＤＮＡ配列を得ること、（ｉｉ）複数のミトコンドリアｃＤＮＡ配列を対になった増幅プライマーと接触させ、それによって適合ミトコンドリアｃＤＮＡ配列の集団を生成させること、この際、対になった増幅プライマーのうちの少なくとも１つが１つの鎖上に内部ｄＵを含むアダプタ配列を含み、少なくとも１ラウンドの増幅を行い、（ｉｉｉ）適合ミトコンドリアｃＤＮＡ配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する適合ミトコンドリアｃＤＮＡ配列の集団を形成すること、（ｉｖ）一本鎖末端を有する適合ミトコンドリアｃＤＮＡ配列の集団をリガーゼと接触させ、それによりアレイ核酸配列を形成すること、（ｖ）アレイ核酸配列から配列情報を取得すること（任意選択で、配列は、ロングリード配列決定によって得られる）

、及び（ｖｉ）線状アレイ核酸配列から得られた配列情報を分析してミトコンドリア系統を追跡すること、を含み、それによって入力ミトコンドリアｃＤＮＡ配列の集団に対してミトコンドリア系統追跡を実施する。本開示の更なる態様は、核酸配列の線状アレイのアレイを調製する方法を提供し、方法は、（ｉ）本明細書に開示されるＣＡｓｅｑ方法によって入力核酸配列の第１の集団から第１の線状アレイを調製すること、（ｉｉ）本明細書に開示されるＣＡｓｅｑ法によって入力核酸配列の第２の集団から第２の線状アレイを調製すること、この際、第１の線状アレイ及び第２の線状アレイがそれぞれ適合する相補的フランキング配列を有し、（ｉｉｉ）第１の線状アレイ及び第２の線状アレイを溶液中で組み合わせること、及び（ｉｖ）溶液中の第１の線状アレイ及び第２の線状アレイをリガーゼと接触させること、を含み、それにより、核酸配列の線状アレイのアレイを形成する。

【００３０】

ある実施形態では、第１の線状アレイ若しくは第２の線状アレイ、又はその両方は、線状アレイのアレイを含む。

【００３１】

いくつかの実施形態では、方法は更に、（ｖ）本明細書に開示されるＣＡｓｅｑ法によって入力核酸配列の第３の集団から第３の線状アレイを調製すること、この際、線状アレイ及び第３の線状アレイのアレイはそれぞれ、適合する相補的フランキング配列を有する、（ｖｉ）線状アレイ及び第３の線状アレイのアレイを溶液中で組み合わせること、及び、（ｖｉｉ）溶液中の線状アレイのアレイ及び第３の線状アレイをリガーゼと接触させること、を含み、それにより、核酸配列の線状アレイのより大きなアレイを形成する。任意選択的に、工程（ｖ）～（ｖｉｉ）は、第４の線状アレイ、第５の線状アレイ、及び／又はより多くの線状アレイを線状アレイのより大きなアレイに組み込むために繰り返される。

【００３２】

本開示の別の態様は、アレイ核酸配列を調製する方法を提供し、方法は、（ａ）複数の入力核酸配列を得ること、この際、各入力配列は、約３００キロベース以下の長さであり、（ｂ）複数の核酸配列を、一本の鎖上の内部ｄＵを含むアダプタ配列及びリガーゼと接触させ、それによって適合核酸配列の集団を生成すること、（ｃ）適合核酸配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する核酸配列の集団を形成すること、及び（ｄ）一本鎖末端を有する核酸配列の集団をリガーゼと接触させること、を含み、それによりアレイ核酸配列を形成する。

【００３３】

更なる態様では、本開示は、アレイ核酸配列を調製するための方法を提供し、方法は、（ｉ）複数の入力核酸配列を得ること、この際、各入力配列が約３００キロベース以下の長さである；（ｉｉ）複数の核酸配列を、１つの鎖上に内部ｄＵを有するアダプタ配列と接触させ、少なくとも１回の増幅を行い、それにより適合核酸配列の集団を生成すること；（ｉｉｉ）適合核酸配列の集団をウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩと接触させ、それにより一本鎖末端を有する核酸配列の集団を形成すること；及び（ｉｖ）一本鎖末端を有する核酸配列の集団をリガーゼと接触させること、を含み、それにより線状アレイ核酸配列を形成する。

【００３４】

実施形態では、複数の入力配列内の各入力核酸配列は、約３０キロベース以下の長さである。

【００３５】

本開示の更なる態様は、複数の核酸配列を含む組成物を提供し、複数の核酸配列の少なくとも２つは、配列番号１～１８から選択されるアダプタ配列を含む。

【００３６】

本開示の別の態様は、配列番号１～１８から選択される複数のアダプタ配列、及びその

使用説明書を含むキットを提供する。

【００３７】

本開示の更なる態様は、核酸配列リードの集団の個々の核酸配列リード内の別個の配列要素を同定するための方法を提供し、個々の核酸配列リードは、配列要素の線状アレイを有し、配列要素の線状アレイの各々は、高複雑度のライブラリから引き出された２つ以上の核酸配列要素を含み、高複雑度のライブラリから引き出された各核酸配列要素は、低複雑度のライブラリから引き出された１つ若しくは複数の予想される核酸配列の、又は低複雑度のライブラリから引き出された１つ若しくは複数の予想される核酸配列及び配列リード終端のいずれかに隣接（ｆｌａｎｋｅｄ）し、前記方法は：（ａ）核酸配列リードの集団の配列データに１つ又は複数の統計的アノテーションモデルを適用して、高複雑度のライブラリから引き出された個々の核酸配列要素の領域及び低複雑度のライブラリから引き出された核酸配列リードの領域を集団内で予測すること、この際、前記１つ又は複数の統計的アノテーションモデルは、ｉ）核酸配列リード全体に散在する１つ又は複数の予想される核酸配列を認識するための生成統計的アライメントモデル（ｇｅｎｅｒａｔｉｖｅ ｓｔａｔｉｓｔｉｃａｌ ａｌｉｇｎｍｅｎｔ ｍｏｄｅｌ）、及び、ｉｉ）既知ではない配列又は高複雑度の配列の辞書から引き出された配列を認識するためのランダム統計的アライメントモデル（ｒａｎｄｏｍ ｓｔａｔｉｓｔｉｃａｌ ａｌｉｇｎｍｅｎｔ ｍｏｄｅｌ）を含み、予測された転位部位は各モデルの末端に配置され、生成統計的アライメントモデルの内部位置内では許容されず；（ｂ）複数の核酸配列リードに対して工程（ａ）を繰り返し、それによって前記１つ又は複数の統計的モデルを複数の核酸配列リードの各核酸配列リードに順相補性配向及び逆相補性配向の両方で適用し、最大対数尤度値を有するモデルを同定することによって選択された最大事後状態経路の最終的リード当たりのモデル（ｍａｘｉｍｕｍ ａ ｐｏｓｔｅｒｉｏｒｉ ｓｔａｔｅ ｐａｔｈ Ｆｉｎａｌ ｐｅｒ－ｒｅａｄ ｍｏｄｅｌ）選択を決定すること；及び、（ｃ）複数の核酸配列リードの各核酸配列リードを、工程（ｂ）の最大事後状態経路の最終的リード当たりのモデル選択によって同定される転位部位によって区画された別個の配列要素にセグメント化すること、を含み、それにより核酸配列リードの集団内の別個の配列要素を同定する。

【００３８】

一実施形態では、高複雑度のライブラリは、１，０００を超える異なる要素を含むか、又は潜在的に含む。任意選択的に、高複雑度のライブラリは、１０，０００を超える異なる要素を含むか、又は潜在的に含む。

【００３９】

別の実施形態では、高複雑度のライブラリ及び／又は先験的に知られていない配列、又は高複雑度の配列の辞書から引き出された配列は、ｃＤＮＡ転写物配列、バーコード配列、及び／又は固有の分子識別子である要素を含む。

【００４０】

ある実施形態において、低複雑度のライブラリは、１００個以下の異なる配列を含む。任意選択で、低複雑度のライブラリは、５０個以下の異なる配列を含む。任意選択で、低複雑度のライブラリは、２５個以下の異なる配列を含む。任意選択で、低複雑度のライブラリは、１５個以下の異なる配列を含む。

【００４１】

いくつかの実施形態において、低複雑度のライブラリは、アダプタ及び／又はリンカー配列を含む。

【００４２】

実施形態では、先験的に予想される核酸配列は、アダプタ及び／又はリンカー配列を含む。

【００４３】

ある実施形態において、先験的に知られていない配列又は高複雑度の配列の辞書から引き出された配列は、以下のタイプの配列：ｃＤＮＡ配列、バーコード配列及び／又は固有の分子識別子（ｕｎｉｑｕｅ ｍｏｌｅｃｕｌａｒ ｉｄｅｎｔｉｆｉｅｒ）配列のうち

の１つ又は複数を含む。任意選択的に、バーコード配列は、単一細胞バーコード配列を含む。

【００４４】

本開示の別の態様は、複数の核酸配列リードの個々の配列リード内の別個の配列要素を同定し、配列要素データを保存するためのシステムを提供し、システムは、ネットワークと通信するための１つ又は複数のネットワークインターフェース；ネットワークインターフェースに結合され、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び、プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリを含み、プロセスは、実行されると、（ａ）配列要素の線状アレイを有する個々の核酸配列リードを含む複数の核酸配列リードを得るように、この際、配列要素の線状アレイを有する各リードは、高複雑度のライブラリから引き出された２つ以上の個々の核酸配列要素を含み、高複雑度のライブラリから引き出された各核酸配列要素は、低複雑度の１つ若しくは複数の予想される核酸配列に、又は低複雑度の１つ若しくは複数の予想される核酸配列及び配列リード終端のいずれかに隣接する；（ｂ）高複雑度のライブラリから引き出された個々の核酸配列要素の複数の領域及び低複雑度のライブラリから引き出された核酸配列の領域の核酸配列リード内で予測するために、１つ又は複数の統計的アノテーションモデルを複数の核酸配列リードの配列データに適用するように、この際、１つ又は複数の統計的アノテーションモデルは、ｉ）核酸配列リード全体に散在する１つ又は複数の予想される核酸配列を認識するための生成統計的アライメントモデル、及び、ｉｉ）既知ではない配列又は高複雑度の配列の辞書から引き出された配列を認識するためのランダム統計的アライメントモデル含み、予測された転位部位が各モデルの末端に配置され、生成統計的アライメントモデルの内部位置内では許容されない；（ｃ）複数の核酸配列リードに対して工程（ａ）を繰り返し、それにより、１つ又は複数の統計的モデルを複数の核酸配列リードの各核酸配列リードに順相補性配向及び逆相補性配向の両方で適用し、モデルを最大対数尤度値で同定することによって選択された最大事後状態経路の最終的リード当たりのモデル選択を決定し、それにより、核酸配列リード内の既知のセグメントを標識するように；及び、（ｄ）複数の核酸配列リードの各核酸配列リードを、工程（ｃ）の最大事後状態経路の最終的リード当たりのモデル選択によって同定される転位部位によって区画された（標識された既知のセグメントの）個別の配列要素にセグメント化し、それにより、複数の核酸配列リード内の個別の配列要素を同定するように；及び、（ｅ）複数の核酸配列リード内で同定された別個の配列要素を配列要素データファイルに保存するように、構成される。

【００４５】

本開示の更なる態様は、複数の核酸配列リードの個々の配列リードを低品質として識別し、除去し、配列データを保存するためのシステムを提供し、システムは、ネットワークと通信するための１つ又は複数のネットワークインターフェース；ネットワークインターフェースに結合され、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び、プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリ、を含み、プロセスが実行される場合、ｉ）複数の核酸配列リードの個々の配列リードに対して上記の工程（ａ）～（ｅ）を実施するように；ｉｉ）ライブラリ調製により予想される順序では生じない別個の配列要素を有する任意のリードを同定及び除去するように、この際、最初の別個の配列要素の後に始まるが、残りの別個の配列要素が順番になっているリード、及び最終的に予想される別個の配列要素の前に終わるが、前のセクションが全て順番になっているリード、並びにこれらの場合の組合わせは除去されず；及び、ｉｉｉ）低品質リードが除去された複数の核酸配列リードを配列データファイルに保存するように、構成される。

【００４６】

ある実施形態において、Ｃｉｒｃｕｌａｒ Ｃｏｎｓｅｎｓｕｓ Ｓｅｑｕｅｎｃｉｎｇソフトウェアが高品質であると識別した個々の配列リードは、この方法によって低品質であると識別される。

【００４７】
本開示の別の態様は、更なる分析のために十分に高い品質の個々の配列リードを同定し、複数の核酸配列リードの個々の配列リードを配列データに付加し、配列データを保存するためのシステムを提供し、システムは：ネットワークと通信するための１つ又は複数のネットワークインターフェース；ネットワークインターフェースに結合され、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び、プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリ、を含み、プロセスは実行される場合、ｉ）複数の核酸配列リードの個々の配列リードに対して上記の工程（ａ）～（ｅ）を実施し、最初に予想される別個の配列要素の後に始まるが、残りの別個の配列要素が順番になっているリード、及び最後に予想される別個の配列要素の前に終わるが、以前の別個の配列要素が順番になっているリード、並びにこれらの場合の任意の組合わせを含む、ライブラリ調製より出現すると予想される順序で別個の配列要素を有する任意のリードを、更なる分析のために十分に高品質であると識別するように；及び、ｖ）更なる分析のために十分に高品質であると識別された核酸配列リードを配列データファイルに保存するように、構成される。
【００４８】
ある実施形態において、Ｃｉｒｃｕｌａｒ Ｃｏｎｓｅｎｓｕｓ Ｓｅｑｕｅｎｃｉｎｇソフトウェアが低品質であると識別した個々の配列リードは、この方法によって高品質であると識別される。
【００４９】
本開示の最終態様は、新たに識別された高品質及び低品質リードの品質を概算し、推定品質スコアをデータに追加し、データを保存するためのシステムを提供し、システムは、ネットワークと通信するための１つ又は複数のネットワークインターフェース；ネットワークインターフェースに結合され、１つ又は複数のプロセスを実行するように構成されたプロセッサ；及び、プロセッサによって実行可能なプロセスを保存するように構成された非一時的メモリ、を含み、プロセスは、実行されると：（ｉ）各新しく識別された高品質又は低品質のリード内の各別個の配列要素について、別個の配列要素内のヌクレオチド間の観察されたアライメントスコア及び別個の配列要素に対する予想される配列を計算し、別個の配列要素内のヌクレオチドと別個の配列要素に対する予想される配列のヌクレオチドとの間の最良のアライメントスコアを計算するように；（ｉｉ）任意選択的に、工程（ｉ）で計算されたアライメントスコアを最良のアライメントスコアで除算して、各セクションの品質スコアを得るように；及び、（ｉｉｉ）工程（ｉ）で計算された全ての観察されたアライメントスコアを合計して、全体的な観察されたアライメントスコアを得、工程（ｉ）で計算された全ての最良の可能なアライメントスコアを合計して、全体的な最良のアライメントスコアを得；全体的な観察されたアライメントスコアと全体的な最良のアライメントスコアとの比を得ることによって、核酸配列リードの推定品質スコアを計算するように；及び、（ｉｖ）核酸配列リードについての推定品質スコアをデータファイルに保存するように、構成される。
【００５０】
ある実施形態では、アライメントスコアは、動的プログラミングアルゴリズムを直接使用して、又は別個の配列要素と予想される配列との間のレーベンシュタイン距離を計算し、その距離を予想される配列の長さから減算することによって直接、工程（ａ）で計算される。任意選択で、動的プログラミングアルゴリズムは、Ｓｍｉｔｈ－Ｗａｔｅｒｍａｎ（ローカル）アルゴリズム、Ｎｅｅｄｌｅｍａｎ－Ｗｕｎｓｃｈ（グローバル）アルゴリズム、又は類似／同等のアライメントアルゴリズム（例えば、ペア隠れマルコフモデル（Ｐａｉｒ Ｈｉｄｄｅｎ Ｍａｒｋｏｖ Ｍｏｄｅｌ））のうちの１つ又は複数を含む。
【００５１】
いくつかの実施形態において、最良のアライメントスコアは、予想される配列とそれ自体との間のアライメントスコアを計算することによって得られる。
【００５２】

定義
本明細書で使用される場合、特に明記されない限り、又は文脈から明らかでない限り、「約」という用語は、当技術分野における通常の許容範囲内、例えば平均の２標準偏差以内であると理解される。「約」は、記載された値の１０％、９％、８％、７％、６％、５％、４％、３％、２％、１％、０．５％、０．１％、０．０５％、又は０．０１％以内と理解することができる。
【００５３】
ある実施形態では、「およそ」又は「約」という用語は、特に明記しない限り、又は文脈から明らかでない限り（そのような数が可能な値の１００％を超える場合を除き）、記載された基準値のいずれかの方向（より大きい又はより小さい）において２５％、２０％、１９％、１８％、１７％、１６％、１５％、１４％、１３％、１２％、１１％、１０％、９％、８％、７％、６％、５％、４％、３％、２％、１％、又はそれ未満に入る値の範囲を指す。
【００５４】
文脈から明らかでない限り、本明細書で提供される全ての数値は、「約」という用語によって修飾される。
【００５５】
「対照」又は「参照」とは、比較の基準を意味する。対照試料を選択及び試験する方法は、当業者の能力の範囲内である。統計学的有意性の決定は、当業者の能力の範囲内であり、例えば、陽性結果を構成する平均からの標準偏差の数である。
【００５６】
本明細書で使用される場合、「異なる」という用語は、核酸に関して使用される場合、核酸が互いに同じではないヌクレオチド配列を有することを意味する。２つ以上の核酸は、それらの全長に沿って異なるヌクレオチド配列を有することができる。あるいは、２つ以上の核酸は、それらの長さのかなりの部分に沿って異なるヌクレオチド配列を有することができる。例えば、２つ以上の核酸は、２つ以上の分子について異なる標的ヌクレオチド配列部分を有することができるが、２つ以上の分子上で同じであるユニバーサル配列部分も有することができる。
【００５７】
本明細書で使用される場合、「各」という用語は、アイテムの集合に関して使用される場合、集合内の個々のアイテムを識別することを意図しているが、必ずしも集合内の全てのアイテムを指すとは限らない。明示的な開示又は文脈がそうでないことを明確に指示する場合、例外が発生する可能性がある。
【００５８】
本明細書で使用される場合、単一細胞核酸配列決定は、試料中の細胞又は他の種類の核酸の配列を測定し、その細胞及び／又は試料核酸が得られた個々の細胞及び／又は供給源を同定する方法を指す。同様に、単一細胞ＲＮＡ配列決定は、細胞ＲＮＡ（任意選択で、転写物）の配列を測定し、その細胞ＲＮＡが得られた個々の細胞を同定する方法を指す。
【００５９】
本明細書で使用される場合、「アンプリコン」という用語は、核酸に関して使用される場合、核酸を複製する産物を意味し、産物は、核酸のヌクレオチド配列の少なくとも一部と同じ又は相補的なヌクレオチド配列を有する。アンプリコンは、例えば、ポリメラーゼ伸長、ポリメラーゼ連鎖反応（ＰＣＲ）、ローリングサークル増幅（ＲＣＡ）、多重置換増幅（ＭＤＡ）、ライゲーション伸長、又はライゲーション連鎖反応を含む、核酸又はそのアンプリコンを鋳型として使用する様々な増幅方法のいずれかによって産生され得る。アンプリコンは、特定のヌクレオチド配列の単一コピー（例えば、ＰＣＲ産物）又はヌクレオチド配列の複数コピー（例えば、ＲＣＡのコンカテマー生成物）を有する核酸分子であり得る。標的核酸の第１アンプリコンは、典型的には相補的コピーである。後続のアンプリコンは、第１のアンプリコンの生成後に、標的核酸又は第１のアンプリコンから作製されるコピーである。後続のアンプリコンは、標的核酸に実質的に相補的であるか又は標

的核酸と実質的に同一である配列を有することができる。

【００６０】

本明細書で使用される場合、「アレイ」という用語は、相対的な位置によって互いに区別することができる特徴又は部位の集団を指す。アレイの異なる部位にある異なる分子は、アレイ内の部位の位置に応じて互いに区別することができる。アレイの個々の部位は、特定の種類の１つ又は複数の分子を含むことができる。例えば、部位は、特定の配列を有する単一の核酸分子を含むことができ、又は部位は、いくつかの核酸分子を含むことができる。ある実施形態では、「線状アレイ（ｌｉｅａｒ　ａｒｒａｙ）」という用語は、より大きな線状の核酸分子に沿ったアレイの別個の位置における配列要素の線状の集合体を指すために使用される。

【００６１】

本明細書で使用される場合、「バーコード配列」という用語は、核酸、核酸の特徴（例えば、同一性）、又は核酸に対して行われた操作を識別するために使用することができる核酸中の一連のヌクレオチドを意味することを意図している。バーコード配列は、天然に存在する配列又はバーコード化核酸が得られた生物に天然には存在しない配列であり得る。バーコード配列は、集団中の単一の核酸種に固有であり得るか、又はバーコード配列は、集団中のいくつかの異なる核酸種によって共有され得る。更なる例として、集団中の各核酸プローブは、集団中の他の全ての核酸プローブとは異なるバーコード配列を含むことができる。あるいは、集団中の各核酸プローブは、集団中のいくつか又はほとんどの他の核酸プローブからの異なるバーコード配列を含み得る。例えば、集団中の各プローブは、共通のバーコードを有するプローブがそれらの長さに沿った他の配列領域において互いに異なる場合であっても、集団中のいくつかの異なるプローブについて存在するバーコードを有することができる。特定の実施形態では、生物学的検体（例えば、組織試料）と共に使用される１つ又は複数のバーコード配列は、生物学的検体のゲノム、トランスクリプトーム又は他の核酸には存在しない。例えば、バーコード配列は、特定の生物学的検体中の核酸配列に対して８０％、７０％、６０％、５０％又は４０％未満の配列同一性を有し得る。

【００６２】

本明細書で使用される場合、「伸長する」という用語は、核酸に関して使用される場合、核酸への少なくとも１つのヌクレオチド又はオリゴヌクレオチドの付加を意味することを意図している。特定の実施形態では、１つ又は複数のヌクレオチドを、例えばポリメラーゼ触媒作用（例えば、ＤＮＡポリメラーゼ、ＲＮＡポリメラーゼ又は逆転写酵素）を介して核酸の３’末端に付加することができる。化学的又は酵素的方法を使用して、核酸の３’又は５’末端に１つ又は複数のヌクレオチドを付加することができる。１つ又は複数のオリゴヌクレオチドを、例えば、化学的又は酵素的（例えばリガーゼ触媒）方法によって、核酸の３’末端又は５’末端に付加することができる。核酸は、鋳型指向的に伸長することができ、それによって伸長産物は、伸長される核酸にハイブリダイズする鋳型核酸に相補的である。

【００６３】

本明細書で使用される場合、「逆転写酵素」という用語は、ＲＮＡ鋳型から相補的ＤＮＡ（ｃＤＮＡ）を生成するために使用される酵素を指す。当技術分野で一般的に使用される逆転写酵素（ＲＴ）には、非鎖置換転写酵素ＲＴＸ、及びウイルス逆転写酵素Ｍ－ＭＬＶが含まれる。

【００６４】

本明細書で使用される場合、「増幅する」、「増幅」又は「増幅反応」及びそれらの派生語は、一般に、核酸分子の少なくとも一部が少なくとも１つの更なる核酸分子に複製又はコピーされる任意の作用又はプロセスを指す。追加の核酸分子は、任意選択で、鋳型核酸分子の少なくとも一部と実質的に同一又は実質的に相補的な配列を含む。鋳型核酸分子は一本鎖又は二本鎖であり得、追加の核酸分子は独立して一本鎖又は二本鎖であり得る。増幅は、任意選択で、核酸分子の線状の又は指数関数的複製を含む。いくつかの実施形態

において、そのような増幅は、等温条件を用いて行うことができ、他の実施形態では、そのような増幅は熱サイクリングを含むことができる。いくつかの実施形態において、増幅は、単一の増幅反応における複数の標的配列の同時増幅を含む多重増幅である。増幅反応は、当業者に公知の増幅プロセスのいずれかを含むことができる。いくつかの実施形態では、増幅反応は、１つ又は複数の核酸配列を増幅するポリメラーゼ連鎖反応（ＰＣＲ）を含む。そのような増幅は、線状又は指数関数的であり得る。いくつかの実施形態では、増幅条件は、等温条件を含むことができ、あるいは熱サイクリング条件、又は等温条件と熱サイクリング条件との組合わせを含むことができる。いくつかの実施形態では、１つ又は複数の核酸配列を増幅するのに適した条件は、ポリメラーゼ連鎖反応（ＰＣＲ）条件を含む。典型的には、増幅条件は、ユニバーサル配列に隣接する１つ又は複数の標的配列等の核酸を増幅するために、又は１つ又は複数のアダプタに連結された増幅標的配列を増幅するために十分な反応混合物を指す。一般に、増幅条件は、増幅又は核酸合成のための触媒、例えばポリメラーゼ；増幅される核酸に対してある程度の相補性を有するプライマー；及び、核酸にハイブリダイズするとプライマーの伸長を促進するための、ヌクレオチド、例えばデオキシリボヌクレオチド三リン酸及びリボヌクレオチド三リン酸を含む。増幅条件は、プライマーの核酸へのハイブリダイゼーション又はアニーリング、プライマーの伸長、及び伸長されたプライマーが増幅を受ける核酸配列から分離される変性工程を必要とし得る。本明細書で使用される場合、「ポリメラーゼ連鎖反応」（「ＰＣＲ」）という用語は、目的のポリヌクレオチドのセグメントの濃度を増加させるための方法を記載している、Ｍｕｌｌｉｓの米国特許第４，６８３，１９５号及び同第４，６８３，２０２号の方法を指す。本明細書で使用される場合、「増幅標的配列」及びその派生語は、一般に、標的特異的プライマー及び本明細書で提供される方法を使用して標的配列を増幅することによって生成される核酸配列を指す。増幅された標的配列は、標的配列に関して同じセンス（すなわち、プラス鎖）又はアンチセンス（すなわち、マイナス鎖）のいずれかであり得る。

【００６５】

本明細書で使用される場合、「サーキュラーコンセンサスシーケンシング（Ｃｉｒｃｕｌａｒ　Ｃｏｎｓｅｎｓｕｓ　Ｓｅｑｕｅｎｃｉｎｇ）ソフトウェア低品質リード」という用語は、サーキュラーコンセンサスシーケンシングソフトウェアが０．９９未満のリード品質スコアを割り当てる配列決定リード、又はサーキュラーコンセンサスシーケンシングソフトウェアが「ＺＭＷパスフィルタ」以外のカテゴリにリードを割り当てるリードを指す。

【００６６】

本明細書で使用される場合、「サーキュラーコンセンサスシーケンシングソフトウェア高品質リード」という用語は、サーキュラーコンセンサスシーケンシングソフトウェアが「ＺＭＷパスフィルタ」カテゴリにリードを割り当てる配列リードを指す。ある実施形態では、ＣＣＳソフトウェア高品質リードは、ＣＣＳソフトウェアが０．９９以上のリード品質スコアを割り当てたリードである。

【００６７】

本明細書で使用される場合、「高複雑度のライブラリ」という用語は、特定のライブラリ要素が所与の位置に存在するかどうかの先験的な予測を統計的に不確実にする（例えば、所与の場所における特定のライブラリ要素の可能性は１％未満、所与の場所における特定のライブラリ要素の可能性は０．１％未満等である）のに十分な数の異なる要素（異なる配列、サイズ、長さ等を有する要素）を含むか、又は潜在的に含むライブラリを指す。ある実施形態では、「高複雑度のライブラリ」は、１００を超える別個の要素、任意選択で１０００を超える別個の要素、任意選択的に１０，０００を超える別個の要素、及び／又は任意選択的に１００，０００を超える別個の要素を含むか、又は潜在的に含む。実施形態では、「高複雑度のライブラリ」はｃＤＮＡ配列ライブラリ、任意選択でゲノムｃＤＮＡ配列ライブラリを指す。いくつかの実施形態において、「高複雑度のライブラリ」は、後の処理工程（例えば、バーコード配列（任意選択で、単一細胞バーコード配列、ビー

ズバーコード配列等）、固有の分子識別子等）において異なる検討に値するほど大きな配列の辞書から引き出されたライブラリを指す。

【００６８】

本明細書で使用される場合、「低複雑度のライブラリ」という用語は、特定のライブラリ要素が所与の位置に存在するかどうかの先験的予測を、限られた統計的不確実性のみで可能にするために（例えば、特定のライブラリ要素が所与の場所で発生する可能性は１％超、特定のライブラリ要素が所与の場所で発生する可能性は５％超、特定のライブラリ要素が所与の場所で発生する可能性は２０％超等である）、十分に少数の別個の要素（異なる配列、サイズ、長さ等を有する要素）を含むか、又は潜在的に含むライブラリを指す。ある実施形態では、「低複雑度のライブラリ」は、１００個未満の異なる要素、任意選択的に５０個未満の異なる要素、任意選択的に３０個未満の異なる要素、及び／又は任意選択的に１５個未満の異なる要素を含むか、又は潜在的に含む。実施形態では、「低複雑度のライブラリ」は、リンカー及び／又はアダプタ配列ライブラリを指す。

【００６９】

本明細書中で使用されるとき、用語「ライゲーションすること」、「ライゲーション」及びそれらの派生語は、一般に、２つ以上の分子を互いに共有結合的に連結するためのプロセス、例えば、２つ以上の核酸分子を互いに共有結合的に連結するためのプロセスのことを指す。いくつかの実施形態において、ライゲーションは、核酸の隣接するヌクレオチド間にニックをつなぐことを含む。いくつかの実施形態では、ライゲーションは、第１の核酸分子の末端と第２の核酸分子の末端との間に共有結合を形成することを含む。いくつかの実施形態では、ライゲーションは、１つの核酸の５’リン酸基と第２の核酸の３’ヒドロキシル基との間に共有結合を形成し、それによってライゲーションされた核酸分子を形成することを含み得る。一般に、本開示の目的のために、ライブラリ配列（任意選択で増幅されたライブラリ配列）をアダプタ配列にライゲーションして（又はそうでなければプライマー媒介増幅を介して付着させて）アダプタ連結配列を生成することができ、次いで、これを更に操作して、異なる配列要素を線状アレイ核酸に連結することができる。

