(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6777966号
(P6777966)
(45)発行日　令和2年10月28日(2020.10.28)　(24)登録日　令和2年10月13日(2020.10.13)

(51)Int.Cl.　F I
G16B 20/30 (2019.01)　G16B 20/30

請求項の数 38 (全 85 頁)

(21)出願番号　特願2017-561612(P2017-561612)
(86)(22)出願日　平成28年2月17日(2016.2.17)
(65)公表番号　特表2018-512092(P2018-512092A)
(43)公表日　平成30年5月10日(2018.5.10)
(86)国際出願番号　PCT/US2016/018295
(87)国際公開番号　WO2016/134034
(87)国際公開日　平成28年8月25日(2016.8.25)
審査請求日　平成31年2月14日(2019.2.14)
(31)優先権主張番号　62/117,256
(32)優先日　平成27年2月17日(2015.2.17)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　62/294,208
(32)優先日　平成28年2月11日(2016.2.11)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　517286593
ダブテイル　ゲノミクス　エルエルシー
アメリカ合衆国　９５０６６　カリフォルニア州　スコッツバレー　エンタープライズ・ウェイ　１００　スイート　エー１０１
(74)代理人　100082072
弁理士　清原　義博
(72)発明者　パトナム，ニコラス　エイチ.
アメリカ合衆国　０２４５３　マサチューセッツ州　ウォルサム　リバー・ストリート　３６　アパートメント　３６２

最終頁に続く

(54)【発明の名称】核酸配列アセンブリ

(57)【特許請求の範囲】
【請求項１】
プロセッサを含む、核酸配列情報のコンティグにスキャフォールドするためのコンピュータにより実施されるシステムであって、該プロセッサは、
初期の構成を有するコンティグ配列のセットを受け取り、
ペアエンドリードのセットを受け取り、
標準的なペアエンドリード距離度数データを受け取り、
少なくとも１つのペアエンドリード中に共存する配列を共有するコンティグペアがグループ分けされるように、コンティグペアを処理し、
コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データが、初期の構成のコンティグ配列のリードペア度数データに対して標準的なペアエンドリード距離度数データにより近く近似するように、グループ化されたコンティグ配列にスキャフォールドし、ならびに、
ネットワーク、スクリーン、またはサーバーへ処理されたコンティグスキャフォールドを出力するように構成される、コンピュータにより実施されるシステム。
【請求項２】
ペアエンドリードのセットは、核酸内で内部二本鎖切断を生じさせるためにサンプルＤＮＡを消化し、複数の再連結接合部を形成するために二本鎖切断を無作為に再連結させ、および複数の再連結接合部にわたって配列決定することにより得られる、請求項１に記載のコンピュータにより実施されるシステム。

【請求項３】
標準的なペアエンドリード度数は、両方のリードが共通のコンティグにマッピングされるペアエンドリードから得られる、請求項１に記載のコンピュータにより実施されるシステム。
【請求項４】
標準的なペアエンドリード度数は、事前に生成された曲線から得られる、請求項１に記載のコンピュータにより実施されるシステム。
【請求項５】
初期の構成は無作為の構成である、請求項１に記載のコンピュータにより実施されるシステム。
【請求項６】
コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データは、リードペア距離尤度が増大するときにペアエンドリード距離度数データにより近く近似する、請求項１に記載のコンピュータにより実施されるシステム。
【請求項７】
リードペア距離尤度は最大化される、請求項６に記載のコンピュータにより実施されるシステム。
【請求項８】
コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データは、リードペアの距離度数データと標準的なペアエンドリード距離度数データとの間の差の統計的尺度が減少する際に、ペアエンドリード距離度数データにより近く近似する、請求項１に記載のコンピュータにより実施されるシステム。
【請求項９】
リードペア距離度数データと標準的なペアエンドリード距離度数データとの間の距離の統計的尺度は、ＡＮＯＶＡ、ｔ検定、およびＸ２乗検定の少なくとも１つを含む、請求項８に記載のコンピュータにより実施されるシステム。
【請求項１０】
コンティグを分離するためにマッピングされるリードペアのリードペア距離は、標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差が減少するときに、ペアエンドリード距離度数データとより近く一致する、請求項１に記載のコンピュータにより実施されるシステム。
【請求項１１】
標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差は最小限に抑えられる、請求項１０に記載のコンピュータにより実施されるシステム。
【請求項１２】
ペアエンドリードのセットは、核酸内で内部二本鎖切断を生じさせるためにサンプルＤＮＡを消化し、少なくとも１つの再連結接合部を形成するために二本鎖切断を再連結し、および少なくとも１つの再連結接合部にわたって配列決定することにより得られる、請求項１に記載のコンピュータにより実施されるシステム。
【請求項１３】
ＤＮＡは少なくとも１つのＤＮＡ結合剤に架橋される、請求項１２に記載のコンピュータにより実施されるシステム。
【請求項１４】
ＤＮＡは、再構成された染色質へ再度アセンブルされる分離された裸のＤＮＡである、請求項１２に記載のコンピュータにより実施されるシステム。
【請求項１５】
生体サンプルから得られる核酸配列情報のコンティグを、プロセッサを用いることによってスキャフォールドする方法であって、該方法は：
（ａ）初期の構成を持つ一連のコンティグ配列を得る工程であって、コンティグ配列は生

体物質からＤＮＡを抽出することおよび生体物質を配列決定することにより得られる、工程；
（ｂ）一連のペアエンドリードを得る工程であって、一連のペアエンドリードは、核酸内に内部の二本鎖切断を生成するためにサンプルＤＮＡを消化すること、複数の再連結結合部を形成するために二本鎖切断が無作為に再連結するのを可能にすること、及び、複数の再連結結合部にわたり配列決定を行うことにより得られる、工程；
（ｃ）標準のペアエンドリードの距離度数データを得る工程；
（ｄ）少なくとも１つのペアエンドリードに共存する配列を共有するコンティグのペアをグループ分けする工程であって、それによってグループ分けされたコンティグを生成する工程；及び、
（ｅ）コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データが、初期の構成におけるグループ分けされたコンティグのリードペア度数データに比べて少なくとも５％、標準のペアエンドリードの距離度数データとより近くに近似するように、グループ分けされたコンティグをスキャフォールドする工程、
を含む、方法。
【請求項１６】
サンプルＤＮＡが、少なくとも１つのＤＮＡ結合剤に架橋される、請求項１５に記載の方法。
【請求項１７】
サンプルＤＮＡは、再構成された染色質へ再度アセンブルされる分離された裸のＤＮＡを含む、請求項１５に記載の方法。
【請求項１８】
再構成された染色質が架橋される、請求項１７に記載の方法。
【請求項１９】
再構成された染色質がＤＮＡ結合タンパク質を含む、請求項１７に記載の方法。
【請求項２０】
再構成された染色質がナノ粒子を含む、請求項１７に記載の方法。
【請求項２１】
一連のペアエンドリードは、核酸内に内部の二本鎖切断を生成するためにサンプルＤＮＡを消化すること、複数の再連結結合部を形成するために二本鎖切断が無作為に再連結するのを可能にすること、及び、複数の再連結結合部の各側での配列決定を行うことにより得られる、請求項１５に記載の方法。
【請求項２２】
標準的なペアエンドリード距離度数データが、両方のリードが共通のコンティグにマッピングされるペアエンドリードから得られる、請求項１５に記載の方法。
【請求項２３】
標準的なペアエンドリード距離度数データが、事前に生成された曲線から得られる、請求項１５に記載の方法。
【請求項２４】
前記スキャフォールドする工程が、前記グループ分けされたコンティグの推定上の隣接するコンティグの第１のセットを選択する工程、前記リードペアに関してリードペアの距離の統計的尺度を減少させる推定上の隣接するコンティグの前記第１のセットの最小限の距離順序を判定する工程、および、リードペア距離の前記統計的尺度を減少させるために、推定上の隣接するコンティグの前記第１のセットをスキャフォールドする工程を含む、請求項１５に記載の方法。
【請求項２５】
前記最小限の距離順序を判定する工程が、推定上の隣接するコンティグの第１のセットの２つのコンティグにマッピングされているリードを含む少なくとも１つのリードペアのためのすべての可能なコンティグ構成についての予想されるリードペア距離を判定する工程を含む、請求項２４に記載の方法。

【請求項２６】
全ての可能なコンティグ構成からの他の可能な構成と比較してリードペア距離尤度を最も改善する、全ての可能なコンティグ構成からのコンティグ配向を選択する工程をさらに含む、請求項２５に記載の方法。
【請求項２７】
生体サンプルはゲノムを含む、請求項１５に記載の方法。
【請求項２８】
生体サンプルは、複数のゲノムを含むヘテロ接合サンプルである、請求項１５に記載の方法。
【請求項２９】
Ｎ５０値は、初期の構成におけるグループ分けされたコンティグに比べて少なくとも１％増加する、請求項１５に記載の方法。
【請求項３０】
核酸配列情報のコンティグを、プロセッサを用いることによってスキャフォールドする方法であって、該方法は：
（ａ）初期の構成を持つ一連のコンティグ配列を得る工程；
（ｂ）一連のペアエンドリードを得る工程；
（ｃ）標準のペアエンドリードの距離度数データを得る工程；
（ｄ）少なくとも１つのペアエンドリードに共存する配列を共有するコンティグのペアをグループ分けする工程であって、それによってグループ分けされたコンティグを生成する工程；及び、
（ｅ）コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データが、初期の構成におけるグループ分けされたコンティグのリードペア度数データに比べて少なくとも５％、標準のペアエンドリードの距離度数データとより近くに近似するように、グループ分けされたコンティグをスキャフォールドする工程、
を含む、方法。
【請求項３１】
スキャフォールドする工程が、グループ分けされたコンティグを順序付けする工程、グループ分けされたコンティグを配向する工程、少なくとも２つのコンティグの端と端を融合する工程、第２のコンティグに１つのコンティグを挿入する工程、および、少なくとも２つの成分のコンティグへ、コンティグを切断する工程、を含む、請求項３０に記載の方法。
【請求項３２】
リードペア距離尤度が、初期の構成におけるグループ分けされたコンティグのリードペア度数データに比べて少なくとも５％増加する時、コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データが、ペアエンドリードの距離度数データとより近くに近似する、請求項３０に記載の方法。
【請求項３３】
リードペア距離尤度は最大化される、請求項３２に記載の方法。
【請求項３４】
コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データと標準的なペアエンドリード距離度数データとの間の差の統計的尺度が、初期の構成におけるグループ分けされたコンティグのリードペア度数データに比べて少なくとも５％減少する時、コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データが、標準的なペアエンドリードの距離度数データとより近くに近似する、請求項３０に記載の方法。
【請求項３５】
コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データと標準的なペアエンドリード距離度数データとの間の距離の統計的尺度は、ＡＮＯＶＡ、ｔ－検定、及びＸ二乗検定の少なくとも１つを含む、請求項３４に記載の方法。

【請求項３６】
コンティグを分離するためにマッピングされるリードペアのリードペア距離が、標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差が減少する時、ペアエンドリード距離度数データとより近く一致する、請求項３５に記載の方法。
【請求項３７】
標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差が最小化される、請求項３６に記載の方法。
【請求項３８】
Ｎ５０値は、初期の構成におけるグループ分けされたコンティグに比べて少なくとも１％増加する、請求項３０に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本出願は、全体において参照によって本明細書に明確に組み込まれる、２０１５年２月１７日出願の米国仮特許出願第６２／１１７，２５６号の利益を主張し、且つ、全体において参照によって本明細書に明確に組み込まれる、２０１６年２月１１日出願の米国仮特許出願第６２／２９４，２０８号の利益も同様に主張する。
【背景技術】
【０００２】
現時点でアクセス可能且つ手頃な高スループット配列決定方法は、短距離の配列の近接性及びゲノム変異の特性付けに最も良く適している。長距離の結合及びハプロタイプの位相合せの達成は、長い（例えば、数十キロベース）配列を直接且つ正確に読み取る能力、又は、ペアとなった或いはグループ分けされた配列のリードを通じた結合及び位相の関係の捕捉の何れかを要する。しかし、配列決定情報のグループ分け、及び、長距離の結合とハプロタイプ位相合せの達成に必要な配列決定情報のアセンブリの生成は、計算的に集約的なものであり、且つ時間を消費してしまう。本明細書には、ペアになった又はグループ分けされた配列のリードによって通知された配列情報から、染色体規模の近接性を持つアセンブリを得るための、計算上効果的な方法及びシステムが開示される。
【発明の概要】
【０００３】
本明細書には、核酸データのスキャフォールディングに関連する、方法、組成物、アルゴリズム、及びシステムが開示される。本明細書における手法は、リードペア（ｒｅａｄ　ｐａｉｒ）のリードがデータセットにおいてマッピングされるコンティグの位相又は物理的結合の情報に関する情報を推測するために、リードペアを利用する。核酸データセットにおけるコンティグは、それらの相対的な位置決めに関連するスコア又はパラメータに対するそのような活性の影響を考慮すると、端から端まで順序づけられ、配向され、又は融合され、或いは、場合によっては一方が他方に挿入される（総体的に、「スキャフォールドされる」）。
【０００４】
場合によっては、スコア又はパラメータは、１つ又は別のコンティグ構成のリードペアのデータセットに関する集合的なリードペアの分離に対するコンティグの再配置の結果として生じるインパクトの尺度である。それを生成するのに使用される手法に依存して、リードペアのデータセットは、特定のリードペア分離の分布曲線を持つこともある。度数に応じてリードペア分離としてマッピングされると、与えられたリードペアのデータセットに関して予期されたリードペアの距離分布を判定することができる。その後、核酸サンプル及びリードペアの生成方法を考慮して予期されるリードペアの距離分布に、データセットに関するリードペアの距離分布を一致、接近、又はより近くに接近させるように、一連のコンティグに対してリードペアをマッピングして、コンティグを位置決めしてもよい（順に、配向、又はその他）。

【０００５】
コンティグの位置決めは、コンティグ又はスキャフォールドを互いに対して順序付けること、コンティグ又はスキャフォールドを互いに対して配向すること、コンティグ又はスキャフォールドを端から端まで結合すること、コンティグ又はコンティグのスキャフォールドにおける間隙に１以上のコンティグを挿入すること、又はデータセットにおいて誤ってアセンブルされるコンティグ又はスキャフォールドを分割することを多様に含む。場合によっては、このプロセスは、最適な又は最適化された構成を得るまで継続される一方、代替的な場合においては、このプロセスは、初期のコンティグ又はスキャフォールドの構成にわたる改善を達成するためにのみ実行される。代替的に、前記プロセスは、サンプルのコンティグのセットの一部の画分が、例えば７０％、７５％、８０％、８５％、９０％、９５％、９９％、９９．９％、又はそれ以上、正確にスキャフォールドされるまで継続される。多くの場合、配列データセットは、ヒト又は倍数体植物サンプルのゲノム或いはトランスポゾンが豊富なゲノムサンプルなどの、複雑なゲノムサンプルさえも表示し、コンティグの順序づけ、配向、端から端までの組み合わせ、別のスキャフォールド内の１つのスキャフォールドの組み合わせ、或いは、スキャフォールド又はコンティグを壊す（総体的に「スキャフォールドする」）ことによる、データセットの構成及びデータセットの改善の計算上の評価は、わずか８時間、７時間、６時間、５時間、４時間、又は４時間未満で完了する。

【０００６】
スコア評価は、隣接したコンティグ又はスキャフォールドのサブセットを一度に調べることにより、全体的又は局所的に、或いは全体的且つ局所的に行われる。局所的に行なわれると、例えば、２、３、４、５、６、又は６より多くのコンティグのサブセットが、最適化されたスコアを判定するために調べられ、その後、以前のウィンドウについて判定される最適化された構成を頻繁に考慮して、「ウィンドウ」が１つのコンティグに移り、プロセスが繰り返される。代替的に、サブセットは、０．０１％、０．１％、１％、又は５％などの、合計の核酸配列セット（例えば、１つ又は複数のゲノム）の画分として一度に定義される。場合によっては、「ウィンドウ」のサイズは変動し、それにより、容易にアセンブルされた領域は、より大きなウィンドウを割り当てられ、一方で、より多くの負荷のある領域、又は、より高密度のリード、或いは矛盾且つ複雑な分析であるより高密度のリードを持つ領域は、より小さなウィンドウのサイズを割り当てられる。

【０００７】
本明細書には、核酸配列情報のコンティグをスキャフォールドする方法が提供され、該方法は、初期の構成を持つ一連のコンティグ配列を得る工程；一連のペアエンドリードを得る工程；標準のペアエンドリードの距離度数データを得る工程；少なくとも１つのペアエンドリードに共存する配列を共有するコンティグのペアをグループ分けする工程；及び、コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データが、初期の構成におけるコンティグ配列のリードペア度数データに比べて、標準のペアエンドリードの距離度数データとより近くに近似するように、グループ分けされたコンティグ配列をスキャフォールドする工程を含む。スキャフォールドする工程は、一連のコンティグのセットを順序付けること、一連のコンティグを配向すること、少なくとも２つのコンティグを端から端まで融合すること、１つのコンティグを別のコンティグに挿入すること、及び１つのコンティグを少なくとも２つの構成要素であるコンティグへと切断することの、少なくとも１つを含む。幾つかの方法において、標準のペアエンドリード度数は、両方のリードが共通のコンティグに対してマッピングされる、ペアエンドリードから得られる。代替的に又は組み合わせにおいて、標準のペアエンドリード度数は、以前に生成された曲線から得られる。初期の構成は、ランダムな構成であるか、又は予め構成されている。好ましい実施形態において、コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データは、リードペア距離の尤度が増大する場合に、ペアエンドリードの距離度数データとより近くに近似する。多くの場合、リードペア距離の尤度は最大限にされる。コンティグを分離するためにマッピングされるリードペ

アに関するリードペアの距離度数データは、リードペアの距離度数データと標準のペアエンドリードの距離度数データとの間の差の統計的な尺度が減少する場合に、ペアエンドリードの距離度数データとより近くに近似する。多くの統計的な尺度が利用可能である。例えば、リードペアの距離度数データと標準のペアエンドリードの距離度数データとの間の距離の統計的な尺度は、様々な場合において、ＡＮＯＶＡ、ｔ－検定、及びＸ二乗検定（Ｘ－ｓｑｕａｒｅｄ　ｔｅｓｔ）の少なくとも１つを含む。コンティグを分離するためにマッピングされるリードペアに関するリードペアの距離度数データは、得られる順序づけたコンティグの中のリードペアの距離分布の偏差が、標準のペアエンドリードの距離度数と比べて減少する場合に、ペアエンドリードの距離度数データとより近くに一致する。代替的に又は組み合わせにおいて、得られる順序付けたコンティグ中のリードペアの距離分布の偏差は、標準のペアエンドリードの距離度数と比べて最小化される。幾つかのスキャフォールドの評価において、第１のクラスター及び第２のクラスターに関連した、ペアエンドリードにおける配列を共有するコンティグは、より多くの共有されたエンドのリードを持つクラスターに割り当てられる。クラスター化は頻繁に、生体における染色体の数以上である多くのグループに、コンティグを配することを含む。頻繁に、ほんの１つのペアエンドリードを、クラスターの１つのコンティグと共有するコンティグは、当該クラスターに含まれていない。反復配列を含む、少なくとも１つのペアエンドリードのみをクラスターと共有するコンティグは頻繁に、前記クラスターに含まれていない。同様に、低品質の配列を含む、少なくとも１つのペアエンドリードのみをクラスターと共有するコンティグは頻繁に、前記クラスターに含まれていない。幾つかの方法において、一連のペアエンドリードは、核酸内に内部の二本鎖切断を生成するためにサンプルＤＮＡを消化すること、少なくとも１つの再連結結合部を形成するために二本鎖切断が再連結するのを可能にすること、及び、少なくとも１つの再連結結合部にわたり配列決定を行うことにより、得られる。ＤＮＡは、核タンパク質又はナノ粒子などの少なくとも１つのＤＮＡ結合剤に、幾つかの手法においてはペアリード生成に架橋結合される。ＤＮＡは、再構成された染色質へと再びアセンブルされる、分離された裸のＤＮＡであるが、特にＤＮＡ分子が互いに結合しない場合、結合タンパク質を持つＤＮＡは一部の状況下で適切となる。頻繁に、再構成された染色質は架橋結合される。再構成された染色質はＤＮＡ結合タンパク質を含む。代替的に又は組み合わせにおいて、再構成された染色質はナノ粒子を含む。好ましくは、場合によっては、コンティグのクラスター化は、生体に関する染色体の数とは無関係である。第１のクラスター及び第２のクラスターに関連した、ペアエンドリードにおける配列を共有するコンティグは、多くの場合に、より多くの共有されたエンドのリードを持つクラスターに割り当てられる。代替的に又は組み合わせにおいて、第１のクラスター及び第２のクラスターに関連した、ペアエンドリードにおける配列を共有するコンティグは、より大きなリードペアの距離の尤度値を有しているクラスターに割り当てられ、又は、第１のクラスター及び第２のクラスターに関連した、ペアエンドリードにおける配列を共有するコンティグは、標準のリードペアの距離分布に比べて、そのリードペアの分布における偏差が低いクラスターに割り当てられる。代替的に、第１のクラスター及び第２のクラスターに関連した、ペアエンドリードにおける配列を共有するコンティグは、各クラスターから除外される。頻繁に、クラスター化は、生体における染色体の数以上である多くのグループに、コンティグを配することを含む。幾つかのスキャフォールディングは、前記クラスター化されたコンティグの推定上の隣接するコンティグの第１のセットを選択すること、前記リードペアに関するリードペアの距離の総計的な尺度を減らす、推定状の隣接するコンティグの第１のセットの最小の距離順序を判定すること、及び、リードペアの前記総計的な尺度を減らすように推定状の隣接するコンティグの第１のセットをスキャフォールドすることを含む。推定上の隣接したコンティグの第１のセットは２つのコンティグから成る。代替的に、推定上の隣接したコンティグの第１のセットは３つのコンティグから成る。代替的に、推定上の隣接したコンティグの第１のセットは４つのコンティグから成る。代替的に、推定上の隣接したコンティグの第１のセットは４つのコンティグを含む。幾つかのスキャフォールディングは、推定上の隣接したコンティグの第１のセットにおけ

る各コンティグの順序と配向を判定することを含む。最小の距離順序の判定は、場合によっては全ての起こり得るコンティグ構成に関して前記セットの２つのコンティグに対しマッピングされるリードを含む、少なくとも１つのリードペアに関する予期されるリードペアの距離を比較することを含む。スキャフォールディングは頻繁に、前記リードペアに関して最小のリードペア距離に対応するコンティグの配向を選択することを含む。幾つかの方法は、最大尤度のリードペア距離分布に対応するコンティグの配向を選択することを更に含む。幾つかの方法は、前記コンティグクラスターのリードペアの総計的な尺度に関して最小のリードペア距離に対応するコンティグ配向を選択することを更に含む。幾つかの方法において、予期されたリードペアの距離は、ペアエンドリードの距離度数データと比較される。幾つかの方法において、前記ペアエンドリードの距離度数データとの比較は、式１を使用することを含む。幾つかの方法は、前記クラスター化されたコンティグの推定上の隣接したコンティグの第２のセットを選択することを含み、当該第２のセットは、前記第１のセットの終端のコンティグの１つを除いて全て、及び前記クラスター化されたコンティグの１つの追加のコンティグを含み、及び、前記方法は、リードペアの距離の前記総計的な尺度を減らすように推定状の隣接したコンティグの第２のセットをスキャフォールドすることを含む。幾つかの方法は、前記クラスター化されたコンティグの推定上の隣接したコンティグの第３のセットを選択することを含み、当該第３のセットは、前記第２のセットの終端のコンティグの１つを除いて全て、及び前記第１のセット及び第２のセットに含まれていない、前記クラスター化されたコンティグの１つの追加のコンティグを含み、及び、前記方法は、リードペアの距離の前記総計的な尺度を減らすように推定状の隣接したコンティグの第３のセットをスキャフォールドすることを含む。この後に、多くの場合において、前記クラスター化されたコンティグの大多数が順序付けられるまで、少なくとも１つの追加のセットが反復して選択される。選択は頻繁に、前記クラスター化されたコンティグがそれぞれ順序付けられるまで少なくとも１つの追加のセットを反復して選択することを必要とする。核酸配列は、１つのゲノム、又は場合によっては複数のゲノムなどのサンプルに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【０００８】

本明細書には、クラスターにおけるコンティグをスキャフォールドする方法が提供され、該方法は：コンティグの各ペアに関して対数尤度比スコアを割り当てる工程；対数尤度比スコアにより接続を選別する工程；及びアセンブリの合計スコアを増やすように、対数尤度比スコアの減少する順序でコンティグの接続を許容又は拒絶する工程を含む。幾つかの方法において、スキャフォールディングは、一連のコンティグのセットを順序付けること、及び／又は一連のコンティグを配向すること、及び／又は２つのコンティグを端から端まで融合すること、及び／又は１つのコンティグを別のコンティグに挿入すること、及び／又は１つのコンティグを少なくとも２つの構成要素であるコンティグへと切断することを含む。多くの場合、コンティグは１つ又は複数のゲノムを含む。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【０００９】

本明細書には、クラスター内の複数のコンティグの局所的に最適なコンティグ構成を判定する方法が提供される。幾つかのそのような方法は、ａ）コンティグのクラスターに沿って位置ｉにて始まるサイズｗのコンティグの配列ウィンドウを識別する工程；ｂ）ウィンドウにおける各位置ｉの互換的な順序及び配向のスコアを調べることにより、ウィンドウのコンティグについてのｗ！２^{w}の順序付け及び配向の選択肢を考慮する工程；ｃ）最