【００７０】

本明細書で使用される場合、「リガーゼ」及びその派生語は、一般に、２つの基質分子のライゲーションを触媒することができる任意の薬剤を指す。いくつかの実施形態において、リガーゼは、核酸の隣接するヌクレオチド間のニックの連結を触媒することができる酵素を含む。いくつかの実施形態では、リガーゼは、１つの核酸分子の５’リン酸と別の核酸分子の３’ヒドロキシルとの間の共有結合の形成を触媒し、それによってライゲーションされた核酸分子を形成することができる酵素を含む。適切なリガーゼには、Ｔ４　ＤＮＡリガーゼ、Ｔ７　ＤＮＡリガーゼ、Ｔａｑ　ＤＮＡリガーゼ、及び大腸菌（Ｅ．ｃｏｌｉ）ＤＮＡリガーゼが含まれ得るが、これらに限定されない。

【００７１】

本明細書で使用される場合、「ライゲーション条件」及びその派生語は、一般に、２つの分子を互いに連結するのに適した条件を指す。

【００７２】

本明細書中で使用されるとき、用語「次世代配列決定」又は「ＮＧＳ」とは、従来の配列決定方法（例えば、標準的なサンガー又はマクサム－ギルバート配列決定法）を使用したときには前例のない速度でポリヌクレオチドを配列決定する能力を有する配列決定技術のことを指し得る。これらの前例のない速度は、数千から数百万の配列決定反応を並行して実行し、読み出すことによって達成される。ＮＧＳ配列決定プラットフォームとしては、限定されないが、以下が挙げられる：Ｍａｓｓｉｖｅｌｙ　Ｐａｒａｌｌｅｌ　Ｓｉｇｎａｔｕｒｅ　Ｓｅｑｕｅｎｃｉｎｇ（Ｌｙｎｘ　Ｔｈｅｒａｐｅｕｔｉｃｓ）；４５４　ｐｙｒｏ－ｓｅｑｕｅｎｃｉｎｇ（４５４　Ｌｉｆｅ　Ｓｃｉｅｎｃｅｓ／Ｒｏｃｈｅ　Ｄｉａｇｎｏｓｔｉｃｓ）；ｓｏｌｉｄ－ｐｈａｓｅ，ｒｅｖｅｒｓｉｂｌｅ　ｄｙｅ－ｔｅｒｍｉｎａｔｏｒ　ｓｅｑｕｅｎｃｉｎｇ（Ｓｏｌｅｘａ／Ｉｌｌｕｍｉｎａ（商標））；ＳＯＬｉＤ（商標）技術（Ａｐｐｌｉｅｄ　Ｂｉｏｓｙｓｔｅｍｓ）；Ｉｏｎ　ｓ

emiconductor sequencing（Ion Torrent（商標））；及びDNA nanoball sequencing（Complete Genomics）。あるNGSプラットフォームの説明は、以下に見出すことができる：Shendure, et al.,''Next-generation DNA sequencing,''Nature, 2008, vol. 26, No. 10, 135-1 145;Mardis,''The impact of next-generation sequencing technology on genetics,''Trends in Genetics, 2007, vol. 24, No. 3, pp. 133-141;Su, et al.,''Next-generation sequencing and its applications in molecular diagnostics''Expert Rev Mol Diagn, 2011, 11(3):333-43;及びZhang et al.,''The impact of next-generation sequencing on genomics'', J Genet Genomics, 201, 38(3):95-109。

【００７３】

本明細書で使用される場合、「核酸」及び「ヌクレオチド」という用語は、当技術分野でのそれらの使用と一致し、天然に存在する種又はその機能的類似体を含むことを意図している。核酸の特に有用な機能的類似体は、配列特異的な様式で核酸にハイブリダイズすることができるか、又は特定のヌクレオチド配列の複製のための鋳型として使用することができる。

【００７４】

天然に存在する核酸は、一般に、ホスホジエステル結合を含む骨格を有する。類似体構造は、当技術分野で公知の様々なもののいずれかを含む代替の骨格連結を有することができる。天然に存在する核酸は、一般に、デオキシリボース糖（例えば、デオキシリボ核酸（ＤＮＡ）に見られる）又はリボース糖（例えば、リボ核酸（ＲＮＡ）に見られる）を有する。核酸は、当技術分野で公知のこれらの糖部分の様々な類似体のいずれかを有するヌクレオチドを含むことができる。核酸は、天然又は非天然ヌクレオチドを含むことができる。これに関して、天然デオキシリボ核酸は、アデニン、チミン、シトシン又はグアニンからなる群から選択される１つ又は複数の塩基を有することができ、リボ核酸は、ウラシル、アデニン、シトシン又はグアニンからなる群から選択される１つ又は複数の塩基を有することができる。核酸又はヌクレオチドに含めることができる有用な非天然塩基は、当技術分野で公知である。「プローブ」又は「標的」という用語は、核酸又は核酸の配列に関して使用される場合、本明細書に記載の方法又は組成物の文脈における核酸又は配列の意味的識別子として意図され、核酸又は配列の構造又は機能を、他に明示的に示されるものを超えて必ずしも限定しない。

【００７５】

本明細書で使用される場合、「プライマー」という用語及びその派生語は、一般に、目的の標的配列にハイブリダイズすることができる任意の核酸を指す。典型的には、プライマーは、ポリメラーゼによってヌクレオチドを重合することができるか、又はインデックス等のヌクレオチド配列をライゲーションすることができる基質として機能するが、いくつかの実施形態では、プライマーは、合成された核酸鎖に組み込まれ、別のプライマーがハイブリダイズして、合成された核酸分子に相補的な新しい鎖の合成を開始することができる部位を提供することができる。プライマーは、ヌクレオチド又はその類似体の任意の組み合わせを含むことができる。いくつかの実施形態において、プライマーは、一本鎖オリゴヌクレオチド又はポリヌクレオチドである。「ポリヌクレオチド」及び「オリゴヌクレオチド」という用語は、任意の長さのヌクレオチドのポリマー形態を指すために本明細書で互換的に使用され、リボヌクレオチド、デオキシリボヌクレオチド、それらの類似体、又はそれらの混合物を含み得る。この用語は、等価物として、ＤＮＡ、ＲＮＡ又はｃＤＮＡのいずれかの類似体及び二本鎖ポリヌクレオチドを含むと理解されるべきである。本明細書で使用される用語はまた、例えば逆転写酵素の作用によってＲＮＡ鋳型から産生さ

れる相補的又はコピーＤＮＡであるｃＤＮＡを包含する。この用語は、分子の一次構造のみを指す。

【００７６】

例として与えられるが、説明されるある実施形態のみに本開示を限定することを意図するものではない以下の詳細な説明は、添付の図面と併せて最もよく理解され得る。

【図面の簡単な説明】

【００７７】

【図１】図１Ａ～１Ｃは、アイソフォーム配列決定を効果的に実行するための核酸リード長及びスループット要件を実証し、本明細書に開示される「ＣＡｓｅｑ」アプローチを提示するグラフを示す。図１Ａは、以前に記載された配列決定アプローチがアイソフォーム配列決定領域にギャップを残したことを実証するプロットを示す。具体的には、組み合わせたハイスループット（＞２０Ｍリード）及び中間リード長（０．５～５ｋｂ）配列決定アプローチは存在せず、本ＣＡｓｅｑアプローチは対処するために本明細書で提供されている。図１Ｂは、本明細書に開示される線状核酸アレイが、ロングリードプラットフォーム上で配列決定され、配列決定されたＤＮＡ分子の全出力をアレイあたりの断片の数に多重化して、それらの個々の全長ＤＮＡ断片に逆多重化され得ることを示す（現在のグラフに示すように３倍であるが、有効配列出力の１０倍以上の多重化を容易に達成することができる）。図１Ｃは、デオキシウラシル（ｄＵ）消化を用いて断片の協調的アセンブリを駆動する技術によって、アレイへのＤＮＡアンプリコンの制御された不偏ライゲーションがどのようにして本明細書において達成されたかの描写を示す。例示されるように、ＤＮＡライブラリを、５’「相補配列」とそれに続くｄＵとを含有するプライマーを用いて増幅する。増幅後、ｄＵ含有アンプリコンをウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩで消化すると、ｄＵが除去され、ＤＮＡの残りの上流鎖が融解し、それによって一本鎖「相補配列」が露出する。次いで、これらのｄＵ消化アンプリコンは、相補的な「相補配列」を含むアンプリコンとハイブリダイズして、標的化されたアセンブリを駆動することができる。アレイ長は、生成される「オーバーラップ配列」断片の数によって単純に変調される。

【図２】図２Ａ及び２Ｂは、本開示のＣＡｓｅｑプロセスを使用して、１．２ｋｂの平均断片サイズを有するｃＤＮＡライブラリからの８断片多重化アセンブリについて得られた結果を示す。図２Ａは、例示されたようなＣＡｓｅｑプロセスが、表示されたｃＤＮＡサイズ分布（開始、ライゲーション及び配列決定／逆多重化ｃＤＮＡ）に従って、ライゲーション時に約１０ｋｂの多重化断片をもたらしたことを示す。図２Ｂは、ＳｅｑｕｅｌＩＩで配列決定された多重化ライブラリについて得られた結果を示す図であり、これは、逆多重化後に約２３Ｍの転写物を伴う合計約２．５Ｍのリードをもたらし、これは、これまでに知られているアプローチを超えるスループットの約９倍の増加を表したことを示す。逆多重化されたリードの分析により、元のｃＤＮＡライブラリと同様のサイズ分布が確認された（図２Ａに見られるように）。

【図３】図３Ａ及び図３Ｂは、本開示のキメラアレイの完全な配列内容を、そのようなキメラアレイに存在する構造を利用する様式で解明することに関連する、ヒトゲノムにわたる遺伝子及び転写物の長さの分布を示す。図３Ａは、ヒトゲノムにわたるタンパク質コード遺伝子転写物（左側の緑色の点）及び遺伝子（黒色の点、右の分布）についてのカウント及び長さの分布を示す。ヒトタンパク質をコードする遺伝子転写物の大部分は１０ｋｂ未満の長さであり、事実上全てのタンパク質をコードする転写物は１００ｋｂ未満の長さであるが、遺伝子のかなり大部分は１０ｋｂの長さを超え、かなりの数の遺伝子は１００ｋｂの長さを超え、数は１Ｍｂの長さを超える。図３Ｂは、長さが増加するにつれて累積頻度をより明確に示すように表される、タンパク質コード遺伝子転写産物長（左側の緑色の点）及び遺伝子（黒色の点、右の分布）のヒトゲノムにおける累積分布（頻度）を示す。８０％のヒトタンパク質コード遺伝子転写物は、５０００塩基未満を含有すると特異的に認められた。

【図４】図４は、それぞれＳｐｉｋｅ－Ｉｎ　ＲＮＡバリアント（ＳＩＲＶ）で実施した

場合の、長鎖リード配列分析のための既存の「Ｓｍａｒｔ－ｓｅｑ３」プロセスと本開示のキメラアンプリコンアレイ配列決定分析との混同行列の比較を示す。ＳＩＲＶは、ヒト遺伝子と同様に選択的にスプライシングされる７つのＳＩＲＶ遺伝子（ＳＩＲＶ１～ＳＩＲＶ７）に分けられる。各遺伝子の転写物群は、四角で囲まれた領域によって示される。影付きの四角は、データ間の類似性を示す。対角線（左上から右下）は、ＳＩＲＶ転写物の自己類似性を示す。Ｓｍａｒｔ－ｓｅｑ３で生成されたデータは、各ＳＩＲＶ遺伝子の個々の転写物を区別することが困難であることが観察されたが、本開示のキメラアンプリコンアレイ配列決定法及び分析によって生成されたデータは、配列決定されたＳＩＲＶ転写物にほぼ完全にマッピングして戻された。

【図５】図５は、本開示のキメラアンプリコンアレイ配列決定方法及びヒトＴ細胞試料に対して行われた分析の全体的な収率のサンキーダイアグラムを示す。本開示の計算的逆多重化方法及び低品質リード再生方法を組み合わせたライブラリ調製は、既存のＣＣＳ補正ＨｉＦｉリードプロセス（すなわち、「Ｓｍａｒｔ－ｓｅｑ３」）のみを使用する方法と比較して、データ収率の全体的な２１．８５倍の増加をもたらした。

【図６】図６は、本開示のキメラアンプリコンアレイ配列決定法を用いて調製したヒトＴ細胞試料中のアダプタライゲーションのヒートマップを示す。カウントは、各列に示されるオーバーハングアダプタから各行に示されるオーバーハングアダプタまでのライゲーションの数を示す。逆方向に補完された配列は、’記号で示されている。この特定のライブラリでは、アレイサイズは１５であり、予想されるライゲーション順序はＡ－＞Ｂ－＞Ｃ－＞Ｄ－＞Ｅ－＞Ｆ－＞Ｇ－＞Ｈ－＞Ｉ－＞Ｊ－＞Ｋ－＞Ｌ－＞Ｍ－＞Ｎ－＞Ｏ－＞Ｐであった。対角線に沿った高いカウント（１つ下にシフト）は、調製されたライブラリ全体にわたって予想されるライゲーションの極めて高い割合を示す。中央の切れ目は、プロットが向きを切り替える場所である（逆相補ライゲーションを別々に示すため）。「ホット対角」上にない正方形のほとんどのカウントは０であり、予想外に検出されたライゲーションを示す正方形の最大のカウントでさえ、「ホット対角」のカウントよりも最大で３桁小さい。

【図７】図７は、予想されるライゲーション順序Ａ－＞Ｂ－＞Ｃ－＞Ｄ－＞Ｅ－＞Ｆ－＞Ｇ－＞Ｈ－＞Ｉ－＞Ｊ－＞Ｋ－＞Ｌ－＞Ｍ－＞Ｎ－＞Ｏ－＞Ｐでの長さ１５アレイライブラリ調製の上位２０のライゲーションプロファイル（有病率による）を示す。逆の相補的アダプタは’記号で示されている。これらのデータは、本明細書に現在開示されているキメラアレイの分析方法によって未だフィルタリングされていない。

【図８】図８は、２つのヒトＴ細胞試料にわたる、直接配列決定と、本開示のキメラアンプリコンアレイの配列決定方法及び分析の使用との比較を示す。

【図９】図９Ａ及び図９Ｂは、本開示の方法によって調製及び分析されたキメラアンプリコンアレイについての高品質及び低品質のアダプタライゲーションのヒートマップをそれぞれ示す。図９Ａは、本開示のキメラアンプリコンアレイ配列決定法を用いて調製したヒトＴ細胞試料中の高品質アダプタライゲーションのヒートマップを示す。カウントは、各列に示されるオーバーハングアダプタから各行に示されるオーバーハングアダプタまでのライゲーションの数を示す。逆方向に補完された配列は、’記号で示されている。この特定のライブラリでは、アレイサイズは１５であり、予想されるライゲーション順序はＡ－＞Ｂ－＞Ｃ－＞Ｄ－＞Ｅ－＞Ｆ－＞Ｇ－＞Ｈ－＞Ｉ－＞Ｊ－＞Ｋ－＞Ｌ－＞Ｍ－＞Ｎ－＞Ｏ－＞Ｐであった。高品質のデータを、本開示のキメラアンプリコンアレイ配列決定分析プロセス（「Ｌｏｎｇｂｏｗ」と呼ばれる）によって決定した。図９Ｂは、本開示のキメラアンプリコンアレイ配列決定法を用いて調製したヒトＴ細胞試料中の低品質アダプタライゲーションのヒートマップを示す。カウントは、各列に示されるオーバーハングアダプタから各行に示されるオーバーハングアダプタまでのライゲーションの数を示す。逆方向に補完された配列は、’記号で示されている。この特定のライブラリでは、アレイサイズは１５であり、予想されるライゲーション順序はＡ－＞Ｂ－＞Ｃ－＞Ｄ－＞Ｅ－＞Ｆ－＞Ｇ－＞Ｈ－＞Ｉ－＞Ｊ－＞Ｋ－＞Ｌ－＞Ｍ－＞Ｎ－＞Ｏ－＞Ｐであった。低品質のデータを、本開示のキメラアンプリコンアレイ配列決定分析プロセス（「Ｌｏｎｇｂｏｗ」）に

よって決定した。対角線上に生じないライゲーションは多数存在するが、低品質のデータであってもほぼ全てのライゲーションが予想通りに生じた。
【図１０】図１０Ａ～図１０Ｄは、ＣＯＶＩＤ－１９患者と健康な対照（ＨＣ）との間で行われた比較から得られた転写物データのクラスタリング評価を提示するｔ分布型確率的近傍埋め込み（ｔ－ｄｉｓｔｒｉｂｕｔｅｄ Ｓｔｏｃｈａｓｔｉｃ Ｎｅｉｇｈｂｏｒ Ｅｍｂｅｄｄｉｎｇ：ｔ－ＳＮＥ）プロットを示し、これにより、健康な患者と軽度及び重度のＣＯＶＩＤ－１９を有する患者との間の単球区画における著しい転写の違いが識別された。ｔ－ＳＮＥプロットは、健康な人及びＣＯＶＩＤ－１９患者の血液試料の評価から得られ、本明細書に開示のＣＡｓｅｑプロセスを介して得られた遺伝子アイソフォーム情報をショートリードデジタル遺伝子発現データにどのように補足できるかを実証している。図１０Ａは、表現型によってクラスター化されたｔ－ＳＮＥ分析プロットを示す。図１０Ｂは、試料によってクラスタリングされたｔ－ＳＮＥ分析プロットを示す。図１０Ｃは、ライデンクラスタリング（ｌｅｉｄｅｎ ｃｌｕｓｔｅｒｉｎｇ）を使用して実行されたｔ－ＳＮＥ分析のプロットを示す。図１０Ｄは、細胞タイプによってクラスター化されたｔ－ＳＮＥ分析プロットを示す。
【図１１】図１１Ａ～図１１Ｃは、末梢血単核球（ＰＢＭＣ）試料から得られた結果を示す。図１１Ａは、免疫細胞型を同定するために使用される、ＰＢＭＣ試料からの標準的なショートリード遺伝子発現データのクラスタリングの結果を示す。図１１Ｂは、同じ試料からの遺伝子（ショートリード）及びアイソフォーム（ロングリード）発現データの統合を示す。図１１Ｃは、図１１Ｂに示される遺伝子（ショートリード）及びイソ型（ロングリード）の発現データの統合により、カノニカルＣＤ４５（ＰＴＰＲＣ）アイソフォームの細胞型特異的発現が明らかにされたことを示す。
【図１２】図１２は、本開示のシステムを示す。
【図１３】図１３は、本開示の１つ又は複数の実施形態による最大の状態経路を決定するための例示的な手順を示す。
【発明を実施するための形態】
【００７８】
本開示は、少なくとも部分的には、核酸配列の入力集団に見出され得る不偏であり、及び／又はバイアスを最小化する方法で、ロングリード配列決定プラットフォームのスループット及び／又は収率を増強するための方法及び組成物に関する。したがって、ある態様において、特に、ロングリード配列決定プラットフォームを使用してキメラ核酸に対して核酸配列決定を行うための方法が提供される。ある実施形態において、本方法の核酸の線状キメラアレイは、ロングリード配列決定プラットフォームへの適用に有用である。そのような線状キメラアレイは、以前は不明瞭であった遺伝的特徴の解明、例えば選択的スプライシングの検出；腫瘍クローン進化等のクローン進化の改善された検出；例えば、疾患診断及び疾患病因の解明のための、ゲノム組成の忠実な再構成；体細胞モザイク現象の特徴付け；及びより一般的には改良されたゲノムハプロタイプ評価を可能にする。
【００７９】
本開示は、特に、ロングリードプラットフォームの固有の特徴を利用して、複数の共通配列決定ライブラリの出力を増強するための一般化可能なワークフローを提供する。ロングリードシーケンサは、非常に大きな配列決定出力を有するが（例えば、ＰａｃＢｉｏ（登録商標）Ｓｅｑｕｅｌ ＩＩは約３００ＧＢである）、ラン当たりのリードの総数は限られている（例えば、ＰａｃＢｉｏ（登録商標）Ｓｅｑｕｅｌ ＩＩは約４Ｍである）。出力を最大化するために、より小さい断片のライブラリをアレイにアセンブルし、ロングリードシーケンサで効率的に配列決定し、配列決定されたライブラリメンバーの数をアレイ中の断片の数に対して線状に増加させることができる。したがって、本開示のある態様は、単一細胞遺伝子発現試料からのハイスループット完全転写物配列決定を可能にするという本開示の主な利点を有する、高効率ロングリード配列決定のためのアレイの組立てのための合理化され、一般化可能な方法を詳述する。
【００８０】

近年、単一細胞遺伝子発現研究の劇的な増加が見られているが、そのような研究の注目すべき欠点は、これまで、そのような試みにおいてアイソフォーム組成又は遺伝的変異を解決することができなかったことである。ハイスループット単一細胞配列決定／発現分析における完全長転写物情報の捕捉における制限は、これらのワークフローにおけるハイスループットショートリード配列決定への依存を反映している。ショートリードアプローチは、転写産物の５’末端又は３’末端からの小さな約１００ｂｐのスナップショットを効果的に配列決定し、１×１０^{8}を超える転写産物から遺伝子カウントを効率的に取得するのに十分であるが、遺伝子アイソフォーム組成又は遺伝的変異を捕捉するには短すぎる（約５ｋｂ以上のリード長を必要とする）。ロングリード配列決定技術における印象的な最近の進歩があるが、それらのスループットは、単一細胞遺伝子発現試料から完全長転写物を適切にサンプリングするには依然として不十分である。したがって、ある態様において、本明細書中に提供されるのは、これらの制限を克服するための合理化された方法であり、この方法は、ある態様では、ロングリード配列決定プラットフォームのための核酸配列の精密に設計された線状アレイを作製することに依存し、それにより、本方法は、単一細胞遺伝子発現試料からのハイスループット完全転写物配列決定を可能にする。

【００８１】

上記のように、ＰａｃＢｉｏ（登録商標）及びＯｘｆｏｒｄ Ｎａｎｏｐｏｒｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ（「Ｎａｎｏｐｏｒｅ」）によって製造された２つの先駆的なロングリード配列決定プラットフォームにおける最近の著しい進歩は、ロングリード配列決定のリード長、スループット、及び精度を劇的に増加させ、単一細胞アイソフォーム配列決定の目標をほぼ手の届くところに置いた。最近の試みはこの２つのロングリード配列決定プラットフォーム（１～３）を活用してきたが、それらのワークフローは、大量のアーチファクト及びスループットの欠如に関連する著しい制限を受ける。これらの非効率性の合計は結果として、トランスクリプトーム内容物のスパース（ｓｐａｒｓｅ）サンプリングをもたらしており、これは今日まで、ロングリード配列決定分析の能力を厳しく制限してきた。例えば、Ｎａｎｏｐｏｒｅアイソフォーム配列決定法であるＲ２Ｃ２（Ｒｏｌｌｉｎｇ Ｃｉｒｃｌｅ Ａｍｐｌｉｆｉｃａｔｉｏｎ ｔｏ Ｃｏｎｃａｔｅｍｅｒｉｃ Ｃｏｎｓｅｎｓｕｓ）は、フィルタを通過する転写物の５２％しか達成しないことが観察されており、これはＮａｎｏｐｏｒｅフローセルあたり約３００，０００個の配列決定された転写物（約７９０ドル）に相当する（２）。ＰａｃＢｉｏ（登録商標）方法、ＳｃＩＳＯｒ－ｓｅｑも同様にアーチファクトによって制限されており、リードの約３６％のみがフィルタを通過し、ＰａｃＢｉｏ（登録商標）１Ｍフローセル当たり約３６０，０００個の全長転写物（約＄６４０）になる（１）。これらの欠点は、既知の配列決定技術（図１Ａ）間にこれまで存在していたギャップ、具体的にはハイスループット（＞２０Ｍリード）及び中間リード長（０．５～５ｋｂ）配列決定が存在しないことを強調している。本開示のある態様は、配列決定アーチファクトを＞９０％（図１Ａ）減少させながら、ロングリード配列決定プラットフォームのスループットを１０倍超増加させることができる方法、キメラアレイ配列決定（ＣＡｓｅｑ）を提供する。

【００８２】

本明細書に開示されるＣＡｓｅｑ方法は、これらのプラットフォームの固有の特徴に対処することによってロングリードシーケンサの分子配列決定出力を増強する特殊な多重化ワークフローである。特定のリード長さを有するＩｌｌｕｍｉｎａ（登録商標）のショートリード配列決定ワークフローとは対照的に、ロングリードプラットフォームは、フローセル中で約２０ｋｂから莫大な２Ｍｂ／ポア（ＭｉｎＩＯＮ，Ｏｘｆｏｒｄ Ｎａｎｏｐｏｒｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ）又はウェル（Ｓｅｑｕｅｌ ＩＩ、ＰａｃＢｉｏ（登録商標））までの範囲の不確定なリード長さを有する。これらの大量のリード長は、バルク全ゲノム配列決定等の試みには最適であるが、全長転写物等の中間の長さの標的（５００ｂｐ－１０ｋｂ）には過剰である。

【００８３】

個々のロングリード（図１Ａ）からの複数のＤＮＡ標的の配列決定を可能にするキメラ

アレイ配列決定（ＣＡｓｅｑ）は、中程度の長さの標的の拡張性のある捕捉のために、本明細書において、ロングリード配列決定プラットフォームをより良好に適合させるために開発された。本ＣＡｓｅｑ法では、ＤＮＡ断片の多重化は、マルチフラグメントアレイへの所定数の断片のプログラムされたライゲーションの制御されたプロセスを介して行われる。本明細書に開示される線状核酸アレイは、ロングリードプラットフォーム上で配列決定され、配列決定されたＤＮＡ分子の全出力をアレイあたりの断片の数に多重化して、それらの個々の全長ＤＮＡ断片に逆多重化され得る（図１Ｂ）。デオキシウラシル（ｄＵ）消化を用いて断片の協調的アセンブリを駆動する技術によって、アレイへのＤＮＡアンプリコンの制御された不偏ライゲーションが本明細書において達成される。簡潔には、ＤＮＡライブラリを、５’「相補配列（ｃｏｍｐｌｅｍｅｎｔ　ｓｅｑｕｅｎｃｅ）」とそれに続くｄＵとを含有するプライマーを用いて増幅する。増幅後、ｄＵ含有アンプリコンをウラシルＤＮＡグリコシラーゼ及びエンドヌクレアーゼＶＩＩＩで消化すると、ｄＵが除去され、ＤＮＡの残りの上流鎖が融解し、それによって一本鎖「相補配列」が露出する。次いで、これらのｄＵ消化アンプリコンは、相補的な「相補配列」を含むアンプリコンとハイブリダイズして、標的化されたアセンブリを駆動することができる。アレイ長は、生成される「オーバーラップ配列」断片の数によって単純に変調される（図１Ｃ）。ひとたび組み立てられると、これらの多重化断片は、その後の配列決定のための標準的なＮａｎｏｐｏｒｅ又はＰａｃＢｉｏ（登録商標）ライブラリの準備ワークフローに入ることができる。非常に長い又は分子密度の高いアレイを生成するために、アレイを互いに連結してアレイのアレイを作製するようにプログラムすることもできる。特に、相補的配列の最小セットを有する非常に大きい又は高密度の多重化アレイを生成するために、アレイ自体をアレイに連結することができることが明確に企図される。実際には、これは、内部相補的配列の共通のコアセットを有する多数の一次アレイを最初に生成することによって達成することができる。したがって、これらの一次アレイの隣接断片は、一次アレイ間のプログラムされたライゲーションを駆動する固有の相補的配列を含むように設計することができる（一次アレイの初期形成と同様）。

【００８４】

本明細書に開示されるＣＡｓｅｑプロセスは、限定するものではないが、（１）１０Ｘ　Ｇｅｎｏｍｉｃｓ（登録商標）のもの等の単一細胞遺伝子発現ワークフロー、例えば発現された核酸のバーコード化集団を構築し、任意選択でゲルビーズに分配することができるプロセス（例えば、ＰＣＴ／ＵＳ２０１８／１６０１９を参照されたい）、（２）１０Ｘ　Ｇｅｎｏｍｉｃｓ（登録商標）Ｖｉｓｉｕｍ　ｓｐａｔｉａｌ　ｇｅｎｏｍｉｃｓプロセス（特殊化された組織スライド上の捕捉領域内のスポットにグループ化された空間的にバーコード化されたｍＲＮＡ結合オリゴヌクレオチドを使用するＶｉｓｉｕｍ　Ｓｐａｔｉａｌ　Ｇｅｎｅ　Ｅｘｐｒｅｓｓｉｏｎ；ｍＲＮＡが処理された組織切片から放出されると、それは近傍の捕捉オリゴに結合し、次いで、これらの空間バーコードを組み込み、空間情報を保存するｃＤＮＡライブラリを、このｍＲＮＡから調製することができ、この遺伝子発現データは、その後、組織切片の高解像度顕微鏡画像上に重ねられ、どの遺伝子が発現されているか、及び組織試料全体のどこで発現されているかを視覚化することを可能にする）及び例えば米国特許第２０２１／０１２３０４０号に開示されている「Ｓｌｉｄｅ－Ｓｅｑ」空間トランスクリプトームプロファイリング手法等の空間配列決定ワークフロー、（３）ＣＡｓｅｑを使用する単一細胞遺伝子発現ワークフローから、例えば１０Ｘ　Ｇｅｎｏｍｉｃｓ（登録商標）試料からのミトコンドリア遺伝子の標的化増幅によって行うことができるミトコンドリア系統追跡、及び（４）とりわけ、Ｂ細胞受容体（ＢＣＲ）及びＴ細胞受容体（ＴＣＲ）の高効率の自然対ロングリード配列決定と組み合わせることができるＣＡｓｅｑを含む、任意の数の当技術分野で認識されている技術と組み合わせて使用することもできることが明確に企図される。

【００８５】

ある態様では、本開示のＣＡｓｅｑ法は、配列又はライブラリの偏りなしに、ＤＮＡ断片を制御可能かつ効率的に規定の断片番号のアレイに連結する能力を提供する。実施形態

では、本アプローチは、一方の鎖上に内部ｄＵを有する定義された配列（例えば、６～１６ｂｐ長であるが、他の配列長、例えば、５～２５ｂｐ又はそれを超える長さも実現性があると考えられる）で標的ＤＮＡの末端を修飾する（例えば、５’－Ｎ６－１６＿ｄＵ＿ｔａｒｇｅｔ－ＤＮＡ－３’）。配列の末端は、ウラシルＤＮＡグリコシラーゼ（ＵＤＧ）及びＤＮＡグリコシラーゼリアーゼエンドヌクレアーゼＶＩＩＩ（ＮＥＢ（登録商標）からのＵＳＥＲ酵素カクテル）を用いたｄＵの塩基切除によって一本鎖にされ、ハイブリダイゼーションのための定義された配列を明らかにする。これらの断片の複数のファミリーを作製及び処理して、ハイブリダイゼーション及びその後のライゲーションを指示することができる。次いで、長い配列断片をロングリードプラットフォームで配列決定することができ、それによって配列決定された分子のそれらの出力を増加させる。アレイ化された配列を調製するための現在の相補的配列媒介方法が本明細書において例示されているが、アレイを作製するための他の経路もまた、線状キメラアレイを作製するために使用され得ることが明確に企図され、例えば、ギブソンアセンブリ、重複伸長（例えば、遺伝子ＳＯＥ）等である。そのような用途のために、それぞれの反応に対する相補的末端配列を含有する増幅された断片がインキュベートされ、必要に応じて適切な条件でサイクルされ、それにより、キメラアレイが作製される。キメラ配列のアセンブリのために制限酵素を使用した、ロングリード核酸配列を作製するための１つの以前に開示されたアプローチも留意されるが、制限エンドヌクレアーゼ媒介アプローチは、現在のＣＡｓｅｑプロセスが克服する制限であるライブラリ多様性（Ｐｒａｂａｋａｒ　ｅｔ　ａｌ．，Ｇｅｎｏｍｅ　Ｂｉｏｌｏｇｙ　２０：１３４の「ＳＭＵＲＦ－ｓｅｑ」）の保持において有意な制限を示した。

【００８６】

本開示のＣＡｓｅｑプロセスは、配列決定の分野にわたって広範な適用性を有する。ゲノム配列決定のためには、リード長が長いほど配列再構成がより容易かつより正確になるので、リード長が重要である。ゲノムから０．５～２０ｋｂ断片を増幅し、次いで高効率ロングリード配列決定のためのアンプリコンアレイを生成する能力は、ゲノム再構成及びフェージングの精度及び忠実度を高める。ＣＡｓｅｑはまた、このアプローチがＤＮＡのより長い領域からのＳＮＰのフェージングを可能にするため、全エクソーム及び他の標的捕捉配列決定法にも有用である。更に、このＣＡｓｅｑは、本明細書の他の箇所で更に詳細に説明されるように、アイソフォームのＲＮＡ配列決定に適用可能である。ショートリードシーケンサは、従来のＲＮＡｓｅｑワークフローからアイソフォームを捕捉するのにあまり適していない。ロングリードの最近の試みはスループットが低く、したがってパワー不足である。本開示のＣＡｓｅｑプロセスは、ロングリード配列決定の出力を有意に増加させ、それによってＣＡｓｅｑを試料中のアイソフォーム組成を理解するための実行可能なアプローチ、特にアイソフォームｓｃＲＮＡｓｅｑにする。本開示のＣＡｓｅｑプロセスはまた、ＴＣＲα：ＴＣＲβ及びV_H：V_L対の天然にペアリングした（ｎａｔｉｖｅｌｙ　ｐａｉｒｅｄ）配列決定に有用であり、抗原特異的タグの組込みに適していると考えられる。例えば、本開示のＣＡｓｅｑプロセスは、全ゲノム及びエクソーム配列決定のためのＴＣＲ及びＩｇレパートリー並びにライブラリアセンブリのハイスループットによる天然にペアリングした配列決定のための既存のプロセスに適用することができる。具体的には、本開示のＣＡｓｅｑプロセスは、Ｔａｎｎｏ　ｅｔ　ａｌ．（Ｓｃｉｅｎｃｅ　Ａｄｖａｎｃｅｓ．６（１７）：ｅａａｙ９０９３；ＤＯＩ：１０．１１２６／ｓｃｉａｄｖ．ａａｙ９０９３）に記載されているように、現在のワークフローに代わるロングリード配列決定として提供される。Ｔａｎｎｏ　ｅｔ　ａｌ．は、ＴＣＲα：ＴＣＲβ対又はV_H：V_L対に対して実施されるエマルジョン系のＯｖｅｒｌａｐ　Ｅｘｔｅｎｓｉｏｎ　ＲＴ－ＰＣＲによって天然にペアリングした配列決定が達成され、それによってそれらを１つの天然にペアリングした断片にスティッチングする方法を記載している。本明細書では、例えば、そのようなペアリングしたアンプリコンのプールをＣＡｓｅｑワークフローの入力配列として使用することができ、それによってそのようなペアの拡張性のあるロングリード配列決定を可能にすることが特に企図される。更に、そのようなキ

メラアレイの設計中に他の断片を重複伸長ＲＴ－ＰＣＲに組み込むことができ、それにより、個々の細胞からのより多くの情報をそのようなＴＣＲα：ＴＣＲβ対及び／又はV_H：V_L対と対にし、そのようなアレイからの全ての配列情報の捕捉に必要なロングリード配列決定を与えることができると考えられる。

【００８７】

ある実施形態では、本開示のＣＡｓｅｑプロセスは、アレイに組み立てられるＤＮＡ分子を生成するための上流処理を最大化するように適合される。例としては、適切なアダプタを有するより大きなサイズの断片（０．５～２０ｋｂ）を生成するためのＤＮＡの断片化及び増幅の様式の最適化、断片化されたＤＮＡからの特定の配列の誘引、及び／又は標的化されたロングリード配列決定を可能にするためのＤＮＡ若しくはＲＮＡからの標的化された増幅が挙げられる。標的化ＤＮＡ又はＲＮＡは、標的核酸のパネルを使用して配列決定の試みを指示することができるので、特に有利であると考えられる。例えば、標的化は、ゲノムの特定の領域のフェージングに特別な注意を払うため、ゲノムの複雑な／反復的な特徴を解決するため、標的化アイソフォーム増幅のため、及び／又は本明細書の他の箇所でも論じられるように、単一細胞遺伝子発現／エピゲノム（ＡＴＡＣ）／ゲノム試料からの腫瘍ミトコンドリア系統追跡のために使用することができる。

【００８８】

本開示の特定の方法及び組成物の様々な明示的に企図される成分は、以下で更に詳細に考慮される。

【００８９】

核酸ライブラリ

本開示のＣＡｓｅｑプロセスは、ＲＮＡ、ｃＤＮＡ及びゲノムＤＮＡライブラリを含む任意の核酸ライブラリに効果的に適用することができる。本ＣＡｓｅｑ法を介して検出及び整列され得るＲＮＡとしては、ｍＲＮＡ、ｓｎＲＮＡ、ｌｎｃＲＮＡ、ｓｉＲＮＡ及びｇＲＮＡが挙げられ、現在のアプローチでは、ＣＡｓｅｑプロセスを介した整列及び配列決定のために、そのようなＲＮＡの安定化された形態及び／又は対応するＤＮＡ配列を任意選択で使用／産生する。

【００９０】

プライマー／アダプタ

本ＣＡｓｅｑプロセスの例示された態様では、アダプタ配列を入力核酸集団に付着させるためにテールプライマー（ｔａｉｌｅｄ　ｐｒｉｍｅｒ）が使用される。使用されるアダプタ配列は、最終的に、個々の入力核酸配列の一本鎖「粘着末端」のアニーリングを介してキメラアレイライゲーションを進行させることを可能にし、それぞれ末端に１つ又は２つのアダプタ配列が互いに結合している。任意選択で、アダプタ配列内の相補的一本鎖配列の設計は、各キメラアレイが正確な線形順序を有するように行うことができ、又はアダプタ配列の使用は、各キメラアレイ内の線形順序のより大きな柔軟性を可能にし得る。ある例示的な実施形態では、多重ライゲーションのために、１５塩基対（ｂｐ）の相補的配列を増幅し、全長ｃＤＮＡライブラリに付加するためのｄＵ含有プライマーのファミリーが設計されている。アーチファクト配列の主要な供給源に対処するために、例示されたプロセスは、全長ｃＤＮＡアンプリコンの精製を可能にするためにビオチン化プライマーを使用した。効率的な多重化アセンブリを駆動し、不適切なライゲーション事象を軽減するために、本明細書に例示される１５ｂｐ相補的配列は、全ての配列が互いに少なくとも１１ハミング距離単位離れていることを確実にすることによって、最小の類似性を有するように設計された。そのような品質を有するアダプタ配列の例示的な表を以下の表１に示す。

【表１】

表１－使用されるアダプタ配列の例示的なリスト

配列番号	相補的配列
1	AGCACCATAATGTGT
2	CTTGTAAGCTGTCTA
3	CTCTGTCAGGTCCGA
4	CCTCCTCCTCCAGAA
5	TCGCTGGTATTCCAA
6	GCTTACTTGTGAAGA
7	TAACCGTATGGTTGA
8	GATGGCGCTATCTCA
9	CTACCAGTGAGGAAG
10	GAGTCCAATTCGCAG
11	ATCAAGGCTTAACGG
12	TGTTGAATCCTAGCG
13	GTGCGTTGCGAATTG
14	CGGTAATGTACCGGC
15	ATTGCGTAGTTGGCC
16	CACTTGGTCGCAATC
17	GTAAGCCTTCGTGTC
18	CCTAGATCAGAGCCT

【００９１】

ＣＡｓｅｑプロセスにおける入力配列へのアダプタ配列の付加は、テール増幅プライマーを使用して本明細書で例示されているが、アダプタ配列を入力配列の集団に付加するための他の当技術分野で認識されている方法も使用することができることが明確に企図されている。例えば、特に断片の増幅（例えば、長さに、又は修正を維持することに起因する）を回避することが有利な場合、線状アレイの構築のための本明細書に開示されるＣＡｓｅｑプロセスの残りの部分の実施前に、入力配列（例えば、平滑末端入力配列）へのアダプタの直接ライゲーションを行うことができる。

【００９２】

入力核酸の長さ（例えばｃＤＮＡ）

入力核酸配列の長さは、本開示の具体的な用途に応じて、サイズの範囲が広くなり得る。入力核酸としてのｃＤＮＡ集団の場合、長さは一般に０．５ｋｂ～２０ｋｂに分布する。しかしながら、本方法は、２０ヌクレオチド以下という短い入力核酸配列長に、又は最大約メガベース以上の長さを有する入力核酸配列／断片に適用することができることが明確に企図される。実際、本開示のＣＡｓｅｑ法は、例えば、ＣＩＴＥｓｅｑタグ又は他の生物学的に関連する情報等のライブラリからの捕捉のために、１００ｂｐ未満の小さい断片に適用することができることが明確に企図される。上記のように、本開示のＣＡｓｅｑプロセスは、約３５０ｂｐ～１０ｋｂの標準サイズｃＤＮＡにも適用することができる。更に、ロングリード配列決定長が増加し続けると、ＣＡｓｅｑを適用して、多くの大きな（＞１０ｋｂ）核酸配列／断片の線状アレイを作製できることが明確に企図される。

【００９３】

ウラシルＤＮＡグリコシラーゼ

本開示のある態様は、ウラシルＤＮＡグリコシラーゼを使用する。ウラシル－ＤＮＡグリコシラーゼ（ＵＤＧ）は、ＤＮＡの突然変異を復帰させる酵素である。最も一般的な突然変異は、ウラシルへのシトシンの脱アミノ化である。ＵＤＧはこれらの突然変異を修復

する。ＵＤＧはＤＮＡ修復において非常に重要であるが、それがなければ、これらの突然変異はがんをもたらし得る（Ｐｅａｒｌ，ＬＨ．Ｍｕｔａｔ　Ｒｅｓ．４６０：１６５－８１）。

【００９４】

既知のウラシル－ＤＮＡグリコシラーゼ及び関連するＤＮＡグリコシラーゼ（ＥＣ）としては、ウラシル－ＤＮＡグリコシラーゼ（Ｍｏｌ　ｅｔ　ａｌ．Ｃｅｌｌ．８０：８６９－７８）、好熱性ウラシル－ＤＮＡグリコシラーゼ（Ｓａｎｄｉｇｕｒｓｋｙ　ａｎｄ　Ｆｒａｎｋｌｉｎ．Ｃｕｒｒ．Ｂｉｏｌ．９：５３１－４）、Ｇ：Ｔ／Ｕミスマッチ特異的ＤＮＡグリコシラーゼ（Ｍｕｇ）（Ｂａｒｒｅｔｔ　ｅｔ　ａｌ．Ｃｅｌｌ．９２：１１７－２９）、及び一本鎖選択的単機能性ウラシル－ＤＮＡグリコシラーゼ（ｓｉｎｇｌｅ－ｓｔｒａｎｄ　ｓｅｌｅｃｔｉｖｅ　ｍｏｎｏｆｕｎｃｔｉｏｎａｌ　ｕｒａｃｉｌ－ＤＮＡ　ｇｌｙｃｏｓｙｌａｓｅ）（ＳＭＵＧ１；Ｂｕｃｋｌｅｙ　ａｎｄ　Ｅｈｒｅｎｆｅｌｄ．Ｊ．Ｂｉｏｌ．Ｃｈｅｍ．２６２：１３５９９－６０６）が挙げられる。

【００９５】

ウラシルＤＮＡグリコシラーゼは、シトシンの自発的な脱アミノ化によって又はＤＮＡ複製中のｄＡとは反対のｄＵの誤組込みによってのいずれかにより生じ得るＤＮＡからウラシルを除去する。このファミリーの原型メンバーは、最初に発見されたグリコシラーゼの１つである大腸菌（Ｅ．ｃｏｌｉ）ＵＤＧである。ＵＮＧ、ＳＭＵＧ１、ＴＤＧ、及びＭＢＤ４を含む４つの異なるウラシル－ＤＮＡグリコシラーゼ活性が哺乳動物細胞で同定されている。それらは、基質特異性及び細胞内局在性が異なる。ＳＭＵＧ１は、基質として一本鎖ＤＮＡを好むが、二本鎖ＤＮＡからもＵを除去する。非修飾ウラシルに加えて、ＳＭＵＧ１は、環Ｃ５に酸化基を有する５－ヒドロキシウラシル、５－ヒドロキシメチルウラシル及び５－ホルミルウラシルを切除することができる（Ｍａｔｓｕｂａｒａ　ｅｔ　ａｌ．Ｎｕｃｌｅｉｃ　Ａｃｉｄｓ　Ｒｅｓ．３２：５２９１－５３０２）。ＴＤＧ及びＭＢＤ４は、二本鎖ＤＮＡに厳密に特異的である。ＴＤＧは、対向するグアニンが存在する場合、チミングリコール、並びに炭素５に修飾を有するＵの誘導体を除去することができる。現在の証拠は、ヒト細胞では、ＴＤＧ及びＳＭＵＧ１が、自発的なシトシン脱アミノ化によって引き起こされるＵ：Ｇミスペアの修復に関与する主要な酵素であることを示唆しているが、ｄＵの誤組込みによってＤＮＡに生じるウラシルは主にＵＮＧによって処理される。ＭＢＤ４は、ＣｐＧ部位における５－メチルシトシンのチミンへの脱アミノ化から生じるＴ：Ｇミスマッチを修正すると考えられている（Ｗｕ　ｅｔ　ａｌ．Ｊ．Ｂｉｏｌ．Ｃｈｅｍ．１４：５２８５－５２９１．）。ＭＢＤ４変異マウスは正常に発達し、癌感受性の増加又は生存率の低下を示さない。しかし、それらは小腸の上皮細胞のＣｐＧ配列でより多くのＣ　Ｔ変異を獲得する（Ｗｏｎｇ　ｅｔ　ａｌ．ＰＮＡＳ．９９：１４９３７－１４９４２）。制限酵素を使用して（相補的な末端配列を他の断片とアニーリングすることによって）キメラアレイを調製することができることが更に企図される。しかしながら、ＣＡｓｅｑプロセスにおける制限酵素の使用は、特定の断片の消化を介してライブラリを偏らせる可能性が非常に高い。

【００９６】

エンドヌクレアーゼＶＩＩＩ

本開示の例示されたある態様は、エンドヌクレアーゼＶＩＩＩ酵素を使用する。大腸菌（Ｅ．ｃｏｌｉ）由来のエンドヌクレアーゼＶＩＩＩは、Ｎ－グリコシラーゼ及びＡＰ－リアーゼの両方として作用する。Ｎ－グリコシラーゼ活性は、損傷したピリミジンを二本鎖ＤＮＡから放出し、アプリン（ＡＰ部位）を生成する。ＡＰ－リアーゼ活性は、ＡＰ部位に対し３’及び５’を切断して、５’リン酸及び３’リン酸を残す。エンドヌクレアーゼＶＩＩＩによって認識され除去される損傷を受けた塩基には、尿素、５，６－ジヒドロキシチミン、チミングリコール、５－ヒドロキシ－５－メチルヒダントイン、ウラシルグリコール、６－ヒドロキシ－５，６－ジヒドロチミン及びメチルタルトロニル尿素が含まれる。エンドヌクレアーゼＶＩＩＩはエンドヌクレアーゼＩＩＩと類似しているが、エンドヌクレアーゼＶＩＩＩはβ及びδリアーゼ活性を有し、エンドヌクレアーゼＩＩＩはβリア

ーゼ活性のみを有する。

【００９７】

リガーゼ

ある態様において、アダプタのオーバーハング末端がＣＡｓｅｑプロセスにおいて互いにアニールすると、リガーゼが投与されて、キメラアレイ要素を固定し、要素を線状に取り付ける。リガーゼは、一般に、新しい化学結合を形成することによって、通常、大きな分子のうちの１つの上の小さなペンダント基（ｐｅｎｄａｎｔ ｃｈｅｍｉｃａｌ ｇｒｏｕｐ）の加水分解を伴って、２つの大きな分子の結合を触媒することができる酵素、又は２つの化合物の一緒に結合することを触媒する酵素、例えば、Ｃ－Ｏ、Ｃ－Ｓ、Ｃ－Ｎ等の結合を触媒する酵素を指す。一般に、リガーゼは以下の反応、Ａｂ＋Ｃ→Ａ－Ｃ＋ｂ又は場合によってはＡｂ＋ｃＤ→Ａ－Ｄ＋ｂ＋ｃ＋ｄ＋ｅ＋ｆを触媒し、小文字はその従属する小さい基を示し得る。リガーゼは、核酸の２つの相補的断片を結合し、複製中に二本鎖ＤＮＡに生じる一本鎖切断を修復することができる。一般的に使用されるリガーゼには、とりわけ、Ｔ４ ＤＮＡリガーゼ、Ｔ７ ＤＮＡリガーゼ、Ｔａｑ ＤＮＡリガーゼ、及び大腸菌（Ｅ．ｃｏｌｉ）ＤＮＡリガーゼが含まれるが、これらに限定されない。

【００９８】

ロングリード配列決定プラットフォーム

本開示のある態様は、ロングリード配列決定を使用する核酸の調製を使用するか、又は含む。ロングリード配列決定（ＬＲＳ）は、現在活発に開発されているＤＮＡ配列決定方法の一種である（Ｂｌｅｉｄｏｒｎ，Ｃｈｒｉｓｔｏｐｈ．Ｓｙｓｔｅｍａｔｉｃｓ ａｎｄ Ｂｉｏｄｉｖｅｒｓｉｔｙ １４：１－８）。ロングリード配列決定は、ＤＮＡの長い鎖を小さなセグメントに切断し、次いで増幅及び合成によってヌクレオチド配列を推測することを必要とする既存の方法とは対照的に、単一分子レベルでヌクレオチド配列を読み取ることによって機能する（「Ｉｌｌｕｍｉｎａ配列決定技術」ＰＤＦ）。

【００９９】

上で定義したＮＧＳは、その開発以来、ＤＮＡ配列決定分野で影響を及ぼしてきた。これにより、ゲノム全体にわたって非常に高いカバレッジで多数のリードをもたらすことができる超並列アプローチを可能にし、ＤＮＡ配列決定のコストを劇的に削減した（Ｔｒｅａｎｇｅｎ ａｎｄ Ｓａｌｚｂｅｒｇ．Ｎａｔｕｒｅ Ｒｅｖｉｅｗｓ Ｇｅｎｅｔｉｃｓ １３：３６－４６）。

【０１００】

ＮＧＳは、最初にＤＮＡ分子を増幅し、次いで合成によって配列決定を行うことによって機能する。多数の増幅された同一のＤＮＡ鎖の合成の結果得られる集合的な蛍光シグナルは、ヌクレオチド同一性の推論を可能にする。しかしながら、ランダムエラーのために、増幅されたＤＮＡ鎖間のＤＮＡ合成は、次第に同期しなくなる。急速に、信号品質は、リード長が増大するにつれて劣化する。リード品質を維持するためには、長いＤＮＡ分子を小さなセグメントに分割しなければならず、ＮＧＳ技術の重大な制限をもたらす（Ｔｒｅａｎｇｅｎ及びＳａｌｚｂｅｒｇ）。この課題を克服するための計算の試みは、正確なアセンブリをもたらさない可能性がある近似的なヒューリスティックスに依存することが多い。

【０１０１】

単一のＤＮＡ分子の直接配列決定を可能にすることによって、ロングリード配列決定技術は、第２世代配列決定よりも実質的に長いリードを生成する能力を有する（Ｂｌｅｉｄｏｒｎ）。このような利点は、ゲノム科学及び生物学全般の研究の両方に重大な意味を有する。しかしながら、ロングリード配列決定データは、以前の技術よりもはるかに高いエラー率を有し、下流ゲノム構築及び得られたデータの分析を複雑にする可能性がある（Ｇｕｐｔａ．Ｔｒｅｎｄｓ ｉｎ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ ２６：６０２－６１１）。これらの技術は活発に開発されており、高いエラーレートの改善が期待されている。構造的変異体コーリング等のエラーレートに対してより耐性がある用途では、ロングリード配列決定が既存の方法よりも優れていることが見出されている。

【０１０２】
いくつかの企業、すなわちＰａｃｉｆｉｃ Ｂｉｏｓｃｉｅｎｃｅｓ、Ｏｘｆｏｒｄ Ｎａｎｏｐｏｒｅ Ｔｅｃｈｎｏｌｏｇｙ、Ｑｕａｎｔａｐｏｒｅ（ＣＡ－ＵＳＡ）、及びＳｔｒａｔｏｓ（ＷＡ－ＵＳＡ）が現在、ロングリード配列決定技術開発の中心にある。これらの企業は、単一のＤＮＡ分子を配列決定するために根本的に異なるアプローチを取っている。
【０１０３】
ＰａｃＢｉｏ（登録商標）は、ゼロモード導波路の特性に基づいて、単一分子リアルタイム配列決定（ＳＭＲＴ）の配列決定プラットフォームを開発した。シグナルは、ｚＬウェルの底部に結合したＤＮＡポリメラーゼによって組み込まれた各ヌクレオチドからの蛍光発光の形態である。本明細書で使用されるＰａｃＢｉｏ（登録商標）ロングリード配列決定プラットフォームの現在の例は、ＳｃＩＳＯｒ－ｓｅｑである。
【０１０４】
Ｏｘｆｏｒｄ Ｎａｎｏｐｏｒｅの技術は、ＤＮＡ分子をナノスケール細孔構造に通過させ、次いで細孔を取り囲む電場の変化を測定することを含み、一方、Ｑｕａｎｔａｐｏｒｅは異なる独自のナノポアアプローチを有する。Ｓｔｒａｔｏｓ Ｇｅｎｏｍｉｃｓは、ナノポアｓｓＤＮＡ読取りのノイズチャレンジに対するシグナルを回避するために、ポリマーインサート「Ｘｐａｎｄｏｍｅｒｓ」を用いてＤＮＡ塩基を離間させる。Ｒ２Ｃ２（Ｒｏｌｌｉｎｇ Ｃｉｒｃｌｅ Ａｍｐｌｉｆｉｃａｔｉｏｎ ｔｏ Ｃｏｎｃａｔｅｍｅｒｉｃ Ｃｏｎｓｅｎｓｕｓ）は、例示的なナノポアアイソフォーム配列決定方法として注目されている。
【０１０５】
ある実施形態では、ナノポア配列決定が使用される（例えば、参照により組み込まれるＡｓｔｉｅｒ ｅｔ ａｌ．，Ｊ．Ａｍ．Ｃｈｅｍ．Ｓｏｃ．２００６ Ｆｅｂ ８；１２８（５）：１７０５－１０を参照されたい）。ナノポア配列決定の背後にある理論は、ナノポアが導電性流体に浸漬され、それに電位（電圧）が印加されたときに起こるものと関係がある。これらの条件下では、ナノポアを通るイオンの伝導によるわずかな電流を観察することができ、電流量はナノポアのサイズに非常に敏感である。核酸の各塩基がナノポアを通過すると（又はエキソヌクレアーゼベースの技術の場合には個々のヌクレオチドがナノポアを通過すると）、これにより、４つの塩基のそれぞれについて異なるナノポアを通る電流の大きさが変化し、それによってＤＮＡ分子の配列を決定することが可能になる。
【０１０６】
本開示のある態様は、１つ又は複数のｄＵ残基で終結し、それぞれの配列要素の線状タンデムアレイを調製するために使用することができる別個の相補的配列を有するように設計された特殊なオリゴヌクレオチドプライマーを使用するが、追加の核酸プライマー／配列／アダプタも本開示の核酸ライブラリに付加することができることも企図される。そのような明示的に企図される更なるプライマー／配列／アダプタとしては、他の識別子及び／又はアダプタ配列の中でも、例えば、ＣＩＴＥ－Ｓｅｑプロセス（Ｓｔｏｅｃｋｉｕｓ ｅｔ ａｌ．Ｎａｔｕｒｅ Ｍｅｔｈｏｄｓ．１４：８６５－８６８）、ＲＥＡＰ－Ｓｅｑプロセス（Ｐｅｔｅｒｓｏｎ ｅｔ ａｌ．Ｎａｔｕｒｅ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ．３５：９３６－９３９）又は他のプロセス、Ｓｍｉｔｈ ｅｔ ａｌ．（Ｓｍｉｔｈ，Ａ．Ｍ．Ｇｅｎｏｍｅ Ｒｅｓｅａｒｃｈ １９：１８３６－１８４２）及び他の場所で用いられているもの等の、固有の分子識別子（ＵＭＩ）において使用されるもの等の配列バーコードが挙げられるが、これらに限定されない。そのような配列は、ＣＡｓｅｑプロセスのライゲーション工程の前の任意の時点でライブラリ配列に任意選択的に付加することができ、これにより、それぞれの線状キメラアレイ配列要素の順序が、ロングリード配列決定の実施に先立って固定される。
【０１０７】
バーコード配列及び他の識別配列は、任意の様々な長さであり得る。より長い配列、例

えば、本ＣＡｓｅｑプロセスによって調製されたものは、一般に、集団に対するより多くの数及び多様なバーコードを収めることができる。一般に、キメラアレイ内の複数の個々の要素は、（異なる配列を有するにもかかわらず）同じ長さのバーコードを有するが、単一のアレイの異なる要素に対して、又は異なるＣＡｓｅｑロングリード配列に対して異なる長さのバーコードを使用することも可能である。バーコード配列は、少なくとも２、４、６、８、１０、１２、１５、２０又はそれを超えるヌクレオチド長であり得る。代替的又は追加的に、バーコード配列の長さは、最大で２０、１５、１２、１０、８、６、４又はそれ未満のヌクレオチドであり得る。使用することができるバーコード配列の例は、例えば、それぞれ参照により本明細書に組み込まれる米国特許出願公開第２０１４／０３４２９２１号及び米国特許第８，４６０，８６５号に記載されている。

【０１０８】

本開示のあるオリゴヌクレオチドはまた、更なるリンカー（任意選択で切断可能なリンカー）、プライミング部位（当該技術分野で知られているように、例えば、国際公開第２０１６／０４０４７６号を参照されたい）ごとに異なる固有の分子識別子（ＵＭＩ）、上記のバーコード配列、及び任意選択で、ＰＣＲ増幅を可能にするための共通配列（「ＰＣＲハンドル」）を含むことができると考えられる。

【０１０９】

単一細胞配列決定／分子プロファイリング

単一細胞（ＳＣ）分子プロファイリング法は、そのような方法が最近主流に変遷し、ＦＡＣＳのような既存のＳＣ感受性アプローチと共に変遷しているため、生物医学研究にすでに大きな影響を与えている。飛躍的な進歩や急速な進歩により、多くの「オミクス」（すなわち、ゲノミクス、プロテオミクス、トランスクリプトミクス等である）レベルでのＳＣ解像度が可能になった。技術の飛躍的な進歩がＳＣ分子プロファイリングの性能及びコストの改善を推進し、それ以前の次世代配列決定（ＮＧＳ）と同様に、ＳＣ解析は現在、患者ケア及び製薬研究に直接適用されることが増えている。

【０１１０】

配列分析及びシステム

本開示は、本明細書で同定されるキメラアンプリコンアレイだけでなく、提供される方法を実施するためのコンピュータ及びシステムも包含する。

【０１１１】

試料を得るための一般的な方法、配列決定リードを生成するための一般的な方法、及び本開示を実施するために有用な様々なタイプの配列決定がここで記載される。これらの例示的な方法は限定的ではなく、当業者によって必要に応じて変更され得ることが理解されるべきである。