適なスコアを得るために前記ウィンドウの前記ｗのコンティグを配向且つ順序付ける工程；ｄ）ウィンドウを位置ｉ＋１へと移す工程；及びｅ）最適なスコアを判定するために、前記ｗのコンティグの順序付け及び配向を用いて、位置ｉ＋１にて前記ウィンドウに関して工程（ａ）、（ｂ）、及び（ｃ）を繰り返す工程を含み；それにより、スコアに対して、局所的に最適な構成における前記複数のコンティグを配向且つ順序付ける。幾つかの方法において、クラスターにおける複数のコンティグに対してマッピングされるリードペアのデータが得られ、標準のペアエンドリード度数データが得られ、及び、前記ｗのコンティグの配向及び順序付けに関するスコアは、クラスターにおける複数のコンティグに対してマッピングされるリードペアのデータに関するリードペアの距離データのセットが、どのくらい近くに標準のペアエンドリード度数データのセットに一致するのかについての尺度である。幾つかの方法において、クラスターにおける複数のコンティグに対してマッピングされるリードペアのデータが得られ、スコアは合計のリードペアの距離であり、スコアは合計のリードペアの距離が最小化される場合に最適化される。ウィンドウのサイズｗは３であり、又は代替的にｗは４であり、又は代替的にｗは５であり、又は代替的にｗは６である。場合によっては、ｗは第１のクラスターに関する第１の値を有し、ｗは第２のクラスターにて第２の値を有している。ｗは、幾つかの方法において、セットのコンティグの１％、又は代替的にセットのコンティグの５％、或いは代替的にセットのコンティグの１０％を含むように選択される。多くの方法において、スコアはリードペアの距離尤度スコアであり、スコアは、与えられたウィンドウのサイズについて最大限にされる場合に最適となる。スコアは、幾つかの例示的実施形態において式１を使用して計算される。スコアは、予期されたリードペアの分布からの偏差であり、場合によってスコアが与えられたウィンドウのサイズについて最小化される場合に最適となる。複数のコンティグは、１つ又は複数のゲノム、或いは非ゲノム核酸ソースを含む。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１０】

本明細書には、核酸配列アセンブリのための方法が提供され、該方法は：精製されたＤＮＡを得る工程；ＤＮＡ／染色質複合体を形成するために精製されたＤＮＡをＤＮＡ結合剤と結合させる工程；粘着末端を残すために制限酵素によりＤＮＡ染色質複合体をインキュベートする工程；ＤＮＡの端部を結合するために連結を行なう工程；ペアエンドリードを生成するために連結されたＤＮＡ結合にわたり配列決定を行う工程；及び精製されたＤＮＡの配列を表すコンティグを含む核酸のデータセットをスキャフォールドするためにペアエンドリードを使用する工程を含む。幾つかの方法において、精製されたＤＮＡは、１つ又は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１１】

本明細書には、繰返しのコンティグ領域に対しマッピングされるリードペア配列のリードを識別する方法が提供され、該方法は：核酸サンプルのためにコンティグのデータセットを得る工程；隣接していない物理的に結合した配列情報に対応する少なくとも１つのリードペアの配列のリードを得る工程；及び、リードペアの配列のリードの少なくとも１つのリードがコンティグのデータセットの２つの別個の座位に対してマッピングされる場合にリードペアの配列のリードを除外する工程を含む。幾つかの方法において、繰返しの領域は、第１の閾値を越える、ショットガンリード深さを持つ配列を含む。幾つかの方法において、繰返しの領域は、第２の閾値を越えるリードの深さを持つ塩基の位置を含む。頻繁に、第１の閾値及び第２の閾値は、リード深さの全体的な分布に対して固定される。第

１の閾値は、多くの場合にリードの深さの全体的な分布の３倍である。代替的に、第１の閾値は、このセット内で、又は該セットに隣接して、２、２．５、３．５、４、４．５、５、５．５、６、又は非整数の値である。第２の閾値は頻繁に、リードの深さの全体的な分布の３．５倍である。代替的に、第２の閾値は、このセット内で、又は該セットに隣接して、２、２．５、３、４、４．５、５、５．５、６、又は非整数の値である。幾つかの方法において、精製されたＤＮＡは、１つ又は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１２】

本明細書には、コンティグのアセンブリの決定を誘導する方法が提供され、該方法は、第１のコンティグと第２のコンティグとの間の全域（ｓｐａｎｎｉｎｇ）リードペアの配列の数及び示唆された分離を観察する可能性を判定する工程を含み、ここで、コンティグは、セット［＋＋、＋－、－＋、－－］内にｏの相対的な配向を有ししており、間隙の長さにより分離される。幾つかの方法は更に、分離距離にわたるリードペアの配列の分布の可能性を標準化する工程を含み、該工程は、リードペアの配列を、核酸サンプルを独立してサンプリングするノイズペアと比較することを含む。場合によっては、核酸サンプルはゲノムを含む。代替的に、核酸サンプルは複数のゲノム、又は非ゲノムソースを含む。頻繁に、ノイズペアの総数は、コンティグペアのサンプルについての結合の密度を表にすることにより判定される。更に本明細書には、密度の最高及び最低１％が除外される方法が提供される。該方法に対する代案において、最高．５％、．６％、．７％、．８％、．９％、１．１％、１．２％、１．３％、１．４％、１．５％、１．６％、１．７％、１．８％、１．９％、２％、３％、４％、５％、又は５％より多くが除外され、同様に最低．５％、．６％、．７％、．８％、．９％、１．１％、１．２％、１．３％、１．４％、１．５％、１．６％、１．７％、１．８％、１．９％、２％、３％、４％、５％、又は５％より多くが除外される。幾つかの方法はコンティグの順序を判定する工程を含む。幾つかの方法はコンティグの配向を判定する工程を含む。幾つかの方法は、コンティグの順序と配向の両方を判定する工程を含む。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１３】

本明細書には、コンティグの誤った結合の矯正のための方法が提供され、該方法は、初期の構成を持つ一連のコンティグ配列を得る工程；一連のペアエンドリードを得る工程；標準のペアエンドリードの距離度数データを得る工程；少なくとも１つのペアエンドリードに共存する配列を共有するコンティグのペアをグループ分けする工程；コンティグのグループ分けのためのリードペアの度数データを、標準のペアエンドリード距離度数データと比較する工程；グループ分けしたコンティグに分裂を導入することにより、コンティグのグループ分けのためのリードペアの度数データが、標準のペアエンドリードの距離度数データと近くに近似するかどうかを判定する工程；及び、コンティグのグループ分けのためのリードペアの度数データが標準のペアエンドリードの距離度数データとより近くに近似する場合に、コンティグに分裂を導入する工程を含む。幾つかの方法において、第１の位置は、分裂を導入する前に前記閾値より下の前記対数尤度を持つ少なくとも１つの隣接した第２の位置と融合される。第２の隣接した位置は、第１の位置から３００以下の塩基対である。代替的に、第２の位置は、第１の位置から１０００以下の塩基対の位置を含んでいない。代替的に、第２の隣接した位置は、５０、１００、１５０、２００、２５０、３５０、４００、４５０、５００、５５０、６００、６５０、７００、７５０、８００、

８５０、９００、９５０、１１００、１２００、１３００、１４００、１５００、１６００、１７００、１８００、１９００、又は２０００以下、或いは、列挙した値に及ぶ範囲内の整数の値である。更に本明細書には、対数尤度の変化の判定が、コンティグの密度をマッピングする平均のペアエンドの識別、密度をマッピングする平均のペアエンドの、少なくとも３ｘの密度をマッピングするペアエンドを持つコンティグのセグメントの識別、及び、密度をマッピングする平均のペアエンドの、少なくとも３ｘの密度をマッピングするペアエンドを持つコンティグのセグメントの除外を含む、方法が提供される。代替的に、１．５ｘ、１．６ｘ、１．７ｘ、１．８ｘ、１．９ｘ、２．０ｘ、２．１ｘ、２．２ｘ、２．３ｘ、２．４ｘ、２．５ｘ、２．６ｘ、２．７ｘ、２．８ｘ、２．９ｘ、３．１ｘ、３．２ｘ、３．３ｘ、３．４ｘ、３．５ｘ、３．６ｘ、３．７ｘ、３．８ｘ、３．９ｘ、４ｘ、４．１ｘ、４．２ｘ、４．３ｘ、４．４ｘ、４．５ｘ、４．６ｘ、４．７、４．８ｘ、４．９ｘ、５ｘ、又は５ｘより多くの閾値が使用される。更に本明細書には、コンティグ配列のセットが１つのゲノムに由来する方法が提供される。更に本明細書には、コンティグ配列のセットが複数のゲノムに由来する方法が提供される。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１４】

本明細書には、コンティグアセンブリのための方法が提供され、該方法は：出発アセンブリの分裂したコンティグを示す工程であって、分裂したコンティグはノードであり、分裂したコンティグのエッジは整数の順序付けられたペアのリストにより標識され、分裂したコンティグのエッジはマッピングされたリードペアの配列に対応する、工程；及びマッピングされた接続の閾値の数よりも少ないエッジを排除する工程を含む。幾つかの方法において、閾値の数は５％未満である。代替的に、閾値の数は、２０％、１５％、１４％、１３％、１２％、１１％、１０％、９％、８％、７％、６％、４％、３％、２％、１％、又はそれ以下である。場合によっては、閾値の数はt_Lリンクよりも少ない。幾つかの方法において、コンティグは、コンティグの長さに対する、対応するノードのグラフにおける程度の比率が、全ての値の分布のハイエンド（ｈｉｇｈ　ｅｎｄ）の約５％を超える塩基対であるエッジを含む。幾つかの方法において、コンティグは１つのゲノムに由来する。幾つかの方法において、コンティグは複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１５】

本明細書には、少なくとも１つのスキャフォールドへとコンティグ配列の情報をアセンブルする方法が提供され、該方法は、複数のコンティグに対応する配列情報を得る工程、複数のコンティグにより表される核酸サンプルからペアエンドリードを得る工程、及び、予測されたリードペアの距離データのセットからのリードペアの距離パラメータの偏差が最小化されるように複数のコンティグを構成する工程を含み、ここで、構成する工程は８時間未満で生じる。予測されたリードペアの距離データのセットは、多くの好ましい実施形態においてリードペアの距離尤度曲線を含む。場合によっては、リードペアの距離パラメータは、リードペアの距離尤度曲線に対して最大の距離尤度である。代替的に、リードペアの距離パラメータは、リードペアの距離尤度曲線に対して最小の変異である。コンティグの局所的に隣接したセットは２つのコンティグを含む。代替的に、コンティグの局所的に隣接したセットは３つのコンティグを含む。代替的に、コンティグの局所的に隣接したセットは４つのコンティグを含む。代替的に、コンティグの局所的に隣接したセットは５つのコンティグを含む。代替的に、コンティグの局所的に隣接したセットは６つのコンティグを含む。好ましくは、構成する工程は７時間で生じる。代替的に、構成する工程は

、６時間未満、５時間、４時間、３時間、２時間、１時間、又は１時間未満で生じる。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１６】

本明細書には、一連のコンティグ配列をスキャフォールドする方法が提供され、該方法は、核酸サンプルを表す一連のコンティグ配列を得る工程、核酸サンプルに関するリードペアのデータを得る工程、及び、核酸サンプルに関するリードペアのデータが予期されたリードペアの分布とより近くに近似するように、一連のコンティグ配列を順序付け且つ配向する工程を含み、ここで、一連のコンティグ配列の７０％が、８時間以下で核酸サンプルにおける配列の相対的な順序及び配向と一致するように順序付けられ且つ配向される。スキャフォールディングは、一連のコンティグのセットを順序付けること、一連のコンティグを配向すること、少なくとも２つのコンティグを端から端まで融合すること、１つのコンティグを別のコンティグに挿入すること、及び／又は１つのコンティグを少なくとも２つの構成要素であるコンティグへと切断することの、少なくとも１つを含む。幾つかの方法において、一連のコンティグ配列の８０％が、８時間以下で核酸サンプルにおける配列の相対的な順序及び配向と一致するように順序付けられ且つ配向される。代替的に、一連のコンティグ配列の９０％が、８時間以下で核酸サンプルにおける配列の相対的な順序及び配向と一致するように順序付けられ且つ配向される。代替的に、一連のコンティグ配列の９５％が、８時間以下で核酸サンプルにおける配列の相対的な順序及び配向と一致するように順序付けられ且つ配向される。場合によっては、一連のコンティグ配列の７０％が、４時間以下、又は代替的に２時間以下、或いは代替的に１時間以下で核酸サンプルにおける配列の相対的な順序及び配向と一致するように順序付けられ且つ配向される。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。

【００１７】

本明細書には、一連の核酸配列データを構成する方法が提供され、該方法は：スキャフォールドを含む複数のコンティグに対応する配列情報を得る工程、ペアエンドリードの情報を得る工程、及び、ペアエンドリードの情報に関するペアエンドリードの距離分布が基準のペアエンドリードの距離分布に近似するよう全体的に最適化されるように複数のコンティグを構成する工程を含み、ここで、構成する工程は８時間未満で生じる。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１８】

本明細書には、スキャフォールドアセンブリを改善する方法が提供され、該方法は、各ノードが少なくとも１つのコンティグ配列を含む複数の結合されたノードペアを含んでいるスキャフォールドのセットを得る工程、複数の結合されたノードペアに対してマッピングされたペアエンドリードの情報を得る工程、結合されたノードペアにより共有されるリードペアの数を計数する工程、前記数を閾値と比較する工程、及び、前記数が閾値より下に落ち込む場合にノードペアを結合されていないノードへと切断する工程を含む。場合によっては、固有のコンティグ配列に対してマッピングされるリードペアのみが計数される。更に本明細書には、別個のリードペアエンドの閾値の数がマッピングされるコンティグ配列セグメントに対してマッピングされるリードペアが破棄される方法が提供される。閾

値の数は、多くの場合、繰返しでない配列に関して３ｘの平均数である。代替的に、１．５ｘ、１．６ｘ、１．７ｘ、１．８ｘ、１．９ｘ、２ｘ、２．１ｘ、２．２ｘ、２．３ｘ、２．４ｘ、２．５ｘ、２．６ｘ、２．７ｘ、２．７ｘ、２．８ｘ、２．９ｘ、３．１ｘ、３．２ｘ、３．３ｘ、３．４ｘ、３．５ｘ、３．６ｘ、３．７ｘ、３．８ｘ、３．９ｘ、４ｘ、４．５ｘ、５ｘ、又は５ｘより多くの閾値が利用される。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００１９】

本明細書には、スキャフォールドアセンブリを改善する方法が提供され、該方法は、各ノードが少なくとも１つのコンティグ配列を含む複数の結合されたノードペアを含んでいるスキャフォールドのセットを得る工程、複数の結合されたノードに対してマッピングされたペアエンドリードの情報を得る工程、標準のペアエンドリードの距離度数データを得る工程；複数の結合されたノードに対してマッピングされたペアエンドリードの情報に関するペアエンドリードの度数データを、標準のペアエンドリードの距離度数データと比較する工程；及び、結合されたノードの切断が、結果として、標準のペアエンドリードの距離度数データとより近くに近似するように複数の結合されたノードに対してマッピングされたペアエンドリードの情報に関する、ペアエンドリードの度数データをもたらす場合に、少なくとも１つの結合されたノードを切断する工程を含む。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２０】

本明細書には、スキャフォールドアセンブリの方法が提供され、該方法は、一連のコンティグ配列を得る工程、及び一連のペアエンドリードを含む入力データを得る工程を含み、ここで、ペアエンドリードの少なくとも１％が少なくとも１ｋｂのリードペアの距離を含み、一連のペアエンドリードは自然配向でペアエンドリードを含み、リードペアの配列決定のエラー率は０．１％以下であり、入力データのＲＮ５０はアセンブルされたスキャフォールドの２０％未満であり、前記方法は更に、スキャフォールドを出力する工程を含み、スキャフォールドのＲＮ５０は入力のＲＮ５０の少なくとも２ｘである。随意に、エラー率は、１２％。１１％、１０％、９％、８％、７％、６％、５％、４％、３％、２％、１％、０．９％、０．８％、０．７％、０．６％、０．５％、０．４％、０．３％、０．２％、０．１％、０．０９％、０．０８％、０．０７％、０．０６％、０．０５％、０．０４％、０．０３％、０．０２％、０．０１％、０．００１％、０．０００１％未満、又は０．００００１％以下である。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２１】

本明細書には、スキャフォールドアセンブリのための方法が提供され、該方法は：T_0コンティグ配列を含む一連のコンティグ配列を得る工程、一連のペアエンドリードを得る工程を含み、ここで、ペアエンドリードの少なくとも１％は少なくとも１ｋｂのリードペアの距離を含み、一連のペアエンドリードは自然な飛行でペアエンドリードを含み、リー

ドペアの配列決定のエラー率は０．１％未満であり、及び前記方法は更に、T_1を含むスキャフォールドを出力する工程を含み、T_1は＜T_0である。場合によっては、T_1は３未満である。随意に、エラー率は、１２％。１１％、１０％、９％、８％、７％、６％、５％、４％、３％、２％、１％、０．９％、０．８％、０．７％、０．６％、０．５％、０．４％、０．３％、０．２％、０．１％、０．０９％、０．０８％、０．０７％、０．０６％、０．０５％、０．０４％、０．０３％、０．０２％、０．０１％、０．００１％、０．０００１％未満、又は０．０００００１％以下である。代替的に、T_1は、１０、９、８、７、６、５、又は４未満となるように選択される。場合によっては、T_1が２であり、場合によっては、T_1は１つのコンティグである。T_1は、多くの場合においてT_1の５０％、４０％、３０％、２０％、１０％、５％、３％、２％、又は１％未満である。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２２】

本明細書には、核酸配列データ処理の方法が提供され、該方法は：リードペアを含む入力データを受け取る工程を含み、該リードペアの少なくとも１％は、少なくとも１ｋｂまで分離され且つ自然な配向にある２つの核酸セグメントからの配列データを含み、入力データのＲＮ５０はアセンブルされたスキャフォールドの２０％未満であり、前記入力データのエラー率は０．１％未満であり；及び前記方法は更にスキャフォールドを含む出力データを出力する工程を含み、出力データのＲＮ５０は入力データのＲＮ５０の少なくとも２ｘである。幾つかの方法において、出力データのＲＮ５０は、入力データのＲＮ５０の少なくとも１０ｘ、又は代替的に３ｘ、４ｘ、５ｘ、６ｘ、７ｘ、８ｘ、９ｘ、１１ｘ、１２ｘ、１３ｘ、１４ｘ、１５ｘ、１６ｘ、１７ｘ、１８ｘ、１９ｘ、２０ｘ、３０ｘ、４０ｘ、５０ｘ、６０ｘ、７０ｘ、８０ｘ、９０ｘ、１００ｘ、５００ｘ、１０００ｘ、又は１０００ｘより上である。更に本明細書には、スキャフォールドが正確な順序及び配向で標的ゲノムサンプル配列の少なくとも９０％を含む方法が、提供される。更に本明細書には、スキャフォールドが正確な順序及び配向で標的ゲノムサンプル配列の少なくとも９９％を含む方法が、提供される。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２３】

本明細書には、核酸配列データ処理の方法が提供され、該方法は：リードペアを含むデータセットを出力する工程を含み、該リードペアの少なくとも１％は、少なくとも１ｋｂまで分離され且つ自然配向にある２つの核酸セグメントからの配列データを含み、出力データのＲＮ５０はアセンブルされたスキャフォールドの２０％未満であり、前記出力データのエラー率は０．１％未満であり；及び前記方法は更にスキャフォールドを含むデータセットを受け取る工程を含み、出力データのＲＮ５０は入力データのＲＮ５０の少なくとも２ｘである。幾つかの方法において、出力データのＲＮ５０は、入力データのＲＮ５０の少なくとも１０ｘ、又は代替的に３ｘ、４ｘ、５ｘ、６ｘ、７ｘ、８ｘ、９ｘ、１１ｘ、１２ｘ、１３ｘ、１４ｘ、１５ｘ、１６ｘ、１７ｘ、１８ｘ、１９ｘ、２０ｘ、３０ｘ、４０ｘ、５０ｘ、６０ｘ、７０ｘ、８０ｘ、９０ｘ、１００ｘ、５００ｘ、１０００ｘ、又は１０００ｘより上である。更に本明細書には、スキャフォールドが正確な順序及び配向で標的ゲノムサンプル配列の少なくとも９０％を含む方法が、提供される。更に本明細書には、スキャフォールドが正確な順序及び配向で標的ゲノムサンプル配列の少なくとも９９％を含む方法が、提供される。幾つかの実施形態において、前記方法は、本明細書

において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２４】

本明細書には、核酸配列データ処理の方法が提供され、該方法は：リードペアを含む入力データを受け取る工程を含み、該リードペアの少なくとも１％は、少なくとも１ｋｂまで分離され且つ自然な配向にある２つの核酸セグメントからの配列データを含み、入力データのＮ５０はアセンブルされたスキャフォールドの２０％未満であり、前記出力データのエラー率は０．１％未満であり；及び前記方法は更にスキャフォールドを含む出力データを出力する工程を含み、出力データのＮ５０は入力データのＲＮ５０の少なくとも２ｘである。随意に、エラー率は、１２％。１１％、１０％、９％、８％、７％、６％、５％、４％、３％、２％、１％、０．９％、０．８％、０．７％、０．６％、０．５％、０．４％、０．３％、０．２％、０．１％、０．０９％、０．０８％、０．０７％、０．０６％、０．０５％、０．０４％、０．０３％、０．０２％、０．０１％、０．００１％、０．０００１％未満、又は０．００００１％以下である。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２５】

本明細書には、核酸配列データ処理の方法が提供され、該方法は：リードペアを含む出力データを出力する工程を含み、該リードペアの少なくとも１％は、少なくとも１ｋｂまで分離され且つ自然配向にある２つの核酸セグメントからの配列データを含み、出力データのＮ５０はアセンブルされたスキャフォールドの２０％未満であり、前記出力データのエラー率は０．１％未満であり；及び前記方法は更にスキャフォールドを含む入力データを受け取る工程を含み、出力データのＮ５０はアセンブルされたスキャフォールドの２０％未満である。コンティグの情報は多くの場合において１つのゲノムに由来する。随意に、エラー率は、１２％。１１％、１０％、９％、８％、７％、６％、５％、４％、３％、２％、１％、０．９％、０．８％、０．７％、０．６％、０．５％、０．４％、０．３％、０．２％、０．１％、０．０９％、０．０８％、０．０７％、０．０６％、０．０５％、０．０４％、０．０３％、０．０２％、０．０１％、０．００１％、０．０００１％未満、又は０．００００１％以下である。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２６】

本明細書には、少なくとも１つのペアエンドリードを共有する２つの拡散コンティグを結合する尤度を評価する方法が提供され、該方法は：第１のコンティグに対しマッピングされたショットガンリードの密度を判定する工程、第２のコンティグに対しマッピングされたショットガンリードの密度を判定する工程、第１のコンティグと第２のコンティグを結合するための尤度スコアを判定する工程、及び、第１のコンティグに対しマッピングされたショットガンリードの密度が、第２のコンティグに対しマッピングされたショットガンリードの密度と著しく異なる場合に、尤度スコアを減らす工程を含む。幾つかの方法において、尤度スコアは対数尤度スコアである。頻繁に、前記スコアは本明細書に示されるように減らされる。頻繁に、前記スコアは、第１のコンティグに対しマッピングされたシ

ョットガンリードの密度と、第２のコンティグに対しマッピングされたショットガンリードの密度の、より小さな密度からより大きな密度の比率として減らされる。コンティグの情報は多くの場合において１つのゲノムに由来する。代替的に、コンティグの配列情報は複数のゲノムに由来する。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００２７】

＜参照による組み込み＞

本明細書で挙げられる全ての刊行物、特許、及び特許出願は、あたかも個々の刊行物、特許、又は特許出願がそれぞれ参照により本明細書に具体的且つ個別に組み込まれるのと同じ程度にまで、参照により本明細書に組み込まれている。

【図面の簡単な説明】

【００２８】

本開示の新規な特徴は、特に添付の請求項、本明細書における概要、及び詳細な説明にて、説明されている。本開示の特徴及び利点のより良い理解は、本開示の原理が用いられる実施形態を説明する以下の詳細な説明と、以下の添付図面とを参照することによって得られる。

【図１】新規の配列決定のための材料を生成するためのフロープロセスを表す。

【図２】基準のヒトゲノムアセンブリｈｇ１９に対しマッピングされた、様々な配列決定ライブラリのためのリードペア分離を示す。

【図３】様々なリードペア分離ビンにおける被覆率（推定されたゲノムサイズに分割されたリードペア分離の合計）のチャートである。

【図４Ａ】リードペアのＧＲＣｈ３８基準配列上のマッピングされた場所が、ＧＭ１２８７８と基準との間の構造差の付近でプロットされることを示す。図４Ａは、隣接している２０ｋｂの繰返し領域による８０Ｋｂの逆位に関するデータを表す。

【図４Ｂ】リードペアのＧＲＣｈ３８基準配列上のマッピングされた場所が、ＧＭ１２８７８と基準との間の構造差の付近でプロットされることを示す。図４Ｂは、段階的なヘテロ接合の欠失に関するデータを表す。

【図５】新たな配列決定のための材料を生成するためのフロープロセスを表す。

【図６】本発明の幾つかの実施形態と一致するコンピュータシステムの例を示す。

【図７】本発明の例となる実施形態と一致するコンピュータシステム（２０００）の第１の実例アーキテクチャを示すブロック図である。

【図８】本発明の例となる実施形態と一致する、複数のコンピュータシステム、複数の携帯電話、及び個人用携帯情報端末、及びネットワーク接続ストレージ（ＮＡＳ）を組み込むように構成されたネットワーク（８００）を実証する図である。

【図９】本発明の例となる実施形態と一致する、共有仮想アドレスメモリ空間を用いるマルチプロセッサーコンピュータシステム（９００）のブロック図である。

【図１０】初期のコンティグアセンブリ（図１０のＡ）から、「ウィンドウ」分析（図１０のＤ）のための直線状に順序付けられたコンティグの生成までの典型的なワークフローを表す。

【図１１】最小の全域木を表す。ノードラベルはコンティグサイズをｋｂで示し、エッジのラベルは、示されたコンティグのペアエッジそれぞれにおけるリンクの数を示す。

【発明を実施するための形態】

【００２９】

短いリードデータからの、長距離且つ非常に正確な新たなアセンブリは、ゲノミクスにおいて最も緊急な難問の１つである。我々は本明細書において、数百キロベースまでのＤＮＡ結合が、近接連結ライブラリの産生のための基質として、生きた染色体よりも再構成された染色質を用いてインビトロで生成されることを実証する。結果として生じるライブ

ラリは、リードペア内の距離とリード数との規則的な関係を含む、長距離のゲノムアセンブリ及び位相合せに有用なＨｉ－Ｃデータの特徴の多くを共有する。このインビトロの長距離のメイトペアライブラリを、標準の全体のゲノムショットガン及びジャンピングライブラリと組み合わせて、我々は、わずかな費用及び労力で、より高価な方法に匹敵する長期にわたる精度及び近接性を持つ長距離の正確且つ新規のヒトゲノムアセンブリを生成した。この方法は、高分子量ＤＮＡの適度な量しか使用せず、通常は任意の種に適用可能である。本明細書において我々は、新たな核酸配列アセンブリ（例えば、ゲノム又は一連の染色体を表すスキャフォールドへの）又はヒト及びワニを用いるスキャフォールドアセンブリ用だけでなく、構造の変異の識別及びヘテロ接合の変形の位相合せのための効果的なツールとして、この配列データの値を実証する。

【００３０】

本明細書には、例示的実施形態においてインビトロの再構成された染色質に基づく配列アセンブリ手法が開示される。本明細書における方法、システム及び組成物を通じて、ゲノム又は他の大きな配列データセットの非常に高度な新たなアセンブリ及びスキャフォールディングが達成され、それによりコンティグは、適切なように相内でグループ分けされ、順序付けられ、配向され、融合され、又は分裂される（ｓｐｉｔ）。同様に、以前に利用可能であったコンティグ及びスキャフォールドの配列情報を再びアセンブルする及びスキャフォールディングすることにより既存のアセンブリを改善するための有用性が、実証される。場合によっては、リードペアを生成するためのＩｌｌｕｍｉｎａ　ＨｉＳｅｑ配列決定の１つのライブラリ及び１つのレーンで、スキャフォールドＮ５０は約５００ｋｂｐから１０Ｍｂｐに増大される。本明細書に開示される方法は、任意の核酸サンプル（例えば、１つ又は複数のゲノム）を分析するために使用され、特に、アセンブルが困難な、トランスポゾンの、又は他の反復要素の豊富な反復又は倍数体のゲノムを含むゲノムサンプル、或いは、特に８、７、６、５、４、３、２時間以下、又は２未満時間でアセンブルを行うのに計算上集中的であるサンプルリードデータセットを結果としてもたらす他のサンプルに適切である。幾つかの実施形態において、前記方法は、本明細書において議論されるようなコンティグ及びリードペアのデータを受け取り、上記で議論されるようなデータを処理し、及び、上記で議論されるようなパラメータが改善された、スキャフォールドされたコンティグのデータを出力するように構成されるプロセッサを含む、コンピュータにより実施されたシステム上で行われる。

【００３１】

本明細書及び添付の請求項で使用されるように、単数形「ａ」、「ａｎ」、及び「ｔｈｅ」は、他に内容が明確に指示しない限り、複数の指示対象を含む。故に、例えば、「コンティグ」に対する言及は、そのようなコンティグを複数含み、「染色体の物理レイアウトをプローブする」に対する言及は、１つの実態を指すための文脈により示されない限り、当業者に既知の染色体及びその同等物の物理レイアウトをプローブする１以上の方法に対する言及を含んでいる。また、「及び」の使用は、特に明記しない限り、「及び／又は」を意味する。同様に、「含む（ｃｏｍｐｒｉｓｅ）」、「含む（ｃｏｍｐｒｉｓｅｓ）」、「含むこと（ｃｏｍｐｒｉｓｉｎｇ）」、「含む（ｉｎｃｌｕｄｅ）」、「含む（ｉｎｃｌｕｄｅｓ）」、及び「含むこと（ｉｎｃｌｕｄｉｎｇ）」は互換的なものであり、区別するようには意図されていない。

【００３２】

様々な実施形態の記載が用語「含むこと（ｃｏｍｐｒｉｓｉｎｇ）」を使用する場合、当業者は、幾つかの特定の例において、言語「～から実質的に成る」又は「～から成る」を使用して代替的に記載される追加の別個の実施形態が示唆されると理解することを、更に理解されたい。

【００３３】

用語「リード」又は「配列決定のリード」は、本明細書で使用されるように、配列が判定されたＤＮＡのセグメントの配列情報を指す。

【００３４】

用語「コンティグ」は、本明細書で使用されるように、ＤＮＡ配列の隣接する領域を指す。「コンティグ」は、当該技術分野で既知である任意の数の方法、例えば、配列を重複させるために配列決定のリードを比較すること、及び／又は、どの配列決定のリードが隣接する可能性が高いかを識別するために既知の配列のデータベースに対して配列決定のリードを比較することにより判定され得る。コンティグは頻繁に、個々の配列のリード、或いは、重複するエンド又はエッジ配列を持つ配列のリードと組み合わせて以前にアセンブルされた配列情報からアセンブルされる。一般的ではあるが、非排他的に、コンティグは、多くの場合は判定されていない配列の間隙又は領域に介入することなく、又は代替的に、既知の配列及び道の長さの領域に介入することなく、より大きな配列のグループ分けに集まる、重複する配列のリードを含む。

【００３５】

用語「スキャフォールド」は、本明細書で使用されるように、１つの物理的分子に対応する少なくとも１つのコンティグ又は配列のリードからの配列情報をさせ、それにより、スキャフォールドの全ての配列情報は、共通の位相を共有し、或いは配列情報が表す核酸が物理的に結合されることを反映する。場合によっては、スキャフォールド配列は、１つのコンティグへとアセンブルされないが、未知の配列の、未知の長さの、或いは未知の配列と長さを持つ、その構成要素であるコンティグ又は配列のリードの間に、少なくとも１つの間隙を持つ場合がある。幾つかのそのような場合において、間隙のある配列はそれにもかかわらず、構成要素である配列が位相にあり又は１つの物理的分子に対してマッピングされると見出されるという事実により、１つのスキャフォールドを構成する。場合によっては、スキャフォールドは１つのコンティグを含み、つまり、場合によっては、スキャフォールドは何の間隙も無い配列の隣接する範囲を含む。

【００３６】

動詞として、用語「スキャフォールドする」は、コンティグ又はスキャフォールドを順序付ける、配向する、端から端まで融合する、１つのコンティグ又はスキャフォールド別のコンティグ又はスキャフォールド中で融合する、及び切断することの少なくとも１つを指し、本明細書に提示される方法により知らされるように、コンティグ又はスキャフォールドを順序付ける、配向する、端から端まで融合する、１つの別の中で融合する、及び切断することを全て含む。スキャフォールディングは、真核生物の染色体上でのコンティグのマッピングから生じ、又は不均一なサンプルにおける複数の生体のゲノムに対応し得るように、１つの分子の１つの位相の上で、複数のスキャフォールドの上で、複数のコンティグをアセンブルするように行われる。

【００３７】

本明細書で使用されるように、「自然配向」は、ペアリードの文脈において、それらが由来する核酸分子セグメントを表す配向又はその配向においてペア配列が生じる、ペアリードを指す。

【００３８】

用語「被験体」は、本明細書で使用されるように、任意の真核又は原核の（真正細菌又は古細菌の）生物又はウイルスを指す場合がある。被験体は代替的に、複数の生物及び／又はウイルスからの核酸材料を含む環境サンプルなどの、その生物の起源とは独立したサンプルを指す場合もある。例えば、被験体はヒトなどの哺乳動物であり得るか、又は、ヒト及び実質的に非ヒトの核酸配列の両方を含むと予想される、例えばヒトの腸から得られるサンプルであり得る。

【００３９】

用語「核酸」又は「ポリヌクレオチド」は、本明細書で使用されるように、一本鎖又は二本鎖の何れかの形態である、デオキシリボヌクレオチド（ＤＮＡ）又はリボヌクレオチド（ＲＮＡ）のポリマーを指す場合がある。具体的に限定されない場合に限り、この用語は、自然に生じる核酸の既知のアナログを含有する核酸分子を包含しており、前記アナログは、基準のヌクレオチドと同様の結合特性を有し、及び／又は自然に生じるヌクレオチドと同様の様式で代謝される。

【００４０】
用語「裸のＤＮＡ」は、本明細書で使用されるように、複合化されたタンパク質又はナノ粒子が実質的に無いＤＮＡを指す場合がある。
【００４１】
用語「再構成された染色質」は、本明細書で使用されるように、単離された核タンパク質、又は裸のＤＮＡに対する他の核酸の結合部分を複合化することにより形成された染色質を指す場合がある。場合によっては、再構成された染色質は実際に、ヒストンなどの核酸及び染色質の構成要素を含み、一方で代替的な実施形態において、「再構成された染色質」は、タンパク質、ナノ粒子、又は、例えば特異的又は非特異的に核酸を結合するスペルミジン又はスペルミンなどの非タンパク質分子などの、少なくとも１つの核酸結合部分と組み合わせて裸のＤＮＡ又は抽出されたＤＮＡから形成された任意の複合体を指すために、より口語的に使用される。
【００４２】
用語「ナノ粒子」は、本明細書で使用されるように、ＤＮＡを結合するために修飾され得るナノメートル規模の球体を指す場合がある。場合によっては、ナノ粒子は、（例えば、アミン含有分子で覆うことにより）それらの表面上で正に荷電される。その全体において引用により本明細書に組み込まれる、Ｚｉｎｃｈｅｎｋｏ，　Ａ．　ｅｔ　ａｌ．（２００５）“Ｃｏｍｐａｃｔｉｏｎ　ｏｆ　Ｓｉｎｇｌｅ－Ｃｈａｉｎ　ＤＮＡ　ｂｙ　Ｈｉｓｔｏｎｅ－Ｉｎｓｐｉｒｅｄ　Ｎａｎｏｐａｒｔｉｃｌｅｓ”　Ｐｈｙｓｉｃａｌ　Ｒｅｖｉｅｗ　Ｌｅｔｔｅｒｓ，　９５（２２），　２２８１０１を参照。幾つかの実施形態において、再構成された染色質は、裸のＮＤＡに対する結合ナノ粒子により合成される。
【００４３】
用語「リードペア（“ｒｅａｄ　ｐａｉｒ”　ｏｒ　“ｒｅａｄ－ｐａｉｒ”）」は、本明細書で使用されるように、自然の核酸サンプル中で隣接していないが本明細書又は他の場所で開示されるような化学操作又は酵素操作の結果として隣接して共有結合され、且つ、単一の配列決定のリードとして配列される、核酸配列の２以上のスパン（ｓｐａｎｓ）を指す場合がある。場合によっては、「リードペア」は、人為的に結合される２つの核酸領域にわたり配列決定することにより得られた、配列情報を指す。場合によっては、リードペアの数は、マッピング可能なリードペアの数を指す場合がある。他の場合において、リードペアの数は、生成されたリードペアの総数を指す場合がある。
【００４４】
本明細書で使用されるように、「サンプル」は、スキャフォールド情報が生成又は改善されることとなる核酸材料を指す。幾つかのサンプルは、細胞単一培養物（ｃｅｌｌ　ｍｏｎｏｃｕｌｔｕｒｅ）、又は単一の多細胞個体からの組織といった均質なソースに由来する。場合によっては、サンプルは、個体の腫瘍サンプルに発生し得る変異といった配列変異を含む。場合によっては、サンプルは不均一なソースに由来するものであり、それによりサンプルは、ヒトの腸又は排泄物のサンプル、環境サンプル、又は生物の混合物などの、複数の生物からの核酸を含む。
【００４５】
本明細書で使用されるように、用語「約」の付く数は、その数を明確に列挙することに加えて、その数のプラス又はマイナス１０％の数量を指すために使用される。
【００４６】
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、本開示が属する分野における当業者に共通して理解されるような意味を有する。本明細書に記載されるものと同様又は同等の方法及び試薬が、開示された方法及び組成物の実施に使用され得るが、ここでは例示的な方法及び材料が記載されている。
【００４７】
本明細書には、核酸の配列相又は物理的結合を示すリードペア情報などの、リードペア配列情報といった配列グループ化情報の使用により、より大きなコンティグ及びスキャフ

ォールドへの単一リード及びコンティグの核酸配列アセンブリなどの配列アセンブリに関する組成物、システム、及び方法が開示される。

【００４８】

ゲノミクスの主な目標は、少ない労力及びコストによる、完全長のハプロタイプ分解された（ｒｅｓｏｌｖｅｄ）染色体配列の正確な再構成である。現時点でアクセス可能且つ手頃な高スループット配列決定方法は、短距離の配列の近接性及びゲノム変異の特性付けに最も良く適している。長距離の結合及びハプロタイプの位相合せの達成は、長い（例えば、数十キロベース）配列を直接且つ正確に読み取る能力、又は、ペアとなった或いはグループ分けされた配列のリードを通じた結合及び位相の関係の捕捉の何れかを要する。これらの方法は、ゲノムのサンプルのための完全なサンプルのハプロタイプのマップ情報を生成するのに必要な配列情報の慣例的又は商業的な計算上の分析が排除されるように、技術的に困難であり、且つ計算上集約的なものである。

【００４９】

高スループット配列決定方法は、ゲノミクスの分野における変革を誘発した。ＤＮＡの何百万もの短い断片からデータを直ちに作成することにより、ゲノムの再配列決定のコストは劇的に低下、即ち１つのヒトゲノム当たり１，０００ドルに急速に近づき、且つ、また更に低下するものと予想されている。

【００５０】

しかし、短いリード配列の、長い隣接するゲノムアセンブリへの形質転換には、実質的な障害が残っている。低コストの配列データから基準品質アセンブリを作り出すことの難問は、今日の技術により生成されたアセンブリの品質と、ヒト基準アセンブリとの比較において明らかである（Ａｌｋａｎ　ｅｔ　ａｌ．，　２０１１）。

【００５１】

ＢＡＣクローン配列決定、物理的マップ、及びＳａｎｇｅｒ配列決定を含む多くの技術が、３８．５Ｍｂｐ　Ｎ５０の長さ、及び１００，０００の塩基当たり１のエラー率を持つ、高品質且つ高度に隣接したヒトの参照標準を作り出すために使用された（Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｈｕｍａｎ　Ｇｅｎｏｍｅ　Ｓｅｑｕｅｎｃｉｎｇ　Ｃｏｎｓｏｒｔｉｕｍ，　２００４）。対照的に、複数のインサートサイズを持つライブラリからの非常に高い被覆率のデータセット上で各々が作成者により実行される、全ゲノムショットガン（ＷＧＳ）アセンブリのソフトウェアパイプラインのパフォーマンスの近年の比較により、魚のゲノム上では４．５Ｍｂｐまでに及ぶＮ５０スキャフォールドの長さを持ち、且つヘビのゲノムでは４．０Ｍｂｐに及ぶ長さを持つアセンブリが作成された（Ｂｒａｄｎａｍ　ｅｔ　ａｌ．，　２０１３）。

【００５２】

リードが短い配列の高い被覆率が、高品質且つ高度に隣接するアセンブリを獲得するのに十分なことは、稀である。これは主に、大小両方の規模での反復的な内容物によるものであり、セントロメア及びテロメアの近くの反復構造、ジンクフィンガー遺伝子のような大きなパラロガス遺伝子ファミリー、及びＬＩＮＥとＳＩＮＥなどの散在した核要素の分布を含む。そのようなアセンブルが困難な内容物は、多くの真核生物のゲノム、例えばヒトゲノムの６０－７０％で大部分を構成する（ｄｅ　Ｋｏｎｉｎｇ　ｅｔ　ａｌ．，　２０１１）。そのような繰り返しが入力配列データに及ぶことができない場合、断片化され且つ不正確なアセンブリが結果として生じる。通常、新たにアセンブリのための出発点は、深い被覆率（最小５０Ｘ－２００Ｘ）、２～８ｋｂｐの間のインサートサイズを持つ中距離の「メイトペア」ライブラリを備えた短距離（３００－５００ｂｐ）のペアエンドの「ショットガン」データ、及び頻繁に、より長い距離（３５ｋｂｐ）のフォスミドエンドペアを組み合わせる（Ｇｎｅｒｒｅ　ｅｔ　ａｌ．，　２０１１；　Ｓａｌｚｂｅｒｇ　ｅｔ　ａｌ．，　２０１２）。

【００５３】

低コスト配列データは有用であるが、位相情報が起因し得るより大きなコンティグ又はスキャフォールドへと仕上げることは困難である。従って、ゲノム再編成に関する有益な

情報、又は、単一の遺伝子座（シス又はトランスであり、２つの独立して突然変異した対立遺伝子又は二重突然変異対立遺伝子と組み合わせた単一の野生型対立遺伝子に対応する）内に分散した多数の突然変異の相に関するより単純な情報は頻繁に、幾つかの低コストの配列アセンブリから利用することはできない。

【００５４】

新たにアセンブリの近接性及び正確性を増大させるための多くの方法が、近年開発されている。広くは、そのような方法は、配列決定から生成されるリードの長さを増大させること、又はペアの短いリードの間のインサートサイズを増大させることを試みている。例えば、ＰａｃＢｉｏ ＲＳ ＩＩは、長さ２３ｋｂｐ（中間で２ｋｂｐ）までの生のリードを生成することができる。しかし、この手法は、～１５％もの高さのエラー率に悩まされると報告されており、高スループットの短いリードよりも、～１００倍高価なままである（Ｋｏｒｅｎ ｅｔ ａｌ．， ２０１２； Ｑｕａｉｌ ｅｔ ａｌ．， ２０１２）。Ｏｘｆｏｒｄ Ｎａｎｏｐｏｒｅから市販で入手可能な長いリードは有望なものではあるが、頻繁により高いエラー率及びより低いスループットを有している。Ｉｌｌｕｍｉｎａ’ｓ ＴｒｕＳｅｑ Ｓｙｎｔｈｅｔｉｃ Ｌｏｎｇ－Ｒｅａｄの技術（以前はＭｏｌｅｃｕｌｏ）は現在、最大１０ｋｂｐのリードに制限されている（Ｖｏｓｋｏｂｏｙｎｉｋ ｅｔ ａｌ．， ２０１３）。

【００５５】

多くの改善にもかかわらず、フォスミドのライブラリの作成（Ｗｉｌｌｉａｍｓ ｅｔ ａｌ．， ２０１２； Ｗｕ ｅｔ ａｌ．， ２０１２）は、時間がかかり、高価なままである。

【００５６】

現在まで、配列決定のコミュニティーは、何千もの脊椎のある種（Ｈａｕｓｓｌｅｒ ｅｔ ａｌ．， ２００９）又は何十万ものヒト（Ｔｏｒｊｅｓｅｎ，２０１３）の配列決定のような、大規模の計画に必要とされる規模及びコストにおいて利用可能な、大きなインサート又は長いリードのための一貫して優れた技術に決定していなかった。

【００５７】

本明細書には、核酸配列アセンブリのための方法、及びコンピュータにより実施されるシステムが開示される。本明細書に開示される方法は特に、ペアエンドリードを用いた分析及び配列の改善に適している。ペアエンドリード又はリードペアは多くの異なる手法を使用して生成される。幾つかの手法は、数百キロベースまで分離される長距離のリードペアを生成するためのインビトロの方法、及び、リードペアにおける各リードがマッピングされるコンティグへの共通の相又は物理的結合の情報を割り当てる際のそれらの使用を含む。本明細書における開示の幾つかの実施形態の中心となるのは、新たにスキャフォールドアセンブリを改善し且つ位相合せするのに役立つ長距離のリードペアデータを生成するための、距離とリード数との関係を利用するＨｉ－Ｃの予想外に有効な改善である。その前身となるＨｉ－Ｃの方法とは異なり、本明細書に開示される幾つかの手法は、固定、ペアエンドの形成、及び後の工程のための基質としてインビトロの再構成された染色質を使用する。結果として生じるデータは、リードペア数の間の関係及びリード間の距離を含む、Ｈｉ－Ｃデータの特徴の多くを共有する。しかし、多くの実施形態において、それにより生成されたペアエンドリードは、宿主細胞内の三次元の染色質又は他の核酸の構成に関する情報、Ｈｉ－Ｃ技術の特に有益な目的である情報を欠いているが、これは、位相内に無い又は互いに物理的に連結していない分子からの核酸配列の近接情報を含み得る。このような新たなインビトロのデータは、単離に使用され得るか、又は、ペアエンドのショットガンリード又は他の以前に生成されたコンティグ情報と組み合わせることで、価格及び時間の一部に関してフォスミドベースのアセンブリに匹敵する正確性と近接性を持つ新規のスキャフォールドアセンブリを生成することができる。そのような進展の分析に関連した方法、組成物、及びコンピュータ実装システムは、新規のアセンブリの品質を改善し、ハプロタイプを位相合せし、及び構造変異種を識別するための有用性として、本明細書に開示される。

【００５８】
本明細書における開示は、わずか約１０，０００、約２０，０００、約５０，０００、約１００，０００、約２００，０００、約５００，０００、約１００万、約２００万、約５００万、約１０００万、約２０００万、約３０００万、約４０００万、約５０００万、約６０００万、約７０００万、約８０００万、約９０００万、約１億、約２億、約３億、約４億、約５億、約６億、約７億、約８億、約９億、約１０億のリードペア、又は１０億より多くの塩基対を用いて、高度に隣接し且つ正確なヒトゲノムアセンブリを生成することができる方法及び計算システムを提供する。場合によっては、本開示は、約５０％、６０％、７０％、７５％、８０％、８５％、９０％、９１％、９２、９３％、９４％、９５％、９６％、９７％、９８％、９９％、又はそれより多くの精度で、ヒトゲノムにおけるヘテロ接合型変異体の約５０％、６０％、７０％、７５％、８０％、８５％、９０％、９１％、９２％、９３％、９４％、９５％、９６％、９７％、９８％、９９％、又はそれより多くを位相合せする、方法及び計算システムを提供する。
【００５９】
リードペアのエンドの生成及び配列決定に関連する組成物及び方法、並びにタグを付けたエンドのリードを生成する代替的な方法は、例えば、その全体において引用により本明細書に組み込まれる、２０１４年８月７日公開の米国特許公開ＵＳ２０１４０２２０５８７に見出される。
【００６０】
本明細書に開示される実施形態は、以下の工程の各々までを含む以下の工程の少なくとも１つを、単独で、又は本明細書に開示される或いは当業者に既知の付加的な工程と合わせて、含んでいる：
１．入力の前処理
２．コンティグ間連結グラフの構築
３．スキャフォールド構築物にシード値をもたらす
４．局所の順序、配向、及び間隙の大きさの改善
５．ペアのスキャフォールドの混合に関するスコアの計算
６．接合の貪欲な受け入れ
【００６１】
幾つかの実施形態において、一連の前述の工程は組み合わせて実行される。場合によっては、工程４、５、及び６の少なくとも１つが、例えばアセンブリの品質の持続した改善を達成するために反復して実行される。
【００６２】
本明細書に開示される幾つかの方法及びシステムは、３つの入力を考慮する：
１．開始アセンブリの配列（随意にＦＡＳＴＡのフォーマットにある）；
２．開始アセンブリへのペアリードの位置合わせ（随意にＢＡＭのフォーマットで、分類され、且つインデックスを付けられる）；
３．開始アセンブリへのショットガンリードの位置合わせ（随意にＢＡＭのフォーマットで、分類され、且つインデックスを付けられる）。
【００６３】
代替的なデータフォーマットが考慮され、方法及びシステムは特定のデータフォーマット又はインデクシングには限定されない。開始アセンブリは、配列のリードからアセンブルされた少なくとも１つのコンティグを含み、及び場合によっては、配列フォーマットとは独立して、少なくとも１つのコンティグを含む少なくとも１つのスキャフォールドを含む。
【００６４】
ペアリード、ペアエンドリード、又はリードペアは、代替的に言及されるように、配列フォーマットとは独立して、隣接していない標的サンプル配列に対応する配列情報を含む。多くの場合、ペアリードは、サンプル核酸配列における共通の分子の長距離まで離れた位置から、単一の物理的分子のリード配列に対応する。

【００６５】
本明細書には、コンティグデータを、互いに対するコンティグの相対位置、相対配向の少なくとも１つを表すスキャフォールドへとアセンブルするための方法と計算システムが開示される。場合によっては、個々のコンティグは、比較可能な大きさであるが改善された配列品質を持つコンティグ又はスキャフォールドを形成するため、その位置及び／又は配向が決定された隣接するコンティグを結合することにより、又は、少なくとも１つのコンティグを、間隙、或いは第２のスキャフォールド又はコンティグのアセンブルされていない領域へと挿入することにより、前記プロセスにおいて融合される。
【００６６】
サンプルに関するペアエンドリード情報は、本明細書に開示され、本明細書に組み込まれ、又は当業者に既知の方法を用いて生成される。場合によっては、ペアのタグを付けたエンドのリードなどの、タグを付けたエンドのリードは、本明細書のあらゆる場所で議論されるようなペアエンドリードに置き換わる。幾つかの例において、ペアエンドリード又は他の情報は、ショットガン配列決定コンティグ情報などのコンティグ情報を組み合わせて使用され、場合によっては同時に生成され、及び、他の場合においては、例えば平行して実行される以前の配列決定の労力又はショットガン配列決定の労力から独立して獲得される。場合によっては、コンティグ情報は、配列データベース又は以前の配列決定の労力から獲得される。
【００６７】
ペアエンドリードは、本明細書に開示されるように、又は、当業者に既知の或いは本明細書における開示に適合する他の方法を介して生成される。ペアリードの生成における軽微且つ主要な変異が考慮される。ペアリードは、処理の前にサンプル材料において隣接しない、１対のリード配列を含む。大半の場合、ペアリードは、単一の物理的な分子に対しマッピングされるが、互いに幾つか距離が空いている。場合によっては、別個の物理的な分子からのペアを含む、ペアリードが生成される。本明細書に開示される方法において、そのようなペアエンドリードは比較的稀であり、これらは頻繁に、本明細書に開示されるアセンブリプロセスにおいて早期に分析から除外される。
【００６８】
両方のリードが単一の物理的な分子から生じる、ペアエンドリードの中で、幾つかのペアリードが、わずか１００、２００、３００、４００、５００、又は５００より多くの塩基対である領域から離れて生じ、一方で他の幾つかのペアリードは、単一の物理的な分子上で１、２、３、４、５、６、７、８、９、１０、１５、２０、２５、３０、３５、４０、４５、５０、６０、７０、８０、９０、１００ｋｂ、又は１００ｋｂより多くによって分離される配列から生じる。頻繁に、一連のペアエンドリードは、幾つかのリードペアがそれらの自然の配向において核酸セグメントを表示するように、ペアエンドリードの距離の分布を含む。
【００６９】
幾つかの実施形態において、本明細書に開示される方法は、生体に関する染色体の数とは無関係の方法でコンティグをグループ分けする。結果として生じるより小さなコンティグクラスターをスキャフォールドへとアセンブルするための幾つかの関連技術におけるよりも、シングルリンクのクラスタリングのためのコンティグ間リンク上の保存的な閾値が適用され、後のスキャフォールディングの連結は本明細書に開示される様々な方法により可能である。このような実施形態における利益は、予期された合計の染色体の長さ又は数を引き出す（ｆｏｒｃｅ）ために相殺誤差によってアセンブリ誤差が「広がらない」ということである。正確なコンティグ、又はその位置でアセンブルされ又はスキャフォールドされる誤って配されたコンティグにより変位されたスキャフォールドが、配列精度を犠牲にして全体的な染色の数又は長さを保持するように第２の位置へと「引き出され」なければならないため、染色体のサイズ又は数に依存した方法を使用すると、１つの誤って配されたコンティグは頻繁に複数の誤差へと繋がる。
【００７０】