【０１１２】

複数の配列リードを得ることは、配列リードを生成するために試料から核酸を配列決定することを含み得る。複数の配列リードを得ることはまた、シーケンサから配列決定データを受け取ることを含み得る。試料中の核酸は、例えば、組織試料中のゲノムＤＮＡ、実験室試料中の特定の標的から増幅されたｃＤＮＡ、複数の生物由来の混合ＤＮＡ、合成核酸配列（例えば、バーコード及び固有の分子識別子（ＵＭＩ））等を含む任意の核酸であり得る。一実施形態では、核酸鋳型分子（例えば、ＤＮＡ又はＲＮＡ）は、タンパク質、脂質、及び非鋳型核酸等の様々な他の成分を含有する生物学的試料から単離される。核酸鋳型分子は、動物、植物、細菌、真菌、又は任意の他の細胞生物から得られる任意の細胞材料から得ることができる。本開示で使用するための生物学的試料には、ウイルス粒子又は調製物も含まれる。核酸鋳型分子は、生物から直接得ることができ、又は生物から得られた生物学的試料、例えば血液、尿、脳脊髄液、精液、唾液、痰、糞便及び組織から得ることができる。任意の組織又は体液検体（例えば、体液標本のヒト組織）を、本開示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、初代細胞培養物又は細胞株等の培養細胞から単離することもできる。鋳型核酸が得られる細胞又は組織は、ウイルス又は他の細胞内病原体に感染し得る。試料はまた、生物学的検体、ｃ

ＤＮＡライブラリ、ウイルス又はゲノムＤＮＡから抽出された全ＲＮＡであり得る。試料はまた、非細胞起源から単離されたＤＮＡ、例えばフリーザからの増幅／単離されたＤＮＡであり得る。

【０１１３】

一般に、核酸は、Ｇｒｅｅｎ ａｎｄ Ｓａｍｂｒｏｏｋ，Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ：Ａ Ｌａｂｏｒａｔｏｒｙ Ｍａｎｕａｌ（Ｆｏｕｒｔｈ Ｅｄｉｔｉｏｎ），Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ Ｐｒｅｓｓ，Ｗｏｏｄｂｕｒｙ，Ｎ．Ｙ．２，０２８ ｐａｇｅｓ（２０１２）に記載されているような、又は米国特許第７，９５７，９１３号、第７，７７６，６１６号、第５，２３４，８０９号、米国特許出願公開第２０１０／０２８５５７８号、及び米国特許出願公開第２００２／０１９０６６３号に記載されているような様々な技術によって抽出、単離、増幅又は分析することができる。

【０１１４】

生物学的試料から得られた核酸を断片化して、分析に適した断片を生成することができる。鋳型核酸は、様々な機械的、化学的、及び／又は酵素的方法を使用して、所望の長さに断片化又は剪断され得る。ＤＮＡは、例えば、Ｃｏｖａｒｉｓによって販売されている超音波処理機（Ｗｏｂｕｒｎ，Ｍａｓｓ．）、ＤＮａｓｅへの短時間の曝露、あるいは１つ又は複数の制限酵素の混合物、あるいはトランスポザーゼ又はニッキング酵素を使用して、超音波処理によってランダムに剪断され得る。ＲＮＡは、ＲＮａｓｅ、熱＋マグネシウムへの短時間の曝露によって、又は剪断によって断片化され得る。ＲＮＡをｃＤＮＡに変換することができる。断片化が使用される場合、ＲＮＡは、断片化の前又は後にｃＤＮＡに変換され得る。一実施形態では、核酸は超音波処理によって断片化される。別の実施形態では、核酸は、水素化剪断装置によって断片化される。一般に、個々の核酸鋳型分子は、約２ｋｂ塩基～約４０ｋｂであり得る。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、二本鎖、又は一本鎖領域を有する二本鎖（例えば、ステム構造及びループ構造）であり得る。

【０１１５】

生物学的試料は、必要に応じて洗剤又は界面活性物質の存在下で溶解、ホモジナイズ又は分画され得る。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又はＮ－ラウロイルサルコシン）又は非イオン性界面活性剤（例えば、商標ＴＷＥＥＮでＵｎｉｑｅｍａ Ａｍｅｒｉｃａｓ（Ｐａｔｅｒｓｏｎ，Ｎ．Ｊ．）により販売されているポリソルベート８０又はＴＲＩＴＯＮ Ｘ－１００として知られている$C_{14}H_{22}O(C_2H_4)_n$を含み得る。核酸が試料から抽出又は単離されると、それは増幅され得る。

【０１１６】

増幅は、核酸配列の更なるコピーの産生を指し、一般にポリメラーゼ連鎖反応（ＰＣＲ）又は当技術分野で公知の他の技術を使用して行われる。増幅反応は、ＰＣＲ等の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得る。他の増幅反応には、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、鎖置換増幅及び制限断片長多型、転写ベースの増幅システム、ローリングサークル増幅、及び超分岐ローリングサークル増幅、定量ＰＣＲ、定量蛍光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、ピコチターＰＣＲ、エマルジョンＰＣＲ、転写増幅、自立配列複製、コンセンサス配列プライムＰＣＲ、任意プライムＰＣＲ、縮重オリゴヌクレオチド－プライムＰＣＲ、及び核酸ベースの配列増幅（ＮＡＢＳＡ）が含まれる。使用され得る増幅方法としては、米国特許第５，２４２，７９４、第５，４９４，８１０号、第４，９８８，６１７号、及び第６，５８２，９３８号に記載されているものが挙げられる。ある実施形態では、増幅反応は、例えば、米国特許第４，６８３，１９５及び第４，６８３，２０２号に記載されるようなＰＣＲであり、参照により本明細書に組み込まれる。ＰＣＲ、配列決定、及

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、及び当技術分野で公知の他の方法によって調製することができる。プライマーは、Ｅｕｒｏｆｉｎｓ ＭＷＧ Ｏｐｅｒｏｎ（Ｈｕｎｔｓｖｉｌｌｅ，Ａｌａ．）又はＬｉｆｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｃａｒｌｓｂａｄ，Ｃａｌｉｆ．）等の商業的供給元から入手することもできる。

【０１１７】

バーコード配列は、各配列が核酸の特定の部分に相関するように設計することができ、配列リードをそれらが由来する部分に相関させることができる。バーコード配列のセットを設計する方法は、例えば、米国特許第６，２３５，４７５号に示されており、その内容は参照によりその全体が本明細書に組み込まれる。ある実施形態では、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。バーコード配列のセットを設計するための方法及びバーコード配列を取り付けるための他の方法は、米国特許第７，５４４，４７３号、第７，５３７，８９７号、第７，３９３，６６５号、第６，３５２，８２８号、第６，１７２，２１８号、第６，１７２，２１４号、第６，１５０，５１６号、第６，１３８，０７７号、第５，８６３，７２２号、第５，８４６，７１９号、第５，６９５，９３４、及び第５，６０４，０９７号に示されており、それぞれ参照によりに組み込まれる。

【０１１８】

配列決定は、当技術分野で公知の任意の方法によるものであり得る。ＤＮＡ配列決定技術には、標識ターミネータ又はプライマー及びスラブ又はキャピラリでのゲル分離を使用する古典的なジデオキシ配列決定反応（サンガー法）、可逆的に末端化された標識ヌクレオチドを使用する合成による配列決定、パイロシーケンシング、４５４配列決定、Ｉｌｌｕｍｉｎａ／Ｓｏｌｅｘａ配列決定、標識オリゴヌクレオチドプローブのライブラリへの対立遺伝子特異的ハイブリダイゼーション、標識クローンのライブラリへの対立遺伝子特異的ハイブリダイゼーションを使用した合成とそれに続くライゲーションによる配列決定、重合工程中の標識ヌクレオチドの取込みのリアルタイムモニタリング、ポロニー配列決定、及びＳＯＬｉＤ配列決定が含まれる。分離された分子の配列決定は、最近になって、ポリメラーゼ又はリガーゼを用いた連続的又は単一の伸長反応、並びにプローブのライブラリとの単一又は連続的な差次的ハイブリダイゼーションによって実証された。

【０１１９】

使用され得る配列決定技術としては、例えば、Ｒｏｃｈｅ（Ｂｒａｎｆｏｒｄ，Ｃｏｎｎ．）の４５４ Ｌｉｆｅ Ｓｃｉｅｎｃｅｓによって商標ＧＳ ＪＵＮＩＯＲ、ＧＳ ＦＬＸ＋及び４５４ ＳＥＱＵＥＮＣＩＮＧとして販売されており、内容は、参照によりその全体が本明細書に組み込まれる、Ｍａｒｇｕｌｉｅｓ，Ｍ．ｅｔ ａｌ．，Ｇｅｎｏｍｅ ｓｅｑｕｅｎｃｉｎｇ ｉｎ ｍｉｃｒｏ－ｆａｂｒｉｃａｔｅｄ ｈｉｇｈ－ｄｅｎｓｉｔｙ ｐｉｃｏｔｉｔｅｒ ｒｅａｃｔｏｒｓ，Ｎａｔｕｒｅ，４３７：３７６－３８０（２００５）、米国特許第５，５８３，０２４号、第５，６７４，７１３号、及び第５，７００，６７３号に記載されている合成による配列決定システムの使用が挙げられる。４５４配列決定は二段階を含む。これらのシステムの第１の工程では、ＤＮＡを約３００～８００塩基対の断片に剪断し、断片を平滑末端化する。次いで、オリゴヌクレオチドアダプタを断片の末端にライゲーションする。アダプタは、断片の増幅及び配列決定のためのプライマーとして機能する。断片は、例えば５’－ビオチンタグを含むアダプタＢを使用して、ＤＮＡ捕捉ビーズ、例えばストレプトアビジン被覆ビーズに結合させることができる。ビーズに付着した断片は、油－水エマルジョンの液滴内でＰＣＲ増幅される。結果は、各ビーズ上のクローン増幅ＤＮＡ断片の複数のコピーである。第２の工程では、ビーズをウェル（ピコリットルサイズ）に捕捉する。パイロシーケンシングは、各ＤＮＡ断片に対して並行して行われる。１つ又は複数のヌクレオチドの付加は、配列決定機器においてＣＣＤカメラによって記録される光信号を生成する。シグナル強度は、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ホスファート（ＰＰｉ）を利用する。ＰＰｉは、アデノシン５’ホスホスルフ

ェートの存在下でＡＴＰスルフリラーゼによってＡＴＰに変換される。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に変換し、この反応は検出及び分析される光を生成する。

【０１２０】

使用することができるＤＮＡ配列決定技術の別の例は、Ｌｉｆｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ Ｃｏｒｐｏｒａｔｉｏｎ（Ｃａｒｌｓｂａｄ，Ｃａｌｉｆ．）のＡｐｐｌｉｅｄ ＢｉｏｓｙｓｔｅｍｓによるＳＯＬｉＤ技術である。ＳＯＬｉＤ配列決定では、ゲノムＤＮＡを断片に剪断し、断片の５’及び３’末端にアダプタを結合させて断片ライブラリを生成する。あるいは、アダプタを断片の５’末端及び３’末端にライゲーションし、断片を環状化し、環状化された断片を消化して内部アダプタを生成し、得られた断片の５’末端及び３’末端にアダプタを付着させて、メイト・ペア（ｍａｔｅ－ｐａｉｒｅｄ）ライブラリを生成することによって、内部アダプタを導入することができる。次に、クローンビーズ集団を、ビーズ、プライマー、鋳型及びＰＣＲ成分を含有するマイクロリアクタ中で調製する。ＰＣＲの後、鋳型を変性させ、ビーズを濃縮して、伸長した鋳型を有するビーズを分離する。選択されたビーズ上の鋳型は、スライドガラスへの結合を可能にする３’修飾に供される。配列は、部分的にランダムなオリゴヌクレオチドと、特定のフルオロフォアによって識別される中央の決定された塩基（又は塩基対）との連続的なハイブリダイゼーション及びライゲーションによって決定することができる。色を記録した後、ライゲーションしたオリゴヌクレオチドを除去し、次いでプロセスを繰り返す。

【０１２１】

使用され得るＤＮＡ配列決定技術の別の例は、例えば、Ｌｉｆｅ Ｔｅｃｈｎｏｌｏｇｉｅｓ（Ｓｏｕｔｈ Ｓａｎ Ｆｒａｎｃｉｓｃｏ，Ｃａｌｉｆ．）によってＩｏｎ ＴｏｒｒｅｎｔによりＩＯＮ ＴＯＲＲＥＮＴの商標で販売されているシステムを使用するイオン半導体配列決定である。イオン半導体配列決定は、例えば、Ｒｏｔｈｂｅｒｇ，ｅｔ ａｌ．，Ａｎ ｉｎｔｅｇｒａｔｅｄ ｓｅｍｉｃｏｎｄｕｃｔｏｒ ｄｅｖｉｃｅ ｅｎａｂｌｉｎｇ ｎｏｎ－ｏｐｔｉｃａｌ ｇｅｎｏｍｅ ｓｅｑｕｅｎｃｉｎｇ，Ｎａｔｕｒｅ ４７５：３４８－３５２（２０１１）、米国特許公開第２０１０／０３０４９８２号、米国特許公開第２０１０／０３０１３９８号、米国特許公開第２０１０／０３００８９５号、米国特許公開第２０１０／０３００５５９号、及び米国特許公開第２００９／００２６０８２号に記載されており、これらの各々の内容は、参照によりその全体が組み込まれる。

【０１２２】

使用され得る配列決定技術の別の例は、Ｉｌｌｕｍｉｎａ配列決定である。Ｉｌｌｕｍｉｎａ配列決定は、フォールドバック（ｆｏｌｄ－ｂａｃｋ）ＰＣＲ及び固定プライマーを用いた固体表面上のＤＮＡの増幅に基づく。ゲノムＤＮＡを断片化し、断片の５’及び３’末端にアダプタを付加する。フローセルチャネルの表面に付着したＤＮＡ断片は伸長され、ブリッジ増幅される。断片は二本鎖になり、二本鎖分子は変性する。固相増幅とそれに続く変性の複数のサイクルは、フローセルの各チャネルに同じ鋳型の一本鎖ＤＮＡ分子の約１，０００コピーの数百万のクラスターを作製することができる。プライマー、ＤＮＡポリメラーゼ及び４つのフルオロフォア標識された可逆的に終結するヌクレオチドを使用して、順次配列決定（ｓｅｑｕｅｎｔｉａｌ ｓｅｑｕｅｎｃｉｎｇ）を行う。ヌクレオチド取込み後、レーザーを使用してフルオロフォアを励起し、画像を取り込み、第１の塩基のアイデンティティーを記録する。組み込まれた各塩基からの３’ターミネータ及びフルオロフォアを除去し、組込み、検出及び同定工程を繰り返す。この技術による配列決定は、米国特許第７，９６０，１２０号、第７，８３５，８７１号、第７，２３２，６５６号、第７，５９８，０３５号、第６，９１１，３４５号、第６，８３３，２４６号、第６，８２８，１００号、第６，３０６，５９７号、第６，２１０，８９１号、米国特許公開第２０１１／０００９２７８号、米国特許公開第２００７／０１１４３６２号、米国特許公開第２００６／０２９２６１１号、及び米国特許公開第２００６／００２４６８１号に記載されており、これらの各々は、参照によりその全体が組み込まれる。

【０１２３】
使用され得る配列決定技術の別の例としては、Ｐａｃｉｆｉｃ Ｂｉｏｓｃｉｅｎｃｅｓ（Ｍｅｎｌｏ Ｐａｒｋ，Ｃａｌｉｆ．）の単一分子リアルタイム（ＳＭＲＴ）技術が挙げられる。ＳＭＲＴでは、４つのＤＮＡ塩基のそれぞれが、４つの異なる蛍光色素のうちの１つに結合している。これらの色素は、リン連結されている。単一のＤＮＡポリメラーゼは、ゼロモード導波路（ＺＭＷ）の底部に鋳型一本鎖ＤＮＡの単一分子で固定化される。成長中の鎖にヌクレオチドを組み込むのに数ミリ秒かかる。この間、蛍光標識が励起され、蛍光シグナルが発生し、蛍光タグが切断される。色素の対応する蛍光の検出は、どの塩基が組み込まれたかを示す。このプロセスを繰り返す。

【０１２４】
使用され得る配列決定技術の別の例は、ナノポア配列決定である（Ｓｏｎｉ＆Ｍｅｌｌｅｒ，２００７，Ｐｒｏｇｒｅｓｓ ｔｏｗａｒｄ ｕｌｔｒａｆａｓｔ ＤＮＡ ｓｅｑｕｅｎｃｅ ｕｓｉｎｇ ｓｏｌｉｄ－ｓｔａｔｅ ｎａｎｏｐｏｒｅｓ，Ｃｌｉｎ Ｃｈｅｍ ５３（１１）：１９９６－２００１）。ナノポアは、直径１ナノメートル程度の小さな孔である。ナノポアを導電性流体に浸漬し、ナノポアの両端に電位を印加すると、ナノポアを通過するイオンの伝導に起因してわずかな電流が生じる。流れる電流の量は、ナノポアのサイズに対し感度を有する。ＤＮＡ分子がナノポアを通過するとき、ＤＮＡ分子上の各ヌクレオチドは、ナノポアを異なる程度で遮る。したがって、ＤＮＡ分子がナノポアを通過するときにナノポアを通過する電流の変化は、ＤＮＡ配列の読取りを表す。

【０１２５】
使用され得る配列決定技術の別の例は、化学感応性電界効果トランジスタ（ｃｈｅｍＦＥＴ）アレイを使用してＤＮＡを配列決定することを含む（例えば、米国特許出願公開第２００９／００２６０８２号）。この技術の一例では、ＤＮＡ分子を反応チャンバ内に配置することができ、鋳型分子をポリメラーゼに結合した配列決定プライマーにハイブリダイズさせることができる。１つ又は複数の三リン酸の、配列決定プライマーの３’末端での新しい核酸鎖への組込みは、ｃｈｅｍＦＥＴによる電流の変化によって検出することができる。アレイは、複数のｃｈｅｍＦＥＴセンサを有することができる。別の例では、単一の核酸をビーズに付着させることができ、核酸をビーズ上で増幅させることができ、個々のビーズをｃｈｅｍＦＥＴアレイ上の個々の反応チャンバに移送することができ、各チャンバはｃｈｅｍＦＥＴセンサを有し、核酸を配列決定することができる。

【０１２６】
使用することができる配列決定技術の別の例は、例えばＭｏｕｄｒｉａｎａｋｉｓ，Ｅ．Ｎ．ａｎｄ Ｂｅｅｒ Ｍ．，ｉｎ Ｂａｓｅ ｓｅｑｕｅｎｃｅ ｄｅｔｅｒｍｉｎａｔｉｏｎ ｉｎ ｎｕｃｌｅｉｃ ａｃｉｄｓ ｗｉｔｈ ｔｈｅ ｅｌｅｃｔｒｏｎ ｍｉｃｒｏｓｃｏｐｅ，ＩＩＩ．Ｃｈｅｍｉｓｔｒｙ ａｎｄ ｍｉｃｒｏｓｃｏｐｙ ｏｆ ｇｕａｎｉｎｅ－ｌａｂｅｌｅｄ ＤＮＡ，ＰＮＡＳ ５３：５６４－７１（１９６５）によって記載されているような電子顕微鏡を使用することを含む。この技術の一例では、個々のＤＮＡ分子は、電子顕微鏡を用いて識別可能な金属標識を使用して標識される。次いで、これらの分子を平らな表面に伸ばし、電子顕微鏡を使用して画像化して配列を測定する。

【０１２７】
本開示の実施形態による配列決定は、複数のリードを生成する。本開示によるリードは、一般に、約１５０塩基長未満、又は約９０塩基長未満のヌクレオチドデータの配列を含む。ある実施形態において、リードは、約８０～約９０塩基長、例えば、約８５塩基長である。いくつかの実施形態において、本開示の方法は、非常に短いリード、すなわち、約５０又は約３０塩基長未満の長さに適用される。配列リードデータは、配列データ及びメタ情報を含み得る。配列リードデータは、当業者に知られているように、例えば、ＶＣＦファイル、ＦＡＳＴＡファイル又はＦＡＳＴＱファイルを含む任意の適切なファイルフォーマットで保存することができる。

【０１２８】

ＦＡＳＴＡは、元々、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、標準ファイルフォーマットも指すようになった。Ｐｅａｒｓｏｎ＆Ｌｉｐｍａｎ，１９８８，Ｉｍｐｒｏｖｅｄ ｔｏｏｌｓ ｆｏｒ ｂｉｏｌｏｇｉｃａｌ ｓｅｑｕｅｎｃｅ ｃｏｍｐａｒｉｓｏｎ，ＰＮＡＳ ８５：２４４４－２４４８を参照されたい。ＦＡＳＴＡフォーマットの配列は、１行の記述で始まり、その後に配列データの行が続く。記述行は、１列目の大なり（「＞」）記号によって配列データと区別される。「＞」記号に続く単語は配列の識別子であり、行の残りは記述である（両方とも任意である）。「＞」と識別子の最初の文字との間にスペースがあってはならない。テキストの全ての行が８０文字未満であることが推奨される。「＞」で始まる別の行が現れると配列は終了し、これは別の配列の開始を示す。

【０１２９】

ＦＡＳＴＱフォーマットは、生物学的配列（通常はヌクレオチド配列）及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。これはＦＡＳＴＡフォーマットに類似しているが、配列データに続く品質スコアを有する。配列文字及び品質スコアの両方は、簡潔にするために単一のＡＳＣＩＩ文字で符号化される。ＦＡＳＴＱフォーマットは、Ｉｌｌｕｍｉｎａ Ｇｅｎｏｍｅ Ａｎａｌｙｚｅｒ．Ｃｏｃｋｅｔ ａｌ．，２００９，Ｔｈｅ Ｓａｎｇｅｒ ＦＡＳＴＱ ｆｉｌｅ ｆｏｒｍａｔ ｆｏｒ ｓｅｑｕｅｎｃｅｓ ｗｉｔｈ ｑｕａｌｉｔｙ ｓｃｏｒｅｓ，ａｎｄ ｔｈｅ Ｓｏｌｅｘａ／Ｉｌｌｕｍｉｎａ ＦＡＳＴＱ ｖａｒｉａｎｔｓ，Ｎｕｃｌｅｉｃ Ａｃｉｄｓ Ｒｅｓ ３８（６）：１７６７－１７７１．等のハイスループット配列決定装置の出力を保存するための事実上のスタンダードである。

【０１３０】

ＦＡＳＴＡ及びＦＡＳＴＱファイルの場合、メタ情報は記述行を含み、配列データの行を含まない。いくつかの実施形態では、ＦＡＳＴＱファイルの場合、メタ情報は品質スコアを含む。ＦＡＳＴＡ及びＦＡＳＴＱファイルの場合、配列データは、記述行の後に始まり、典型的には、任意選択的に「－」を有するＩＵＰＡＣ多義性符号のいくつかのサブセットを使用して存在する。好ましい実施形態では、配列データは、必要に応じて「－」又はＵを（例えば、間隙又はウラシルを表すために）任意選択的に含むＡ、Ｔ、Ｃ、Ｇ、及びＮ文字を使用する。

【０１３１】

上記及び他の場所で説明したように、ＮＧＳ機器の出力量は増加している。例えば、Ｐｉｎｈｏ＆Ｐｒａｔａｓ，２０１３，ＭＦＣｏｍｐｒｅｓｓ：ａ ｃｏｍｐｒｅｓｓｉｏｎ ｔｏｏｌ ｆｏｒ ＦＡＳＴＡ ａｎｄ ｍｕｌｔｉ－ＦＡＳＴＡ ｄａｔａ，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ３０（１）：１１７－８；Ｄｅｏｒｏｗｉｃｚ＆Ｇｒａｂｏｗｓｋｉ，２０１３，Ｄａｔａ ｃｏｍｐｒｅｓｓｉｏｎ ｆｏｒ ｓｅｑｕｅｎｃｉｎｇ ｄａｔａ，Ａｌｇ Ｍｏｌ Ｂｉｏ ８：２５；Ｂａｌｚｅｒ ｅｔ ａｌ．，２０１３，Ｆｉｌｔｅｒｉｎｇ ｄｕｐｌｉｃａｔｅ ｒｅａｄｓ ｆｒｏｍ ４５４ ｐｙｒｏｓｅｑｕｅｎｃｉｎｇ ｄａｔａ，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ２９（７）：８３０－８３６；Ｘｕ ｅｔ ａｌ．，２０１２，ＦａｓｔＵｎｉｑ：Ａ ｆａｓｔ ｄｅ ｎｏｖｏ ｄｕｐｌｉｃａｔｅｓ ｒｅｍｏｖａｌ ｔｏｏｌ ｆｏｒ ｐａｉｒｅｄ ｓｈｏｒｔ ｒｅａｄｓ，ＰＬｏＳ Ｏｎｅ ７（１２）：ｅ５２２４９；Ｂｏｎｆｉｅｌｄ ａｎｄ Ｍａｈｏｎｅｙ，２０１３，Ｃｏｍｐｒｅｓｓｉｏｎ ｏｆ ＦＡＳＴＱ ａｎｄ ＳＡＭ ｆｏｒｍａｔ ｓｅｑｕｅｎｃｉｎｇ ｄａｔａ，ＰＬｏＳ Ｏｎｅ ８（３）：ｅ５９１９０；ａｎｄ Ｖｅｅｎｅｍａｎ ｅｔ ａｌ．，２０１２，Ｏｃｕｌｕｓ：ｆａｓｔｅｒ ｓｅｑｕｅｎｃｅ ａｌｉｇｎｍｅｎｔ ｂｙ ｓｔｒｅａｍｉｎｇ ｒｅａｄ ｃｏｍｐｒｅｓｓｉｏｎ，ＢＭＣ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ １３：２９７を参照されたい。ＮＧＳ技術によって生成されるデータの量は、そのような配列決定情報を含むファイルを保存及び転送する際の困難を引き起こす。したがって、本開示の方法及びシステムは、核酸配列決定技術に由来するＦＡＳＴＡ又はＦＡＳＴＱファイル（ＦＡＳＴＡ／Ｑファイル）に含まれる大量の配列データ等の情報を保存するために使

用することができる。

【０１３２】

いくつかの実施形態において、配列リードファイル及び／又は配列出力ファイルは、プレーンテキストファイル（例えば、ＡＳＣＩＩ、ＩＳＯ／ＩＥＣ６４６、ＥＢＣＤＩＣ、ＵＴＦ－８、又はＵＴＦ－１６等の符号化を使用する）として保存される。本開示によって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できるテキストエディタプログラムを含むことができる。テキストエディタプログラムは、コンピュータ画面上にテキストファイル（プレーンテキストファイル等）の内容を提示し、人がテキストを編集することを可能にすることができるコンピュータプログラムを指すことができる（例えば、モニタ、キーボード、及びマウスを使用する）。例示的なテキストエディタには、Ｍｉｃｒｏｓｏｆｔ Ｗｏｒｄ、ｅｍａｃｓ、ｐｉｃｏ、ｖｉ、ＢＢＥｄｉｔ、及びＴｅｘｔＷｒａｎｇｌｅｒが含まれるが、これらに限定されない。好ましくは、テキストエディタプログラムは、コンピュータ画面上にプレーンテキストファイルを表示することができ、メタ情報及び配列リードを人が読める形式（例えば、バイナリエンコードされていない）で示すことができる。

【０１３３】

いくつかの実施形態では、本開示の工程のいずれか又は全ては自動化される。例えば、Ｐｅｒｌスクリプト又はシェルスクリプトを記述して、上述の様々なプログラムのいずれかを呼び出すことができる（例えば、Ｔｉｓｄａｌｌ，Ｍａｓｔｅｒｉｎｇ Ｐｅｒｌ ｆｏｒ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，Ｏ’Ｒｅｉｌｌｙ＆Ａｓｓｏｃｉａｔｅｓ，Ｉｎｃ．，Ｓｅｂａｓｔｏｐｏｌ，Ｃ Ａ ２００３；Ｍｉｃｈａｅｌ，Ｒ．，Ｍａｓｔｅｒｉｎｇ Ｕｎｉｘ Ｓｈｅｌｌ Ｓｃｒｉｐｔｉｎｇ，Ｗｉｌｅｙ Ｐｕｂｌｉｓｈｉｎｇ，Ｉｎｃ．，Ｉｎｄｉａｎａｐｏｌｉｓ，Ｉｎｄ．２００３を参照されたい）。あるいは、本開示の方法は、１つ又は複数の専用プログラムで全体的又は部分的に具体化されてもよく、例えば、それぞれ任意選択的にＣ＋＋等のコンパイル型言語で記述され、次いでコンパイルされ、バイナリとして配布される。本開示の方法は、既存の配列分析プラットフォーム内のモジュールとして、又は既存の配列分析プラットフォーム内の機能を呼び出すことによって、全体的又は部分的に実施され得る。ある実施形態では、本開示の方法は、単一の開始キュー（例えば、人の活動、別のコンピュータプログラム、又は機械から供給されるトリガーイベントの１つ又は組合わせ）に応答して全て自動的に呼び出される多数の工程を含む。したがって、本開示は、任意の工程又は工程の任意の組合わせがキューに応答して自動的に行われ得る方法を提供する。人の入力、影響、又は相互作用を介在させることのない自動的な一般的手段である（すなわち、元の又はプレキューの人の活動にのみ応答性である）。

【０１３４】

本開示はまた、対象核酸の正確かつ高感度な解釈を含む様々な形態の出力を包含する。出力は、コンピュータファイルの形式で提供することができる。ある実施形態では、出力は、ＦＡＳＴＡファイル、ＦＡＳＴＱファイル、又はＶＣＦファイルである。出力を処理して、テキストファイル、又は参照ゲノムの配列にアライメントされた核酸の配列等の配列データを含むＸＭＬファイルを生成することができる。他の実施形態では、処理は、参照ゲノムに対する対象核酸中の１つ又は複数の突然変異を記述する座標又は文字列を含む出力をもたらす。当技術分野で公知のアライメントストリングとしては、Ｓｉｍｐｌｅ ＵｎＧａｐｐｅｄ Ａｌｉｇｎｍｅｎｔ Ｒｅｐｏｒｔ（ＳＵＧＡＲ）、Ｖｅｒｂｏｓｅ Ｕｓｅｆｕｌ Ｌａｂｅｌｅｄ Ｇａｐｐｅｄ Ａｌｉｇｎｍｅｎｔ Ｒｅｐｏｒｔ（ＶＵＬＧＡＲ）、及びＣｏｍｐａｃｔ Ｉｄｉｏｓｙｎｃｒａｔｉｃ Ｇａｐｐｅｄ Ａｌｉｇｎｍｅｎｔ Ｒｅｐｏｒｔ（ＣＩＧＡＲ）（Ｎｉｎｇ，Ｚ．，ｅｔ ａｌ．，Ｇｅｎｏｍｅ Ｒｅｓｅａｒｃｈ １１（１０）：１７２５－９（２００１））が挙げられる。これらの文字列は、例えば、Ｅｕｒｏｐｅａｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｈｉｎｘｔｏｎ，ＵＫ）のＥｘｏｎｅｒａｔｅ配列アライメントソフトウェアに実装されている。