ペアエンドの配列は、コンティグ情報、及び場合によっては、完全なヒトゲノムプロジェクトに利用可能な配列情報などの以前の既存のスキャフォールド情報へとマッピングされる。ペアエンドは、場合によっては、１対の１つ又は両方の配列がスキャフォールドの単一のコンティグ又は固有の領域へと一意的にマッピングされるように選択される。コンティグ又はスキャフォールド上で互いに関連して判定され得る別個の位置へと、ペアの両方のリードが一意的にマッピングされるペアエンドリードのために、ペアにおけるリード間の距離が判定される。リードペアの距離度数曲線はこのデータから計算され、及び場合によっては、広範囲のリードペア距離にわたって与えられたリードペア距離に関する度数予測値をもたらすように外挿される。結果的に、例えば１つのコンティグの別のコンティグへの２以上の相対的な配向に対応する、２以上の別個のリードペア距離が可能である単一のリードペアのために、どのコンティグ配向がリードペア距離へとより可能に対応するのかを判定することができる。同様に、複数のリードペアが１対のコンティグにマッピングされる場合、個々のリードペアのための、及びセットに関する総計の距離のための最も可能なリードペア距離を判定することができる。場合によっては、複数のリードペア距離に対するコンティグ配向の相対的な影響を評価することにより、及び、全体でより可能な又は最も可能なリードペア距離のセットに通じる配向を選択することにより、与えられたコンティグペアの「総計の最も好ましい配向」を判定することができる。

【００７１】

一般的に、個々のリードペアに関して、及び大半のリードペア距離分布に関して、より短いリードペア距離が、より可能となる。しかし、複数のリードペア距離が、１対のコンティグ又は複数のコンティグについて予測されると、最も可能なリードペア分布は、場合によっては短い及び長いリードペア距離の両方を含み、それにより、リードペア距離の分布は、全体のリードペア長さの単純な最小化よりも、予測されたリードペア分布をより近くに反映する。

【００７２】

図２において、本明細書に記載されるようなリードペア距離度数曲線の例を見ることができる。データは、リードペア距離に応じて出現度数として示され、及び、「指数関数的に」又は「対数的に」減少して、リードペア距離を増大させると観察される。代替的なデータの描写は、本明細書における開示と一致している。

【００７３】

場合によっては、一連のリードペアデータがサンプルのために生成されると、リードペア距離の曲線が独立して判定される。代替的に又は組み合わせにおいて、以前に判定されたリードペアデータは、リードペア距離の曲線を生成するために使用される。場合によっては、以前に生成された、又は独立して生成されたリードペア曲線が使用される。

【００７４】

順序情報又は配向情報、或いは順序と配向の情報を持つスキャフォールドへとコンティグを組み込むために、ペアリードが選択され、ペアの両方のリードが別個の位置へと一意的にマッピングされるが、リードペアの２つのリードの別個の位置は、コンティグ又はスキャフォールド上で互いに対して判定することができない。この状況は例えば、リードペアの別個のリードが、共通のスキャフォールド上で確信をもって配されない別々のコンティグにマッピングされる時に、又は、互いに対するコンティグの距離、配向、又は距離と配向が分からないように、別々のデータが別のコンティグに対する１つのコンティグの位置に疑いを抱かせる場合に、生じる。

【００７５】

品質の低いリードが除外されるように、随意に、リードペア配列がコンティグ分析の前に選別される。場合によっては、単一のコンティグに一意にマッピングしない少なくとも１つのリードを有するリードペアが除外されるように、リードペア配列がコンティグ分析の前に選別される。場合によっては、リードペア配列がマッピングするコンティグごとの単一の位置に一意にマッピングしない少なくとも１つのリードを有するリードペアが除外されるように、リードペア配列がコンティグ分析の前に選別される。場合によっては、リ

ードペアの配列ペアの少なくとも１つが、不均衡な数のリードペア配列がマッピングする領域にマッピングする場合に、例えば、あらかじめスキャフォールドされた配列データセット、または配列アセンブリの対象であるデータセットの叙述にわたり、リードペア「ヒット」におけるスパイクによって同定される場合に、リードペアは選別される。いくつかの選別手法では、ステップサイズは、例えば１０００ｂｐと定義され、各ステップにおいては、連結のための大部分の支持を占める可能性のある、連結された領域に隣接するリードペアのビンが有る状態および無い状態で、連結されたコンティグのための支持が計算される。代替的なステップサイズは、１００、２００、３００、４００、５００、６００、７００、８００、９００、１５００、２０００、３０００、４０００、５０００、１００００または１００００を超えるサイズも検討される。ビンが、ヒット分布の平均値、中央値または代表値に基づいた閾値を超えるリードペアの数を有する場合は、リードペアは除外される。場合によっては、閾値は、平均値、中央値または代表値の１．５ｘ、１．６ｘ、１．７ｘ、１．８ｘ、１．９ｘ、２ｘ、２．１ｘ、２．２ｘ、２．３ｘ、２．４ｘ、２．５ｘ、２．６ｘ、２．７ｘ、２．８ｘ、２．９ｘ、３ｘ、３．１ｘ、３．２ｘ、３．３ｘ、３．４ｘ、３．５ｘ、３．６ｘ、３．７ｘ、３．８ｘ、３．９ｘ、４ｘ、４．５ｘ、５ｘ、または５ｘを超えるように選択される。配列リード選別のためのパラメータおよび基準は、当業者に知られており、多くの選別パラメータが本明細書で検討される。場合によっては、最多数のリードがマッピングするビンを備える少なくとも単一の塩基によって重複するリードが、分析から除外される。

【００７６】

コンティグは、初期のコンティグの位置決めを生成するために互いに対してグループ分けされる。コンティグの位置決めのための多数の手法が本明細書で検討され、そして代替的な実施形態は、初期のコンティグの順序および／または配向の選択において異なる。例えば、場合によっては、コンティグは、以前に生成されたドラフト、完全なゲノムスキャフォールドまたは染色体マップ上にマッピングされる。そのようなマップは、標的種の以前の配列決定から、重要なゲノム規模のシンテニーが予期されるもしくは知られるための関連した種のような密接に関連した種の以前の配列決定から、または、代表的なゲノムが配列決定されている種の別個の集団のために、得られる。例えば、ネアンデルタールのゲノムは、ヒトゲノムスキャフォールドに対してマッピングされ、野生ナス科（ｗｉｌｄ Ｓｏｌｅｎａｃｅａｅ）のファミリーメンバーは栽培されたトマト（Ｌｙｃｏｐｅｒｓｉｃｏｎ ｅｓｃｕｌｅｎｔｕｍ）のゲノムに対してマッピングされ、または特定の植物品種は近縁種の配列決定が最も近いゲノムに従って整列される。当技術分野で知られている位相決定の代替的な方法または計算システムを用いた順序付けなどの他の初期のコンティグの位置決めの方法は、本明細書で検討され、一貫している。

【００７７】

いくつかの場合、コンティグは、共有リードペアの数に従ってグループ分けされる。すなわち、場合によっては、より多くのリードペアを共有するコンティグは、互いに近接してグループ分けされるが、少数のリードペアしか共有しないコンティグは、互いに比較的離れて位置決めされる。

【００７８】

いくつかの局所的なグループ分けについて、コンティグは直線的に整列される。しかしながら、初期のグループ分けは、大量の分枝したコンティグまたは環状コンティグの配向をしばしば表し、場合によっては、初期のグループ分けは例えば、遺伝的または物理的結合グループに関する既知の遺伝情報、または染色体数に関する既知の細胞学的情報に基づいて予測されるよりも実質的に大きいグループを含む。

【００７９】

場合によっては、初期のグループ分けは、共有ペアエンドリードによって表されるようなコンティグ間の関係の強度について評価される。例えば、共有ペアエンドリードの数が２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、または、場合によっては１５を超える閾値数である場合、ペアエンドリードの閾値未満のコンティグは

、場合によっては分離される。

【００８０】

少なくとも１つのリードペアを共有するコンティグは同定され、単一の物理結合グループへとグループ分けされる。場合によっては、コンティグは少なくとも２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０、または２０よりも多いペアエンドリード共有する場合、単一の物理的結合グループへとグループ分けされる。場合によっては、この閾値は、配列決定されている核酸サンプルに従って変動する。

【００８１】

場合によっては、コンディグは、例えば、前もって生成されたまたは前もって得られたスキャフォールド情報、コンティグ間で共有されるリードペアの総数、またはコンティグの長さもしくは固有のコンティグ配列によって標準化されたコンティグ間で共有されるリードペアの総数に基づいて、互いに対してあらかじめ順序付けされている。

【００８２】

共有リードペアの生の数（ｒａｗ　ｎｕｍｂｅｒ）を数える代わりに、場合によっては、コンティグは各リードペアのリードペア距離に関して評価され、確率値が閾値より下となる距離を有するリードペアが、一般的なリードペアの評価から除外される。

【００８３】

したがって、初期の評価においてコンティグのペアが近位のグループ分けを確実にするという判定（すなわち、それらが一般的な「エッジ」に一時的に配置されるかどうか）は、共有ペアリードの総数か、または閾値距離確率を有する共有ペアリードの総数によって判定される。場合によっては、コンティグエッジの判定と関係があるペアエンドリードの代替的な評価が用いられる。

【００８４】

コンティグは、局所的に直線状の配置へと位置決めされる。場合によっては、局所的に直線状の配置は、１つのエンドで他のコンティグにそれらを連結するためには不十分なリードペアを有する切断点によって境界が定められる。場合によっては、局所的に直線状の配置は分枝点によって１つのエンドにおいて境界が定められ、それによって、単一のコンティグは、１つのエンドで２つの別々のコンティグに等しく強力に結合される。場合によっては、分枝点によって１つのエンドで境界が定められたコンティグの直線状の配置は分枝点で切断され、別々の直線状のグループ分けとして扱われる。場合によっては、代替的に、分枝点の接合部は切断され、代替的なアセンブリ条件下で再アセンブルされ、結果として、変更されたアセンブリ条件またはアセンブリ条件の第２のセットの下、分枝はもはや同程度に可能ではなく、単一の分枝はコンティグの直線状の配置を継続するために選択される。

【００８５】

いくつかの場合、特に短い分枝または短いコンティグを含む分枝の場合、分枝は、分枝点における他のエッジコンティグ間にマッピングするコンティグまたは直線状のコンティグのセットを表す。小さなコンティグは、偶然でもそうでなくても、多数のリードペアエンドのマッピング目標になる可能性が低い。結果として、短いコンティグは、それらにマッピングする１つまたは少数のリードペアエンドを有するだけでもよく、結果として１つの隣接するコンティグへの予測される結合をもたらす。この状態は、分枝点の１つのメンバーが集合的に比較的長い配列を表さない、シングルトンコンティグまたはコンティグの直線状のグループ分け、もしくはクラスターである時に、しばしば観察される。場合によっては、小さい分枝点コンティグまたはコンティグの直線状の配列は、隣接する直線状のコンティグシリーズの様々なアセンブリに対するリードペア距離およびリードペア距離確率に対する影響について評価され、その結果、小さな分枝点コンティグまたはコンティグの直線状の配列は、リードペア距離に対する局所的な影響が良好であれば分枝点に挿入される。場合によっては、リードペア距離に対するコンティグまたはコンティグシリーズ挿入の影響は、挿入部位から遠く離れている位置に挿入されている、挿入されたコンティグ

またはコンティグのセットに対して挿入されるような、同一のコンティグまたはコンティグシリーズのリードペアエッジスコアに対する相対的な影響を比較することによって評価される。

【００８６】

初期のコンティグの直線状のグループ分けの概略図は、図１０のＡ～Ｄに示されている。コンティグは円として描かれ、コンティグペアエッジは線として描かれる。図１０のＡには、初期のグループ分けが示される。場合によっては、初期のグループ分けは比較的あまり厳重ではなく、全てのコンティグが単一のクラスターに結合される。図１０のＢでは、コンティグエッジが評価され、コンティグ、スキャフォールドまたはノード間のあまり厳重ではない結合が除去される。続く高度に厳重なグループ分けは、多くの場合、結合グループ（染色体または染色体部分）に対応する、またはマッピングする。図１０のＣでは、グラフ内のサイクルは、分枝ツリーを形成する可能性がより弱いまたはより低い残りのエッジを切断することによって除去される。図１０のＤでは、分枝ツリーは、直線状のスキャフォールドをもたらすために分枝点で切断される。場合によっては、図１０のＤの直線状のスキャフォールドに対して回帰ウィンドウ分析を行う。

【００８７】

図１１は、模範的な最小全域木を示す。サイクルは切断されているが、複数の分枝がクラスター内に示される。コンティグはノードとして示され、長さは各楕円内に示され、そして各コンティグペアエッジの数値はそのエッジをサポートするリードペアの数を示す。この段階での枝の大部分が小さなコンティグまたはコンティググループの短いグループ分けに対応することが、観察される。多くの場合、これらの分枝は、さらなる分析によって、それらが隣接してマッピングされるノードのような、隣接するノード間でマッピングすること、またはノード内の不確定配列の間隙または領域にマッピングすることが分かった。小さなスキャフォールド、コンティグまたはノードがしばしば分枝として現れる１つの理由は、その長さが比較的短く、それぞれ隣接するノードにマッピングするリードペア配列によってヒットされる傾向が低いためである。

【００８８】

本明細書に開示される方法または順序付けの代替的な方法を使用して、コンティグを局所的に直線状の順序にグループ分けした上で、コンティグは、リードペア距離確率の局所的に網羅的な評価を使用して再び位置決めされる。

【００８９】

コンティグの直線状のグループ分けが同定され、コンティグの直線状のグループの１つのエンドのサブグループのような、コンティグの「ウィンドウ」または局所的に隣接するサブグループが選択される。代替的に、グループ分けが直線的ではなくても、局所的なクラスターへとグループ分けされたコンティグ、またはターゲット「ウィンドウサイズ」に分割可能なグループ分けは、以下で分析される。多くのウィンドウサイズは、本明細書の開示に対応しており、そのウィンドウサイズは、２、３、４、５、６、７、８、９、１０、または１０よりも上であり、コンティグの直線状のグループ分けの全長を含んでいる。いくつかの模範的な実施形態では、ウィンドウサイズは、３または４である。一般に、より大きなウィンドウサイズは、評価されたコンティグリードペアの特性または指標（予測されたリードペア距離の最大尤度、予想される尤度距離からの最小全体偏差、最小全長距離、または代替指標など）に関して、最終的な順序が最適であるという、より大きな計算上の信頼性を提供する。しかしながら、より大きいウィンドウサイズは、実質的により長い計算時間を必要とし、場合によっては、またはいくつかのコンピュータシステムでは、より大きなウィンドウサイズは、時間またはコストに関して計算上極端なものである。３から４のウィンドウサイズがが、追加されるウィンドウサイズを網羅的に調査するために必要な追加の計算時間無しに、より大きなウィンドウの精度に近似するスキャフォールディング精度を達成することが分かった。

【００９０】

しばしば、単一のウィンドウサイズは、望まれる計算負荷または計算時間に適応するよ

うに選択され、スキャフォールドまたはコンティグデータセット全体に繰り返し使用される。しかしながら、代替案は本明細書に開示される方法とも一致する。場合によっては、ウィンドウサイズはデータセットの異なる領域に従って変動し、結果として、ウィンドウは、ウィンドウ共通のマッピングされたリードペアの最小、最大または平均目標数を含むように定義される。場合によっては、ウィンドウサイズは、代替的に、または組み合わせて、基礎となる配列の計算複雑性を反映するために選択され、結果として、計算上負荷のある領域のような反復領域または不十分な配列品質の配列に、より簡単な配列の領域と比較してより大きなまたはより少ない計算能力が提供される。場合によっては、ウィンドウサイズは、単独または本明細書の他の箇所に列挙された因子と組み合わせて、最小または最大範囲のコンティグ配列長さによって与えられる。よく特徴付けられたゲノムに対し、ウィンドウサイズは、トランスポゾンの豊富な領域、テロメア、セントロメア、リボソーム反復領域、および／または複合遺伝子座または反復遺伝子座に適応するために、随意に変更される。

【００９１】

ウィンドウ内にあるコンティグセットの各リードペアまたはリードペアのサブセットに対し、リードペアが互いに対してマッピングする２つのコンティグのための４つの配向のそれぞれは、２つのコンティグ間で共有される各リードペアを含む少なくとも１つのリードペアの間の距離に対する与えられた配向を有する距離の影響について評価される。場合によっては、距離の組み合わせの確率は、与えられた配向について、コンティグの他の配向と比較して与えられた配向の総計の相対確率が評価され得るように判定される。本明細書に開示されたいくつかの方法では、最も高い総計の距離確率をもたらす配向は、互いに対するコンティグ配列の物理的位置の最も可能な反映として選択される。場合によっては、予期されるリードペア分布から最も低い偏差を表すリードペア距離をもたらす配向が選択される。他の順序付けの基準は、全リードペアの長さを最小化すること、閾値長さまたは代替的な基準を超えるリードペアの数を最小化することのような代替的な実施形態において検討される。

【００９２】

本明細書の他の箇所で論じられるように、いくつかの実施形態では、２つより多いコンティグのグループが、それらの順序、配向、または順序および配向に関して評価される。場合によっては、２、３、４、５、６、７、８、９、１０、または１０より多いコンティグの「ウィンドウ」が、それらの配向、順序、または互いに対する配向および順序を判定するために同時に評価される。実際に計算時間は、与えられたウィンドウ内の約４つのコンティグよりも大きいグループ分けの分析を妨げる。しかしながら、上記された、および本明細書の他の箇所で論じられた他の形態と同じく、単一のウィンドウ内で一度に多数のコンティグを分析することが検討される。

【００９３】

ウィンドウ分析をより詳細に見ると、コンティグを結合するリードペアの分離の確率モデルを考慮して、コンティグの構成を改善するための本明細書に開示された方法および計算システムが開発され、試験された。いくつかの方法は、コンティグのクラスターの初期の順序付けにわたってサイズｗのウィンドウをスライドさせる、動的プログラミングアルゴリズムに関する。複数の位置ｉまで、場合によっては各位置ｉでは、ウィンドウ内のコンティグを順序付けおよび配向付する全てのｗ！２^{w}の方法を考慮し、ウィンドウ内のコンティグの現在の構成で終わる現在のウィンドウ位置の端部までの全てのコンティグの最適な順序付けおよび配向を表すスコアを記憶する。そのためには、位置ｉ－１、ｉ－２．．．ｉ－ｗでウィンドウ内の「適合性のある」順序および配向のスコアを見て、現在の構成を用いてそれらの順序付けの拡張を評価する。ｗ！２^{w}は、そのような急勾配関数であるため、本方法は、実際にはｗが小さい値になるように限定されている。いくつかのデータの試験では、ｗ＝３は、いくつかのデータセットのための構成精度を大幅に向上させることができる。不変的なおよび可変的なウィンドウサイズ選択のための代替的な論理的根拠として、代替的なウィンドウサイズが検討される。ウィンドウのための最適なスコアま

たは局所的に最大のスコアを発見した上で、コンティグはスキャフォールドされ、それによって前記最適なスコアまたは局所的な最大スコアが達成される。その後、ウィンドウは1つの位置へと進められ、分析が繰り返される。

【００９４】

ウィンドウ分析は、初期コンティグの順序付けプロセスのいくつかの実施形態とは異なるが、なぜならば、初期の順序付けとは異なり、ウィンドウ分析は、コンティグ順序およびコンティグ配向の両方を含むすべての可能なコンティグ構成、ならびリードペア長さに対する、および計算されたまたは前もって判定されたリードペア長さ確率曲線に関するリードペア長さ確率に対する各構成の影響をしばしば局所的に網羅するからである。すなわち、いくつかのウィンドウ分析では、ウィンドウ内の全ての可能なコンティグ構成が評価され、場合によっては、リードペア長さに対するコンティグ構成の全ての可能な影響は、コンティグまたはスキャフォールドの順序付けおよび配向において反映されたリードペア長さの組み合わせが、リードペア長さ確率曲線の観点から局所的に最も可能性が高いことについて判定され、評価される。

【００９５】

多くの方法およびアルゴリズムは、２つのコンティグ間で共有されるリードペアのセットのための相対的な総計の確率の計算に対して利用可能である。本明細書に開示される１つの模範的なアルゴリズムは以下の通りである。

【００９６】

【数１】

$$L(l1,l2,g,o) = \frac{N!}{(N-n)!}(1-p_o)^{N-n}\prod_{i=1}^{n} f(d_i)$$

（式１）

【００９７】

この尤度関数は、コンティグが相対的な配向ｏ

【００９８】

【数２】

$$o \in ++,+-,-+,--$$

を有し、長さｇの間隙によって分離されると仮定して、数ｎ、およびコンティグ１と２との間のスパニングリードペアｄｉの示唆される分離を観察する確率を与える。

【００９９】

代替的なアルゴリズムが検討され、本明細書の開示と一致することが理解される。いくつかのアルゴリズムは、相対的なコンティグの配向に対する全リードペア距離を評価することを含み、その結果、最小の全リードペア距離が評価され、最小の全リードペア長さの合計に対応する配向が選択される。代替的に、または組み合わせて、予期されるリードペア分布パターンから与えられたコンティグ配向セットのためのリードペア距離の変動が評価され、その変動を最小にするコンティグ配向が選択される。

【０１００】

上に述べたように、いくつかの実施形態では、２、３、４、５、６または６より大きいコンティグのウィンドウが、その成分コンティグの相対的な順序、配向または構成などのスキャフォールディングに関して評価される。グループ内のコンティグの２つに共通するリードペアのための総計の相対距離度数を、外的制約の観点から最小化するまたは減少させる２、３、４、５、６、または６より多いコンティグのセット、または予測されるリードペア距離の発生の尤度を最大にする順序付けのための順序、配向または構成を含むスキャフォールドを判定する上で、そのグループ内のコンティグは適宜に順序付けられ、配向さられる。

【０１０１】

与えられたウィンドウ内のコンティグのセットを順序付け、配向、または順序付けおよ

び配向する上で、最適に判定された順序における１つの末端コンティグが除外され、残りのコンティグは本明細書で論じられるように、コンティグの予備的な順序付けに基づいて、１つの付加的な新しいコンティグと組み合わせて再度分析され、隣接するコンティグを推定的に表すものとして再度同定される。

【０１０２】

このプロセスは、例えば、推定のコンティグ順序の一端から他端まで一方向に繰り返される。代替的な実施形態では、ウィンドウの配向は、順序付けられたコンティグセット内のランダムな位置から始まり、コンティグエッジに向かって一方向にもしくは両方のエッジに向かって双方向に段階的に進むか、またはスキャフォールドのいずれかの端部から始まり、内部合流点に進む。

【０１０３】

代替的な実施形態では、ウィンドウは、１つの工程で２つまたは３つのコンティグにより「移動」され、その結果、単一のコンティグが同時に排除されそして追加されるのではなく、むしろ、２つまたは３つの隣接するコンティグが、コンティグの順序付けされたウィンドウの片方から除去され、そして対応する数がもう片方に追加される。

【０１０４】

場合によっては、グループの全てのコンティグに至るまでの大部分が、本明細書で選択されたとおりにパラメータを最小化または最大化するように順序付けられるまで、ウィンドウ分析が継続される。ウィンドウ分析の多くの実施形態の共通の特徴は、与えられたウィンドウ内のコンティグに対して各コンティグを含むウィンドウコンティググループのサブセットについて局所的な最大または最小をもたらす配向が網羅的に調査されるように全ての配向が網羅的に分析され、および場合によっては、リードペア距離に関する局所的な最大または最小のパラメータが、コンティグのグループ分けにおいて、のために同定される。

【０１０５】

部分的に完全な、または完全なウィンドウ分析から結果として生じるコンティグの配向は、場合によっては、単一のスキャフォールドまたはデータセットサイズの「ウィンドウ」内の全てのコンティグを網羅的に分析することによって生成されるような、全体的に最適化されたコンティグの配向と同一または実質的に類似するコンティグの配向を表す。しかしながら、本明細書のウィンドウ分析を用いることにより、実質的に少ないコンティグの配向の空間は、局所的に最適化されたコンティグの配向に達するために分析される必要があり、実質的により少ない計算時間または計算能力は、局所的に最適な配向に達するために占有される必要がある。

【０１０６】

いくつかの実施形態では、（ｉ）挿入分布のモデルは、既知の配列ライブラリからリードペアの観測されたデータ分布に基づいて構築され、（ｉｉ）生成されたリードペアは、モデルに対して個別に順位付けされ、（ｉｉｉ）モデルとの比較に基づいて、リードペアにスコアが与えられ、そして（ｉｖ）高いスコアを有するリードペアは、モデルに最も適合するリードペアである。