【０１３５】
いくつかの実施形態では、配列アライメントは、例えば配列アライメントマップ（ＳＡＭ）又はバイナリアライメントマップ（ＢＡＭ）ファイル等、ＣＩＧＡＲ文字列（ＳＡＭ形式は、例えば、Ｌｉ　ｅｔ　ａｌ．，Ｔｈｅ　Ｓｅｑｕｅｎｃｅ　Ａｌｉｇｎｍｅｎｔ／Ｍａｐ　ｆｏｒｍａｔ　ａｎｄ　ＳＡＭｔｏｏｌｓ，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，２００９，２５（１６）：２０７８－９に記載されている）を含むものとして作成される。いくつかの実施形態では、ＣＩＧＡＲは、ギャップのあるアライメントをラインごとに表示又は含む。ＣＩＧＡＲは、ＣＩＧＡＲ文字列として報告される圧縮されたペアワイズアライメントフォーマットである。ＣＩＧＡＲ文字列は、長い（例えば、ゲノム）ペアワイズアライメントを表すのに有用である。ＣＩＧＡＲ文字列は、参照ゲノム配列に対するリードのアライメントを表すためにＳＡＭ形式で使用される。

【０１３６】
ＣＩＧＡＲ文字列は、確立されたモチーフに続く。各文字の前に数字が付けられ、イベントの塩基カウントが与えられる。使用される文字は、Ｍ、Ｉ、Ｄ、Ｎ、及びＳ（Ｍ＝マッチ；Ｉ＝挿入；Ｄ＝欠失；Ｎ＝ギャップ；Ｓ＝置換）を含むことができる。ＣＩＧＡＲ文字列は、マッチ／ミスマッチ及び欠失（又はギャップ）の配列を定義する。例えば、ＣＩＧＡＲ文字列２ＭＤ３Ｍ２Ｄ２Ｍは、アライメントが２つのマッチ、１つの欠失（いくらかのスペースを節約するために番号１は省略されている）、３つのマッチ、２つの欠失及び２つのマッチを含むことを意味する。

【０１３７】
本開示によって企図されるように、上述の機能は、ソフトウェア、ハードウェア、ファームウェア、ハード配線、又はこれらの任意の組合わせを含む本開示のシステムを使用して実施することができる。機能を実装する特徴はまた、機能の一部が異なる物理的位置に実装されるように分散されることを含む、様々な位置に物理的に配置され得る。

【０１３８】
当業者であれば、本開示の方法の実行に必要又は最も適していると認識するように、本開示のコンピュータシステム又はマシンは、バスを介して互いに通信する１つ又は複数のプロセッサ（例えば、中央プロセッシングユニット（ＣＰＵ）、グラフィックスプロセッシングユニット（ＧＰＵ）、又はその両方）、メインメモリ、及びスタティックメモリを含む。

【０１３９】
図１２は、本開示の方法を実行するのに適したシステム７０１を示す。図１２に示されるように、システム７０１は、サーバコンピュータ７０５、端末７１５、シーケンサ７１５、シーケンサコンピュータ７２１、コンピュータ７４９、又はそれらの任意の組合わせのうちの１つ又は複数を含み得る。そのようなコンピュータデバイスの各々は、ネットワーク７０９を介して通信することができる。シーケンサ７２５は、任意選択的に、それ自体の、例えば専用のシーケンサコンピュータ７２１（任意の入力／出力機構（Ｉ／Ｏ）、プロセッサ、及び、例えばダイナミックランダムアクセスメモリＤＲＡＭ又はＤＡＭ７２９等のメモリを含む）を含むか、又はそれに動作可能に結合されてもよい。追加的又は代替的に、シーケンサ７２５は、ネットワーク７０９を介してサーバ７０５又はコンピュータ７４９（例えば、ラップトップ、デスクトップ、又はタブレット）に動作可能に結合されてもよい。コンピュータ７４９は、１つ又は複数のプロセッサ、メモリ、及びＩ／Ｏを含む。本開示の方法がクライアント／サーバアーキテクチャを使用する場合、本開示の方法の任意の工程は、データ、命令等を取得するか、又はインターフェースモジュールを介して結果を提供するか、又はファイルとして結果を提供することができる、プロセッサ、メモリ、及びＩ／Ｏのうちの１つ又は複数を含むサーバ７０５を使用して実行され得る。サーバ７０５は、コンピュータ７４９又は端末７１５によりネットワーク７０９を介して係合されてもよく、又はサーバ７０５は、端末７１５に直接接続されてもよい。端末７１５は、好ましくはコンピュータデバイスである。本開示によるコンピュータは、好ましくは、Ｉ／Ｏ機構及びメモリに結合された１つ又は複数のプロセッサを含む。

【０１４０】
プロセッサは、例えば、シングルコア又はマルチコアプロセッサ（例えば、ＡＭＤ Ｐｈｅｎｏｍ ＩＩ Ｘ２、Ｉｎｔｅｌ Ｃｏｒｅ Ｄｕｏ，ＡＭＤ Ｐｈｅｎｏｍ ＩＩ Ｘ４、Ｉｎｔｅｌ Ｃｏｒｅ ｉ５、Ｉｎｔｅｌ Ｃｏｒｅ ｉ＆Ｅｘｔｒｅｍｅ Ｅｄｉｔｉｏｎ ９８０Ｘ、又はＩｎｔｅｌ Ｘｅｏｎ Ｅ７－２８２０）のうちの１つ又は複数を含む１つ又は複数のプロセッサによって提供され得る。
【０１４１】
Ｉ／Ｏ機構は、ビデオ表示ユニット（例えば、液晶ディスプレイ（ＬＣＤ）又は陰極線管（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）、ディスク駆動ユニット、信号生成デバイス（例えば、スピーカ）、加速度計、マイクロフォン、セルラー無線周波数アンテナ、及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、セルラーモデム、データジャック、イーサネットポート、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）、光／画像投影デバイス、又はそれらの組合わせを含むことができる。
【０１４２】
本開示によるメモリは、１つ又は複数の有形デバイスによって提供される非一時的メモリを指し、有形デバイスは、本明細書に記載の方法又は機能のいずれか１つ又は複数を具現化する１つ又は複数の命令セット（例えば、ソフトウェア）が格納された１つ又は複数の機械可読媒体を含むことが好ましい。ソフトウェアはまた、システム５０１内のコンピュータによる実行中に、メインメモリ、プロセッサ、又はその両方内に完全に又は少なくとも部分的に存在してもよく、メインメモリ及びプロセッサはまた、機械可読媒体を構成する。ソフトウェアは、ネットワークインターフェース装置を介してネットワークにわたって更に送信又は受信することができる。
【０１４３】
機械可読媒体は、例示的な実施形態では単一の媒体であり得るが、「機械可読媒体」という用語は、１つ又は複数の命令セットを格納する単一の媒体又は複数の媒体（例えば、集中型又は分散型データベース、並びに／あるいは関連するキャッシュ及びサーバ）を含むと解釈されるべきである。「機械可読媒体」という用語はまた、機械によって実行するための命令のセットを格納、エンコード、又は搬送することができ、機械に本開示の方法論のうちの任意の１つ又は複数を実行させる任意の媒体を含むと解釈されるべきである。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」保存場所、又はそれらの組合わせのうちの１つ又は複数であってもよい。ある実施形態では、本開示の装置は、メモリ用の有形の非一時的コンピュータ可読媒体を含む。メモリとして使用するための例示的なデバイスには、半導体メモリデバイス（例えば、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ソリッドステートドライブ（ＳＳＤ）、及びフラッシュメモリデバイス、例えば、ＳＤ、マイクロＳＤ、ＳＤＸＣ、ＳＤＩＯ、ＳＤＨＣカード）が含まれ、磁気ディスク（例えば、内蔵ハードディスク又はリムーバブルディスク）、及び光ディスク（例えば、ＣＤ及びＤＶＤディスク）が含まれる。
【０１４４】
コンティグを構築し、コンセンサス配列を生成する様々な方法を以下に論じる。
【０１４５】
コンティグは、一般に、核酸配列、例えばリードの複数のセグメント間、又はその中の関係を指す。配列リードが重複する場合、コンティグは、重複リードの階層化画像として表すことができる。コンティグは、例えば、テキストファイル又はデータベース内の任意の特定の視覚的配置又は任意の特定の配置によって定義されず、それらに限定されない。コンティグは、一般に、配列決定された核酸の一部に対応するように編成された多数のリ

ードからの配列データを含む。コンティグは、表示又は保存された、リードのセット又は互いに対する若しくは参照に対するそれらの位置に関する情報等の組み立て結果を含むことができる。コンティグは、行が個々の配列リードであり、列がその部位にアライメントすると推定される各リードの塩基を含むグリッドとして構成することができる。コンセンサス配列は、アセンブリの各カラム中の優勢な塩基を同定することによって作製することができる。本発明によるコンティグは、それらが互いに重なり合う（又は、重複せず、例えば、単に隣接する）ことを示すリードの視覚的表示を含むことができる。コンティグは、複数のリードに関連付けられ、互いに対するリードの位置を与える座標のセットを含み得る。コンティグは、リードの配列データを変換することによって得られたデータを含み得る。例えば、Ｂｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換をリードに対して行うことができ、コンティグは、リードの非変換配列を必ずしも含まずに変換データを含むことができる。ヌクレオチド配列データのＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ変換は、米国特許出願公開第２００５／００３２０９５号に記載され、その全体が参照により本明細書に組み込まれる。

【０１４６】

リードは、当技術分野で公知の任意の方法によってコンティグに組み立てることができる。複数の配列リードのデノボアセンブリのためのアルゴリズムは当技術分野において公知であるが、そのような公知のアルゴリズムは、本開示で記載されている構造化した配列リード入力のために本明細書において改良されている（個々の配列要素は、長い配列リードのより広い集団の各長配列リード内の反復シリーズ（キメラアレイ）として存在する、低複雑度のリンカー配列に隣接する、高複雑度のライブラリに由来していた）。

【０１４７】

配列リードを組み立てるための１つのアルゴリズムは、オーバーラップコンセンサスアセンブリとして知られている。オーバーラップコンセンサスアセンブリは、配列リード間のオーバーラップを使用してそれらの間のリンクを作成する。リードは、一般に、非ランダムな重複が想定されるのに十分に重複する領域によって連結される。このようにリードを一緒に連結すると、コンティグ又はオーバーラップグラフが生成され、各ノードはリードに対応し、エッジは２つのリード間のオーバーラップを表す。オーバーラップグラフによるアセンブリは、例えば、米国特許第６，７１４，８７４号に記載されている。

【０１４８】

いくつかの実施形態では、デノボアセンブリは、いわゆるグリーディアルゴリズムに従って進行する。グリーディアルゴリズムに従って組み立てるために、リードの一群のリードのうちの１つが選択され、それは、それがかなりの量の重複を示す別のリードと対にされ、一般に、それは他の全てのリードのうちの最も多くの重複を示すリードと対にされる。これらの２つのリードはマージされて新しいリード配列を形成し、次いでそのリード群に戻され、プロセスが繰り返される。グリーディアルゴリズムによるアセンブリは、例えば、Ｓｃｈａｔｚ，ｅｔ　ａｌ．，Ｇｅｎｏｍｅ　Ｒｅｓ．，２０：１１６５－１１７３（２０１０）及び米国特許出願公開第２０１１／０２５７８８９号に記載され、これらの各々は、参照によりその全体が本明細書に組み込まれる。

【０１４９】

他の実施形態では、アセンブリは、ペアワイズアライメント、例えば網羅的又はヒューリスティック（例えば、網羅的ではない）ペアワイズアライメントによって進行する。アライメントについては、一般に、以下でより詳細に説明する。「力づく（ｂｒｕｔｅ　ｆｏｒｃｅ）」アプローチと呼ばれることもある網羅的なペアワイズアライメントは、セット内の配列の可能な全ての対の間の可能な全てのアライメントについてアライメントスコアを計算する。ヒューリスティック多重配列アライメントによるアセンブリは、数学的にありそうにない特定の組合わせを無視し、計算的に高速であり得る。マルチプル配列アライメントによる組立ての１つのヒューリスティックな方法は、いわゆる「分割統治」ヒューリスティックであり、これは例えば、米国特許出願公開第２００３／０２２４３８４号に記載される。マルチプル配列アライメントによる組立ての別のヒューリスティック方法

は、プログラムＣｌｕｓｔａｌＷによって実施されるプログレッシブアライメントである（例えば、Ｔｈｏｍｐｓｏｎ，ｅｔ ａｌ．，Ｎｕｃｌ．Ａｃｉｄｓ．Ｒｅｓ．，２２：４６７３－８０（１９９４）を参照されたい）。多重配列アライメントによるアセンブリは、一般に、Ｌｅｃｏｍｐｔｅ，Ｏ．，ｅｔ ａｌ．，Ｇｅｎｅ ２７０：１７－３０（２００１）；Ｍｕｌｌａｎ，Ｌ．Ｊ．，Ｂｒｉｅｆ Ｂｉｏｉｎｆｏｒｍ．，３：３０３－５（２００２）；Ｎｉｃｈｏｌａｓ，Ｈ．Ｂ．Ｊｒ．，ｅｔ ａｌ．，Ｂｉｏｔｅｃｈｎｉｑｕｅｓ ３２：５７２－９１（２００２）；ａｎｄ Ｘｉｏｎｇ，Ｇ．，Ｅｓｓｅｎｔｉａｌ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，２００６，Ｃａｍｂｒｉｄｇｅ Ｕｎｉｖｅｒｓｉｔｙ Ｐｒｅｓｓ，Ｎｅｗ Ｙｏｒｋ，Ｎ．Ｙ．に記載されている。

【０１５０】

アライメントによる組立ては、リードを互いにアライメントすることによって、又はリードを参照にアライメントすることによって進行することができる。例えば、各リードを参照ゲノムに順にアライメントすることによって、全てのリードを互いに関連して配置してアセンブリを作製する。

【０１５１】

リードをコンティグに組み立てる１つの方法は、ｄｅ Ｂｒｕｉｊｎグラフを作成することを含む。ｄｅ Ｂｒｕｉｊｎグラフは、リードをｋ－ｍｅｒと呼ばれるＤＮＡのより小さな配列に分割することによって計算労力を削減し、パラメータｋはこれらの配列の塩基長を表す。ｄｅ Ｂｒｕｉｊｎグラフでは、全てのリードがｋ－ｍｅｒ（リード内の長さｋの全ての部分配列）に分割され、ｋ－ｍｅｒ間の経路が計算される。この方法によるアセンブリでは、リードは、ｋ－ｍｅｒを通る経路として表される。ｄｅ Ｂｒｕｉｊｎグラフは、これらのｋ－ｍｅｒ間で長さｋ－１の重複を捕捉し、実際のリード間では捕捉しない。したがって、例えば、配列決定ＣＡＴＧＧＡは、以下の２－ｍｅｒ：ＣＡ、ＡＴ、ＴＧ、ＧＧ、及びＧＡを通る経路として表すことができる。ｄｅ Ｂｒｕｉｊｎグラフ手法は、冗長性をうまく扱い、複雑な経路の計算を扱いやすくする。データセット全体をｋ－ｍｅｒ重複まで削減することにより、ｄｅ Ｂｒｕｉｊｎグラフは、ショートリードデータセットの高い冗長性を削減する。特定のアセンブリの最大効率的なｋ－ｍｅｒサイズは、リード長並びにエラーレートによって決定される。パラメータｋの値は、アセンブリの品質に大きな影響を及ぼす。良好な値の推定は、組み立て前に行うことができ、又は、最適な値は、小さな範囲の値を試験することによって見つけることができる。ｄｅ Ｂｒｕｉｊｎグラフを使用したリードのアセンブリは、米国特許出願公開第２０１１／０００４４１３号、米国特許出願公開第２０１１／００１５８６３号、及び米国特許出願公開第２０１０／００６３７４２号に記載され、これらの各々は、参照によりその全体が本明細書に組み込まれる。

【０１５２】

本発明による、リードをコンティグに組み立てる他の方法も可能である。例えば、リードは、配列決定中に鋳型核酸に挿入されたバーコード情報を含み得る。ある実施形態において、リードは、バーコード情報を参照することによってコンティグにアセンブルされる。例えば、バーコードを識別することができ、バーコードを一緒に配置することによってリードを組み立てることができる。

【０１５３】

リードのコンティグへのアセンブリは、Ｈｕｓｅｍａｎｎ，Ｐ．ａｎｄ Ｓｔｏｙｅ，Ｊ，Ｐｈｙｌｏｇｅｎｅｔｉｃ Ｃｏｍｐａｒａｔｉｖｅ Ａｓｓｅｍｂｌｙ，２００９，Ａｌｇｏｒｉｔｈｍｓ ｉｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ：９ｔｈ Ｉｎｔｅｒｎａｔｉｏｎａｌ Ｗｏｒｋｓｈｏｐ，ｐｐ．１４５－１５６，Ｓａｌｚｂｅｒｇ，Ｓ．、及びＷａｒｎｏｗ，Ｔ．，Ｅｄｓ．Ｓｐｒｉｎｇｅｒ－Ｖｅｒｌａｇ，Ｂｅｒｌｉｎ Ｈｅｉｄｅｌｂｅｒｇで更に論じられている。リードをコンティグに組み立てるためのいくつかの例示的な方法は、例えば、国特許出願公開第６，２２３，１２８号、国特許出願公開第２００９／０２９８０６４号、米国特許出願公開２０１０／００６９２６３号、及び米国特許出願公開第２０１１／０２５７８８９号に記載され、これらの各々は、参照によ

りその全体が本明細書に組み込まれる。
【０１５４】
リードを組み立てるためのコンピュータプログラムは、当技術分野において公知である。そのようなアセンブリプログラムは、単一の汎用コンピュータ上で、コンピュータのクラスター若しくはネットワーク上で、又は配列分析専用の専用コンピューティングデバイス上で実行することができる。
【０１５５】
アセンブリは、例えば、カナダのＭｉｃｈａｅｌ Ｓｍｉｔｈ Ｇｅｎｏｍｅ Ｓｃｉｅｎｃｅｓ Ｃｅｎｔｒｅ（Ｖａｎｃｏｕｖｅｒ，Ｂ．Ｃ．，ＣＡ）からのプログラム「Ｔｈｅ Ｓｈｏｒｔ Ｓｅｑｕｅｎｃｅ Ａｓｓｅｍｂｌｙ ｂｙ ｋ－ｍｅｒ ｓｅａｒｃｈ ａｎｄ ３’ｒｅａｄ Ｅｘｔｅｎｓｉｏｎ’（ＳＳＡＫＥ）」によって実施することができる（例えば、Ｗａｒｒｅｎ，Ｒ．，ｅｔ ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，２３：５００－５０１（２００７）を参照されたい）。ＳＳＡＫＥは、リードのテーブルを循環し、任意の２つの配列間の可能な限り長い重複についてプレフィックスツリーを検索する。ＳＳＡＫＥクラスターはコンティグに読み取る。
【０１５６】
別のリードアセンブリプログラムは、Ｄａｒｒｅｎ Ｐｌａｔｔ及びＤｉｒｋ Ｅｖｅｒｓによって書かれたＦｏｒｇｅ Ｇｅｎｏｍｅ Ａｓｓｅｍｂｌｅｒであり、Ｇｅｅｋｎｅｔ（Ｆａｉｒｆａｘ，Ｖａ．）によって管理されているＳｏｕｒｃｅＦｏｒｇｅウェブサイトを通じて入手可能である（例えば、ＤｉＧｕｉｓｔｉｎｉ，Ｓ．，ｅｔ ａｌ．，Ｇｅｎｏｍｅ Ｂｉｏｌｏｇｙ，１０：Ｒ９４（２００９）を参照されたい）。Ｆｏｒｇｅは、利用可能であれば、その計算及びメモリ消費を複数のノードに分配し、したがって、大きなリードセットを組み立てる可能性を有する。Ｆｏｒｇｅは、並列ＭＰＩライブラリを使用してＣ＋＋で書かれた。Ｆｏｒｇｅは、リードの混合物、例えば、Ｓａｎｇｅｒ、４５４及びＩｌｌｕｍｉｎａによるリードを扱うことができる。
【０１５７】
多重配列アライメントによるアセンブリは、例えば、Ｕｎｉｖｅｒｓｉｔｙ Ｃｏｌｌｅｇｅ Ｄｕｂｌｉｎ（Ｄｕｂｌｉｎ，Ｉｒｅｌａｎｄ）から入手可能なプログラムＣｌｕｓｔａｌ Ｏｍｅｇａ（Ｓｉｅｖｅｒｓ Ｆ．，ｅｔ ａｌ．，Ｍｏｌ Ｓｙｓｔ Ｂｉｏｌ ７（２０１１）），ＣｌｕｓｔａｌＷ，ｏｒ ＣｌｕｓｔａｌＸ（Ｌａｒｋｉｎ Ｍ．Ａ．，ｅｔ ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，２３，２９４７－２９４８（２００７））によって行うことができる。
【０１５８】
当技術分野で知られている別の例示的なリードアセンブリプログラムは、Ｅｕｒｏｐｅａｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｈｉｎｘｔｏｎ，ＵＫ）のウェブサイトを通じて入手可能なＶｅｌｖｅｔである（Ｚｅｒｂｉｎｏ Ｄ．Ｒ．ｅｔ ａｌ．，Ｇｅｎｏｍｅ Ｒｅｓｅａｒｃｈ １８（５）：８２１－８２９（２００８））。Ｖｅｌｖｅｔは、ｄｅ Ｂｒｕｉｊｎグラフに基づく手法を実装し、リード対からの情報を使用し、様々なエラー補正工程を実装する。
【０１５９】
リードアセンブリは、Ｂｅｉｊｉｎｇ Ｇｅｎｏｍｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｂｅｉｊｉｎｇ，ＣＮ）又はＢＧＩ Ａｍｅｒｉｃａｓ Ｃｏｒｐｏｒａｔｉｏｎ（Ｃａｍｂｒｉｄｇｅ，Ｍａｓｓ．）のウェブサイトを通じて入手可能なｐａｃｋａｇｅ ＳＯＡＰからのプログラムを用いて実行することができる。例えば、ＳＯＡＰｄｅｎｏｖｏプログラムは、ｄｅ Ｂｒｕｉｊｎグラフ手法を実装する。ＳＯＡＰＳ／ＧＰＵは、短いリードを参照配列にアライメントする。
【０１６０】
別のリードアセンブリプログラムは、カナダのＭｉｃｈａｅｌ Ｓｍｉｔｈ Ｇｅｎｏｍｅ Ｓｃｉｅｎｃｅｓ Ｃｅｎｔｒｅ（Ｖａｎｃｏｕｖｅｒ，Ｂ．Ｃ．，ＣＡ）（Ｓｉｍｐｓｏｎ，Ｊ．Ｔ．，ｅｔ ａｌ．，Ｇｅｎｏｍｅ Ｒｅｓ．，１９（６）：１１１７

－２３（２００９））である。ＡＢｙＳＳは、ｄｅ　Ｂｒｕｉｊｎグラフ手法を使用し、並列環境で実行される。

【０１６１】

リードアセンブリは、Ｒｏｃｈｅ　４５４シーケンサ（例えば、Ｋｕｍａｒ，Ｓ．ｅｔａｌ．，Ｇｅｎｏｍｉｃｓ　１１：５７１（２０１０）及びＭａｒｇｕｌｉｅｓ，ｅｔａｌ．，Ｎａｔｕｒｅ　４３７：３７６－３８０（２００５）に記載されている）からのリードをアセンブルするように設計された、ｇｓＡｓｓｅｍｂｌｅｒ又はＮｅｗｂｌｅｒ（ＮＥＷ　ａｓｓｅｍＢＬＥＲ）として知られるＲｏｃｈｅのＧＳ　Ｄｅ　Ｎｏｖｏ　Ａｓｓｅｍｂｌｅｒによって行うこともできる。Ｎｅｗｂｌｅｒは、４５４のＦｌｘ　Ｓｔａｎｄａｒｄリード及び４５４のＴｉｔａｎｉｕｍリード、並びに単一及びペアエンドリード、並びに任意選択でＳａｎｇｅｒのリードを受け付ける。Ｎｅｗｂｌｅｒは、３２ビット又は６４ビットのいずれかのバージョンでＬｉｎｕｘ（登録商標）上で実行される。Ｎｅｗｂｌｅｒは、コマンドライン又はＪａｖａベースのＧＵＩインターフェースを介してアクセスすることができる。

【０１６２】

オクスフォード大学のＭａｒｉｏ　Ｃａｃｃａｍｏ及びＺａｍｉｎ　Ｉｑｂａｌによって作成されたＣｏｒｔｅｘは、リードアセンブリを含むゲノム解析のためのソフトウェアフレームワークである。Ｃｏｒｔｅｘは、Ｓｐａｎｕ，Ｐ．Ｄ．，ｅｔ　ａｌ．，Ｓｃｉｅｎｃｅ　３３０（６０１０）：１５４３－４６（２０１０）に記載されているように使用される、コンセンサスゲノムアセンブリのためのｃｏｒｔｅｘ＿ｃｏｎを含む。Ｃｏｒｔｅｘは、Ｉｑｂａｌ，ｅｔ　ａｌ．，Ｄｅ　ｎｏｖｏ　ａｓｓｅｍｂｌｙ　ａｎｄ　ｇｅｎｏｔｙｐｉｎｇ　ｏｆ　ｖａｒｉａｎｔｓ　ｕｓｉｎｇ　ｃｏｌｏｒｅｄ　ｄｅ　Ｂｒｕｉｊｎ　ｇｒａｐｈｓ，Ｎａｔｕｒｅ　Ｇｅｎｅｔｉｃｓ（ｉｎ　ｐｒｅｓｓ）に記載されており、Ｍｉｌｌｓ，Ｒ．Ｅ．，ｅｔ　ａｌ．，Ｎａｔｕｒｅ　４７０：５９－６５（２０１０）に記載されているように使用される、変異及び集団アセンブリのためのｃｏｒｔｅｘ＿ｖａｒを含む。Ｃｏｒｔｅｘは、制作者のウェブサイトを通じて、及びＧｅｅｋｎｅｔ（Ｆａｉｒｆａｘ，Ｖａ．）が管理するＳｏｕｒｃｅＦｏｒｇｅウェブサイトから入手可能である。

【０１６３】

他のリードアセンブリプログラムには、Ｒｅａｌ　Ｔｉｍｅ　Ｇｅｎｏｍｉｃｓ，Ｉｎｃ．（Ｓａｎ　Ｆｒａｎｃｉｓｃｏ，Ｃａｌｉｆ．）からのＲＴＧ　Ｉｎｖｅｓｔｉｇａｔｏｒ；ｉＡｓｓｅｍｂｌｅｒ（Ｚｈｅｎｇ，ｅｔ　ａｌ．，ＢＭＣ　Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ　１２：４５３（２０１１））；ＴｇｉＣＬ　Ａｓｓｅｍｂｌｅｒ（Ｐｅｒｔｅａ，ｅｔ　ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ　１９（５）：６５１－５２（２００３））；Ｇｅｅｋｎｅｔ（Ｆａｉｒｆａｘ，Ｖａ．）が管理するＳｏｕｒｃｅＦｏｒｇｅのウェブサイトを通してダウンロード可能な、Ｈｅｎｇ　ＬｉによるＭａｑ（Ｍａｐｐｉｎｇ　ａｎｄ　Ａｓｓｅｍｂｌｙ　ｗｉｔｈ　Ｑｕａｌｉｔｉｅｓ）；ＭＩＲＡ３（Ｍｉｍｉｃｋｉｎｇ　Ｉｎｔｅｌｌｉｇｅｎｔ　Ｒｅａｄ　Ａｓｓｅｍｂｌｙ）、Ｃｈｅｖｒｅｕｘ，Ｂ．，ｅｔ　ａｌ．，Ｇｅｎｏｍｅ　Ｓｅｑｕｅｎｃｅ　Ａｓｓｅｍｂｌｙ　Ｕｓｉｎｇ　Ｔｒａｃｅ　Ｓｉｇｎａｌｓ　ａｎｄ　Ａｄｄｉｔｉｏｎａｌ　Ｓｅｑｕｅｎｃｅ　Ｉｎｆｏｒｍａｔｉｏｎ，１９９９，Ｃｏｍｐｕｔｅｒ　Ｓｃｉｅｎｃｅ　ａｎｄ　Ｂｉｏｌｏｇｙ：Ｐｒｏｃｅｅｄｉｎｇｓ　ｏｆ　ｔｈｅ　Ｇｅｒｍａｎ　Ｃｏｎｆｅｒｅｎｃｅ　ｏｎ　Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ（ＧＣＢ）９９：４５－５６；ＰＧＡ４ｇｅｎｏｍｉｃｓ（Ｚｈａｏ　Ｆ．，ｅｔ　ａｌ．，Ｇｅｎｏｍｉｃｓ．９４（４）：２８４－６（２００９）に記載；及びＰｈｒａｐ（例えば、ｄｅ　ｌａ　Ｂａｓｔｉｄｅ，Ｍ．ａｎｄ　ＭｃＣｏｍｂｉｅ，Ｗ．Ｒ．，Ｃｕｒｒｅｎｔ　Ｐｒｏｔｏｃｏｌｓ　ｉｎ　Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，１７：１１．４．１－１１．４．１５（２００７）に記載）が含まれる。ＣＬＣ　ｃｅｌｌは、ＣＬＣ　ｂｉｏ　Ｇｅｒｍａｎｙ（Ｍｕｅｈｌｔａｌ，Ｇｅｒｍａｎｙ）から入手可能な、ＮＧＳリードのリードマッピング及びデノボアセンブリのためのｄｅ　Ｂｒｕｉｊｎグラフベースのコンピュータプログラムであ

る。
【０１６４】
リードのアセンブリは、１つ又は複数のコンティグを生成する。ホモ接合又は単一標的配列決定の場合、単一コンティグが生成される。ヘテロ接合性の二倍体標的、稀な体細胞変異又は混合試料の場合、例えば、２つ以上コンティグが生成され得る。各コンティグは、そのコンティグを構成するリードからの情報を含む。
【０１６５】
リードをコンティグに組み立てることは、各コンティグに対応するコンセンサス配列を生成するのに役立つ。ある実施形態において、コンセンサス配列は、アセンブルされたリードの中からの各位置における最も一般的な又は優勢なヌクレオチドを指す。コンセンサス配列は、そのコンティグによって表される核酸の配列の解釈を表すことができる。
【０１６６】
本明細書で使用されるアライメントは、一般に、１つの配列を別の配列に沿って配置すること、各配列に沿って反復的にギャップを導入すること、その２つの配列がどの程度よく一致するかをスコアリングすること、及び、好ましくは参照に沿った様々な位置について繰り返すこと、を含む。最良のスコアリング一致は、アライメントであると見なされ、配列間の歴史的関係性に関する推論を表す。アライメントにおいて、参照中の一致しない塩基と並んだリード中の塩基は、その時点で置換突然変異が起こったことを示す。同様に、一方の配列が他方の配列中の塩基と並んでギャップを含む場合、挿入又は欠失突然変異（「インデル」）が生じたと推測される。１つの配列が互いにアライメントされていることを明示することが望まれる場合、アライメントはペアワイズアライメントと呼ばれることがある。多重配列アライメントは、一般に、例えば、一連のペアワイズアライメントを含む、２つ以上の配列のアライメントを指す。
【０１６７】
いくつかの実施形態では、アライメントのスコアリングは、置換及びインデルの確率の値を設定することを含む。個々の塩基がアライメントされる場合、マッチ又はミスマッチは、置換確率によってアライメントスコアに寄与し、これは、例えば、マッチについては１、ミスマッチについては０．３３であり得る。インデルは、例えば、－１とすることができるギャップペナルティによってアライメントスコアから推定する。ギャップペナルティ及び置換確率は、配列がどのように変異するかについての経験的知識又は先験的仮定に基づくことができる。それらの値は、結果として生じるアライメントに影響を及ぼす。特に、ギャップペナルティと置換確率との間の関係は、得られるアライメントにおいて置換又はインデルが優先されるかどうかに影響を及ぼす。
【０１６８】
形式的に言えば、アライメントは、２つの配列ｘとｙとの間の推測される関係性を表す。例えば、いくつかの実施形態では、配列ｘ及びｙのアライメントＡは、（ｉ）｜ｘ’｜＝｜ｙ’｜；（ｉｉ）ｘ’及びｙ’からスペースを除去すると、それぞれｘ及びｙに戻るはずであり、及び（ｉｉｉ）任意のｉについて、ｘ’［ｉ］及びｙ’［ｉ］は両方のスペースであることはできないようにスペースを含有し得る、別の２つの文字列ｘ’及びｙ’にそれぞれｘ及びｙをマップする。
【０１６９】
ギャップは、ｘ’又はｙ’のいずれかにおける連続スペースの最大部分文字列である。アライメントＡは、以下の３種類の領域を含む：（ｉ）一致した対（例えば、ｘ’［ｉ］＝ｙ’［ｉ］；（ｉｉ）ミスマッチ対、（例えば、ｘ’［ｉ］≠ｙ’［ｉ］であり、両方ともスペースではない）；又は（ｉｉｉ）ギャップ（例えば、ｘ’［ｉ．．ｊ］又はｙ’［ｉ．．ｊ］のいずれかはギャップである）を含むことができる。ある実施形態では、一致した対のみが高い陽性スコアａを有する。いくつかの実施形態では、ミスマッチ対は一般に負のスコアｂを有し、長さｒのギャップも負のスコアｇ＋ｒｓを有し、ここでｇ、ｓ＜０である。ＤＮＡの場合、１つの一般的なスコアリングスキーム（例えば、ＢＬＡＳＴによって使用される）は、スコアａ＝１、スコアｂ＝－３、ｇ＝－５及びｓ＝－２とする。アラ

イメントＡのスコアは、全ての一致した対、不一致の対及びギャップのスコアの合計である。ｘ及びｙのアライメントスコアは、ｘ及びｙの全ての可能なアライメントの中の最大スコアとして定義することができる。

【０１７０】

いくつかの実施形態では、任意の対は、置換確率の４×４マトリックスＢによって定義されるスコアａを有する。例えば、Ｂ（ｉ，ｉ）＝１であり、０＜Ｂ（ｉ，ｊ）ｉ＜＞ｊ＜１が、１つの可能なスコアリングシステムである。例えば、転位（ｔｒａｎｓｉｔｉｏｎ）が転換（ｔｒａｎｓｖｅｒｓｉｏｎ）よりも生物学的に可能性が高いと考えられる場合、マトリックスＢは、Ｂ（Ｃ、Ｔ）＝０．７及びＢ（Ａ、Ｔ）＝０．３、又は当技術分野で公知の方法によって所望又は決定される任意の他の値のセットを含み得る。

【０１７１】

本発明のいくつかの実施形態によるアライメントは、ペアワイズアライメントを含む。ペアワイズアライメントは、一般に、ｍ個の文字及びｎ個の文字の参照ゲノムＴ（標的）を有する配列Ｑ（クエリ）について、ＱとＴとの間の可能な局所アライメントを発見及び評価することを含む。ｈ≦ｉ及びｋ≦ｊである場合の、任意の１≦ｉ≦ｎ及び１≦ｊ≦ｍについて、Ｔ［ｈ．．ｉ］及びＱ［ｋ．．ｊ］の可能な最大アライメントスコアが計算される（すなわち、位置ｉで終了するＴの任意の部分文字列及び位置ｊで終了するＱの任意の部分文字列の最良のアライメントスコア）。これは、ｃｍ文字を有する全ての部分文字列を検査することを含むことができ、ｃは類似性モデルに応じた定数であり、各部分文字列をＱと別々に整列させる。各アライメントはスコア付けされ、好ましいスコアとのアライメントが、アライメントとして受け入れられる。いくつかの実施形態では、網羅的なペアワイズアライメントが実施され、これは一般に、ＱとＴとの間の全ての可能な局所アライメント（任意選択的にいくつかの制限基準を受ける）がスコア付けされる上記のペアワイズアライメントを含む。

【０１７２】

いくつかの実施形態では、ペアワイズアライメントは、ドットマトリクス法、動的プログラミング法、又はワード法に従って進行する。動的プログラミング方法は、一般に、Ｓｍｉｔｈ－Ｗａｔｅｒｍａｎ（ＳＷ）アルゴリズム又はＮｅｅｄｌｅｍａｎ－Ｗｕｎｓｃｈ（ＮＷ）アルゴリズムを実装する。ＮＷアルゴリズムによるアライメントは、一般に、線形ギャップペナルティｄで類似度行列Ｓ（ａ、ｂ）（例えば、前述のマトリックスＢ等）に従ってアライメントされた文字をスコアリングする。行列Ｓ（ａ，ｂ）は一般に置換確率を供給する。ＳＷアルゴリズムはＮＷアルゴリズムと同様であるが、負のスコア行列セルはどれも０に設定される。ＳＷ及びＮＷアルゴリズム、並びにそれらの実装形態は、米国特許出願公開第５，７０１，２５６号及び米国特許出願公開第２００９／０１１９３１３号に更に詳細に記載され、両方ともその全体が参照により本明細書に組み込まれる。これらの方法を実施するための当技術分野で知られているコンピュータプログラムを以下により詳細に説明する。

【０１７３】

本発明によるアライメントは、当技術分野で公知の任意の適切なコンピュータプログラムを使用して実行することができる。

【０１７４】

ＢＷＴ手法を実装する１つの例示的なアライメントプログラムは、Ｇｅｅｋｎｅｔ（Ｆａｉｒｆａｘ，Ｖａ．）によって管理されるＳｏｕｒｃｅＦｏｒｇｅウェブサイトから入手可能なＢｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ　Ａｌｉｇｎｅｒ（ＢＷＡ）である。ＢＷＡは、リード、コンティグ又はコンセンサス配列を参照に対しアライメントすることができる。ＢＷＴは、ヌクレオチドあたり２ビットのメモリを占有し、典型的なデスクトップ又はラップトップコンピュータで４Ｇ塩基対の長さのヌクレオチド配列をインデックスすることを可能にする。前処理は、ＢＷＴの構築（すなわち、参照にインデックスを付ける）及びサポート補助データ構造を含む。

【０１７５】

ＢＷＡは、両方ともＢＷＴに基づいて、２つの異なるアルゴリズムを実装する。ＢＷＡによるアライメントは、約２００ｂｐまでの短いクエリに対して低いエラー率（＜３％）で設計されたアルゴリズムｂｗａ－ｓｈｏｒｔを使用して進行することができる（Ｌｉ Ｈ．及びＤｕｒｂｉｎ Ｒ．Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ、２５：１７５４－６０（２００９））。第２のアルゴリズムであるＢＷＡ－ＳＷは、より多くのエラーを伴うロングリード用に設計されている（Ｌｉ Ｈ．及びＤｕｒｂｉｎ Ｒ．（２０１０）Ｆａｓｔ ａｎｄ ａｃｃｕｒａｔｅ ｌｏｎｇ－ｒｅａｄ ａｌｉｇｎｍｅｎｔ ｗｉｔｈ Ｂｕｒｒｏｗｓ－Ｗｈｅｅｌｅｒ Ｔｒａｎｓｆｏｒｍ．Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，Ｅｐｕｂ．）。ＢＷＡ－ＳＷコンポーネントは、ヒューリスティックなＳｍｉｔｈ－Ｗａｔｅｒｍａｎ様アライメントを実行して、高スコアの局所ヒットを見つける。当業者は、ｂｗａ－ｓｗが「ｂｗａ－ｌｏｎｇ」、「ｂｗａ ｌｏｎｇアルゴリズム」等と呼ばれることがあることを認識するであろう。そのような使用は、一般にＢＷＡ－ＳＷを指す。

【０１７６】

Ｓｍｉｔｈ－Ｗａｔｅｒｍａｎアルゴリズムのバージョンを実装するアライメントプログラムはＭＵＭｍｅｒであり、Ｇｅｅｋｎｅｔ（Ｆａｉｒｆａｘ，Ｖａ．）が管理するＳｏｕｒｃｅＦｏｒｇｅウェブサイトから入手可能である。ＭＵＭｍｅｒは、完全形態又はドラフト形態にかかわらず、全ゲノムを迅速にアライメントするためのシステムである（Ｋｕｒｔｚ，Ｓ．，ｅｔ ａｌ．，Ｇｅｎｏｍｅ Ｂｉｏｌｏｇｙ，５：Ｒ１２（２００４）；Ｄｅｌｃｈｅｒ，Ａ．Ｌ．，ｅｔ ａｌ．，Ｎｕｃｌ．Ａｃｉｄｓ Ｒｅｓ．，２７：１１（１９９９））。例えば、ＭＵＭｍｅｒ ３．０は、２．４ＧＨｚのＬｉｎｕｘ（登録商標）デスクトップコンピュータ上で、７８ＭＢのメモリを使用して、１３．７秒で一対の５メガベースのゲノム間の２０塩基対又はそれより長い完全一致を全て見つけることができる。ＭＵＭｍｅｒはまた、不完全なゲノムをアライメントすることができ、それは、ショットガン配列決定プロジェクトからの１００ｓ又は１０００ｓのコンティグを容易に取り扱うことができ、システムに含まれるＮＵＣｍｅｒプログラムを使用してそれらを別のセットのコンティグ又はゲノムにアライメントする。種が、類似性を検出するためのＤＮＡ配列アライメントにはあまりにも多様である場合、ＰＲＯｍｅｒプログラムは、両方の入力配列の６フレーム翻訳に基づいてアライメントを生成することができる。

【０１７７】

本発明の実施形態による別の例示的なアライメントプログラムは、Ｋｅｎｔ Ｉｎｆｏｒｍａｔｉｃｓ（Ｓａｎｔａ Ｃｒｕｚ，Ｃａｌｉｆ．）からのＢＬＡＴである（Ｋｅｎｔ，Ｗ．Ｊ．，Ｇｅｎｏｍｅ Ｒｅｓｅａｒｃｈ ４：６５６－６６４（２００２））。ＢＬＡＴ（ＢＬＡＳＴではない）は、ＲＡＭ等のメモリに参照ゲノムのインデックスを保持する。インデックスは、全ての重複しないｋ－ｍｅｒ（リピートに大きく関与するものを任意選択的に除く）を含み、デフォルトでｋ＝１１である。ゲノム自体はメモリに保持されない。インデックスは、可能性のある相同性の領域を見つけるために使用され、その領域は、その後、詳細なアライメントのためにメモリにロードされる。

【０１７８】

別のアライメントプログラムは、Ｂｅｉｊｉｎｇ Ｇｅｎｏｍｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｂｅｉｊｉｎｇ，ＣＮ）又はＢＧＩ Ａｍｅｒｉｃａｓ Ｃｏｒｐｏｒａｔｉｏｎ（Ｃａｍｂｒｉｄｇｅ，Ｍａｓｓ．）のＳＯＡＰ２である。ＳＯＡＰ２は、双方向ＢＷＴを実装する（Ｌｉ ｅｔ ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ２５（１５）：１９６６－６７（２００９）；Ｌｉ，ｅｔ ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ２４（５）：７１３－１４（２００８））。

【０１７９】

配列を整列させるための別のプログラムは、Ｂｏｗｔｉｅである（Ｌａｎｇｍｅａｄ，ｅｔ ａｌ．，Ｇｅｎｏｍｅ Ｂｉｏｌｏｇｙ，１０：Ｒ２５（２００９））。Ｂｏｗｔｉｅインデックスは、ＢＷＴを作製することによってゲノムを参照する。

【０１８０】

他の例示的なアライメントプログラムには、以下が挙げられる：Ｅｆｆｉｃｉｅｎｔ

Ｌａｒｇｅ－Ｓｃａｌｅ Ａｌｉｇｎｍｅｎｔ ｏｆ Ｎｕｃｌｅｏｔｉｄｅ Ｄａｔａｂａｓｅｓ（ＥＬＡＮＤ）又はＣｏｎｓｅｎｓｕｓ Ａｓｓｅｓｓｍｅｎｔ ｏｆ Ｓｅｑｕｅｎｃｅ ａｎｄ Ｖａｒｉａｔｉｏｎ（ＣＡＳＡＶＡ）ソフトウェアのＥＬＡＮＤ ｖ２コンポーネント（Ｉｌｌｕｍｉｎａ，Ｓａｎ Ｄｉｅｇｏ，Ｃａｌｉｆ．）；Ｒｅａｌ Ｔｉｍｅ Ｇｅｎｏｍｉｃｓ，Ｉｎｃ．（Ｓａｎ Ｆｒａｎｃｉｓｃｏ，Ｃａｌｉｆ．）からのＲＴＧ Ｉｎｖｅｓｔｉｇａｔｏｒ；Ｎｏｖｏｃｒａｆｔ（Ｓｅｌａｎｇｏｒ，Ｍａｌａｙｓｉａ）からのＮｏｖｏａｌｉｇｎ；Ｅｘｏｎｅｒａｔｅ、Ｅｕｒｏｐｅａｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｈｉｎｘｔｏｎ，ＵＫ）（Ｓｌａｔｅｒ，Ｇ．，ａｎｄ Ｂｉｒｎｅｙ，Ｅ．，ＢＭＣ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ ６：３１（２００５））；ユニバーシティ・カレッジ・ダブリン（Ｄｕｂｌｉｎ，Ｉｒｅｌａｎｄ）からのＣｌｕｓｔａｌ Ｏｍｅｇａ（Ｓｉｅｖｅｒｓ Ｆ．，ｅｔ ａｌ．，Ｍｏｌ Ｓｙｓｔ Ｂｉｏｌ ７，ａｒｔｉｃｌｅ ５３９（２０１１））；ユニバーシティ・カレッジ・ダブリン（Ｄｕｂｌｉｎ，Ｉｒｅｌａｎｄ）からのＣｌｕｓｔａｌ Ｗ又はＣｌｕｓｔａｌ Ｘ（Ｌａｒｋｉｎ Ｍ．Ａ．，ｅｔ ａｌ．，Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ，２３，２９４７－２９４８（２００７））；及び、ＦＡＳＴＡ，Ｅｕｒｏｐｅａｎ Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ Ｉｎｓｔｉｔｕｔｅ（Ｈｉｎｘｔｏｎ，ＵＫ）（Ｐｅａｒｓｏｎ Ｗ．Ｒ．，ｅｔ ａｌ．，ＰＮＡＳ ８５（８）：２４４４－８（１９８８）；Ｌｉｐｍａｎ，Ｄ．Ｊ．，Ｓｃｉｅｎｃｅ ２２７（４６９３）：１４３５－４１（１９８５）。

【０１８１】

図１３は、本開示の１つ又は複数の実施形態による最大状態経路を決定するための例示的な手順を示し、例示する。例えば、非汎用の、具体的に構成されたデバイス（例えば、システム７０１）は、格納された命令を実行することによって手順１２００を実行することができる。手順１２００は、工程１２０５で開始し、工程１２１０に進行し得、ここで、上で詳細に記載されるように、プロセスは、配列要素の線状アレイを有する個々の核酸配列リードを含む複数の核酸配列リードを得てもよい。実施形態において、高複雑度のライブラリから引き出された各核酸配列要素は、低複雑度の１つ若しくは複数の予想される核酸配列の、又は低複雑度の１つ若しくは複数の予想される核酸配列及び配列リード終端のいずれかに隣接し得る。

【０１８２】

工程１２１５において、プロセスは、高複雑度のライブラリ及び低複雑度のライブラリから引き出された個々の核酸配列要素の領域を予測するために、複数の核酸配列リードに１つ又は複数の統計的アノテーションモデルを適用し得る。実施形態において、１つ又は複数の統計的アノテーションモデルは、ｉ）核酸配列リード全体に散在する１つ又は複数の予想される核酸配列を認識するための生成統計的アライメントモデル、又はｉｉ）既知ではない配列又は高複雑度の配列の辞書から引き出された配列を認識するためのランダム統計アライメントモデルを含み得る。実施形態では、予測された転位部位は、各モデルの末端に配置され、生成統計的アライメントモデルの内部位置内では許容されない。

【０１８３】

工程１２２０において、前の２つの工程を複数の核酸配列リードに対して繰り返すことができる。次いで、工程１２２５において、プロセスは、最大対数尤度値を有するモデルを識別することによって選択された最大事後状態経路の最終的リード当たりのモデル選択を決定することができる。このようにして、次いで、プロセスは、１つ又は複数の統計的モデルを複数の核酸配列リードの各核酸配列リードに順相補性配向及び逆相補性配向の両方で適用し、最大対数尤度値を有するモデルを識別することによって選択された最大事後状態経路の最終的リード当たりのモデル選択を決定することができる。

【０１８４】

次いで、工程１２３０において、プロセスは、複数の核酸配列リードの各核酸配列リードを、最大事後状態経路の最終的リード当たりのモデルによって識別される転位部位によって区画された個別の配列要素にセグメント化することができ、これにより、複数の核酸

配列リード内の個別の配列要素を識別することができる。

【０１８５】

次いで、工程１２３５において、プロセスは、複数の核酸配列リード内で同定された別個の配列要素を、配列要素データファイルに保存し得る。簡略化された手順１７００は、例示的に、新しいプロセスが開始されるまで、工程１２４０で終了することができる。

【０１８６】

キット

本開示はまた、本開示の方法で使用するための本開示の薬剤を含有するキットを提供する。本開示のキットは、本開示の薬剤及び／又は組成物を含む１つ又は複数の容器を含み得る。いくつかの実施形態において、キットは、本開示の方法に従って使用するための説明書を更に含む。

【０１８７】

本開示のキットで提供される説明書は、典型的には、ラベル又は添付文書（例えば、キットに含まれる紙のシート）に記載された説明書であるが、機械可読説明書（例えば、磁気又は光ストレージディスク上で実行される命令）も許容される。本明細書に記載の方法のいずれかを実施するための説明書を提供することができる。

【０１８８】

本開示のキットは適切な包装中にある。適切な包装としては、バイアル、ボトル、瓶、フレキシブル包装（例えば、密封されたマイラー又はビニール袋）等が挙げられるが、これらに限定されない。容器は、薬学的に活性な薬剤を更に含み得る。

【０１８９】

キットは、必要に応じて、バッファ及び説明的情報等の追加の構成要素を提供することができる。通常、キットは、容器と、容器上の又は容器に関連するラベル又は添付文書（複数可）とを含む。

【０１９０】

本開示の実施は、特に明記しない限り、当業者の技能の範囲内である化学、分子生物学、微生物学、組換えＤＮＡ、遺伝学、免疫学、細胞生物学、細胞培養及びトランスジェニック生物学の従来の技術を使用する。例えば、Ｍａｎｉａｔｉｓ ｅｔ ａｌ．，１９８２，Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ（Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ Ｐｒｅｓｓ，Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ，Ｎ．Ｙ．）；Ｓａｍｂｒｏｏｋ ｅｔ ａｌ．，１９８９，Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ，２ｎｄ Ｅｄ．（Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ Ｐｒｅｓｓ，Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ，Ｎ．Ｙ．）；Ｓａｍｂｒｏｏｋ ａｎｄ Ｒｕｓｓｅｌｌ，２００１，Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ，３ｒｄ Ｅｄ．（Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ Ｐｒｅｓｓ，Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ，Ｎ．Ｙ．）；Ａｕｓｕｂｅｌ ｅｔ ａｌ．，１９９２），Ｃｕｒｒｅｎｔ Ｐｒｏｔｏｃｏｌｓ ｉｎ Ｍｏｌｅｃｕｌａｒ Ｂｉｏｌｏｇｙ（Ｊｏｈｎ Ｗｉｌｅｙ＆Ｓｏｎｓ，ｉｎｃｌｕｄｉｎｇ ｐｅｒｉｏｄｉｃ ｕｐｄａｔｅｓ）；Ｇｌｏｖｅｒ，１９８５，ＤＮＡ Ｃｌｏｎｉｎｇ（ＩＲＬ Ｐｒｅｓｓ，Ｏｘｆｏｒｄ）；Ａｎａｎｄ，１９９２；Ｇｕｔｈｒｉｅ ａｎｄ Ｆｉｎｋ，１９９１；Ｈａｒｌｏｗ ａｎｄ Ｌａｎｅ，１９８８，Ａｎｔｉｂｏｄｉｅｓ，（Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ Ｌａｂｏｒａｔｏｒｙ Ｐｒｅｓｓ，Ｃｏｌｄ Ｓｐｒｉｎｇ Ｈａｒｂｏｒ，Ｎ．Ｙ．）；Ｊａｋｏｂｙ ａｎｄ Ｐａｓｔａｎ，１９７９；Ｎｕｃｌｅｉｃ Ａｃｉｄ Ｈｙｂｒｉｄｉｚａｔｉｏｎ（Ｂ．Ｄ．Ｈａｍｅｓ＆Ｓ．Ｊ．Ｈｉｇｇｉｎｓ ｅｄｓ．１９８４）；Ｔｒａｎｓｃｒｉｐｔｉｏｎ Ａｎｄ Ｔｒａｎｓｌａｔｉｏｎ（Ｂ．Ｄ．Ｈａｍｅｓ＆Ｓ．Ｊ．Ｈｉｇｇｉｎｓ ｅｄｓ．１９８４）；Ｃｕｌｔｕｒｅ Ｏｆ Ａｎｉｍａｌ Ｃｅｌｌｓ（Ｒ．Ｉ．Ｆｒｅｓｈｎｅｙ，Ａｌａｎ Ｒ．Ｌｉｓｓ，Ｉｎｃ．，１９８７）；Ｉｍｍｏｂｉｌｉｚｅｄ Ｃｅｌｌｓ Ａｎｄ Ｅｎｚｙｍｅｓ（ＩＲＬ Ｐｒｅｓｓ，１９８６）；Ｂ．Ｐｅｒｂａｌ，Ａ Ｐｒａｃｔｉｃａｌ Ｇｕｉｄｅ Ｔｏ Ｍｏｌｅｃｕｌａｒ Ｃｌｏｎｉｎｇ（１９８４）；ｔｈ

e treatise, Methods In Enzymology (Academic Press, Inc., N.Y.); Gene Transfer Vectors For Mammalian Cells (J.H. Miller and M.P. Calos eds., 1987, Cold Spring Harbor Laboratory); Methods In Enzymology, Vols. 154 and 155 (Wu et al. eds.), Immunochemical Methods In Cell And Molecular Biology (Mayer and Walker, eds., Academic Press, London, 1987); Handbook Of Experimental Immunology, Volumes I-IV (D.M. Weir and C.C. Blackwell, eds., 1986); Riott, Essential Immunology, 6th Edition, Blackwell Scientific Publications, Oxford, 1988; Hogan et al., Manipulating the Mouse Embryo, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1986); Westerfield, M., The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio), (4th Ed., Univ. of Oregon Press, Eugene, 2000) を参照されたい。

【0191】

他に定義されない限り、本明細書で使用される全ての技術用語及び科学用語は、本開示が属する技術分野の当業者によって一般的に理解されるのと同じ意味を有する。本明細書に記載の方法及び材料と類似又は同等の方法及び材料を本開示の実施又は試験に使用することができるが、適切な方法及び材料を以下に記載する。本明細書で言及される全ての刊行物、特許出願、特許、及び他の参考文献は、その全体が参照により組み込まれる。矛盾する場合、定義を含む本明細書が優先する。更に、材料、方法、及び例は例示にすぎず、限定することを意図するものではない。

【0192】

ここで、本開示の例示的な実施形態を詳細に参照する。本開示は、例示的な実施形態に関連して説明されるが、本開示をそれらの実施形態に限定することを意図するものではないことが理解されよう。逆に、添付の特許請求の範囲によって定義される本開示の趣旨及び範囲内に含まれ得る代替、修正、及び均等物を網羅することが意図されている。当技術分野で周知の標準的な技術又は以下に具体的に記載される技術を利用した。

【実施例】

【0193】

実施例１：ＣＡｓｅｑプロセス

最近の試みは、単一細胞遺伝子発現試料からアイソフォーム配列決定を行うためにロングリード配列決定プラットフォームを活用してきたが、それらのワークフローは、これまで、不十分なスループット及び実質的な配列決定アーチファクトに悩まされており、リードの約３５～５０％しかフィルタを通過せず、フローセル当たり約３００，０００個の配列決定された転写物に相当する（約６５０～８００ドル）。ある態様において、本開示は、例えばＰａｃｉｆｉｃ Ｂｉｏｓｃｉｅｎｃｅｓ（ＰａｃＢｉｏ（登録商標））からの最近更新されたＳｅｑｕｅｌ ＩＩプラットフォーム上で、１０ｘ単一細胞遺伝子発現試料からのハイスループット完全転写配列決定を可能にする「ＣＡｓｅｑ」プロセスを提供する。本開示のＣＡｓｅｑプロセスの使用は、観察される配列決定アーチファクトの割合を１０％未満に減少させることを可能にし、一方で、全長配列決定出力をフローセルあたり約２５Ｍ全長転写物に増強することも可能にする。これを達成するために、多重ライゲーションのために、１５塩基対（ｂｐ）の相補的配列を増幅し、全長ｃＤＮＡライブラリに付加するためのｄＵ含有プライマーのファミリーが設計されている。アーチファクト配列の主要な供給源に対処するために、例示されたプロセスは、全長ｃＤＮＡアンプリコン

の精製を可能にするためにビオチン化プライマーを使用する。効率的な多重化アセンブリを駆動し、不適切なライゲーション事象を軽減するために、本明細書に例示される１５ｂｐ相補的配列は、全ての配列が互いに少なくとも１１ハミング距離単位離れていることを確実にすることによって、最小の類似性を有するように設計された（Ｂｕｓｃｈｍａｎｎ，Ｔ．Ｂｉｏｃｏｎｄｕｃｔｏｒ ｖｅｒｓｉｏｎ：Ｒｅｌｅａｓｅ（３．１１）．ＤＯＩ：１０．１８１２９／Ｂ９．ｂｉｏｃ．ＤＮＡＢａｒｃｏｄｅｓ）。更なる設計上の考慮事項は、１５～２０ｋｂの多重化アレイの生成、すなわち、Ｓｅｑｕｅｌ ＩＩの出力及び塩基呼出し精度のバランスをとるための現在の最適な長さを保証することであった。適切なサイズのライブラリは、ｃＤＮＡのサイズ分布に基づいて、集められた断片の数をプログラムすることによって構築される。多重ロングリード及び単一細胞遺伝子発現データを処理及び統合するために、分析パイプラインも調製する。

【０１９４】

実施例２：予備実験で効率的に生成された線状キメラアレイのＣＡｓｅｑ

予備的なＣＡｓｅｑランにおいて、１．２ｋｂの平均断片サイズを有するｃＤＮＡライブラリからの８断片多重化アセンブリを行い、ライゲーション時に約１０ｋｂの多重化断片を得た（図２Ａ）。多重化ライブラリをＳｅｑｕｅｌ ＩＩで配列決定し、これにより、合計約２．５Ｍのリードが得られ、逆多重化後に約２３Ｍの転写物が得られ、これはスループットの約９倍の増加を表した（図２Ｂ）。逆多重化されたリードの分析により、元のｃＤＮＡライブラリと同様のサイズ分布が確認された（図２Ａ）。

【０１９５】

例示されたｃＤＮＡライブラリサイズ分布は、効果的な線状キメラアレイを形成することを可能にしたが、サイズ選択はまた、ある特定の状況下では、キメラアレイからの効果的な配列収率を増加させることが予想される入力核酸ライブラリ（例えば、キメラアレイライゲーションプロセスの実施前に、電気泳動又は入力核酸ライブラリの他の分離を介して）に対して行うこともでき、特に個々のリード長がメガベースである場合、配列された別個の配列の総数が多く、及び／又は核酸サイズ範囲の元の分布が分散していることが更に企図される。