【０１０７】

いくつかの実施形態では、（ｉ）挿入分布のモデルは、前もってマッピングされていない配列ライブラリからリードペアの観測データ分布に基づいて構築され、（ｉｉ）その後、生成されたリードペアは、モデルに対して個別に順位付けされ、（ｉｉｉ）モデルとの比較に基づいて、リードペアにスコアが与えられ、（ｉｖ）高いスコアを有するリードペアは、モデルに最も適合するリードペアであり、そして（ｖ）モデルは、スキャフォールド内のコンティグの、より可能性のある配置を作成するスコアを生成するために、変更される。

【０１０８】

場合によっては、１つのウィンドウ（ｗ）は一度に１、２、３、４、またはそれよりも多くのコンティグを確認し、かつ重複するウィンドウのみ一度に表示されるように、回帰

アルゴリズムが使用される。アルゴリズムの回帰的な性質の結果、ウィンドウは、ウィンドウが終点に進入するかスキャフォールド内の開始地点に戻るまで、スキャフォールド上のまたは他のグループ分けされた核酸データセット上のコンティグを通って徐々に進行する。

【０１０９】

多くのプロトコルによって生成されたペアエンドは、本明細書の分析およびコンピュータにより実施されるシステムと一致する。いくつかの実施形態では、当技術分野で知られているＤＮＡ抽出の方法を含むＤＮＡ抽出プロトコルが使用される。いくつかの実施形態では、市販のＤＮＡ抽出キットが使用される。模範的な市販のＤＮＡ抽出キットには、Ｑｉａｇｅｎ　ＢｌｏｏｄおよびＣｅｌｌ　Ｍｉｄｉ　ｋｉｔｓが含まれる。いくつかの実施形態において、ＤＮＡ抽出のための開始組織は、ヒト由来の体液または組織である。いくつかの実施形態では、ＤＮＡ抽出のためのサンプルは、非ヒト動物、植物または真菌由来である。いくつかの実施形態では、ＤＮＡのソースは微生物またはウイルス由来である。

【０１１０】

いくつかの実施形態において、細胞核は、細胞溶解および遠心分離プロトコル、または当技術分野で知られる任意の他の細胞単離方法を用いて、細胞から単離される。いくつかの実施形態では、非ＤＮＡ細胞成分を分解するために、核を酵素で消化する。模範的な酵素は、プロテイナーゼＫおよびＲＮＡｓｅＡが挙げられる。いくつかの実施形態において、細胞またはウイルスＤＮＡは、当技術分野で良く知られている方法を用いて精製および単離される。模範的なキットは、ＤＮＡを洗浄し、溶出し、イソプロパノール中で沈殿させ、遠心分離によってペレット状にするＱｉａｇｅｎゲノムカラムである。乾燥後、ペレットを２００μＬ　ＴＥ（Ｑｉａｇｅｎ）内で再懸濁する。代替的な実施形態では、核酸は、全細胞から、または複数の細胞型もしくは複数のソース由来の核酸を含むサンプルから単離される。いくつかの実施形態において、ＤＮＡは新規に合成される。

【０１１１】

場合によっては、二本鎖切断中に物理的結合情報を保存するように、ペアエンドが、再アセンブルされた染色質の切断されたクラスターの再連結によって生成される。染色質は、当技術分野で知られている任意の方法を用いて、精製ＤＮＡと共に再アセンブルされる。例えば染色質は、Ａｃｔｉｖｅ　Ｍｏｔｉｆ　ｉｎ　ｖｉｔｒｏ　Ｃｈｒｏｍａｔｉｎ　Ａｓｓｅｍｂｌｙ　ｋｉｔを使用して、ゲノムＤＮＡから２７℃で一晩にわたりアセンブルされる。さらなる実施形態では、成功した染色質アセンブリを確認するために、インキュベーション後に試験を行う。試験の一例は、成功した染色質アセンブリを確認するためにＭＮａｓｅ消化のためのサンプルの１０％を使用することである。「再アセンブルされた染色質」は非常に広く使用され、ヒストン、ヌクレオソーム、または、例えば転写因子、ＤＮＡ結合タンパク質、トランスポゼース、もしくは核酸結合に関与する他の核タンパク質などの他の核タンパク質のような生物学的な染色質成分の再構成物、および核酸分子への非タンパク質ナノ粒子の添加によって生成されるような人工的に再構成された染色質の両方を指す。

【０１１２】

染色質は、幾つかの場合において選別試薬への染色質の結合および結合した対の分離を促進する試薬で標識される。一例では、標識試薬はビオチンである。より具体的な例では、染色質は、ヨードアセチル－ＰＥＧ－２－ビオチン（ＩＰＢ）でビオチン化される。幾つかの実施形態では、ＤＮＡおよび染色質の複合体は、固定試薬で固定される。幾つかの場合では、ＤＮＡ－染色質複合体は、１５分間室温（ＲＴ）で１％のホルムアルデヒド中で固定され、続いて、２．５Ｍのグリシンの２倍のモル過剰量でクエンチされる。幾つかの実施形態では、ＤＮＡ－染色質複合体は、ＤＮＡ断片を作り出す反応にさらされる。幾つかの実施形態では、ＤＮＡ－染色質複合体は、制限酵素での消化にさらされる。幾つかの場合では、ＤＮＡは、ＭｂｏＩまたはＭｌｕＣＩのいずれかで消化される。

【０１１３】

非結合のストレプトアビジン部位は、随意に、ＲＴで１５分間遊離ビオチンの存在下でビーズをインキュベートすることよって占められる。幾つかの実施形態では、ｄＮＴＰ、例えば、－Ｓ－ｄＧＴＰおよびビオチン化されたｄＣＴＰでインキュベートすることによって、粘着末端が充填される。幾つかの実施形態では、ｄＮＴＰ充填（ｆｉｌｌ－ｉｎ）によって生成された平滑末端を連結するために、連結工程が実行される。幾つかの実施形態では、ＤＮＡは、ビオチン化された遊離末端を除去するためにエキソヌクレアーゼで消化される。幾つかの場合では、エキソヌクレアーゼは、エキソヌクレアーゼＩＩＩ（＃Ｍ０２０６Ｓ、ＮＥＢ）である。

【０１１４】

幾つかの実施形態では、ＤＮＡは剪断力にさらされる。さらなる実施形態では、剪断されたＤＮＡは、クレノウポリメラーゼおよびＴ４ ＰＮＫとともに充填される。幾つかの実施形態では、充填反応に続いて、ＤＮＡは、例えば、プルダウンされた反応によって濃縮される。

【０１１５】

幾つかの実施形態では、制限酵素の消化認識部位（ｃｉｔｅｓ）の選択によって生成された接合部によって、配列リードが定義される。例えば、ＭｂｏＩおよびＭｌｕＣＩが使用される場合、配列リードは、接合部が存在するときはいつでも切り詰められる（ＭｂｏＩに対してＧＡＴＣＧＡＴＣ、ＭｌｕＣＩに対してＡＡＴＴＡＡＴＴ）。幾つかの実施形態では、リードは、その後、順方向および逆方向のリードを独立してアライメントさせる－ｘオプションとともにＳＭＡＬＴ［ｈｔｔｐ：／／ｗｗｗ．ｓａｎｇｅｒ．ａｃ．ｕｋ／ｒｅｓｏｕｒｃｅｓ／ｓｏｆｔｗａｒｅ／ｓｍａｌｔ／］を使用してアライメントされる。幾つかの実施形態では、Ｐｉｃａｒｄ－ｔｏｏｌｓ ＭａｒｋＤｕｐｌｉｃａｔｅｓ［ｈｔｔｐ：／／ｂｒｏａｄｉｎｓｔｉｔｕｔｅ．ｇｉｔｈｕｂ．ｉｏ／ｐｉｃａｒｄ／］を使用して、ＰＣＲ複製物がマークされる。幾つかの実施形態では、非重複のリードペアは、両方のリードが、マッピングされ、１０を超えるマッピング品質を有する分析に使用される。

【０１１６】

＜スキャフォールディング － 入力の前処理＞

例えばコンピュータにより実施されるシステム上での入力の前処理は、随意に、ウィンドウ分析前に又はそれとは無関係に利用される。アセンブリまたはコンティグセットまたはスキャフォールドセットの高反復領域にマッピングされるペアリードは、さらなる分析から除去され、それによって、一意にマッピングする配列を含むようにリードペア集団をきれいにする（ｃｌｅａｎ）。幾つかの場合では、アセンブリへの全ゲノムショットガンリードのアライメントが、これらの領域を特定するために使用される。代替的に又は加えて、閾値を超えるマッピングされたショットガンリード深さを有している開始アセンブリの区間にマッピングされるリードペアは除外される。幾つかの領域がリードペアのための「ホットスポット」であること、およびそのようなホットスポット領域からのペアデータの包含が、ダウンストリーム分析を偏らせ（ｂｉａｓ）得ることが観察されている。幾つかの場合では、２閾値法が使用され、その結果、区間が「トリガー」ｔ２を超えるマッピングされたリード深さを有する少なくとも１つの塩基を含む場合、「カットオフ」ｔ１を超える深さを有するその区間にマッピングされるリードがすべて除外される。幾つかの場合では、二重閾値戦略が使用され、それによって、ｔ１またはｔ２を超えるマッピングされたリード深さを有する少なくとも１つの塩基を含む、ｔ１を超えるマッピングされたショットガンリード深さを有する開始アセンブリの区間はすべて、特定され、除外される。幾つかの典型的な実施形態では、アセンブリの約０．５％がマスクされるように、またはアセンブリの０．０５、０．１、０．２、０．３、０．４、０．５、０．６、０．７、０．８、０．９、１．０、１．１、１．２、１．３、１．４、１．５、２、３、４、５、６、７、８、９、１０、１５、２０、３０、４０、５０、６０、７０、８０％、またはそれ以上がマスクされるように、ｔ１およびｔ２が選択される。幾つかの実施形態では、アセンブリの約０．５％がマスクされるように、ｔ１およびｔ２が設定される。幾つかの場合

では、マスクされるアセンブリのパーセントは、反復するアセンブリの割合による影響を受ける。幾つかの実施形態では、マッピングされたショットガンリードの深さに関する閾値のセットは、ｔ１が３ｘであり、ｔ２が３．５ｘであるものであり、ここでｘは、深さの分布全体の平均と等しい。例えば、特定のヒトのアセンブリ（ｈｕｍａｎ　ａｓｓｅｍｂｌｙ）の場合には、ｔ１が８７であり、ｔ２が１０２であり、１０２の閾値が達成される場合、８７以上の深さに対するヒット（ｈｉｔｓ）を有している領域がマスクされる。本例では、「ｘ」は２９であり、３ｘは８７であり、および１０２は３．５ｘである。幾つかの実施形態では、ｔ１は、２ｘ、２．０ｘ、２．１ｘ、２．２ｘ、２．３ｘ、２．４ｘ、２．５ｘ、２．６ｘ、２．７ｘ、２．８ｘ、２．９ｘ、３．０ｘ、３．１ｘ、３．２ｘ、３．３ｘ、３．４ｘ、３．５ｘ、３．６ｘ、３．７ｘ、３．８ｘ、３．９ｘ、４．０ｘ未満、または４．０ｘ以上から選択される。幾つかの実施形態では、ｔ２は、２ｘ、２．０ｘ、２．１ｘ、２．２ｘ、２．３ｘ、２．４ｘ、２．５ｘ、２．６ｘ、２．７ｘ、２．８ｘ、２．９ｘ、３．０ｘ、３．１ｘ、３．２ｘ、３．３ｘ、３．４ｘ、３．５ｘ、３．６ｘ、３．７ｘ、３．８ｘ、３．９ｘ、４．０ｘ、４．１ｘ、４．２ｘ、４．３ｘ、４．４ｘ、４．５ｘ、４．６ｘ、４．７ｘ、４．８ｘ、４．９ｘ、５．０ｘ未満、または５．０ｘ以上から選択される。

【０１１７】

幾つかの場合では、リードペアは、少なくとも２つのリードペアリンクによって４つを超える他の入力コンティグ（ｉｎｐｕｔ　ｃｏｎｔｉｇｓ）に結合されるアセンブリ上の１Ｋｂｐのウィンドウ内に該当する場合、除外される。幾つかの場合では、除外ウィンドウは、１００ｂｐ、１００ｂｐ、２００ｐｂ、３００ｐｂ、４００ｐｂ、５００ｐｂ、６００ｐｂ、７００ｐｂ、８００ｐｂ、９００ｐｂ、１ｋｂ、１．１ｋｂ、１．２ｋｂ、１．３ｋｂ、１．４ｋｂ、１．５ｋｂ、２ｋｂ、３ｋｂ、４ｋｂ　５ｋｂ未満、または５ｋｂ以上である。幾つかの場合では、除外ウィンドウは、領域が２、３、４、５、６、７、またはそれ以上の他の入力コンティグに結合される場合に引き起こされる。

【０１１８】

＜スキャフォールディング　－　尤度モデルパラメータの推測＞

幾つかの場合では、尤度モデルパラメータは、入力処理前に推測される。本明細書に開示される方法およびシステムの幾つかの工程は、アセンブリ決定を誘導するために又はスキャフォールディング内のコンティグ構成を最適化するために、リードペアデータの尤度モデルを使用することができる。幾つかの実施形態では、尤度は、アセンブリ決定を誘導するかために又はスキャフォールディング内のコンティグ構成を最適化するために使用される。幾つかの場合では、尤度関数は以下である：

【０１１９】

【数３】

$$L(l1, l2, g, o) = \frac{N!}{(N-n)!}(1-p_o)^{N-n}\prod_{i=1}^{n} f(d_i)$$

（式１）

コンティグが、相対配向ｏ，

【０１２０】

【数４】

$$o \in ++, +-, -+, --$$

を有し、長さｇの間隙によって分離されることを想定すると、この尤度関数は、数ｎを観察する確率およびコンティグ１と２との間のリードペアｄｉの暗示される分離（ｉｍｐｌｉｅｄ　ｓｅｐａｒａｔｉｏｎｓ）を与える。関数ｆ（ｘ）は、リードペアのゲノミクスの分離距離にわたる正常化された確率分布であり、核酸サンプルを独立してサンプリングする「ノイズ」ペアからの寄与度を有すると想定される。ｆ（ｘ）＝ｐ$_{n}$／Ｇ＋（１－ｐ$_{n}$）ｆ’（ｘ）であり、ｆ’（ｘ）（式２）は、指数分布の合計として表わされる。これ

らのパラメータの少なくとも１つの評価を含む代替関数が、本開示と一致するものとして熟考される。

【０１２１】

幾つかの実施形態では、利用可能な開始アセンブリの近接性（ｃｏｎｔｉｇｕｉｔｙ）が限定されるときにＮ、ｐ$_n$、Ｇ、およびｆ’（ｘ）のロバスト推定値を得るために、産物Ｎ　ｐ$_n$の推定値、密度の最高および最低の１％を除く、コンティグペアのサンプルのための（ｎ／ｌ$_1$ｌ$_2$として定義された）リンクの密度を集計することによる「ノイズ」ペアの総数、およびＧの値としての入力コンティグの長さの合計を使用する、セッティング

【０１２２】

【数５】

$$N_n = G^2 \sum n_{ij} / \sum l_i l_j$$

（式３）が判定される。

幾つかの場合では、残りのパラメータは、

【０１２３】

【数６】

$$G(\sum_{i=1}^{N_c} min(0, l_i - x))^{-1}$$

（式４）の倍数補正因子を分離ｘでの平滑化数（ｓｍｏｏｔｈｅｄ　ｃｏｕｎｔｓ）に適用した後に、開始アセンブリコンティグにマッピングされたリードペアの観察された分離のヒストグラムに対する最小自乗によってＮ　ｆ（ｘ）において固定される。本明細書の概念の代替的な方程式および数学的表現は、本明細書の方法およびシステムの幾つかの実施形態と一致して熟考される。

＜スキャフォールディング　－　メタゲノムの尤度モデルパラメータ＞

幾つかの典型的な実施形態では、入力データは、複数のソース（例えばメタゲノムのライブラリ）からの核酸の混合物を含むサンプルから得られる。幾つかのそのような状況では、本明細書に開示される幾つかの方法および計算システムにおいて使用される尤度モデルは、混合された核酸サンプルかメタゲノムのライブラリに由来する入力データを構成する（ａｃｃｏｕｎｔ　ｆｏｒ）ように修正される。幾つかの場合では、尤度スコアが、それぞれの長さ（例えばｌ$_1$およびｌ$_2$）およびカウント（例えばｓ$_1$およびｓ$_2$）を有する２つのフラグメントのために計算されていると想定される。幾つかの場合では、カウントは、フラグメントの長さおよび入力された混合した核酸サンプル中のその相対的存在量の産物にほぼ比例する量である。限定しない例として、ｓ$_1$およびｓ$_2$は、適切な配列決定ライブラリから各コンティグにマッピングされるリードの数であり得る。幾つかの実施形態では、尤度スコアは、ノイズリードおよびリードペアの予測される数を構成するように修正される。幾つかの場合では、尤度スコア

【０１２４】

【数７】

$$S = \sum_i \ln \frac{f(x_i)}{f(\infty)} - (\bar{n} - \bar{n}_0)$$

（式５）は、ノイズリード

【０１２５】

【数８】

$$\bar{n}_o = Np\frac{l_1 l_2}{G^2} \rightarrow Np\frac{s_1 s_2}{S_T^2}$$

（式６）およびリードペア

【０１２６】

【数９】

$$\bar{n} = N\frac{1}{G}\int_A f(x)dxdy$$

（式７）の予測される数のために修正される。幾つかの場合では、リードペアの有効な総数が計算される。幾つかの場合では、リードペアの有効な総数は、

【０１２７】

【数１０】

$$N \to N\left(\frac{s_1+s_2}{l_1+l_2}\right)\left(\frac{G}{S_T}\right)$$

（式８）である。幾つかの場合では、連結される２つのフラグメントにマッピングされたリードペアの密度間の不一致が増加するにつれ尤度スコアを減少させるスコアペナルティーなどの、スコアペナルティーが適用される。限定しない例として、ペナルティーは以下のように計算される：

【０１２８】

【数１１】

$$\Delta \ln \frac{L(H_1)}{L(H_0)} = \frac{Binom(s_1|s_1+s_2, \frac{l_1}{l_1+l_2})}{Poisson(s_1|\lambda=s_1)Poisson(s_2|\lambda=s_2)}$$

（式９）

【０１２９】

＜スキャフォールディング － 入力コンティグにおける低支持の連結部（ｌｏｗ－ｓｕｐｐｏｒｔ　ｊｏｉｎｓ）の切断＞

コンティグ、スキャフォールドまたはアセンブリの情報が部分的に分解されるように、例えば、比較的弱いアセンブリ決定がダウンストリーム分析において永続化されないように、入力アセンブリデータが随意に処理される。例えば、開始アセンブリにおいて候補の誤連結部（ｍｉｓｊｏｉｎｓ）を特定し、切断するために、尤度モデルを使用して、開始アセンブリにおいて各コンティグの各位置ｉの左側および右側を連結することによって獲得された対数尤度変化を計算する（例えば、位置ｉでの切断によって作成される２つのコンティグに対する対数尤度比（ＬＬＲ）L_i＝ｌｎ Ｌ（ｇ＝０）＝Ｌ（ｇ＝∞））。幾つかの実施形態では、この支持が、入力コンティグの最大の内部セグメントにわたって閾値t_b以下にあるときに、セグメントは「低支持」セグメントとして定義される。幾つかの実施形態では、例えば互いに３００ｂｐ以内に位置する、および例えばコンティグエンドの１Ｋｂｐ以内の低支持セグメントを除く低支持セグメントを融合した後に、セグメントのサイズに依存して、追加の修正が行われる。例えば、１０００ｂｐより下のセグメントに対しては、コンティグにおいてセグメントの中点で切断が導入され、セグメントが１０００ｂｐより長い場合は、セグメントの各エンドに切断が導入される。幾つかの場合では、１００ｂｐ、１００ｂｐ、２００ｐｂ、３００ｐｂ、４００ｐｂ、５００ｐｂ、６００ｐｂ、７００ｐｂ、８００ｐｂ、９００ｐｂ、１ｋｂ、１．１ｋｂ、１．２ｋｂ、１．３ｋｂ、１．４ｋｂ、１．５ｋｂ、２ｋｂ、３ｋｂ、４ｋｂ、５ｋｂより長い、または５ｋｂ以上である場合、切断はセグメントの各エンドに導入される。

【０１３０】

＜スキャフォールディング － コンティグ間を結合するグラフ構築＞

アセンブリのプロセスの間に、生成された結合するデータは、開始アセンブリの（破損された）コンティグがノードであり、エッジが整数の順序対のリスト（その各々は、マッピングされたペアからのリードの２つのコンティグにおける位置を表わしている）で標識されるグラフとして随意に表わされる。幾つかの実施形態では、スキャフォールディングの初期工程は、生成されたリンクt_Lの閾値数より少ない閾値数を有するエッジを除外す

ることによってグラフを連結成分へと分割することにより作成されたデータのサブセット上で平行して実行され、ここで最低の整数閾値は、入力コンティグの５％以上を含む連結成分を生じさせなかった。特定のデータセットか分析システムの詳細に依存して、閾値t_Lは、入力コンティグの０．５、０．６、０．７、０．８、０．９、１．０、１．１、１．２、１．３、１．４、１．５、２、３、４、５、６、７、８、９、１０、１５、２０％未満またはおよそのそれらの％未満、あるいは２０％以上を除外するように選択される。

【０１３１】

＜スキャフォールディング － スキャフォールド構築のシード＞

幾つかの実施形態では、スキャフォールド構築の反復位相は、コンティグ間のグラフのエッジをフィルター処理し、それを高信頼度の線形のサブグラフへと分解することによってシード値が与えられる（ｓｅｅｄｅｄ）。幾つかの場合では、コンティグ間のエッジがフィルター処理され、フィルター処理されたグラフの最小スパニングフォレストが発見される（下記の「エッジフィルタリング」を参照）。幾つかの場合では、グラフは、程度１のノードを除去し、続いて２を超える程度のノードを除去する、３回連続のラウンドによって直線化される。幾つかの場合では、結果として生じるグラフの連結成分の各々は、線形トポロジーを有し、入力コンティグのサブセットの順序付けを定義した。幾つかの場合では、初期のスキャフォールドの生成に結果的に達する工程は、各直線成分に対するコンティグ配向の最大尤度の選択を見つけることである。幾つかの実施形態では、グラフは、ノード除去の１、２、４、５、６回、または６回以上の連続のラウンドによって直線化される。幾つかの実施形態では、除去されたノードの程度は様々である。幾つかの場合では、最大尤度の選択は、最大尤度を判定するために特定の方程式を使用して計算される。幾つかの場合では、最大尤度は、最も可能性の高い順序、配向、または順序及び配向の一般的な評価である。

【０１３２】

＜スキャフォールディング － エッジフィルタリング＞

フィルターは、直線化の前にコンティグ間のグラフのエッジに随意に適用される。典型的なフィルターは以下を含む：t_L未満のリンクを有するエッジの除外および「無差別の（ｐｒｏｍｉｓｃｕｏｕｓ）」コンティグからのエッジの除外。「無差別の」コンティグは、塩基対におけるコンティグ長さに対応するノードのグラフにおける程度の比率がt_pを超えるものとして特定されるか、あるいはフィルター（１）をd_mを超える他のコンティグに通すリンクを有する。閾値t_pおよびd_mは、幾つかの場合において対応する値の分布の上部の尾部（ｔａｉｌ）のおよそ５％を除外するように選択される。幾つかの場合では、閾値t_pおよびd_mは、対応する値の分布の上部の尾部のおよそ０．５、０．６、０．７、０．８、０．９、１．０、１．１、１．２、１．３、１．４、１．５、２、３、４、５、６、７、８、９、１０、１５、２０％未満、または２０％以上を除外するために選択された。幾つかの実施形態では、t_Lは、７より下、７、８、９、１０、１１、１２、１３、１４から１５まで、または１５以上の範囲である。幾つかの実施形態では、t_pは、０．０１、０．０２、０．０３、０．０４、０．０５、０．０６、０．０７、０．０８、０．０９から、０．１までの範囲である。幾つかの実施形態では、d_mは、５未満、５、６、７、８、９、１０、１１、１２、１３、１４、１５、または１５以上の範囲である。幾つかの実施形態では、t_L、t_p、およびd_mが、それぞれ１１、０．０４、および１０であるときに、改善されたパフォーマンスが得られることが分かる。

【０１３３】

＜スキャフォールディング － コンティグの配向＞

各入力スキャフォールドは、フォワードおよびリバースの塩基配列、または「Ｗａｔｓｏｎ」および「Ｃｒｉｃｋ」のＤＮＡ鎖に対応する、最終アセンブリにおける２つの配向のうちの１つを有する。本明細書で開示されるように、各線状紐部におけるスキャフォールドに最適な配向は、以下に示されるような再帰関連性を使用して動的プログラミングなどの分析によって見つけられる。長さｎのスキャフォールドの順序付きリストにおいて、スキャフォールドｉ－ｋからｉが特定の配向

【０１３４】
【数１２】

$o_{i-k}, o_{i-k+1}, \ldots o_i$

を有するような、最大でスキャフォールドｉまでのスキャフォールドに対する配向選択の最も高いスコアリング配列のスコアは、以下によって与えられる：
【０１３５】
【数１３】

$$S_m(i, o_{i-k}, o_{i-k+1}, \ldots, o_i) = \max_{o_{i-1-k} \in \{+,-\}} (S_m(i-1, o_{i-1-k}, o_{i-k}, \ldots, o_{i-1}) + \sum_{j=i-1-k}^{j=i-1} \log p(o_j, o_i))$$