【０１９６】

実施例３：改善されたデータアノテーション、逆多重化及びセグメント化方法によるＣＡｓｅｑリード収率の向上

本開示のキメラアンプリコンスアレイの最初の処理は、既存のゲノムリードアライメントソフトウェアに基づく反復アダプタ発見戦略を用いた既存のサーキュラーコンセンサスシーケンシング（ｃｉｒｃｕｌａｒ ｃｏｎｓｅｎｓｕｓ ｓｅｑｕｅｎｃｉｎｇ）（ＣＣＳ）の正確な高忠実度ロングリード（ＨｉＦｉリード）プロセスを用いた。このプロセスは、本キメラアンプリコンアレイのロングリードからの配列データの抽出に最適ではないと確認され、ＣＡｓｅｑリードの分析のための改良された方法の開発が開始された。それによって、以下の実施による、キメラアンプリコンアレイ配列決定リードの統計的配列アノテーション、逆多重化、及びセグメント化を含む「Ｌｏｎｇｂｏｗ」と呼ばれる改良されたＣＡｓｅｑリード解析プロセスが設計された：

（１）アンプリコンアレイ配列及びそれらの間の転位を識別するための１つ又は複数の統計的アノテーションモデル（例えば、複数のリンクされたサブモデルを有するプロファイル隠れマルコフモデル）を使用したキメラアンプリコンアレイ配列決定データのアノテーション。その１つ又は複数の統計的アノテーションモデルは、（ａ）キメラアンプリコンアレイ配列決定リード全体に散在する先験的に予想される核酸配列（すなわち、アダプタ配列）を認識するための生成統計的アライメントモデル；（ｂ）先験的に知られていない配列（例えばｃＤＮＡ転写物配列）を認識するための、又は後の処理工程（例えば、単一細胞バーコード配列、固有の分子識別子）で異なる考慮事項に値するほど大きい配列の辞書からの、ランダム統計アライメントモデルを含み、転位が各モデルの末端に配置され、そのアダプタ配列モデル内の内部位置内では許容されない；

（２）最大対数尤度値を有するモデルを評価し、それによってキメラアンプリコンアレ

イ配列決定リードを逆多重化することによって決定された、最大事後状態経路の最終的リード当たりのモデル選択の決定による順相補配向及び逆相補配向の両方における各長いリードへの上記工程（１）の統計的アノテーションモデルの反復適用；及び
（３）上記の工程（１）及び（２）の実行によって同定された部位におけるキメラアンプリコンアレイ配列決定リードのセグメント化。

【０１９７】

上に開示された「Ｌｏｎｇｂｏｗ」プロセスは、少なくとも、（１）サーキュラーコンセンサスシーケンシング（ＣＣＳ）ソフトウェアによって最初に同定されたリードの集団から実際には低品質の配列リードを、名目上高品質であると識別し、除去すること、（２）サーキュラーコンセンサスシーケンシング（ＣＣＳ）ソフトウェアによって最初に使用不可能な品質であると主張されるリードの集団から高品質の配列リードを救済すること、（３）「Ｌｏｎｇｂｏｗ」プロセスから新たに同定された高品質リードの品質を概算すること、への適用を照らして、本開示のキメラアンプリコンアレイからの品質管理及び配列データ収率の増強に有用であると更に確認された。そのような各用途は、以下で更に詳細に検討される。

【０１９８】

本開示のキメラアンプリコンアレイ配列決定から潜在的に低品質のデータを識別するために、方法は、（ａ）シーケンサによって高品質であると確認されたキメラアンプリコンアレイ配列決定リードに（上記のような）Ｌｏｎｇｂｏｗモデルを適用すること（それによって、これらのリードのそれぞれにおける各ヌクレオチドを、それが由来するライブラリアダプタ配列で標識すること）；（ｂ）等しい隣接するＬｏｎｇｂｏｗのヌクレオチド標識を、その標識された部分全体を含む領域にマージすること；及び（ｃ）全ての標識されたリードにわたって反復し、その順序で生じない標識された部分を有する任意のリードを、そのライブラリ調製により予想される順序で同定すること、を含む。最初の予想されるセグメントの後に始まるが、その残りのセクションが順番通りであるリード、及び最後の予想されるセグメントの前に終わるが、その前のセクションが全て順番通りであるリード、並びにこれらの場合の組合わせはこれから除外される。予想されるライブラリに適合しないリードは、低品質と見なされる。

【０１９９】

シーケンサによって低品質で使用不可能であると報告されたサブセットから高品質の配列決定データを同定するために、方法は以下の工程を含む。（ａ）シーケンサが使用不可能な品質として報告したデータ（すなわち、リード）を同定すること。そのような使用不可能な品質データは、データに非常に低いリード品質スコア（０未満の値、０～０．５の値、及び０．５～１．０の値を含むが、これらに限定されない）を割り当てるサーキュラーコンセンサスシーケンシングのソフトウェアによって、又はリードを「ＺＭＷパスフィルタ」以外の任意のカテゴリに割り当てるサーキュラーコンセンサスシーケンシングのソフトウェアのいずれかによって決定される。（ｂ）使用不可能な品質のこれらのリードに（上記のような）Ｌｏｎｇｂｏｗモデルを適用し、それにより、これらのリードのそれぞれにおける各ヌクレオチドを、それが由来するライブラリアダプタ配列で標識すること。（ｃ）等しい隣接するＬｏｎｇｂｏｗのヌクレオチド標識を、標識された部分全体を含む領域にマージすること。そして、（ｄ）全ての標識されたリードを反復し、最初の予想されるセグメントの後に始まるが、その残りのセクションが順番であるリード、及び最後の予想されるセグメントの前に終わるが、その前のセクションが順番であるリード、並びにこれらの場合の任意の組合わせを含めて、そのライブラリ調製により出現すると予想される順序で標識された部分を有する任意のリードを同定すること。そのようなリードは、そのリードが更なる分析のために十分に高品質であることを示す予想されるライブラリ調製に適合する。前述のプロセスは、サーキュラーコンセンサスシーケンシングのソフトウェアによって、例えば、０．９９未満のリード品質が割り当てられた、又は「ＺＭＷパスフィルタ」以外の任意のカテゴリが割り当てられた、使用不可能なデータに適用するための例示であるが、このプロセスはまた、任意の主張される品質の任意のリード又はリードの

集団にも適用できることが明示的に記載される。

【０２００】

Ｌｏｎｇｂｏｗプロセスの新たに同定された高品質リードの品質を近似するために、方法は以下の工程を含む。（ａ）新たに同定された高品質リードの各々の標識された部分について、標識された部分のヌクレオチドとその部分に対する予想される配列との間のアライメントスコアを計算すること。このアライメントスコアは、Ｓｍｉｔｈ－Ｗａｔｅｒｍａｎ又はＮｅｅｄｌｅｍａｎ－Ｗｕｎｓｃｈアルゴリズム等の動的プログラミングアルゴリズムを使用して直接計算することができ、又は、標識された部分と予想される配列との間のレーベンシュタイン距離を計算して、その距離をその予想される配列の長さから減算することによって直接計算することができる。（ｂ）このアライメントスコアを最良のアライメントスコア（予想される配列とそれ自体との間のアライメントスコアを計算することによって得ることができる）で除算して、各セクションの品質を得ること。そして、（ｃ）（ａ）で計算された全てのアライメントスコアを合計して、全体のアライメントスコアを得ること。（ｂ）で計算された全ての最良のアライメントスコアを合計して、全体の最良のアライメントスコアを得る。全体のアライメントスコアと全体の最良のアライメントスコアとの比は、リードの推定品質である。

【０２０１】

実施例４：ＣＯＶＩＤ－１９患者試料の評価のための拡張性のある単一細胞アイソフォーム配列決定ワークフローにおけるＣＡｓｅｑの実装

単一細胞遺伝子発現研究からの遺伝子アイソフォーム組成の解明は、以前は不可能であった。選択的スプライシングは、転写物の成熟中に差次的なエクソンスプライシングによって内在性タンパク質の構造及び機能を調節するコア調節プロセスである。選択的スプライシングから得られる遺伝子アイソフォームは、細胞のシグナル伝達及び機能の媒介において中心的役割を果たすことが示されている（Ｂａｒａｌｌｅ ａｎｄ Ｇｉｕｄｉｃｅ．Ｎａｔ Ｒｅｖ Ｍｏｌ Ｃｅｌｌ Ｂｉｏｌ １８：４３７－４５１）。細胞発生及び恒常性維持を超えて、遺伝子アイソフォームは、複数の病状又は腫瘍の進行及び耐性を駆動する異常なスプライシングに関連する顕著なアイソフォームを有する複数の病状に関与している（Ｋｉｍ ｅｔ ａｌ．Ｐｆｌｕｇｅｒｓ Ａｒｃｈ－Ｅｕｒ Ｊ Ｐｈｙｓｉｏｌ ４７０：９９５－１０１６；Ｓｃｏｔｔｉ ａｎｄ Ｓｗａｎｓｏｎ．Ｎａｔ Ｒｅｖ Ｇｅｎｅｔ １７：１９－３２）。単一細胞解像度でアイソフォーム組成物を効果的に捕捉することができないことは、不均一な生物学的系を効果的に特徴付けるための上述の方法の能力に重大な欠陥があることを強調する。

【０２０２】

本実施例では、本開示のＣＡｓｅｑプロセスを用いて、単一細胞遺伝子発現試料に対してハイスループットアイソフォーム配列決定を実施する。アイソフォーム及び単一細胞遺伝子発現データを処理及び統合するためのパイプラインは、当技術分野で認識されている分析ツールを使用して開発される。標的化アイソフォーム配列決定のために、遺伝子パネルも開発されている。免疫応答及び感染組織の両方を特徴付けるために、ＣＯＶＩＤ－１９患者を評価する。

【０２０３】

ＣＯＶＩＤ－１９症状は、部分的には、ＳＡＲＳ－ＣｏＶ－２感染に対する過活動免疫応答に起因して生じる。本開示の実施例では、ＣＡｓｅｑをＣＯＶＩＤ－１９試料（３００名のＣＯＶＩＤ－１９患者の血液及び約１０名の剖検由来の組織からの免疫区画の進行中の単一細胞ゲノム研究に由来する）に対し使用し、疾患の重症度に関連する免疫細胞クラスターにおいて差次的に発現されるアイソフォームを発見することを目的とする。

【０２０４】

（非ＣＡｓｅｑ）予備データの初期セットは、健康な患者と軽度及び重度のＣＯＶＩＤ－１９患者との間の単球区画における著しい転写の違いを確認している（図１０Ａ～１０Ｄ）。アイソフォーム分析は、炎症及び単球活性化経路に関連する遺伝子に焦点を当てているが、これらに限定されない（ｄｏｉ．ｏｒｇ／１０．１０９３／ｎａｒ／ｇｋｙ４０

１及びｄｏｉ．ｏｒｇ／１０．１０３８／ｓ４１４６７－０１９－１１０７６－１を参照されたい）。アイソフォーム分析の出力を高めるために、Ｌｅｉｄｅｎクラスターを一緒にグループ化して、クラスター間の差次的なアイソフォーム組成のより堅牢な統計的比較を可能にする。ＳＡＲＳ－ＣｏＶ２感染試料を健康な対照患者と比較して、遺伝子発現の違い及び選択的スプライシングの役割を特徴付けた。ＳＡＲＳ－ＣｏＶ２は、そのゲノムからの転写の複雑な不連続プロセスを利用することが示され、ショートリード配列決定は特にウイルス遺伝子発現の解明に適さないので、ＳＡＲＳ－ＣｏＶ２トランスクリプトームの再構築は洞察力が期待される。感染過程にわたる潜在的な転写動態に光を当てるために、感染細胞でのウイルス転写物の組成及び量との潜在的な関連がそれによって調査される。

【０２０５】

実施例５：単一細胞遺伝子発現試料からのミトコンドリア系統追跡

腫瘍内不均一性及びクローン進化は、腫瘍進行及び治療耐性を可能にする推進力である。クローン動態を追跡する能力は、治療に直面して腫瘍がどのように進化しているかを理解するために重要である。最近のアプローチは、ミトコンドリア変異がクローン同一性を推論するためのマーカとして役立ち得ることを実証している（Ｌｕｄｗｉｇ ｅｔ ａｌ．Ｃｅｌｌ １７６：１３２５－１３３９）。そのようなアプローチは、ミトコンドリアゲノムが核ゲノムと比較してはるかに高い割合（１０～１００倍）で突然変異を起こし、配列決定データに高度に表されているという事実に部分的に依存している。当技術分野で認識されているショートリード単一細胞遺伝子発現ワークフローからのカバレッジは制限されるために、研究者らはこれまで、クローン推論に必要なミトコンドリアゲノムの均一かつ十分なカバレッジを提供するために、単一細胞ＡＴＡＣ（トランスポザーゼアクセシブルクロマチンのアッセイ（Ａｓｓａｙ ｆｏｒ Ｔｒａｎｓｐｏｓａｓｅ Ａｃｃｅｓｓｉｂｌｅ Ｃｈｒｏｍａｔｉｎ））配列決定に依存してきた。本実施例では、本開示のＣＡｓｅｑアプローチを適用して、単一細胞遺伝子発現試料からの完全ミトコンドリア転写物の標的化ロングリード配列決定を実施し、それによって遺伝子発現試料とのクローン同一性の統合を可能にする。現在のミトコンドリア系統追跡バイオインフォマチックパイプラインを適用し、当技術分野で認識されている方法に対してベンチマークを実施して、全長転写物データで動作するように適合させる。次いで、患者の腫瘍試料を、本ＣＡｓｅｑプロセスを使用して評価して、治療の過程にわたるクローン動態を明らかにする。全ミトコンドリア転写物のＣＡｓｅｑ対応標的化ロングリード配列決定によってクローン情報を抽出する能力は、クローン性と同じ試料からの遺伝子発現との連結を提供する。クローン性及び遺伝子発現のこのような協調的評価は、進行及び治療耐性の過程にわたって腫瘍におけるクローン進化の研究を劇的に向上させる。

【０２０６】

実施例６：単一細胞遺伝子発現試料からのミトコンドリア転写物捕捉及び多重ライゲーションの最適化

これまで、単一細胞遺伝子発現ワークフローは、対立遺伝子情報を、個々の細胞からのクローン関係の堅牢な再構築を可能にする程度まで捕捉するには不十分であった。広く使用されている単一細胞遺伝子発現データから得られたクローン関係を明らかにする能力は、深い洞察を促進し、遺伝子発現状態、クローン性及び細胞運命間の連結を同定することを可能にするため、このことは計り知れない機会の損失を表している。これまで単一細胞遺伝子発現試料からのクローン再構築を妨げてきた低いカバレッジに対処するために、本明細書に開示されるＣＡｓｅｑはまた、全長ミトコンドリア転写物配列情報を得ることを標的とする。ミトコンドリア転写物の高効率配列決定は、本明細書の他の箇所に記載されている多重化プライマーを使用してミトコンドリアから発現される１３個の遺伝子の標的化増幅を行うことによって達成される。配列決定出力及び忠実度のバランスをとりながら、１５～２０ｋｂの最適な多重化アレイ長を確保するために、ミトコンドリアｃＤＮＡプールの長さ分布を考慮して、組み立てられた断片の数が確立される。配列決定されると、全長転写物は、マッピング及び塩基品質のために逆多重化及びフィルタリングされる。リ

ード通過フィルタは、ミトコンドリアゲノムのカバレッジを定量化するために使用される。既存のミトコンドリア系統追跡パイプラインはまた、クローン関係の再構築のために全長ミトコンドリア転写物を使用するように適合されている。

【０２０７】

実施例７：全長ミトコンドリア転写物系統追跡のベンチマーキング

完全長ミトコンドリア転写物系統追跡を検証するために、安定に組み込まれたＤＮＡバーコードを有するＨｅＬａ細胞株集団からクローン関係を再構築する能力を定量化し、これは、クローン同一性のためのグラウンドトゥルースを確立するのに役立ち得る。具体的には、ＣｌｏｎＭａｐｐｅｒ発現バーコードシステム（単一細胞ＲＮＡ配列決定によるクローン同定を可能にする以前に開発されたシステム）でタグ付けされた細胞を使用する。更に、Ｌｕｄｗｉｇ　ｅｔ　ａｌ．（Ｃｅｌｌ　１７６：１３２５－１３３９）に記載の方法を、バーコード化集団の並行試料に対して実施し、特異性及びリコールに関連する測定値をクローン同一性の割り当てのために計算し、比較する。

【０２０８】

したがって、本明細書に開示されるＣＡｓｅｑプロセスは、既存のプラットフォームによってこれまで達成できなかった配列決定スループット及びリード長を可能にするため、配列決定の分野における重要な進歩を提供する。更に、本ＣＡｓｅｑプロセスは、高度に適合可能であり、目的の遺伝的特徴を捕捉するために容易に特殊化することができる。本開示に記載されたＣＡｓｅｑの実装形態は、発見のための新しいプラットフォームとして提供され、多くの科学分野に広く適用可能である。本ＣＡｓｅｑアプローチは、ロングリードプラットフォームと共進化する能力を有し、それらのリード長が増加し続けるにつれてそれらの分子出力を更に高めるのに役立つ。

【０２０９】

参考文献

１．Ｉ．Ｇｕｐｔａ　ｅｔ　ａｌ．，Ｓｉｎｇｌｅ－ｃｅｌｌ　ｉｓｏｆｏｒｍ　ＲＮＡ　ｓｅｑｕｅｎｃｉｎｇ　ｃｈａｒａｃｔｅｒｉｚｅｓ　ｉｓｏｆｏｒｍｓ　ｉｎ　ｔｈｏｕｓａｎｄｓ　ｏｆ　ｃｅｒｅｂｅｌｌａｒ　ｃｅｌｌｓ．Ｎａｔ　Ｂｉｏｔｅｃｈｎｏｌ．３６：１１９７－１２０２（２０１８）．

２．Ｒ．Ｖｏｌｄｅｎ　ｅｔ　ａｌ．，Ｉｍｐｒｏｖｉｎｇ　ｎａｎｏｐｏｒｅ　ｒｅａｄ　ａｃｃｕｒａｃｙ　ｗｉｔｈ　ｔｈｅ　Ｒ２Ｃ２　ｍｅｔｈｏｄ　ｅｎａｂｌｅｓ　ｔｈｅ　ｓｅｑｕｅｎｃｉｎｇ　ｏｆ　ｈｉｇｈｌｙ　ｍｕｌｔｉｐｌｅｘｅｄ　ｆｕｌｌ－ｌｅｎｇｔｈ　ｓｉｎｇｌｅ－ｃｅｌｌ　ｃＤＮＡ．Ｐｒｏｃ　Ｎａｔｌ　Ａｃａｄ　Ｓｃｉ　Ｕ　Ｓ　Ａ　１１５：９７２６－９７３１（２０１８）．

３．Ｍ．Ｓｉｎｇｈ　ｅｔ　ａｌ．，Ｈｉｇｈ－ｔｈｒｏｕｇｈｐｕｔ　ｔａｒｇｅｔｅｄ　ｌｏｎｇ－ｒｅａｄ　ｓｉｎｇｌｅ　ｃｅｌｌ　ｓｅｑｕｅｎｃｉｎｇ　ｒｅｖｅａｌｓ　ｔｈｅ　ｃｌｏｎａｌ　ａｎｄ　ｔｒａｎｓｃｒｉｐｔｉｏｎａｌ　ｌａｎｄｓｃａｐｅ　ｏｆ　ｌｙｍｐｈｏｃｙｔｅｓ．Ｎａｔ　Ｃｏｍｍｕｎ．１０：３１２０（２０１９）．

【０２１０】

本明細書で言及される全ての特許及び刊行物は、本開示が関係する当業者の技術レベルを示す。本開示において引用された全ての参考文献は、あたかも各参考文献が個別にその全体が参照により組み込まれたのと同程度に、参照により組み込まれる。

【０２１１】

当業者は、本開示が目的を実行し、言及された目的及び利点、並びにそれらに固有の目的及び利点を得るようによく適合されていることを容易に理解するであろう。本開示の好ましい実施形態の代表例として本明細書に記載される方法及び組成物は例示的なものであり、本開示の範囲に対する限定として意図されるものではない。その中の変更及び他の使用が当業者には思い浮かぶであろうが、それらは本開示の趣旨の範囲内に包含され、特許請求の範囲によって定義される。

【０２１２】

更に、本開示の特徴又は態様がマーカッシュ群又は代替物の他のグループ化に関して記載されている場合、当業者は、本開示がそれによってマーカッシュ群又は他の群の任意の個々のメンバー又はメンバーのサブグループに関しても記載されることを認識するであろう。

【０２１３】

本開示を説明する文脈において（特に以下の特許請求の範囲の文脈において）「ａ」及び「ａｎ」及び「ｔｈｅ」という用語並びに同様の指示対象の使用は、本明細書で特に指示されない限り、又は文脈と明らかに矛盾しない限り、単数及び複数の両方を包含すると解釈されるべきである。「含む（ｃｏｍｐｒｉｓｉｎｇ）」、「有する（ｈａｖｉｎｇ）」、「を含む（ｉｎｃｌｕｄｉｎｇ）」、及び「含有する（ｃｏｎｔａｉｎｉｎｇ）」という用語は、特に明記しない限り、オープンエンド用語（すなわち、「を含むが、限定されない」を意味する）と解釈されるべきである。本明細書における値の範囲の列挙は、本明細書に別段の指示がない限り、範囲内に含まれる各別個の値を個別に参照する簡略方法として作用することを意図しているにすぎず、各別個の値は、本明細書に個別に列挙されているかのように本明細書に組み込まれる。

【０２１４】

本明細書に記載の全ての方法は、本明細書に別段の指示がない限り、又は文脈と明らかに矛盾しない限り、任意の適切な順序で実行することができる。本明細書で提供されるありとあらゆる例又は例示的な言語（例えば、「等」）の使用は、単に本開示をよりよく明らかにすることを意図しており、別段の請求がない限り、本開示の範囲を限定するものではない。本明細書におけるいかなる言語も、特許請求されていない要素を本開示の実施に必須であると示すと解釈されるべきではない。

【０２１５】

開示された発明を実施するための本発明者らに知られている最良の形態を含む、本開示の実施形態を本明細書で説明する。これらの実施形態の変形は、前述の説明を読めば当業者には明らかとなり得る。

【０２１６】

本明細書に例示的に記載された開示は、本明細書に具体的に開示されていない任意の１つ又は複数の要素、１つ又は複数の制限がない状態で適切に実施することができる。したがって、例えば、本明細書の各例では、「を含む（ｃｏｍｐｒｉｓｉｎｇ）」、「から本質的になる（ｃｏｎｓｉｓｔｉｎｇ　ｅｓｓｅｎｔｉａｌｌｙ　ｏｆ）」、及び「からなる（ｃｏｎｓｉｓｔｉｎｇ　ｏｆ）」という用語のいずれかは、他の２つの用語のいずれかと置き換えることができる。使用された用語及び表現は、限定ではなく説明の用語として使用され、そのような用語及び表現の使用において、示され説明された特徴又はその一部の均等物を除外することは意図されていないが、特許請求される発明の範囲内で様々な修正が可能であることが認識される。したがって、本開示は好ましい実施形態を提供するが、本明細書に開示された概念の任意選択の特徴、修正及び変形は当業者によって使用されてもよく、そのような修正及び変形は、説明及び添付の特許請求の範囲によって定義される本開示の範囲内にあると見なされることを理解されたい。

【０２１７】

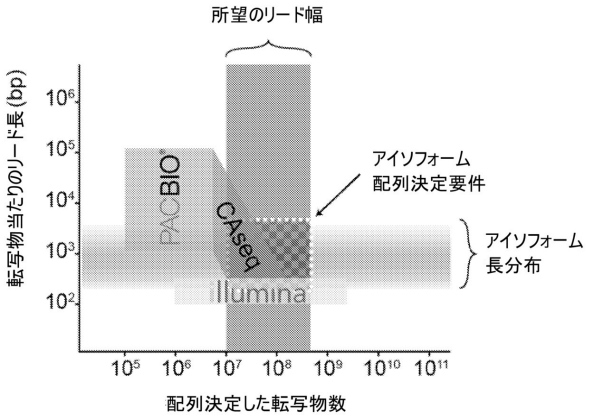
本発明の範囲及び趣旨から逸脱することなく、本明細書に開示された発明に対して様々な置換及び修正を行うことができることは、当業者には容易に明らかであろう。したがって、そのような追加の実施形態は、本開示及び以下の特許請求の範囲の範囲内である。本開示は、改善されたコントラスト、診断及び／又はイメージング活性を有するコンジュゲートを生成するために、本明細書に記載の化学修飾の様々な組合わせ及び／又は置換を試験することを当業者に教示する。したがって、本明細書に記載されるある実施形態は限定的ではなく、当業者は、本明細書に記載される修飾のある組合わせが、改善されたコントラスト、診断及び／又はイメージング活性を有するコンジュゲートを同定するための過度の実験なしに試験され得ることを容易に理解することができる。

【０２１８】

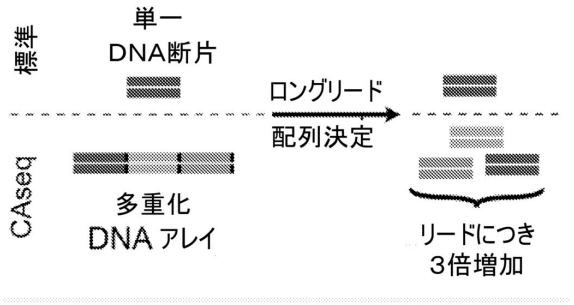
本発明者らは、当業者がそのような変形形態を適切に使用することを期待しており、本発明者らは、本開示が本明細書に具体的に記載されている以外の方法で実施されることを意図している。したがって、本開示は、適用法によって許容されるように、添付の特許請求の範囲に列挙された主題の全ての修正及び均等物を含む。更に、本明細書に別段の指示がない限り、又は文脈と明らかに矛盾しない限り、その全ての可能な変形における上述の要素の任意の組合わせが本開示に含まれる。当業者は、本明細書に記載の開示のある実施形態に対する多くの均等物を認識するか、又は日常的な実験のみを使用して確認することができるであろう。そのような均等物は、以下の特許請求の範囲に包含されることが意図されている。