（式１０）
【０１３６】
随意に、コンティグｋから戻る（ｓｔｅｐｓ ｂａｃｋ）リンクを含むことによって、配向精度の改善がもたらされる。幾つかの場合では、上に議論されるように、分枝したコンティグの順序付けを結果としてもたらす状況に類似した状況において、小さなインターカレートされたスキャフォールドが、片面上に結合と、故に配向情報を有するだけであるかもしれず、隣接するスキャフォールドに関する重要な配向情報が、それを飛び越えるリンクから送られるため、配向精度が改善される。
【０１３７】
＜スキャフォールディング － 成分内のスキャフォールディングの融合＞
コンティグエンドは、スキャフォールドにおけるそれらの位置と相対的な随意に指定された分類である。例えば、コンティグエンドは、スキャフォールドのエンドに位置する場合には「自由」として分類され、スキャフォールドの内部にある場合には「埋設された（ｂｕｒｉｅｄ）」として分類される。幾つかの実施形態では、各連結成分内のすべてのコンティグエンドのペアに関して、それらを連結するためのＬＬＲスコアは、g_0の「標準」間隙寸法で計算される。幾つかの実施形態では、候補連結部は、スコアの降順で分類され、１セットの基準に従って評価される。典型的なセットの基準が続く。両エンドが、自由であり、異なるスキャフォールドからのものである場合、端と端との２つのスキャフォールドの結合を試験する。片側のエンドが埋設されて他エンドが自由であり、両エンドが異なるスキャフォールドからのものである場合、埋設されたエンドに隣接した間隙への自由エンドのスキャフォールドの挿入を試験する。片側のエンドまたは両エンドが埋設されて、両エンドが同じスキャフォールド上にある場合、両エンド間のスキャフォールド部分の逆位を試験する。両エンドが埋設されて、異なるスキャフォールドからのものである場合、端と端でスキャフォールドを連結するすべての４つの方法を試験する。幾つかの実施形態では、すべての場合に関して、考えられ得る連結部、挿入部および反転部が、変化による影響を受けたすべてのコンティグのペア間のＬＬＲスコアを合計することによってＬＬＲスコアの総変化を計算することにより試験される。変更がＬＬＲスコアを増大させた場合、最適な移動が許容される。
【０１３８】
＜スキャフォールディング － 局所的な順序、配向および間隙寸法の改善＞
各スキャフォールドにおけるコンティグの局所的な順序付けおよび配向の両方を改善するために、各スキャフォールドの順序付けされ、配向されたコンティグにわたってサイズｗのウィンドウをスライドさせる、動的プログラミングアルゴリズムが随意に適用される。各位置ｉでは、ウィンドウ内でコンティグを順序付けし配向するすべてのｗ！２^{w}の方法が考慮され、ウィンドウにおいてコンティグの現行の構成で終端する現行のウィンドウ位置のエンドまでのすべてのコンティグの最適な順序付け及び配向を表わすスコアが保存された。位置
【０１３９】

【数１４】
$i\text{-}1, i\text{-}2, \ldots i\text{-}w$

でのウィンドウにおけるすべての「適合性のある」順序および配向のスコア、および現行の構成でそれらの順序付けの拡張のスコアが使用された。ｗ！２^{w}がそのような急勾配関数であるため、方法は、実施時にｗの小さな値に一般に限定される。幾つかの実施形態では、ｗは、２、３、４、５、６、７、８、９、１０、またはそれ以上である。幾つかの典型的なデータに関する試験において、ｗ＝３は、幾つかのデータセットに対する構成精度を劇的に改善する。幾つかの実施形態では、リードペアの結合の分離の確率モデルを与えられたコンティグの構成を改善する方法が提供される。そのような方法は、コンティグの初期の順序付けにわたってサイズｗのウィンドウをスライドさせる動的プログラミングアルゴリズムである。

【０１４０】
＜スキャフォールディング － 繰り返し連結＞
初期のスキャフォールドが各連結成分内に構築された後、結果として生じるスキャフォールドは、単一のプールに戻され、端と端との及びインターカレートするスキャフォールド連結の複数のラウンドが実行される。各ラウンドでは、すべてのスキャフォールドのペアが比較され、尤度スコアが、端と端との及びインターカレートする連結のために平行して計算される。その後、候補連結部は分類され、矛盾しない連結部は、尤度スコアの増加の降順で許容される。

【０１４１】
＜分割の利点＞
本明細書において、高スループットの配列決定データからの新規のアセンブルされたコンティグのスキャフォールディングを劇的に改善することができる、長距離のメイトペアデータのインビトロでの生成のための方法およびコンピュータにより実施されるシステムが記載される。これらのアプローチは、既存の方法以上の幾つかの利点を有する。

【０１４２】
第１に、データライブラリ構築は、生きた生体物質、例えば、一次または形質転換された組織培養物、あるいは生体を必要としない。本明細書に記載されるライブラリは、１０、９、８、７、６、５、４、３、２、１マイクログラム、または１マイクログラム未満などの、わずか５．０マイクログラムまたはそれ以下のインプットＤＮＡから生成される。さらに、インビトロでの染色質再構成は、ヒトのヒストンおよび染色質のアセンブリ因子に基づくが、種々様々な植物、動物、および微生物からのＤＮＡは、記載されるプロトコルを使用してインビトロでの染色質アセンブリ用の基質になり得る。

【０１４３】
第２に、データが、インビボでのソースで得られた染色質よりもむしろインビトロでアセンブルされた染色質の近接連結から生成されるため、アセンブリを混乱させる可能性のある交絡生体信号がない。Ｈｉ－Ｃ及び／又はインビボでの染色質から生成された他の近接連結データは、その内で生物学的関連性のある長距離の近接情報を搬送するが、持続性であり、ゲノムまたはスキャフォールドのアセンブリにとって交絡である可能性がある。幾つかの実施形態では、本明細書に提供される方法は、結果としてノイズの低バックグラウンド速度および持続性且つスプリアスのリードペアの事実上の欠如をもたらす。

【０１４４】
第３に、インビボでのＨｉ－Ｃ方法とは対照的に、生成されたリードペアの最大分離は、インプットＤＮＡの分子量によってのみ限定される。これによって、短いフラグメントのＩｌｌｕｍｉｎａ配列＋生成されたライブラリだけを使用する脊椎動物の核酸スキャフォールドまたはゲノムの高度に隣接するスキャフォールディングの生成が可能になった。

【０１４５】
第４に、これらのライブラリは、長距離の「メイトペア」とフォスミドライブラリの組み合わせを作り出す及び配列決定する必要性を除去し、そのようなライブラリを作り出す

のに通常必要とされる高分子量ＤＮＡを剪断する又はサイズ選択するための高価で専門的な機器の使用を必要としない。

【０１４６】

したがって、既存の方法よりかなり長距離ののコンティグアセンブリのスキャフォールドを生成するバイオインフォマティクス法を使用する、ＤＮＡライブラリの構築方法および計算システムが提供される。幾つかの実施形態では、本明細書に提供されるＤＮＡライブラリの構築方法はまた、核酸サンプルまたはゲノムの変更を特定するために提供される。しばしば、本明細書に提供されるＤＮＡライブラリの構築方法および計算システムは、低い労力およびコストで全長のハプロタイプ分解された染色体配列の正確な再構成をもたらす。

【０１４７】

順序付け及び配向精度の両方を改善するために、順序付け及び配向の問題は、順序付け及び配向の工程を統合することによって解決される。一例では、初期のグラフが構築され、このグラフにおいて、ノードはコンティグエンドであり、各コンティグの２つのエンドノードはエッジによって連結される。具体的な短い間隙寸法の仮定の下でのコンティグ間のエッジの対数尤度比スコアは、計算され、続いて分類された。エッジスコアの降順でリストを当たると、新しいエッジが、アセンブリの合計スコアを増加または減少させるかによって、受理されるか又は拒絶される。既存のスキャフォールドの間隙へのコンティグのインターカレーションを含蓄するエッジの許容が、間隙の両側の結合されたコンティグのペア間の間隙寸法を増大させるため、陽性スコアを有するエッジでさえアセンブリにおいてコンティグのスコアの合計を減少させ得、これが、コンティグに対してより低いスコアを与える可能性があることが留意される。

【０１４８】

さらに、最大尤度の間隙寸法の計算の効率が改善される。報告されたアセンブリの精度は、連続するコンティグ間の未知の配列の長さを推定することによって向上される。ライブラリのリードペア間の分離ｄに関するモデル確率密度関数（ＰＤＦ）を含むライブラリ生成プロセスのモデルを考慮すると、最大尤度の間隙長さは、例えば、間隙に及ぶペアの分離ｄｉの結合尤度を最大限にすることによって見つけられる。微分可能なモデルＰＤＦに関して、効率的な反復最適化方法（例えばＮｅｗｔｏｎ－Ｒａｐｈｓｏｎ）が使用される。

【０１４９】

＜成功を特定するためのパラメータ＞

本明細書における多くの方法および計算システムは、少なくとも２つの順序、配向、スキャフォールド接続状態、コンティグ切断評価、または他の考えられ得る配列の再位置決め（総じて「スキャフォールディング」）の評価を含む。開始または以前のデータに関する改善を表わすスキャフォールディングを数的に評価するために、多くのアプローチが利用可能である。典型的な実施形態では、コンティグまたはスキャフォールドの構成は、（例えば、同じコンティグにマッピングされる両エンドを有するリードペアから）予期されたか、独立して判定されたか、または同時に判定された曲線により近く近似するリードペア分離距離の分布曲線を生成する場合に好ましい。曲線は、例えば、上の式１を使用して評価において有利にスコアを得る場合、予期されたか、独立して判定されたか、または同時に判定された曲線により近く近似する。分散分析（ＡＮＯＶＡ）試験、共変動の評価、または他の試験などの、代替的評価が、当業者にとって利用可能である。

【０１５０】

交互に又は組み合わせて、全体的なスキャフォールディング有効性のための別々のスコアあるいはメトリックスが使用される。幾つかの尺度は、既知の核酸アセンブリに対するパーセントアライメントを含み、その結果、パーセントアライメントの改善につながるスキャフォールディングが好まれる。幾つかの場合では、改善は、少なくとも５０％、６０％、７０％、７５％、８０％、８５％、９０％、９５％、９９％、９９．５％、９９．９％、または９９．９％以上のパーセントアライメントにつながる。

【０１５１】
別の尺度は、Ｎ５０などの全体的なスキャフォールド集団に対するスキャフォールドの効果である。すなわち、幾つかの場合では、スキャフォールディングは、配列データセットに対する結果として生じたＮ５０が増加した場合に好ましい。幾つかの場合では、改善は、少なくとも０．５％、１％　２％　３％、４％、５％、１０％、２０％、３０％、４０％、５０％、６０％、７０％、８０％、９０％、１００％、１５０％、２００％、２５０％、３００％、４００％、５００％、６００％、７００％、８００％、９００％、１０００％、１５００％、２０００％、２５００％、３０００％、４０００％、５０００％、６０００％、７０００％、８０００％、９０００％、１０，０００％、１５，０００％、２０，０００％、２５，０００％、３０，０００％、３５，０００％、４０，０００％、４５，０００％、５０，０００％、５５，０００％、６０，０００％、６５，０００％、７０，０００％、７５，０００％、８０，０００％、８５，０００％、９０，０００％、９５，０００％、１００，０００％、または１００，０００％以上のパーセント増加につながる。
【０１５２】
交互に又は組み合わせて、全体的なスキャフォールド集団に対する効果の追加のメトリックスが使用される。１つのそのようなメトリックスはＲＮ５０測定値である。ＲＮ５０は以下の通りに理解される。１セットの配列Ｓおよび基準配列Ｒに関して、Ｒに関するＳのＲＮ５０または「基準のＮ５０」は、Ｔにおける最短の配列の長さであり、ここでＴは、Ｔにおける配列が、Ｒの長さの少なくとも５０％をカバーするような方法でＲにアライメントされ得るように、Ｓの最小（最小基数）のサブセットである。
【０１５３】
Ｔにおけるすべの配列がＲＮ５０と少なくとも同じくらい大きいため、これは、Ｒのランダムに選択された塩基が、少なくとも長さＲＮ５０であるＴにおける（および故にＳにおける）配列に対するアラインメントによって範囲が及ぶ少なくとも５０％の確率を有することを意味する。
【０１５４】
幾つかの場合では、初期のＲＮ５０は、ゲノム配列などのサンプル配列に、０．５％、１％、２％、３％、４％、５％、１０％、２０％、３０％、４０％、または５０％である値、あるいはこの範囲内の別の数値を有する。幾つかの場合では、最終的なＲＮ５０が、初期のＲＮ５０より大きい値を有する。幾つかの場合では、改善は、少なくとも０．５％、１％、２％、３％、４％、５％、１０％、２０％、３０％、４０％、５０％、６０％、７０％、８０％、９０％、または９０％、１００％以上、２ｘ、５ｘ、１０ｘ、２０ｘ、５０ｘ、１００ｘ、２００ｘ、５００ｘ、１０００ｘ、２０００ｘ、５０００ｘ、１００００ｘ、または１００００ｘ以上のＲＮ５０のパーセント増加につながる。
【０１５５】
本開示と一致するスキャフォールディングの成功の交互のメトリックスが熟考される。
【０１５６】
図面を参照すると、図１のＡ－Ｆは、ＤＮＡライブラリ生成プロトコルの典型的な概要図を例証する。図１のＡでは、染色質（円として描写されるヌクレオソーム）は、インビトロで裸のＤＮＡ（黒鎖（ｂｌａｃｋ　ｓｔｒａｎｄ））上に再構成される。図１のＢでは、染色質は、ホルムアルデヒドで固定される（細線は交差結合である）。固定された染色質は、ストレプトアビジンのビーズ上へと引き下げられ、制限酵素で切断され、自由粘着末端がもたらされる（図１のＣ）。図１のＤでは、粘着末端は、ビオチン化された（より小さな円）およびチオール化された（小さな正方形）ヌクレオチドとともに充填される。図１のＥでは、自由平滑末端は連結される（アスタリスクによって示される連結）。図１のＦでは、交差結合が逆転され、タンパク質が除去されて、ライブラリの断片が得られる。ライブラリの断片は、非連結型の末端からビオチン化されたヌクレオチドを除去するために消化される。ライブラリの断片は、ストレプトアビジンでコーティングしたビーズとともに選択され、その後、配列決定に備えて、アダプターを上に連結させる。

【０１５７】
幾つかの実施形態では、基準アセンブリにマッピングされた幾つかの生成されたライブラリのためのリードペア分離が比較される。例えば、図２は、ｈｇ１９などの、基準ヒトアセンブリにマッピングされた幾つかの生成されたライブラリのためのリードペア分離の例を提供する。当該図では、プロットの右側で下の角括弧によって印されたデータ傾向は、５０Ｋｂｐ入力のヒト配列決定ライブラリに対応している。プロットの右側で下の矢印によって印されたデータ傾向は、１５０Ｋｂｐ入力のヒト配列決定ライブラリに対応している。プロットの右側で中央の矢印によって印されたデータ傾向は、１５０Ｋｂｐ入力のヒト配列決定ライブラリに対応している。プロットの右側で上の矢印によって印されたデータ傾向は、ヒトＨｉ－Ｃライブラリに対応している（Ｋａｌｈｏｒ ｅｔ ａｌ．，２０１２）。暗い垂直線は、代替のメイトペア技術に対する最大の公示された（ａｄｖｅｒｔｉｓｅｄ）または実証された能力を示す。
【０１５８】
推定の隣接する核酸またはゲノムのサイズで割られたリードペア分離の合計が、代替または追加のメトリックスとしての分離範囲によってグループ分けされた様々なリードペアに対して計算される。そのような比較の典型的な概要は、図３に示される。
【０１５９】
図４Ａおよび４Ｂを参照すると、再びアセンブルされた染色質からのＤＮＡの近接連結から生成されたリードペアの基準配列（例えばＧＲＣｈ３８）上のマッピングされた位置の例が提供され、これらはＧＭ１２８７８と基準との間の構造差の近くにプロットされる。生成された各リードペアは、対角線の上および下の両方に表わされる。対角線の上では、陰影は、示されたスケールのマップ品質のスコアを示し、対角線の下では、陰影は、位相合わせされたＳＮＰの重複に基づいた生成されたリードペアの推論されたハプロタイプ相を示す。幾つかの実施形態では、図４Ｂで例証されるように、生成されたプロットは、隣接する反復領域を有する逆位を描写する。幾つかの実施形態では、図４Ｂで例証されるように、生成されたプロットは、位相合わせされたヘテロ接合の欠失に関するデータを描写する。
【０１６０】
図５を参照すると、染色質のリアセンブリおよび処理の例が提供される。幾つかの実施形態では、精製された高分子量ＤＮＡは、（ヒストンおよび染色質のアセンブリ因子を用いて）インビトロでの染色質アセンブリにさらされ、その後、ビオチン化される。幾つかの実施形態では、結果として生じるＤＮＡ－染色質複合体は、その後、固化剤で固定される。図５では、固化剤はホルムアルデヒドである。さらなる実施形態では、ＤＮＡ－染色質複合体は、ストレプトアビジンのビーズを用いて引き下げられる。さらなる実施形態では、ＤＮＡ－染色質複合体は、制限酵素消化で処理される。図５で例証されるように、幾つかの場合では、制限酵素はＭｂｏＩである。幾つかの場合では、制限酵素は、粘着末端を残す、例えば、二本鎖ＤＮＡ中でオーバーハングする酵素である。幾つかの実施形態では、粘着末端は、標識されたヌクレオチドとともに充填される。幾つかの場合では、標識されたヌクレオチドは、ビオチン化または硫酸化される。図５で例証されるように、幾つかの場合では、硫酸化されたｄＧＴＰを用いて内部充填が実行され、ビオチン化されたｄＣＴＰを用いて外部充填が行われる。幾つかの実施形態では、平滑末端の連結工程が実行され、充填末端が連結される。幾つかの実施形態では、ＤＮＡ－染色質複合体は、複合体からＤＮＡを放出するために酵素消化にさらされる。図５で例証されるように、幾つかの実施形態では、酵素はプロテイナーゼＫである。幾つかの実施形態では、ＤＮＡは、標識されたヌクレオチドを除去するために制限酵素で処理される。例えば、図５で例証されるように、ＥｘｏＩＩＩ消化による処理によって、末端上のビオチン化されたシトシンを除去する。幾つかの実施形態では、ＤＮＡ断片は、剪断、プルダウン、およびＩｌｌｕｍｉｎａ互換性のライブラリ生成プロトコルの使用による分析によって調製され、ビーズ上の洗剤および再懸濁の代りに、反応間の正常なクリーンアップ工程が除去される。
【０１６１】

様々な実施形態では、本発明の方法およびシステムは、コンピュータシステムズ上のソフトウエアプログラムおよびそれらの使用をさらに含む。コンピュータシステムズは、ユーザー間でインターフェース接続するようにプログラムされ得る。

【０１６２】

図６で例証されるコンピュータシステム（６００）は、媒体（６１１）、及び／又は固定媒体（６１２）を有するサーバー（６０９）に随意に接続され得るネットワークポート（６０５）からの命令を読み取る論理装置として理解され得る。図６に示されるものなどの、システムは、ＣＰＵ（６０１）、ディスクドライブ（６０３）、キーボード（６１５）及び／又はマウス（６１６）などの随意の入力デバイス、および随意のモニター（６０７）を含むことができる。データ通信は、示された通信媒体を介してローカルまたはリモートの場所でのサーバーへと達成される。通信媒体は、データを送信及び／又は受信する、あらゆるデバイス、装置またはアプローチを含むことができる。例えば、通信媒体は、ネットワーク接続、ワイヤレス接続またはインターネット接続であり得る。そのような接続は、ワールドワイドウェブ上の通信を提供することができる。本開示に関連するデータが、図６に例証されるようにパーティー（６２２）による受信及び／又はレビューのためのそのようなネットワークまたは接続にわたって送信され得ることが予想される。

【０１６３】

図７は、本明細書に記載される実例の実施形態で使用されるコンピュータシステム（７００）の第１の実例のアーキテクチャを例証するブロック図である。図７に描写されるように、実例のコンピュータシステムは、処理命令のためのプロセッサ（７０２）を含む。プロセッサの限定しない例は以下を含む：Ｉｎｔｅｌ ＸｅｏｎTMプロセッサ、ＡＭＤ ＯｐｔｅｒｏｎTMプロセッサ、Ｓａｍｓｕｎｇ ３２－ｂｉｔ ＲＩＳＣ ＡＲＭ １１７６ＪＺ（Ｆ）－Ｓ ｖ１．０TMプロセッサ、ＡＲＭ Ｃｏｒｔｅｘ－Ａ８ Ｓａｍｓｕｎｇ Ｓ５ＰＣ１００TMプロセッサ、ＡＲＭ Ｃｏｒｔｅｘ－Ａ８ Ａｐｐｌｅ Ａ４TMプロセッサ、Ｍａｒｖｅｌｌ ＰＸＡ ９３０TMプロセッサ、または機能的に同等なプロセッサ。実行の複数のスレッドを並列処理に使用することができる。幾つかの実施形態では、複数のコアを備えた複数のプロセッサが、単一のコンピュータシステム中であろうと、クラスター中であろうと、あるいは複数のコンピュータ、携帯電話、及び／又は携帯電話端末装置を含むネットワーク上のシステムにわたって分散されて、使用される。

【０１６４】

図７に例証されるように、高速キャッシュ（７０４）は、プロセッサ（７０２）に接続されるか、またはそこに組み込まれ、プロセッサ（７０２）によって最近または頻繁に使用されている、命令またはデータ用の高速メモリを提供する。プロセッサ（７０２）は、プロセッサバス（７０８）によってノースブリッジ（７０６）に接続される。ノースブリッジ（７０６）は、メモリバス（７１２）によってランダムアクセスメモリ（ＲＡＭ）（７１０）に接続され、プロセッサ（７０２）によってＲＡＭ（７１０）へのアクセスを管理する。ノースブリッジ（７０６）はまた、チップセットバス（７１６）によってサウスブリッジ（７１４）に接続される。サウスブリッジ（７１４）は、順に、周辺バス（７１８）に接続される。周辺バスは、例えばＰＣＩ、ＰＣＩ－Ｘ、ＰＣＩ Ｅｘｐｒｅｓｓ、または他の周辺バスである。ノースブリッジおよびサウスブリッジは、しばしば、プロセッサチップセットと呼ばれ、プロセッサ、ＲＡＭ、および周辺バス（７１８）上の周辺コンポ--ネントの間のデータ転送を管理する。幾つかの代替的なアーキテクチャでは、ノースブリッジの機能は、別々のノースブリッジチップを使用する代わりにプロセッサに組み込まれる。

【０１６５】

幾つかの実施形態では、システム（７００）は、周辺バス（７１８）に付けられたアクセラレータカード（７２２）を含む。アクセラレータは、フィールドプログラマブルゲートアレイ（ＦＰＧＡ）または特定の処理を加速させるための他のハードウェアを含むことができる。例えば、アクセラレータは、幾つかの場合に、適合データ再構成のために、または拡張セットの処理に使用される代数式を評価するために使用される。

【０１６６】
ソフトウェアおよびデータは、外部記憶装置（７２４）に記憶され、プロセッサによる使用のためにＲＡＭ（７１０）及び／又はキャッシュ（７０４）にロードされ得る。システム（２０００）は、システムリソースを管理するためのオペレーティングシステムを含み、オペレーティングシステムの限定しない例は、以下を含む：Ｌｉｎｕｘ（登録商標）、ＷｉｎｄｏｗｓＴＭ、ＭＡＣＯＳＴＭ、ＢｌａｃｋＢｅｒｒｙ ＯＳＴＭ、ｉＯＳＴＭ、および他の機能的に同等なオペレーティングシステムの他に、本発明の実例の実施形態に従ってデータの記憶および最適化を管理するためのオペレーティングシステム上で実行するアプリケーションソフトウェア。
【０１６７】
この例において、システム（７００）はまた、ネットワーク接続ストレージ（ＮＡＳ）、および分散並列処理に使用することができる他のコンピュータシステムなどの、外部記憶装置にネットワークインターフェースを提供するために周辺バスに接続されたネットワークインターフェースカード（ＮＩＣ）（７２０）および（７２１）を含む。
【０１６８】
図８は、複数のコンピュータシステム（８０２ａ）および（８０２ｂ）、複数の携帯電話および携帯電話端末（８００ｃ）、ならびにネットワーク接続ストレージ（ＮＡＳ）（８０４ａ）および（８０４ｂ）を備えたネットワーク（８００）を示す図である。実例の実施形態では、システム（８０２ａ）、（８０２ｂ）、および（８００ｃ）は、データ記憶を管理し、ネットワーク接続ストレージ（ＮＡＳ）（８０４ａ）および（８０４ｂ）に保存されたデータのためのデータアクセスを最適化することができる。数理モデルが、幾つかの場合においてデータに使用され、コンピュータシステム（８０２ａ）および（８０２ｂ）、および携帯電話および携帯電話端末のシステム（８００ｃ）にわたって分散並列処理を使用して評価される。コンピュータシステム（８０２ａ）および（８０２ｂ）、および携帯電話および携帯電話端末のシステム（８００ｃ）はまた、ネットワーク接続ストレージ（ＮＡＳ）（８０４ａ）および（８０４ｂ）に保存されたデータの適合データ再構成のための並列処理を提供することができる。図８は、例のみを例証し、種々様々な他のコンピュータアーキテクチャおよびシステムが、本発明の様々な実施形態に関連して使用され得る。幾つかの例では、並列処理を提供するために、ブレードサーバーが使用される。並列処理を提供するために、プロセッサブレードがバックプレーンを介して接続され得る。記憶装置もバックプレーンに接続され得るか、あるいは別々のネットワークインターフェースを介してネットワーク接続ストレージ（ＮＡＳ）として接続され得る。
【０１６９】
幾つかの例実施形態では、プロセッサは、別々のメモリ空間を維持することができ、ネットワークインターフェース、バックプレーン、または他のプロセッサによる並列処理のための他のコネクターを介してデータを送信することができる。他の実施形態では、プロセッサの幾つか又はすべては、共有仮想アドレスメモリ空間を使用することができる。
【０１７０】
図９は、実例の実施形態に従い共有仮想アドレスメモリ空間を使用する、マルチプロセッサコンピュータシステム（９００）のブロック図である。該システムは、共有メモリサブシステム（９０４）にアクセスすることができる複数のプロセッサ（９０２ａ－ｆ）を含む。該システムは、メモリサブシステム（９０４）に複数のプログラマブルハードウェアメモリのアルゴリズムプロセッサ（ＭＡＰ）（９０６ａ－ｆ）を組み込む。各ＭＡＰ（９０６ａ－ｆ）は、メモリ（９０８ａ－ｆ）および１つ以上のフィールドプログラマブルゲートアレイ（ＦＰＧＡ）（９１０ａ－ｆ）を含むことができる。ＭＡＰは、設定可能な機能ユニットを提供し、特定のアルゴリズムまたはアルゴリズムの部分が、それぞれのプロセッサと密に協働して処理するためのＦＰＧＡ（９１０ａ－ｆ）に提供される。例えば、ＭＡＰは、幾つかの場合において、データモデルに関する代数式を評価するために、および実例の実施形態において適合データ再構成を実行するために使用される。本例において、各ＭＡＰは、これらの目的のためにプロセッサのすべてによって全体的に利用可能で

ある。一構成において、各ＭＡＰは、関連メモリ（９０８ａ－ｆ）にアクセスするためにダイレクトメモリアクセス（ＤＭＡ）を使用することができ、それによって、それぞれのマイクロプロセッサ（９０２ａ－ｆ）とは無関係に及び非同期的にタスクを実行することが可能になる。この構成において、ＭＡＰは、アルゴリズムのパイプライン処理および並列実行のために別のＭＡＰに直接結果を供給することができる。

【０１７１】

上記のコンピュータアーキテクチャおよびシステムは、例のみであり、汎用プロセッサ、コプロセッサ、ＦＰＧＡおよび他のプログラマブルロジックデバイスの任意の組み合わせを使用するシステム、システムオンチップ（ＳＯＣ）、特定用途向け集積回路（ＡＳＩＣ）、および他の処理および論理の素子を含む、種々様々な他のコンピュータ、携帯電話、携帯電話端末のアーキテクチャおよびシステムが、実例の実施形態において使用される。幾つかの実施形態では、コンピュータシステムのすべて又は一部は、ソフトウェアまたはハードウェアにおいて実施される。ランダムアクセスメモリ、ハードドライブ、フラッシュメモリ、テープドライブ、ディスクアレイ、ネットワーク接続ストレージ（ＮＡＳ）、および他のローカルまたは分散型のデータ記憶装置およびシステム含む、あらゆる種類のデータ記憶装置媒体が、実例の実施形態に関連して使用され得る。

【０１７２】

実例の実施形態では、コンピュータシステムは、上記または他のコンピュータアーキテクチャおよびシステムのいずれかで実行するソフトウェアモジュールを使用して実施される。他の実施形態では、システムの機能は、ファームウェア、図９で参照されるようなフィールドプログラマブルゲートアレイ（ＦＰＧＡ）などのプログラマブルロジックデバイス、システムオンチップ（ＳＯＣ）、特定用途向け集積回路（ＡＳＩＣ）、または他の処理および論理の素子において、部分的または完全に実施される。

【０１７３】

本出願の出願時の使用における方法と比べて、本明細書に開示される方法およびシステムは、多くの利点を提供する。

【０１７４】

本明細書に開示される幾つかの方法および計算システムは、生体に対する染色体の数とは無関係な方法でコンティグをクラスター化する。結果として生じるより小さなコンティグのクラスターをスキャフォールドへとアセンブルするために、シングルリンクのクラスタリングのためのコンティグ間のリンク上のより控えめな閾値が適用され、本明細書に開示される様々な方法によって、続くスキャフォールディングの連結が可能になる。

【０１７５】

幾つかの実施形態では、本明細書に開示される方法は、クラスタリングを本質的に含んでいないが、全域木工程に直接関係し、トポロジー的なツリープルーニングが後に続く。幾つかの実施形態では、１つを超えるクラスタリング方法、例えば、Ｍａｒｋｏｖ Ｃｌｕｓｔｅｒ Ａｌｇｏｒｉｔｈｍ（ＭＣＬアルゴリズム）が使用される。理論に限定されることなく、ミスアセンブリは、これらのエッジを特に気を付けて処理し、アセンブリの誤連結を回避することによるトポロジー的なプルーニングによって防ぐことができる。

【０１７６】

スキャフォールドにおいてコンティグの順序を固定した後に、配向は、幾つかの場合において動的プログラミングアルゴリズムの使用によって最適化される。順序付けにおいて隣接するコンティグのペアにマッピングされるそのようなアプローチのみのリードペアは、最適化されているスコアに寄与し、出されてアセンブルされていない（ｏｕｔ ａｎｄ ｕｎａｓｓｅｍｂｌｅｄ）優れたフラグメントペアの最大の分離より短いコンティグは除外される。配向工程を改善するために、最も近隣のコンティグスコアの相互作用に加えて、最も近隣のコンティグスコアの相互作用ではないコンティグは、多くてもｗ－２の干渉コンティグ内でコンティグのペアにマッピングされるすべてのペアからのデータを組み込むアルゴリズムを使用することによって、例えば、２、３、４、５、６、７、８、９、１０または１０以上などの、順序付けにおいて２つ以上のコンティグの値を使用することに

よって、考慮され得る。

【０１７７】

幾つかの実施形態では、インターカレーション工程の精度が改善される。いかなる理論にも縛られることなく、トランクの生成後の優れたリードペア間の最大の分離より短いコンティグを備えたアセンブリにおいて、順序付けに沿ったｗコンティグの近隣内のコンティグからのデータは、トランクからコンティグを除外し、隣接したコンティグ間の結合の量を最大限にする部位でそれに再挿入するときに含まれる。

【０１７８】

配向工程は、幾つかの場合において、最も近隣のコンティグスコアの相互作用以上に考慮することによって改善される。スキャフォールドにコンティグの順序を固定した後に、コンティグ配向は、動的プログラミングアルゴリズムを使用することによって最適化される。順序付けにおいて隣接したコンティグのペアにマッピングされるリードペアだけが、最適化されているスコアに寄与する。幾つかの場合では、順序付けにおいて多くてもｗ－２の干渉コンティグ内でコンティグのペアにマッピングされるすべてのペアからのデータを組み込むアルゴリズムが、優れたフラグメントペアの最大の分離より短いコンティグを備えたアセンブリに使用される。例えば、２、３、４、５、６、７、８、９、１０、または１０以上などの、順序付けにおける２つ以上のコンティグの値を使用する。

【０１７９】

幾つかの実施形態では、順序付け及び配向をさらにより堅く統合することによって順序付け及び配向両方の精度が改善される。ノードがコンティグエンドであり、各コンティグの２つのエンドノードがエッジによって連結される、初期のグラフが構築される。特定の短い間隙寸法を仮定したコンティグ間エッジの対数尤度比スコアが計算され、その後、分類された。エッジスコアの降順でリストを当たると、新しいエッジが、アセンブリの合計スコアを増大させるか又は減少させるかに従って受理または拒絶される。既存のスキャフォールドの間隙へのコンティグのインターカレーションを含蓄するエッジを許容することによって、間隙の両側の結合されたコンティグのペア間の間隙寸法が増大されるため（これはより低いスコアを与える可能性がある）、陽性スコアを有するエッジでさえ、アセンブリにおけるコンティグのシコアの合計を減少させ得ることが留意される。

【０１８０】

さらに、最大尤度の間隙寸法が効率的に計算される。報告されたアセンブリの全体の精度は、連続するコンティグ間の未知の配列の長さを推定することによって増大される。ライブラリのリードペア間の分離ｄに関するモデル確率密度関数（ＰＤＦ）を含むライブラリ生成プロセスのモデルを考慮すると、最大尤度の間隙長さは、間隙に及ぶペアの分離ｄｉの結合尤度を最大限にすることによって見つけられる。微分可能なモデルＰＤＦに関して、効率的な反復最適化方法（例えばＮｅｗｔｏｎ－Ｒａｐｈｓｏｎ）が、幾つかの場合において使用される。

【０１８１】

本明細書に開示される方法および組成物の要素は、コンティグが、８、７、６、５、４、３、２、または２時間未満などの、比較的短い時間で実行可能または入手可能である一方で、例えば、コンティグ順序、配向、または順序及び配向に対する２、３、４、５、６、または６以上のコンティグのコンティグウィンドウの中で局所最適である構成へとアセンブルされることである。したがって、幾つかの場合では、本明細書の方法によって、大量の計算時間を使用せず、また全体的に非常に大規模な計算的空間を探索する必要とすることなく、高度な計算能力が計算集約的な問題に当てられることが可能となる。むしろ、局所的な順序付けは、コンティグの適度に正確な順序付けを達成し、その後、計算集約は、ほとんどの場合一度にすべてのコンティグを全体的に最適化するよりもむしろ、コンティグのローカルウィンドウを最適化して消費される。幾つかの場合では、３、４、５、または６の範囲のウィンドウサイズを使用して、構成最適化が、８、７、６、５、４、３、２、または２時間未満に行われる。より大きなウィンドウサイズについては、構成最適化は、最大１週間で数日で達成される。

【０１８２】
＜デジタル処理装置＞
幾つかの実施形態では、本明細書に記載されるコンティグのアセンブリ方法は、デジタル処理装置、またはその使用を含む。さらなる実施形態では、デジタル処理装置は、装置の機能を実行する１つ以上のハードウェア中央処理装置（ＣＰＵ）を含む。またさらなる実施形態では、デジタル処理装置は、実行可能命令を実行するように構成されたオペレーティングシステムをさらに含む。幾つかの実施形態では、デジタル処理装置は、コンピュータネットワークに随意に接続される。さらなる実施形態では、デジタル処理装置は、ワールドワイドウェブにアクセスするようにインターネットに随意に接続される。またさらなる実施形態では、デジタル処理装置は、クラウド・コンピューティング・インフラストラクチャーに随意に接続される。他の実施形態では、デジタル処理装置は、イントラネットに随意に接続される。他の実施形態では、デジタル処理装置は、データ記憶装置に随意に接続される。
【０１８３】
本明細書の記載に従って、適切なデジタル処理装置は、限定しない例として、サーバーコンピュータ、デスクトップコンピュータ、ラップトップコンピュータ、ノート型コンピュータ、サブノート型コンピュータ、ネットブックコンピュータ、ネットパッドコンピュータ、セットトップコンピュータ、メディアストリーミング装置、ハンドヘルドコンピュータ、インターネットアプライアンス、モバイルスマートフォン、タブレットコンピュータ、携帯情報端末、ビデオゲーム機、および媒体を含む。当業者は、多くのスマートフォンが、本明細書に記載されるシステムにおける使用に適していることを認識する。当業者はまた、随意のコンピュータネットワーク接続性を有する、選択したテレビ、ビデオプレーヤー、およびデジタル音楽プレーヤーが、本明細書に記載されるシステムにおける使用に適していることも認識する。適切なタブレットコンピュータは、当業者に既知の、ブックレット、スレート、および変換可能な構成を備えたものを含む。
【０１８４】
幾つかの実施形態では、デジタル処理装置は、実行可能命令を実行するように構成されたオペレーティングシステムを含む。オペレーティングシステムは、例えば、装置のハードウェアを管理し、アプリケーションの実行のためのサービスを提供する、プログラム及びデータを含むソフトウェアである。当業者は、適切なサーバーのオペレーティングシステムが、限定しない例として、ＦｒｅｅＢＳＤ、ＯｐｅｎＢＳＤ、ＮｅｔＢＳＤ（登録商標）、Ｌｉｎｕｘ（登録商標）、Ａｐｐｌｅ（登録商標）Ｍａｃ ＯＳ Ｘ Ｓｅｒｖｅｒ（登録商標）、Ｏｒａｃｌｅ（登録商標）Ｓｏｌａｒｉｓ（登録商標）、Ｗｉｎｄｏｗｓ Ｓｅｒｖｅｒ（登録商標）、およびＮｏｖｅｌ（登録商標）ＮｅｔＷａｒｅ（登録商標）を含むことを認識する。当業者は、適切なパーソナルコンピュータのオペレーティングシステムが、限定しない例として、Ｍｉｃｒｏｓｏｆｔ（登録商標）Ｗｉｎｄｏｗ（登録商標）、Ａｐｐｌｅ（登録商標）Ｍａｃ ＯＳ Ｘ（登録商標）、ＵＮＩＸ（登録商標）、およびＧＮＵ／Ｌｉｎｕｘ（登録商標）などのＵＮＩＸ（登録商標）様のオペレーティングシステムを含むことを認識する。幾つかの実施形態では、オペレーティングシステムは、クラウドコンピューティングによって提供される。当業者はまた、適切なモバイルスマートフォンのオペレーティングシステムが、限定しない例として、Ｎｏｋｉａ（登録商標）Ｓｙｍｂｉａｎ（登録商標）ＯＳ、Ａｐｐｌｅ（登録商標）ｉＯＳ（登録商標）、Ｒｅｓｅａｒｃｈ Ｉｎ Ｍｏｔｉｏｎ（登録商標）ＢｌａｃｋＢｅｒｒｙ ＯＳ（登録商標）、Ｇｏｏｇｌｅ（登録商標）、Ａｎｄｒｏｉｄ（登録商標）、Ｍｉｃｒｏｓｏｆｔ（登録商標）Ｗｉｎｄｏｗｓ Ｐｈｏｎｅ（登録商標）ＯＳ、Ｍｉｃｒｏｓｏｆｔ（登録商標）Ｗｉｎｄｏｗｓ Ｍｏｂｉｌｅ（登録商標）ＯＳ、Ｌｉｎｕｘ（登録商標）、およびＰａｌｍ（登録商標）ＷｅｂＯＳ（登録商標）を含むことも認識する。
【０１８５】
幾つかの実施形態では、装置は、記憶装置及び／又はメモリ装置を含む。記憶装置及び／又はメモリ装置は、一時的又は恒久的にデータまたはプログラムを保存するために使用

される１以上の物理的な機器である。幾つかの実施形態では、メモリ装置は、揮発性メモリであり、保存した情報を維持するための動力を必要とする。幾つかの実施形態では、メモリ装置は、不揮発性メモリであり、デジタル処理装置に動力が供給されないときに保存した情報を保持する。さらなる実施形態では、不揮発性メモリは、フラッシュメモリを含む。幾つかの実施形態では、不揮発性メモリは、ダイナミック・ランダム・アクセス・メモリ（ＤＲＡＭ）を含む。幾つかの実施形態では、不揮発性メモリは、強誘電体ランダムアクセスメモリ（ＦＲＡＭ）を含む。幾つかの実施形態では、不揮発性メモリは、相変化ランダムアクセスメモリ（ＰＲＡＭ）を含む。随意に、装置は、限定しない例として、ＣＤ－ＲＯＭ、ＤＶＤ、フラッシュメモリ装置、磁気ディスクドライブ、磁気テープドライブ、光ディスクドライブ、およびクラウドコンピューティングベースのストレージを含む、記憶装置である。さらなる実施形態では、記憶装置及び／又はメモリ装置は、本明細書に開示されるものなどの装置の組み合わせである。

【０１８６】

幾つかのデジタル処理装置は、陰極線管（ＣＲＴ）、液晶ディスプレイ（ＬＣＤ）、薄膜トランジスター液晶ディスプレイ（ＴＦＴ－ＬＣＤ）、単純マトリクス方式ＯＬＥＤ（ＰＭＯＬＥＤ）またはアクティブマトリックスＯＬＥＤ（ＡＭＯＬＥＤ）ディスプレイ、プラズマディスプレイ、ビデオプロジェクター、あるいは本明細書に開示されるものなどの装置の組み合わせなどの、有機発光ダイオード（ＯＬＥＤ）ディスプレイなどの、ユーザーに目視情報を送信するためのディスプレイを含む。

【０１８７】

多くの場合、デジタル処理装置は、キーボード、限定しない例として、マウス、トラックボール、トラックパッド、ジョイスティック、ゲームコントローラー、またはスタイラスを含む、ポインティングデバイスなどの、ユーザーから情報を受信するための入力デバイスを含む。幾つかの実施形態では、入力デバイスは、タッチスクリーンまたはマルチタッチスクリーン、音声または他の音入力を捕らえるためのマイクロフォン、あるいは運動または視覚入力を捕らえるためのビデオカメラまたは他のセンサーである。さらなる実施形態では、入力デバイスは、Ｋｉｎｅｃｔ、Ｌｅａｐ Ｍｏｔｉｏｎなどである。多くの場合、入力デバイスは、本明細書に開示されるものなどの装置の組み合わせである。

【０１８８】

＜非一時的コンピュータ可読記憶媒体＞

幾つかの実施形態では、本明細書に開示されるコンティグのアセンブリ方法は、随意にネットワーク化されたデジタル処理装置のオペレーティングシステムによって実行可能な命令を含むプログラムでコードされた１つ以上の非一時的なコンピュータ可読記憶媒体を含む。さらなる実施形態では、コンピュータ可読記憶媒体は、随意に、デジタル処理装置の有形成分である。またさらなる実施形態では、コンピュータ可読記憶媒体は、随意に、デジタル処理装置から取り外し可能である。幾つかの実施形態では、コンピュータ可読記憶媒体は、限定しない例として、ＣＤ－ＲＯＭ、ＤＶＤ、フラッシュメモリ装置、ソリッドステートメモリ、磁気ディスクドライブ、磁気テープドライブ、光ディスクドライブ、クラウドコンピューティングのシステムおよびサービスなどを含む。幾つかの場合では、プログラムおよび命令は、記憶媒体上に、恒久的に、略恒久的に、半恒久的に、または非一時的にコードされる。

【０１８９】

＜コンピュータプログラム＞

幾つかの実施形態では、本明細書に開示されるコンティグのアセンブリ方法は、少なくとも１つのコンピュータプログラム、またはその使用を含む。コンピュータプログラムは、デジタル処理装置のＣＰＵにおいて実行可能であり、タスクを実行するように書き込まれた、命令のシーケンス（ｓｅｑｕｅｎｃｅ）を含む。コンピュータ可読命令は、特定のタスクを実行するか又は特定の抽象データ型を実施する、機能、対象（ｏｂｊｅｃｔｓ）、アプリケーションプログラムインターフェース（ＡＰＩ）、データ構造などのプログラムモジュールとして実行され得る。本明細書で提供される開示に照らして、当業者は、コ

ンピュータプログラムが様々な言語の様々なバージョンで書き込まれ得ることを認識する。

【０１９０】

コンピュータ可読命令の機能性は、様々な環境に応じて組み合わせられるか、または分配され（ｄｉｓｔｒｉｂｕｔｅｄ）得る。幾つかの実施形態では、コンピュータプログラムは、命令の１つのシーケンスを含む。幾つかの実施形態では、コンピュータプログラムは、命令の複数のシーケンスを含む。幾つかの実施形態では、コンピュータプログラムは、１つの場所から提供される。他の実施形態では、コンピュータプログラムは、複数の場所から提供される。様々な実施形態では、コンピュータプログラムは、１以上のソフトウェアモジュールを含む。様々な実施形態では、コンピュータプログラムは、部分的または全体的に、１以上のウェブアプリケーション、１以上のモバイルアプリケーション、１以上の独立型アプリケーション、１以上のウェブブラウザプラグイン、拡張、アドイン、またはアドオン、あるいはそれらの組み合わせを含む。

【０１９１】

＜ウェブアプリケーション＞

幾つかの実施形態では、コンティグのアセンブリ方法を実施するコンピュータプログラムまたはコンピュータにより実施されるシステムは、ウェブアプリケーションを含む。本明細書に提供される開示に照らして、当業者は、ウェブアプリケーションが、様々な実施形態において、１つ以上のソフトウェアフレームワークおよび１つ以上のデータベースシステムを利用することを認識する。幾つかの実施形態では、ウェブアプリケーションは、マイクロソフト（登録商標）．ＮＥＴまたはＲｕｂｙ ｏｎ Ｒａｉｌｓ（ＲｏＲ）などの、ソフトウェアフレームワーク上で作成される。幾つかの実施形態では、ウェブアプリケーションは、限定しない例として、リレーショナル、非リレーショナル、オブジェクト指向、アソシアティブ、およびＸＭＬのデータベースシステムを含む、１つ以上のデータベースシステムを利用する。さらなる実施形態では、適切なリレーショナルデータベースシステムは、限定しない例として、マイクロソフト（登録商標）ＳＱＬ Ｓｅｒｖｅｒ、ｍｙＳＱＬ（商標）、およびＯｒａｃｌｅ（登録商標）を含む。当業者はまた、ウェブアプリケーションが、様々な実施形態において、１つ以上の言語の１つ以上のバージョンで書き込まれることを認識する。ウェブアプリケーションは、１つ以上のマークアップ言語、提示定義言語、クライアント側スクリプト言語、サーバー側コーディング言語、データベース問い合わせ言語、あるいはそれらの組み合わせで書き込まれ得る。幾つかの実施形態では、ウェブアプリケーションは、ハイパーテキストマークアップ言語（ＨＴＭＬ）、拡張可能なハイパーテキストマークアップ（ＸＨＴＭＬ）、または拡張マークアップ言語（ＸＭＬ）などの、マークアップ言語である程度まで書き込まれる。幾つかの実施形態では、ウェブアプリケーションは、カスケーディング・スタイル・シート（ＣＳＳ）などの表現定義言語である程度まで書き込まれる。幾つかの実施形態では、ウェブアプリケーションは、Ａｓｙｎｃｈｒｏｎｏｕｓ ＪａｖａｓｃｒｉｐｔおよびXML（ＡＪＡＸ）、Ｆｌａｓｈ（登録商標）Ａｃｔｉｏｎｓｃｒｉｐ、Ｊａｖａｓｃｒｉｐｔ、またはＳｉｌｖｅｒｌｉｇｈｔ（登録商標）などの、クライアント側スクリプト言語である程度まで書き込まれる。幾つかの実施形態では、ウェブアプリケーションは、Ａｃｔｉｖｅ Ｓｅｒｖｅｒ Ｐａｇｅｓ（ＡＳＰ）、ＣｏｌｄＦｕｓｉｏｎ（登録商標）、Ｐｅｒｌ、Ｊａｖａ（登録商標）、ＪａｖａＳｅｒｖｅｒ Ｐａｇｅｓ（ＪＳＰ）、Ｈｙｐｅｒｔｅｘｔ Ｐｒｅｐｒｏｃｅｓｓｏｒ（ＰＨＰ）、Ｐｙｔｈｏｎ（商標）、Ｒｕｂｙ、Ｔｃｌ、Ｓｍａｌｌｔａｌｋ、ＷｅｂＤＮＡ（登録商標）、またはＧｒｏｏｖｙなどの、サーバー側コーディング言語である程度まで書き込まれる。幾つかの実施形態では、ウェブアプリケーションは、構造化照会言語（ＳＱＬ）などのデータベース問い合わせ言語である程度まで書き込まれる。幾つかの実施形態では、ウェブアプリケーションは、ＩＢＭ（登録商標）Ｌｏｔｕｓ Ｄｏｍｉｎｏ（登録商標）などのエンタープライズサーバー製品を統合する。幾つかの実施形態では、ウェブアプリケーションは、メディアプレイヤー要素を含む。様々なさらなる実施形態では、メディアプレイヤー要素は、限定しない例として、Ａｄｏｂｅ

（登録商標）Ｆｌａｓｈ（登録商標）、ＨＴＭＬ ５、Ａｐｐｌｅ（登録商標）ＱｕｉｃｋＴｉｍｅ（登録商標）、Ｍｉｃｒｏｓｏｆｔ（登録商標）Ｓｉｌｖｅｒｌｉｇｈｔ（登録商標）、Ｊａｖａ（登録商標）、およびＵｎｉｔｙ（登録商標）を含む、多くの適切なマルチメディア技術の１つ以上を利用する。

【０１９２】

＜モバイルアプリケーション＞

幾つかの実施形態では、本明細書に開示されるコンティグのアセンブリ方法を実施するコンピュータプログラムは、モバイルのデジタル処理装置に提供されるモバイルアプリケーションを含む。幾つかの実施形態では、モバイルアプリケーションは、モバイルのデジタル処理装置に、それが製造される時に提供される。他の実施形態では、モバイルアプリケーションは、本明細書に記載されるコンピュータネットワークを介してモバイルのデジタル処理装置に提供される。

【０１９３】

本明細書に提供される開示に照らして、モバイルアプリケーションは、当該技術分野に既知のハードウェア、言語、および開発環境を使用して、当業者に既知の技術によって作成される。当業者は、モバイルアプリケーションが幾つかの言語で書き込まれることを認識する。適切なプログラミング言語は、限定しない例として、Ｃ、Ｃ＋＋、Ｃ＃、Ｏｂｊｅｃｔｉｖｅ－Ｃ、Ｊａｖａ（登録商標）、Ｊａｖａｓｃｒｉｐｔ、Ｐａｓｃａｌ、Ｏｂｊｅｃｔ Ｐａｓｃａｌ、Ｐｙｔｈｏｎ（商標）、Ｒｕｂｙ、ＶＢ．ＮＥＴ、ＷＭＬ、およびＣＳＳを備える又はそれがないＸＨＴＭＬ／ＨＴＭＬ、あるいはそれらの組み合わせを含む。

【０１９４】

適切なモバイルアプリケーションの開発環境は、幾つかのソースから入手可能である。市販の開発環境は、限定しない例として、ＡｉｒｐｌａｙＳＤＫ、ａｌｃｈｅＭｏ、Ａｐｐｃｅｌｅｒａｔｏｒ（登録商標）、Ｃｅｌｓｉｕｓ、Ｂｅｄｒｏｃｋ、Ｆｌａｓｈ Ｌｉｔｅ、．ＮＥＴ Ｃｏｍｐａｃｔ Ｆｒａｍｅｗｏｒｋ、Ｒｈｏｍｏｂｉｌｅ、およびＷｏｒｋＬｉｇｈｔ Ｍｏｂｉｌｅ Ｐｌａｔｆｏｒｍを含む。限定しない例として、Ｌａｚａｒｕｓ、ＭｏｂｉＦｌｅｘ、ＭｏＳｙｎｃ、およびＰｈｏｎｅｇａｐを含む、他の開発環境は、費用をかけずに入手可能である。さらに、モバイルデバイスの製造業者は、限定しない例として、ｉＰｈｏｎｅ（登録商標）およびｉＰａｄ（登録商標）（ｉＯＳ）ＳＤＫ、Ａｎｄｒｏｉｄ（商標）ＳＤＫ、ＢｌａｃｋＢｅｒｒｙ（登録商標）ＳＤＫ、ＢＲＥＷ ＳＤＫ、Ｐａｌｍ（登録商標）ＯＳ ＳＤＫ、Ｓｙｍｂｉａｎ ＳＤＫ、ｗｅｂＯＳ ＳＤＫ、およびＷｉｎｄｏｗｓ（登録商標）Ｍｏｂｉｌｅ ＳＤＫを含む、ソフトウェアディベロッパーキットを提供している。