【図面】

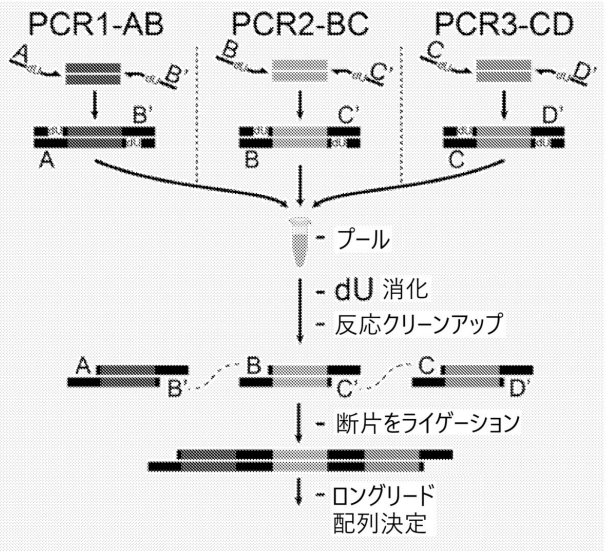
【図１Ａ】



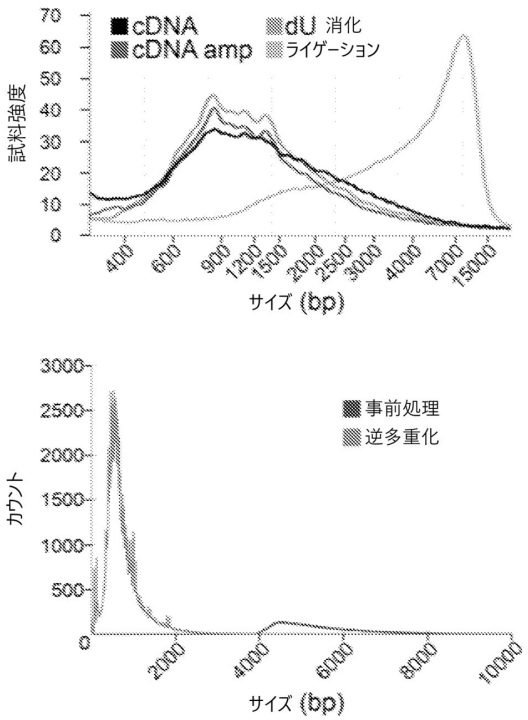
【図１Ｂ】



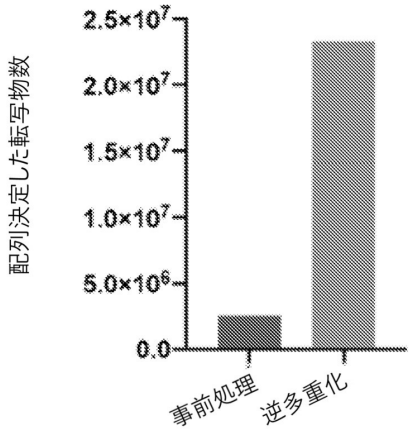
【図１Ｃ】



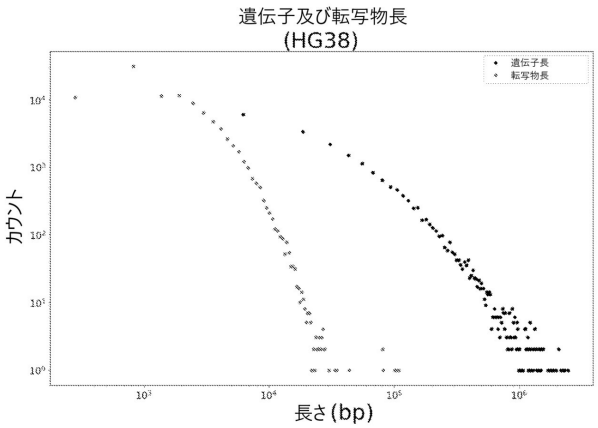
【図２Ａ】



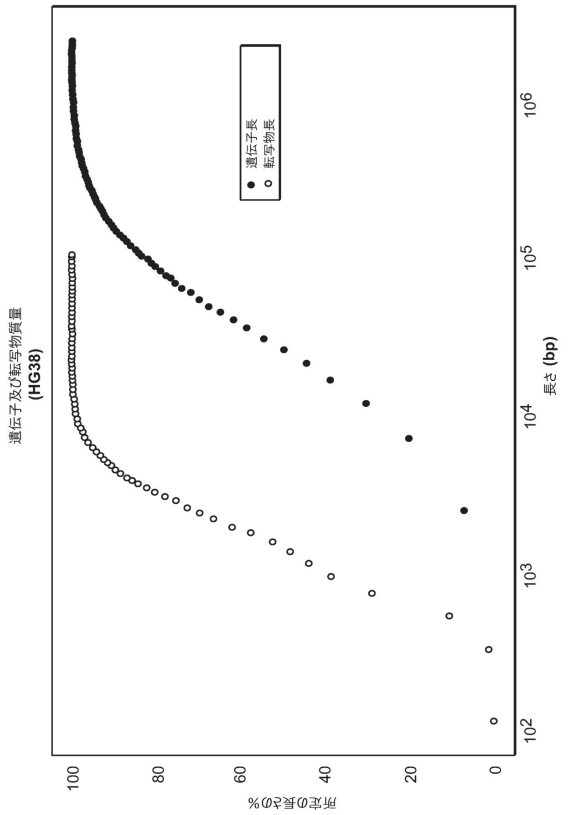
【図２Ｂ】



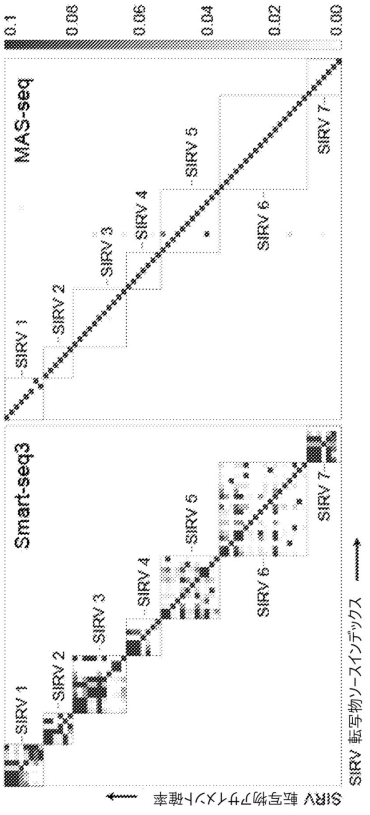
【図３Ａ】



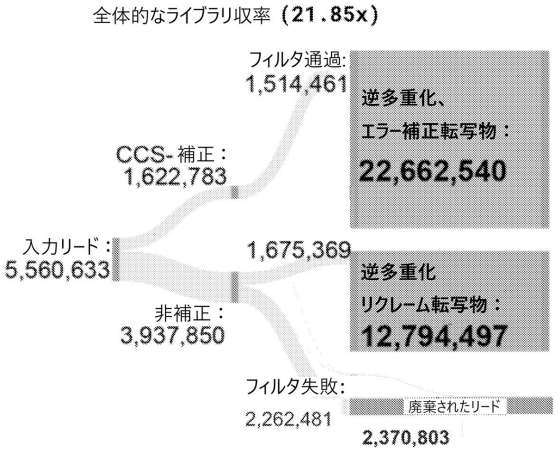
【図３Ｂ】


遺伝子及び転写物質量
(HG38)
遺伝子長
転写物長
長さ (bp)
所定の長さの%

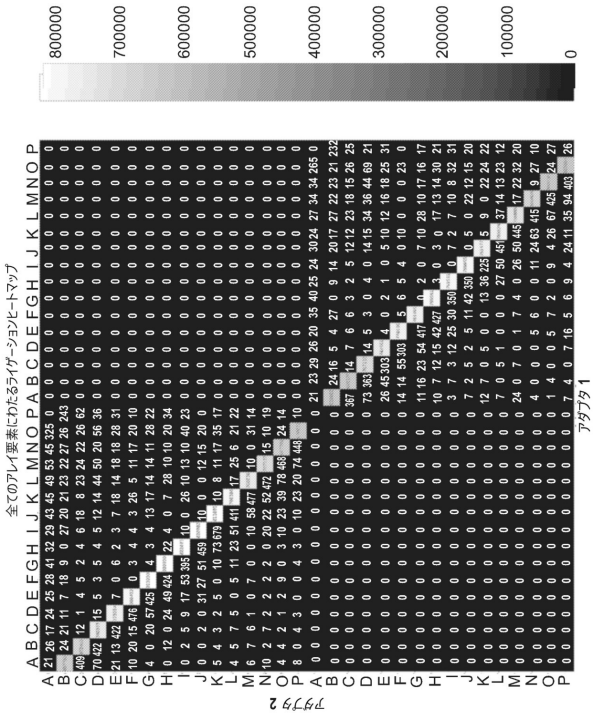
【図４】


MAS-seq
Smart-seq3
SIRV 1
SIRV 2
SIRV 3
SIRV 4
SIRV 5
SIRV 6
SIRV 7
SIRV 転写物ソースインデックス
SIRV 転写物アウトソース確率

【図５】


全体的なライブラリ収率 (21.85x)
フィルタ通過:
1,514,461
逆多重化、
エラー補正転写物：
22,662,540
CCS-補正：
1,622,783
入力リード：
5,560,633
1,675,369
逆多重化
リクレーム転写物：
12,794,497
非補正：
3,937,850
フィルタ失敗:
2,262,481
廃棄されたリード
2,370,803

【図６】


全てのアレイ要素にわたるライゲーションヒートマップ
アダプタ1
アダプタ2

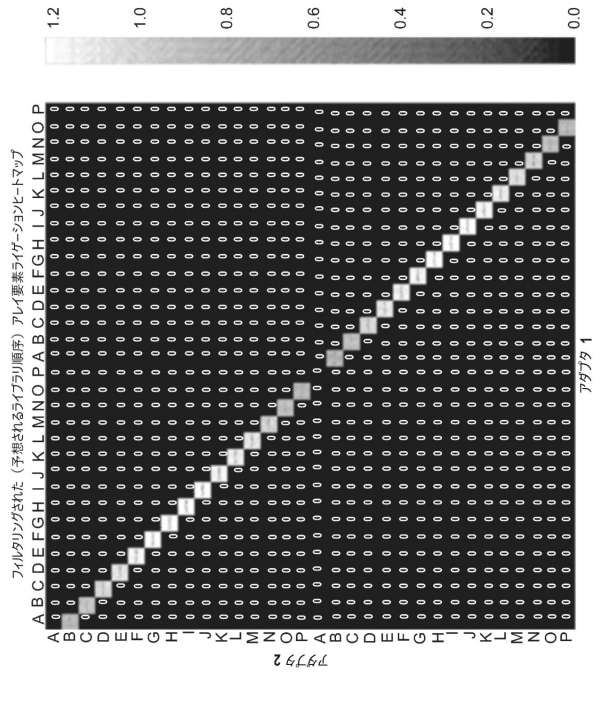
【図７】

	ライゲーションプロファイル	長さ	カウント	全カウントのうちの百分率
0	A B C D E F G H I J K L M N O P	16	484869	29.88%
1	A' B' C' D' E' F' G' H' I' J' K' L' M' N' O' P'	16	476205	29.34%
2	C' D' E' F' G' H' I' J' K' L' M' N' O' P'	14	54879	3.38%
3	C D E F G H I J K L M N O P	14	54471	3.36%
4	A B C D E F G H I J K L M N	14	49074	3.02%
5	A B C D E F G H I J K L M	13	46094	2.84%
6	A' B' C' D' E' F' G' H' I' J' K' L' M' N'	14	40890	2.52%
7	A' B' C' D' E' F' G' H' I' J' K' L' M'	13	39838	2.45%
8	A B C D E F G H I J K L	12	31918	1.97%
9	D' E' F' G' H' I' J' K' L' M' N' O' P'	13	30846	1.90%
10	D E F G H I J K L M N O P	13	29839	1.84%
11	A' B' C' D' E' F' G' H' I' J' K' L'	12	28700	1.77%
12	E' F' G' H' I' J' K' L' M' N' O' P'	12	28590	1.76%
13	E F G H I J K L M N O P	12	28128	1.73%
14	A B C D E F G H I J K	11	14138	0.87%
15	A B C D E F G H I J K L M N O	15	13311	0.82%
16	B' C' D' E' F' G' H' I' J' K' L' M' N' O' P'	15	12297	0.76%
17	A' B' C' D' E' F' G' H' I' J' K'	11	11336	0.70%
18	A' B' C' D' E' F' G' H' I' J' K' L' M' N' O'	15	10711	0.66%
19	B C D E F G H I J K L M N O P	15	10495	0.65%

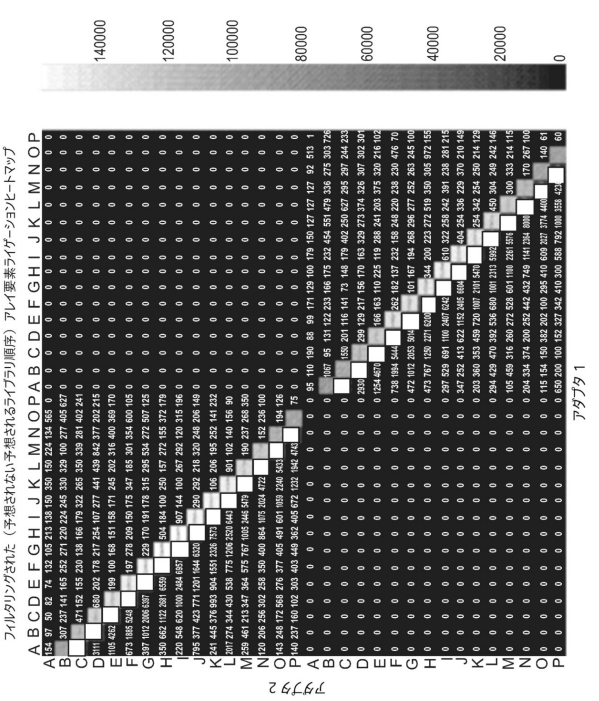
【図８】

HiFiリードの数	キメラアンプリコンアレイ配列決定リードの数	現行のリード分析法によってリクレームされたキメラアンプリコンアレイ配列決定リードの数	キメラアンプリコンアレイ配列決定リードの総数	総キメラアンプリコンアレイ配列決定収率増加
2477400	29789780	10194059	39983839	16.14x
1622783	22662840	12794497	35457337	21.84x

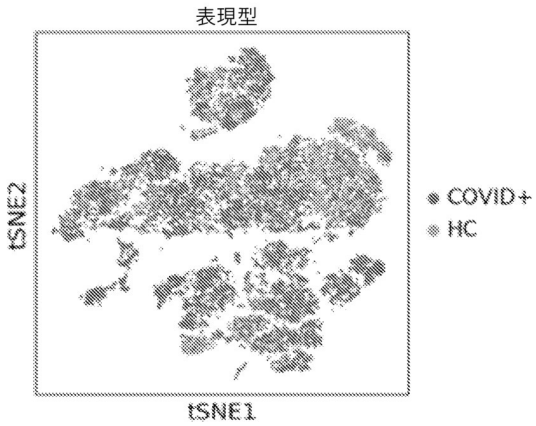
【図９Ａ】



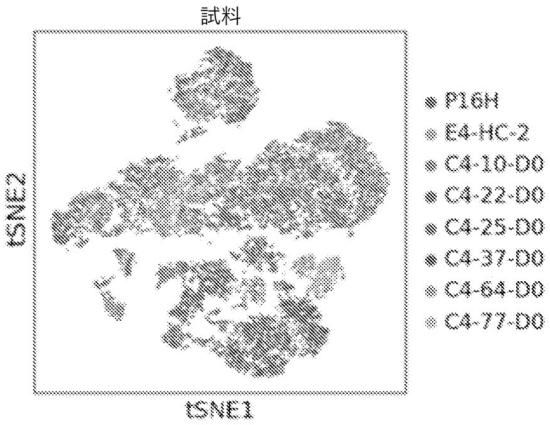
【図９Ｂ】



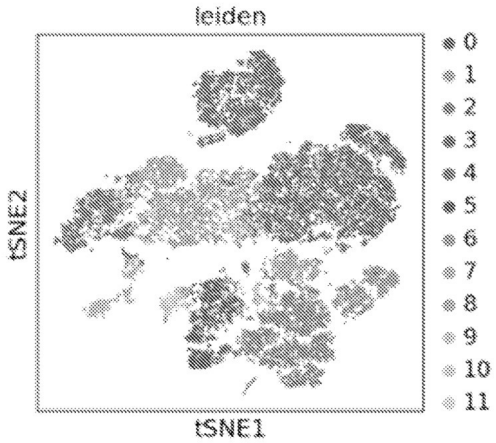
【図１０Ａ】



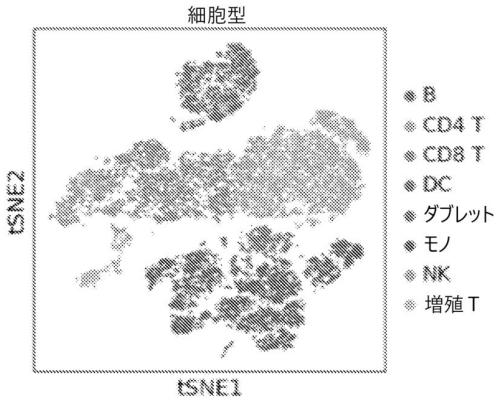
【図１０Ｂ】



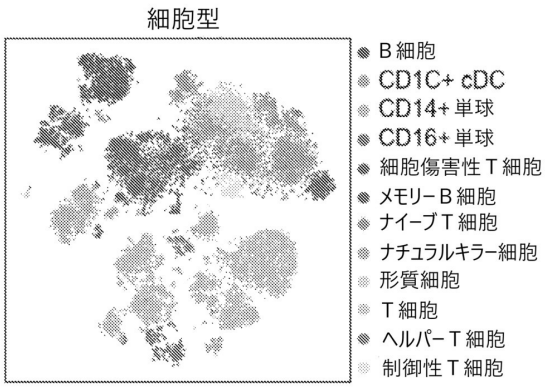
【図１０Ｃ】



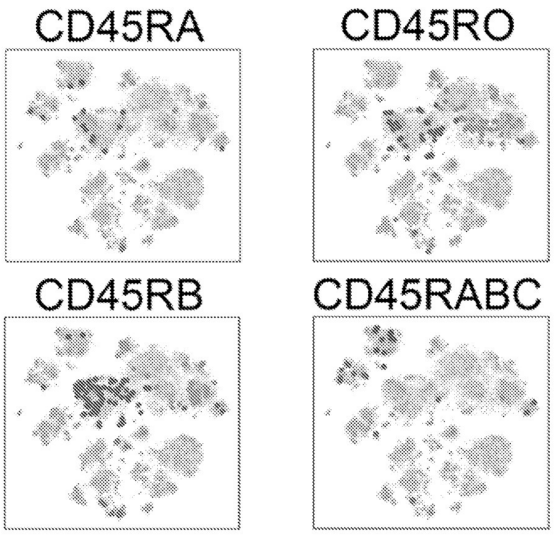
【図１０Ｄ】



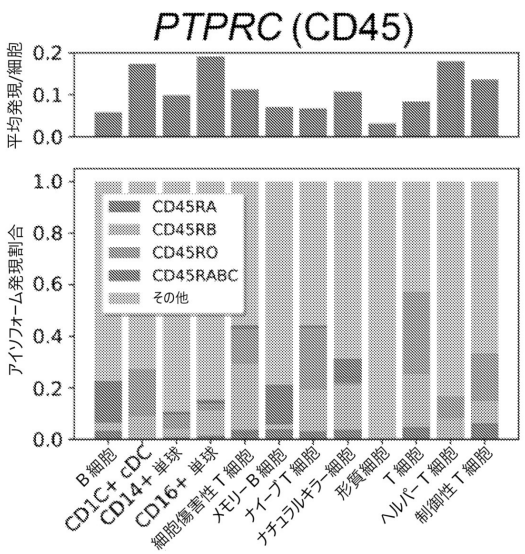
【図１１Ａ】



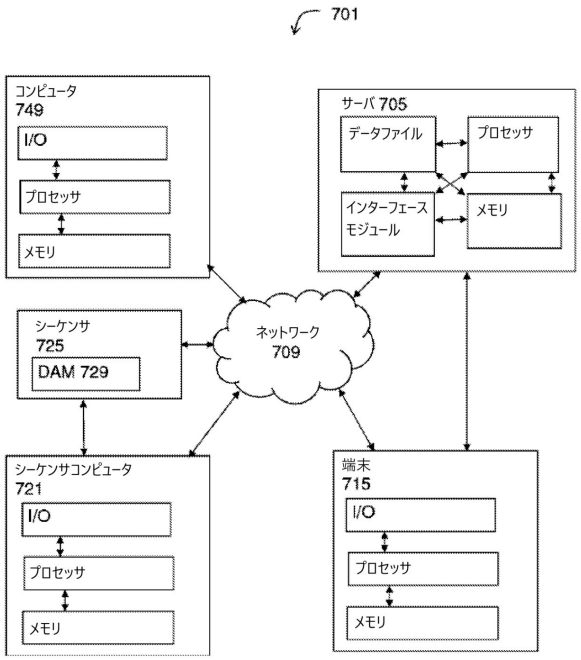
【図１１Ｂ】



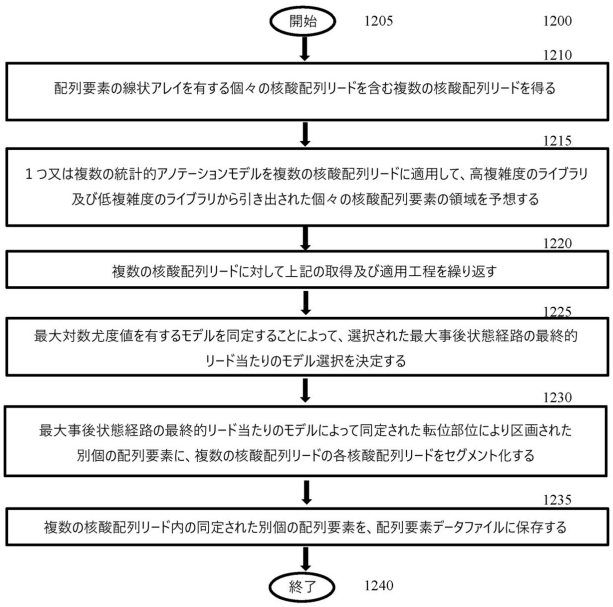
【図１１Ｃ】



【図１２】



【図１３】



【配列表】

2023534882000001.app

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 21/37226

A. CLASSIFICATION OF SUBJECT MATTER
IPC - C12Q 1/68, C12P 19/34 (2021.01)
CPC - C12Q 1/6837, C12Q 1/6844, C12Q 1/6858, C12Q 1/6874, C12Q 1/6806

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X – Y – A	US 2016/0264958 A1 (TWIST BIOSCIENCE CORPORATION) 15 September 2016 (15.09.2016); especially abstract; para [0004]-[0006], [0009]-[0010], [0030]-[0031], [0034]-[0035], [0044], [0072], [0075], [0084], [0127], [0162]	1-4, 6-9, 11-13, 15, 20-24, 27-29 – 5, 14, 16-19 – 10
Y	US 2010/0105052 A1 (DRMANAC et al.) 29 April 2010 (29.04.2010); especially para [0067], [0145], [0258], [0366], [0369], [0511]	5, 14, 16, 18
Y	WO 2018/227025 A1 (ARC BIO, LLC) 13 December 2018 (13.12.2018); especially para [00446]-[00447], [00456]-[00457]	17
Y	GIBSON et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods. May 2009, Vol. 6, No. 5, pg 343-345; especially abstract; Figure 1	19
A	US 2009/0099034 A1 (AHLQUIST et al.) 16 April 2009 (16.04.2009); especially para [0018]; SEQ ID NO: 40	10
A	WO 2008/133643 A2 (MONSANTO TECHNOLOGY, LLC) 6 November 2008 (06.11.2008); especially SEQ ID NO: 6165	10

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 November 2021	DEC 08 2021
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Kari Rodriquez Telephone No. PCT Helpdesk: 571-272-4300

Form PCT/ISA/210 (second sheet) (July 2019)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 21/37226

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

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

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

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

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

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

This International Searching Authority found multiple inventions in this international application, as follows:
This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Continued on Supplemental Page

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

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

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

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-24, 27-29 limited to SEQ ID NO: 1

Remark on Protest

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

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

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 21/37226

Continued from Box No. III Observations where unity of invention is lacking

Groups I+: Claims 1-24, 27-29, drawn to a method for preparing an array nucleic acid sequence, the method comprising: obtaining a plurality of input nucleic acid sequence within each input nucleic acid is of approximately 30 kilobases in length or shorter; attaching one or more adapter sequences to the plurality of input nucleic acid sequences. The method will be searched to the extent that the adapter sequence encompasses SEQ ID NO: 1. It is believed that claims 1-24, 27-29 encompass this first named invention, and thus these claims will be searched without fee to the extent that they encompass SEQ ID NO: 1. Additional adapter sequence(s) will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected adapter sequence(s). Applicants must further indicate, if applicable, the claims which encompass the first named invention, if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched. An exemplary election would be an adapter sequence comprising SEQ ID NO: 2 (Claims 1-24, 27-29).

Group II: Claim 25, drawn to a method for obtaining isoform sequencing information.

Group III: Claim 26, drawn to a method for performing mitochondrial lineage tracing.

Groups IV+: Claims 30-31, drawn to a composition comprising a plurality of adapter sequences. Group IV+ will be searched upon payment of additional fees. The composition may be searched, for example, to the extent that the adapter sequence encompasses SEQ ID NO: 1 for an additional fee and election as such. It is believed that claims 30-31 read on this exemplary invention. Additional adapter sequence(s) will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected adapter sequence(s). Failure to clearly identify how any paid additional invention fees are to be applied to the"+" group(s) will result in only the first claimed invention to be searched. Another exemplary election would be an adapter sequence comprising SEQ ID NO: 2 (Claims 30-31).

Group V: Claims 32-52, drawn to a method/system for identifying discrete sequence elements within individual nucleic acid sequence reads of a population of nucleic acid sequence reads, wherein each of the linear array of sequence elements comprises two or more nucleic acid sequence elements drawn from a library of high complexity flanked nucleic acid sequences drawn from a library of low complexity.

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

Special Technical Features

Groups I+, II and III include the special technical feature of a method which differs from the special technical feature of a composition, as disclosed by Groups IV+.

Groups I+ include the special technical feature of a method of preparing an array nucleic acid sequence wherein each input nucleic acid sequence within the plurality of input nucleic acid sequences is of approximately 30 kilobases in length or shorter, not required by Groups II, III and V.

Group II includes the special technical feature of a method of analyzing the sequence information obtained from the linear array nucleic acid sequence to obtain isoform sequencing information, not required by Groups I+, III and V.

Group III includes the special technical feature of a method of analyzing the sequence information obtained from the array nucleic acid sequence to trace mitochondrial lineage, not required by Groups I+, II and V.

Group V includes the special technical feature of a method for identifying discrete sequence comprising: applying one or more statistical annotation models to sequence data of the population of nucleic acid sequence reads, to predict within the population of nucleic acid sequence reads regions of individual nucleic acid sequence elements drawn from a library of high complexity and regions of nucleic acid sequences drawn from a library of low complexity, not required by Groups I+, II, III and IV+.

No technical features are shared between the polynucleotide sequences of Groups I+ and IV+, accordingly, these groups lack unity a priori.

Common Technical Features

The inventions of Groups I+, II, III, IV+ and V share the technical feature of a linear array of sequence elements.

The inventions of Groups I+, II, III and IV+ share the technical feature of a plurality of nucleic acid sequences, wherein at least two of the plurality of nucleic acid sequences comprise an adapter sequences.

The inventions of Groups I+, II and III share the technical feature of a method for preparing an array nucleic acid sequence.

Continued on Next Supplemental Page

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 21/37226

Continued from Previous Supplemental Page

However, these shared technical features do not represent a contribution over prior art in view of US 2016/0264958 A1 to Twist Bioscience Corporation (hereinafter 'Twist').

Twist discloses (instant claim 27) a method for preparing an array nucleic acid sequence (abstract - "Methods and compositions are provided for assembly of large nucleic acids..."), the method comprising:
i) obtaining a plurality of input nucleic acid sequences (para [0005] - "Methods are provided herein for nucleic acid assembly, comprising: providing a predetermined nucleic acid sequence..."; para [0072] - "In some cases, a precursor nucleic acid sequence is assembled with another precursor nucleic acid sequence via annealing and ligation of complementary sticky ends, followed by additional rounds of sticky end generation and assembly with other precursor fragment(s) to generate a long target nucleic acid sequence"), wherein each input nucleic acid sequence within the plurality of input sequences is of approximately 300 kilobases in length or shorter (para [0005] - "Methods are further provided wherein the predetermined nucleic acid sequence is 1 kb to 100 kb in length.");
ii) contacting the plurality of input nucleic acid sequences with paired amplification primers (para [0005] - "...each precursor double-stranded nucleic acid fragment having two strands, wherein each of the two strands comprises a sticky end sequence... providing primers comprising a nicking endonuclease recognition site and a sequence... corresponding to each of the different sticky end sequences..."),
wherein at least one primer within the paired amplification primers comprises an adapter sequence comprising an internal dU on one strand (para [0005] - "...providing primers comprising a nicking endonuclease recognition site and a sequence comprising (i) 5'-A (Nx) M-3' (SEQ ID NO.: 82) corresponding to each of the different sticky end sequences... wherein M is a non-canonical base... Methods are further provided wherein the uracil is in a deoxyuridine-deoxyadenosine base pair. ... Methods are further provided wherein the precursor double-stranded nucleic acid fragments comprise an adaptor sequence comprising the nicking endonuclease recognition site."), and
performing at least one round of amplification, thereby generating a population of adapted nucleic acid sequences (para [0005] - "...performing a polynucleotide extension reaction to form double-stranded nucleic acid fragments...");
iii) contacting the population of adapted nucleic acid sequences with Uracil DNA glycosylase and Endonuclease VIII (para [0006] - "Methods are further provided wherein the first nicking enzyme comprises uracil-DNA glycosylase (UDG). ... Methods are further provided wherein the first nicking enzyme comprises endonuclease VIII."; para [0084] - "For example a glycosylase, such as UDG, alone or in combination with an AP endonuclease, such as endonuclease VIII, is used to excise uracil and create a gap. "), thereby forming a population of adapted nucleic acid sequences having single-stranded ends (para [0006] - "Methods are further provided wherein the first nicking enzyme comprises endonuclease VIII."); and
iv) contacting the population of adapted nucleic acid sequences having single-stranded ends with a ligase, thereby forming an array nucleic acid sequence (para [0006] - "...and annealing the double-stranded nucleic acid fragments to form a nucleic acid encoding for the predetermined nucleic acid sequence that does not include the nicking endonuclease recognition site. ... Methods are further provided wherein the method further comprises ligating the annealed double-stranded nucleic acid fragments."; para [0072] - "In some cases, a plurality of precursor nucleic acid fragments are prepared with sticky ends, and the sticky ends are annealed and ligated to generate the predetermined target nucleic acid sequence"; para [0127] - "Nucleic Acid Assembly. Provided herein are methods where two or more of the cleavage, annealing and ligation reactions are performed concurrently within the same mixture and the mixture comprises a ligase.").

Twist discloses (instant claim 30) a composition comprising a plurality of nucleic acid sequences, wherein at least two of the plurality of nucleic acid sequences comprise an adapter sequence (para [0005] - "... synthesizing a plurality of precursor double-stranded nucleic acid fragments... Methods are further provided wherein the precursor double-stranded nucleic acid fragments comprise an adaptor sequence comprising the nicking endonuclease recognition site.").

As said technical features were known in the art at the time of the invention, these cannot be considered special technical features that would otherwise unify the groups.

Groups I+, II, III, IV+, V therefore lack unity under PCT Rule 13 because they do not share a same or corresponding special technical feature.

*Note: Claims 31 and 32 run together as there is no whitespace between the end of claim 31 and the claim numbered 32.

フロントページの続き

(51)国際特許分類	FI	テーマコード（参考）
Ｃ１２Ｎ　15/09　(2006.01)	Ｃ１２Ｎ　15/09　Ｚ	

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

（特許庁注：以下のものは登録商標）

１．ＴＲＩＴＯＮ

２．ＪＡＶＡ

アメリカ合衆国 マサチューセッツ州 ボストン フルーツ ストリート ５５

(74)代理人　100189131

弁理士　佐伯 拓郎

(74)代理人　100182486

弁理士　中村 正展

(74)代理人　100147289

弁理士　佐伯 裕子

(72)発明者　ハコーヘン， ニア

アメリカ合衆国 マサチューセッツ州 ０２１１４ ボストン， フルーツ ストリート ５５， ザ ジェネラル ホスピタル コーポレイション内

(72)発明者　アルカファジ， アジズ

アメリカ合衆国 マサチューセッツ州 ０２１４２ ケンブリッジ， メイン ストリート ４１５， ザ ブロード インスティテュート インコーポレイテッド内

(72)発明者　ブレイニー， ポール

アメリカ合衆国 マサチューセッツ州 ０２１４２ ケンブリッジ， メイン ストリート ４１５， ザ ブロード インスティテュート インコーポレイテッド内

(72)発明者　ババディ， メルタシュ

アメリカ合衆国 マサチューセッツ州 ０２１４２ ケンブリッジ， メイン ストリート ４１５， ザ ブロード インスティテュート インコーポレイテッド内

(72)発明者　ガリメラ， キラン ヴイ

アメリカ合衆国 マサチューセッツ州 ０２１４２ ケンブリッジ， メイン ストリート ４１５， ザ ブロード インスティテュート インコーポレイテッド内

(72)発明者　スミス， ジョナサン セオドール

アメリカ合衆国 マサチューセッツ州 ０２１４２ ケンブリッジ， メイン ストリート ４１５， ザ ブロード インスティテュート インコーポレイテッド内

Ｆターム（参考）　4B029　AA07 AA27 BB20 FA15

4B063　QA13 QQ02 QQ04 QQ42 QR10 QR20 QR32 QR55 QS34 QS36