【０１９５】

当業者は、限定しない例として、Ａｐｐｌｅ（登録商標）Ａｐｐ Ｓｔｏｒｅ、Ａｎｄｒｏｉｄ（商標）Ｍａｒｋｅｔ、ＢｌａｃｋＢｅｒｒｙ（登録商標）Ａｐｐ Ｗｏｒｌｄ、Ｐａｌｍデバイス用のＡｐｐ Ｓｔｏｒｅ、ｗｅｂＯＳ用のＡｐｐ Ｃａｔａｌｏｇ、Ｍｏｂｉｌｅ用のＷｉｎｄｏｗｓ（登録商標）Ｍａｒｋｅｔｐｌａｃｅ、Ｎｏｋｉａ（登録商標）デバイス用のＯｖｉ Ｓｔｏｒｅ、Ｓａｍｓｕｎｇ（登録商標）Ａｐｐ、およびＮｉｎｔｅｎｄｏ（登録商標）ＤＳｉ Ｓｈｏｐを含む、モバイルアプリケーションの提供に利用可能である幾つかの商用フォーラムを認識する。

【０１９６】

＜独立型アプリケーション＞

幾つかの実施形態では、本明細書に開示されるコンティグのアセンブリ方法を実施するコンピュータプログラムは、既存プロセスに対するアドオンではなく（例えば、プラグインではない）、独立したコンピュータプロセスとして実行されるプログラムである、独立型アプリケーションを含む。当業者は、独立型アプリケーションがしばしばコンパイルされることを認識する。コンパイラは、プログラミング言語で書き込まれたソースコードを、アセンブリ言語またはマシンコードなどのバイナリー化されたオブジェクトコードに変

換するコンピュータプログラムである。適切なコンパイルされたプログラミング言語は、限定しない例として、Ｃ、Ｃ＋＋、Ｏｂｊｅｃｔｉｖｅ－Ｃ、ＣＯＢＯＬ、Ｄｅｌｐｈｉ、Ｅｉｆｆｅｌ、Ｊａｖａ（登録商標）、Ｌｉｓｐ、Ｐｙｔｈｏｎ（商標）、Ｖｉｓｕａｌ Ｂａｓｉｃ、およびＶＢ．ＮＥＴ、あるいはそれらの組み合わせを含む。コンパイルは、少なくとも部分的に、実行可能プログラムを作成するためにしばしば実行される。幾つかの実施形態では、コンピュータプログラムは、１つ以上の実行可能なコンパイルされたアプリケーションを含む。

【０１９７】

＜ウェブブラウザのプラグイン＞

幾つかの実施形態では、コンティグのアセンブリ方法は、ウェブブラウザのプラグインを含む。コンピューティングにおいて、プラグインは、より大きなソフトウェアアプリケーションに特定の機能性を加える１つ以上のソフトウェアコンポーネントである。ソフトウェアアプリケーションのメーカーは、サードパーティーのディベロッパーが、アプリケーションを拡張する能力を作り出す、容易に加える新しい特徴をサポートする、およびアプリケーションのサイズを縮小することができるプラグインをサポートしている。サポートされるときに、プラグインは、ソフトウェアアプリケーションの機能性をカスタマイズにすることができる。例えば、プラグインは、ビデオを再生する、対話機能を生成する、ウイルスをスキャンする、および特別のファイルタイプを表示するためにウェブブラウザにおいて一般に使用される。当業者は、Ａｄｏｂｅ（登録商標）Ｆｌａｓｈ（登録商標）Ｐｌａｙｅｒ、Ｍｉｃｒｏｓｏｆｔ（登録商標）Ｓｉｌｖｅｒｌｉｇｈｔ（登録商標）、およびＡｐｐｌｅ（登録商標）ＱｕｉｃｋＴｉｍｅ（登録商標）を含む、幾つかのウェブブラウザのプラグインに精通しているだろう。幾つかの実施形態では、ツールバーは、１つ以上のウェブブラウザ拡張、アドイン、またはアドオンを含む。幾つかの実施形態では、ツールバーは、１つ以上のエクスプローラバー、ツールバンド、またはデスクバンドを含む。

【０１９８】

本明細書に提供される開示に照らして、当業者は、限定しない例として、Ｃ＋＋、Ｄｅｌｐｈｉ、Ｅｉｆｆｅｌ、Ｊａｖａ（登録商標）、ＰＨＰ、Ｐｙｔｈｏｎ（商標）、およびＶＢ．ＮＥＴ、あるいはそれらの組み合わせを含む、様々なプログラミング言語でのプラグインの開発を可能にする、幾つかのプラグインフレームワークが利用可能であることを認識する。

【０１９９】

ウェブブラウザ（インターネットブラウザとも呼ばれる）は、ワールドワイドウェブ上で情報リソースを検索、提示、および横断するためのネットワーク接続されたデジタル処理装置との使用のために設計された、ソフトウェアアプリケーションである。適切なウェブブラウザは、限定しない例として、Ｍｉｃｒｏｓｏｆｔ（登録商標）Ｉｎｔｅｒｎｅｔ Ｅｘｐｌｏｒｅｒ（登録商標）、Ｍｏｚｉｌｌａ（登録商標）Ｆｉｒｅｆｏｘ（登録商標）、Ｇｏｏｇｌｅ（登録商標）Ｃｈｒｏｍｅ、Ａｐｐｌｅ（登録商標）Ｓａｆａｒｉ（登録商標）、Ｏｐｅｒａ Ｓｏｆｔｗａｒｅ（登録商標）Ｏｐｅｒａ（登録商標）、およびＫＤＥ Ｋｏｎｑｕｅｒｏｒを含む。幾つかの実施形態では、ウェブブラウザは、モバイルウェブブラウザである。モバイルウェブブラウザ（マイクロブラウザ、ミニブラウザ、およびワイヤレスブラウザとも呼ばれる）は、限定しない例として、ハンドヘルドコンピュータ、タブレットコンピュータ、ネットブックコンピュータ、サブノートブックコンピュータ、スマートフォン、音楽プレーヤー、携帯情報端末（ＰＤＡ）、およびハンドヘルドビデオゲームシステムを含む、モバイルデジタル処理装置上の使用のために設計されている。適切なモバイルウェブブラウザは、限定しない例として、Ｇｏｏｇｌｅ（登録商標）Ａｎｄｒｏｉｄ（登録商標）ブラウザ、ＲＩＭ ＢｌａｃｋＢｅｒｒｙ（登録商標）ブラウザ、Ａｐｐｌｅ（登録商標）Ｓａｆａｒｉ（登録商標）、Ｐａｌｍ（登録商標）Ｂｌａｚｅｒ、Ｐａｌｍ（登録商標）ＷｅｂＯＳ（登録商標）ブラウザ、モバイル用Ｍｏｚｉｌｌａ（登録商標）Ｆｉｒｅｆｏｘ（登録商標）、マイクロソフト（登録商標）Ｉｎｔ

ｅｒｎｅｔ Ｅｘｐｌｏｒｅｒ（登録商標）Ｍｏｂｉｌｅ、Ａｍａｚｏｎ（登録商標）Ｋｉｎｄｌｅ（登録商標）Ｂａｓｉｃ Ｗｅｂ、Ｎｏｋｉａ（登録商標）ブラウザ、Ｏｐｅｒａ Ｓｏｆｔｗａｒｅ（登録商標）Ｏｐｅｒａ（登録商標）Ｍｏｂｉｌｅ、およびＳｏｎｙ（登録商標）ＰＳＰ（商標）ブラウザを含む。

【０２００】

＜ソフトウェアモジュール＞

幾つかの実施形態では、本明細書に開示されるコンティグのアセンブリ方法は、ソフトウェア、サーバー、及び／又はデータベースモジュール、あるいはその使用を含む。本明細書に提供される開示に照らして、ソフトウェアモジュールは、当該技術分野に既知のマシン、ソフトウェア、および言語を使用して、当業者に既知の技術によって作成される。本明細書に開示されるソフトウェアモジュールは、多くの方法で実施される。様々な実施形態では、ソフトウェアモジュールは、ファイル、コードのセクション、プログラミングオブジェクト、プログラミング構造、あるいはそれらの組み合わせを含む。さらに様々な実施形態では、ソフトウェアモジュールは、複数のファイル、複数のコードのセクション、複数のプログラミングオブジェクト、複数のプログラミング機構、あるいはそれらの組み合わせを含む。様々な実施形態では、１つ以上のソフトウェアモジュールは、限定しない例として、ウェブアプリケーション、モバイルアプリケーション、および独立型アプリケーションを含む。幾つかの実施形態では、ソフトウェアモジュールは、１つのコンピュータプログラムまたはアプリケーションにある。他の実施形態では、ソフトウェアモジュールは、１つを超えるコンピュータプログラムまたはアプリケーションにある。幾つかの実施形態では、ソフトウェアモジュールは、１つのマシン上でホストされる。他の実施形態では、ソフトウェアモジュールは、１つを超えるマシン上でホストされる。さらなる実施形態では、ソフトウェアモジュールは、クラウドコンピューティングプラットホーム上でホストされる。幾つかの実施形態では、ソフトウェアモジュールは、１つの位置で１つ以上のマシン上でホストされる。他の実施形態では、ソフトウェアモジュールは、１つを超える位置で１つ以上のマシン上でホストされる。

【０２０１】

＜データベース＞

幾つかの実施形態では、本明細書に開示されるコンティグのアセンブリ方法は、１つ以上のデータベース、またはその使用を含む。本明細書で提供される開示に照らして、当業者は、多くのデータベースが、コンティグ情報の保管および検索に適していることを認識する。様々な実施形態では、適切なデータベースは、限定されないが、リレーショナルデータベース、非リレーショナルデータベース、オブジェクト指向型データベース、オブジェクトデータベース、実体関連モデルデータベース、連想型データベース、およびＸＭＬデータベースを含む。他の実施形態では、データベースは、インターネットベースである。更なる実施形態では、データベースは、ウェブベースである。また更なる実施形態では、データベースは、クラウドコンピューティングベースである。他の実施形態において、データベースは、１つ以上のローカルコンピュータ記憶デバイスベースである。

＜列挙された実施形態＞

【０２０２】

本明細書の開示は、番号付けされた実施形態の非制限的なリストとしてさらに示される。

１．　核酸配列情報のコンティグをスキャフォールドするための方法であって、当該方法は、

初期の構成を有するコンティグ配列のセットを得る工程；ペアエンドリードのセットを得る工程；標準的なペアエンドリードの距離度数データを得る工程；

少なくとも１つのペアエンドリードにおいて共存する配列を共有するコンティグペアをグループ分けする工程；および

コンティグを分離するためにマッピングされるリードペアのためのリードペア距離度数データが、初期の構成におけるコンティグ配列のリードペア度数データに対して標準的な

ペアエンドリード距離度数データにより近く近似するように、グループ化されたコンティグ配列をスキャフォールドする工程、
を含むことを特徴とする、方法である。
２．　スキャフォールドする工程が、一連のコンティグを順序付けする工程を含む、列挙された実施形態１に記載の方法により実施される。
３．　スキャフォールドする工程が、一連のコンティグを配向する工程を含む、列挙された実施形態１に記載の方法により実施される。
４．　スキャフォールドする工程が、少なくとも２つのコンティグの端と端を融合する工程を含む、列挙された実施形態１に記載の方法により実施される。
５．　スキャフォールドする工程が、第２コンティグに１つのコンティグを挿入する工程を含む、列挙された実施形態１に記載の方法により実施される。
６．　スキャフォールドする工程が、少なくとも２つの成分のコンティグへ、コンティグを切断する工程を含む、列挙された実施形態１に記載の方法により実施される。
７．　標準的なペアエンドリード度数が、両方のリードが共通のコンティグにマッピングされるペアエンドリードから得られる、列挙された実施形態１に記載の方法により実施される。
８．　標準的なペアエンドリード度数が、事前に生成された曲線から得られる、列挙された実施形態１に記載の方法により実施される。
９．　初期構成がランダムな構成である、列挙された実施形態１の方法により実施される。
１０．　コンティグを分離するためにマップピングされるリードペアのためのリードペア距離度数データが、リードペア距離尤度が増大するときにペアエンドリード距離度数データにより近く近似する、列挙された実施形態１に記載の方法により実施される。
１１．　リードペア距離尤度が最大化される、列挙された実施形態１０に記載の方法により実施される。
１２．　コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データが、リードペアの距離度数データと標準的なペアエンドリード距離度数データとの間の差の統計的尺度が減少する際に、ペアエンドリード距離度数データにより近く近似する、列挙された実施形態１に記載の方法により実施される。
１３．　リードペア距離度数データと標準的なペアエンドリード距離度数データとの間の距離の統計的尺度が、ＡＮＯＶＡ、ｔ検定、およびＸ２乗検定の少なくとも１つを含む、列挙された実施形態１２に記載の方法により実施される。
１４．　コンティグを分離するためにマッピングされるリードペアのリードペア距離が、標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差が減少するとき、ペアエンドリード距離度数データとより近く一致する、列挙された実施形態１に記載の方法により実施される。
１５．　標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差が最小限に抑えられる、列挙された実施形態１４に記載の方法により実施される。
１６．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグが、より多くの共有エンドリードを有するクラスターに割り当てられる、列挙された実施形態１に記載の方法により実施される。
１７．　前記クラスター化する工程が、生体中の染色体の数以上である多くのグループにコンティグを配置する工程を含む、列挙された実施形態１－１６のいずれか１つに記載の方法により実施される。
１８．　単一のペアエンドリードのみをクラスターの１つのコンティグと共有するコンティグが、そのクラスターには含まれない、列挙された実施形態１１－１７のいずれか１つに記載の方法により実施される。
１９．　反復配列を含む少なくとも１つのペアエンドリードのみをクラスターと共有するコンティグが、そのクラスターには含まれない、列挙された実施形態１－１８のいずれか

１つに記載の方法により実施される。
２０．　低品質の配列を含む少なくとも１つのペアエンドリードのみをクラスターと共有するコンティグが、そのクラスターには含まれない、列挙された実施形態１－１９のいずれか１つに記載の方法により実施される。
２１．　ペアエンドリードのセットが、核酸内で内部二本鎖切断を生じさせるためにサンプルＤＮＡを消化し、少なくとも１つの再連結接合部を形成するために二本鎖切断を再連結し、および少なくとも１つの再連結接合部にわたって配列決定することにより得られる、列挙された実施形態１－２０のいずれか１つに記載の方法により実施される。
２２．　ＤＮＡが、少なくとも１つのＤＮＡ結合剤に架橋される、列挙された実施形態２１に記載の方法により実施される。
２３．　ＤＮＡが単離した裸のＤＮＡである、列挙された実施形態２１に記載の方法により実施される。
２４．　単離したＤＮＡが、再構成された染色質へ再アセンブリされる、列挙された実施形態２２の方法により実施される。
２５．　再構成された染色質が架橋される、列挙された実施形態２４の方法により実施される。
２６．　再構成された染色質がＤＮＡ結合タンパク質を含む、列挙された実施形態２３の方法により実施される。
２７．　再構成された染色質がナノ粒子を含む、列挙された実施形態２３の方法により実施される。
２８．　コンティグの前記クラスター化する工程が、生体のための染色体の数と無関係である、列挙された実施形態１－２７のいずれか１つの方法により実施される。
２９．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグが、より多い共有エンドリードを有するクラスターに割り当てられる、列挙された実施形態１－２８のいずれか１つに記載の方法により実施される。
３０．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグが、より大きなリードペア距離尤度値を有するクラスターに割り当てられる、列挙された実施形態１－２８のいずれか１つに記載の方法により実施される。
３１．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグが、標準的なリードペア距離分布と比較してそのリードペア分布中で偏差が少ないクラスターに割り当てられる、列挙された実施形態１－２８のいずれか１つに記載の方法により実施される。
３２．　第１のクラスターと第２のクラスターに関連したペアエンドリードの中の配列を共有するコンティグが、各クラスターから除外される、列挙された実施形態１－２９のいずれか１つに記載の方法により実施される。
３３．　前記クラスター化する工程が、生体中の染色体の数以上である多くのグループにコンティグを配置する工程を含む、列挙された実施形態１－３２のいずれか１つに記載の方法により実施される。
３４．　前記スキャフォールドする工程が、前記クラスター化されたコンティグの推定上の隣接するコンティグの第１のセットを選択する工程、前記リードペアに関してリードペアの距離の統計的尺度を減少させる推定上の隣接するコンティグの前記第１のセットの最小限の距離順序を判定する工程、および、リードペア距離の前記統計的尺度を減少させるために、推定上の隣接するコンティグの前記第１のセットをスキャフォールドする工程を含む、列挙された実施形態１－３３のいずれか１つに記載の方法により実施される。
３５．　推定上の隣接するコンティグの前記第１のセットが２つのコンティグから成る、列挙された実施形態３４の方法により実施される。
３６．　推定上の隣接するコンティグの前記第１のセットが３つのコンティグから成る、列挙された実施形態３４の方法により実施される。
３７．　推定上の隣接するコンティグの前記第１のセットが４つのコンティグから成る、列挙された実施形態３４の方法により実施される。

３８．　推定上の隣接するコンティグの前記第１のセットが４つのコンティグを含む、列挙された実施形態３４の方法により実施される。
３９．　前記スキャフォールドする工程が、推定上の隣接したコンティグの前記第１のセットにおける各コンティグの順序および配向を判定する工程を含む、列挙された実施形態３４の方法により実施される。
４０．　前記最小限の距離順序を判定する工程が、すべての可能なコンティグ構成についての前記セットの２つのコンティグにマッピングされているリードを含む少なくとも１つのリードペアのための予想されるリードペア距離を比較する工程を含む、列挙された実施形態３４または３５のいずれか１つに記載の方法により実施される。
４１．　前記リードペアの最小限のリードペア距離に対応するコンティグ配向を選択する工程をさらに含む、列挙された実施形態４０に記載の方法により実施される。
４２．　最大尤度のリードペア距離分布に対応するコンティグ配向を選択する工程をさらに含む、列挙された実施形態４０に記載の方法により実施される。
４３．　前記コンティグクラスターのリードペアの統計的尺度のための最小限のリードペア距離に対応するコンティグ配向を選択する工程をさらに含む、列挙された実施形態４０または４１のいずれか１つに記載の方法により実施される。
４４．　予想されるリードペア距離は、前記ペアエンドリード距離度数データと比較される、列挙された実施形態４０－４３のいずれか１つに記載の方法により実施される。
４５．　前記ペアエンドリード距離度数データと比較する工程が、式１を用いる工程を含む、列挙された実施形態４４に記載の方法により実施されるシステムである。
４６．　前記クラスター化されたコンティグの推定上の隣接するコンティグの第２のセットを選択する工程であって、前記第２のセットが前記第１のセットの１つのエンド末端のコンティグを除くすべてと、前記クラスター化されたコンティグの１つの追加のコンティグを含む工程、およびリードペアの距離の前記統計的尺度を減少させるために、推定上の隣接するコンティグの前記第２のセットをスキャフォールドする工程をさらに含む、列挙された実施形態３４－４５のいずれか１つに記載の方法により実施される。
４７．　前記クラスター化されたコンティグの推定上の隣接するコンティグの第３のセットを選択する工程であって、前記第３のセットが前記第２のセットの１つのエンド末端のコンティグを除くすべてと、前記第１のセットと前記第２のセットに含まれない前記クラスター化されたコンティグの１つの追加のコンティグを含む工程、およびリードペアの距離の前記統計的尺度を減少させるために、推定上の隣接するコンティグの前記第３のセットをスキャフォールドする工程をさらに含む、列挙された実施形態４６に記載の方法により実施される。
４８．　前記クラスター化されたコンティグの大多数が順序付けされるまで、少なくとも１つの追加のセットを反復して選択する工程をさらに含む、列挙された実施形態４７に記載の方法により実施される。
４９．　前記クラスター化されたコンティグのそれぞれが順序付けされるまで、少なくとも１つの追加のセットを反復して選択する工程をさらに含む、列挙された実施形態４８に記載の方法により実施される。
５０．　核酸配列がゲノムに由来する、列挙された実施形態１－４９のいずれか１つに記載の方法により実施される。
５１．　核酸配列が複数のゲノムに由来する、列挙された実施形態１－４９のいずれか１つに記載の方法により実施される。
５２．　クラスター内のコンティグをスキャフォールドするための方法であって、当該方法が、
　ａ）コンティグの各ペアのための対数尤度比率スコアを割り当てる工程；
　ｂ）比率スコアにより結合部を分類する工程；および
　ｃ）アセンブリの総合スコアを増大させるために、比率スコアの降順でコンティグ結合部を受理する工程、または拒絶する工程、
を含むことを特徴とする、方法である。

５３．　スキャフォールドする工程が、コンティグのセットを順序付けする工程を含む、列挙された実施形態５２に記載の方法により実施される。
５４．　スキャフォールドする工程が、コンティグのセットを配向する工程を含む、列挙された実施形態５２に記載の方法により実施される。
５５．　スキャフォールドする工程が、少なくとも２つのコンティグの端と端を融合する工程を含む、列挙された実施形態５２に記載の方法により実施される。
５６．　スキャフォールドする工程が、第２コンティグに１つのコンティグを挿入する工程を含む、列挙された実施形態５２に記載の方法により実施される。
５７．　スキャフォールドする工程が、少なくとも２つの成分のコンティグへ、コンティグを切断する工程を含む、列挙された実施形態５２に記載の方法により実施される。
５８．　コンティグがゲノムを含む、列挙された実施形態５２に記載の方法により実施される。
５９．　コンティグが複数のゲノムを含む、列挙された実施形態５２に記載の方法により実施される。
６０．　クラスター中で複数のコンティグの局所的に最適なコンティグ構成を判定するための方法であって、当該方法が：
ａ）コンティグのクラスターに沿って位置ｉから始まるサイズｗコンティグの配列ウィンドウを特定する工程；
ｂ）ウィンドウにおける各位置ｉの中の適合性のある順序と配向のスコアを調べることによりウィンドウｗのコンティグのためのｗ！２^{w}の順序付けと配向のオプションを考慮する工程；
ｃ）最適なスコアを得るために前記ウィンドウ内の前記ｗコンティグを配向および順序付けする工程；
ｄ）ウィンドウを位置ｉ＋１に移動させる工程；および、
ｅ）最適なスコアを判定するための前記ｗコンティグの配向および順序付けを使用して、位置ｉ＋１で前記ウィンドウのための工程（ａ）、（ｂ）および（ｃ）を繰り返す工程であって、それにより、スコアに対する局所的に最適な構成中の前記複数のコンティグを配向および順序付けする、繰り返す工程、
を含むことを特徴とする、方法である。
６１．　クラスター中の複数のコンティグにマッピングされているリードペアデータが得られ、標準的なペアエンドリード度数データセットが得られ、および、前記ｗコンティグの配向および順序付けのためのスコアが、クラスター中の複数のコンティグにマッピングされているリードペアデータのためのリードペア距離データセットが標準的なペアエンドリード度数データセットにどれだけ近く一致しているかの尺度である、列挙された実施形態６０に記載の方法により実施される。
６２．　クラスター中の複数のコンティグにマッピングされるリードペアデータが得られ、スコアは合計のリードペア距離であり、および、合計のリードペア距離が最小限に抑えられるとき、スコアは最適化される、列挙された実施形態６０に記載の方法により実施される。
６３．　ｗが３である、列挙された実施形態６０に記載の方法により実施される。
６４．　ｗが４である、列挙された実施形態６０に記載の方法により実施される。
６５．　ｗが５である、列挙された実施形態６０に記載の方法により実施される。
６６．　ｗが６である、列挙された実施形態６０に記載の方法により実施される。
６７．　ｗが第１クラスターに対する第１値を有し、またｗが第２クラスターで第２値を有する、列挙された実施形態６０に記載の方法により実施される。
６８．　ｗが、セットのコンティグの１％を含むために選択される、列挙された実施形態６０に記載の方法により実施される。
６９．　ｗが、セットのコンティグの５％を含むために選択される、列挙された実施形態６０に記載の方法により実施される。
７０．　ｗが、セットのコンティグの１０％を含むために選択される、列挙された実施形

態６０に記載の方法により実施される。
７１．　スコアがリードペア距離尤度スコアであり、スコアが、与えられたウィンドウサイズに対して最大化されるときに最適である、列挙された実施形態６０に記載の方法により実施される。
７２．　スコアが式１を使用して計算される、列挙された実施形態７０に記載の方法により実施される。
７３．　スコアが予想されるリードペア分布からの偏差であり、スコアが、与えられたウィンドウサイズに対して最小限に抑えられるときに最適である、列挙された実施形態６０に記載の方法により実施される。
７４．　複数のコンティグがゲノムを含む、列挙された実施形態６０－７３のいずれか１つに記載の方法により実施される。
７５．　複数のコンティグが複数のゲノムを含む、列挙された実施形態６０－７３のいずれか１つに記載の方法により実施される。
７６．　核酸配列アセンブリのための方法であって、当該方法が：
ａ）精製ＤＮＡを得る工程；
ｂ）ＤＮＡ／染色質複合体を形成するために精製ＤＮＡとＤＮＡ結合剤を結合する工程；
ｃ）粘着末端を残すために制限酵素でＤＮＡ／染色質複合体をインキュベートする工程；
ｄ）ＤＮＡの末端を連結するためにライゲーションを実施する工程；
ｅ）ペアエンドリードを生成するために連結されたＤＮＡ接合部にわたって配列決定する工程；および、
ｆ）精製ＤＮＡの配列を表わすコンティグを含む核酸データセットをスキャフォールドするためにペアエンドリードを使用する工程、
を含むことを特徴とする、方法である。
７７．　精製ＤＮＡがゲノムに由来する、列挙された実施形態７６に記載の方法により実施される。
７８．　精製ＤＮＡが複数のゲノムに由来する、列挙された実施形態７６に記載の方法により実施される。
７９．　反復コンティグ領域にマッピングされるリードペア配列リードを特定するための方法であって、当該方法が：
核酸サンプルのためのコンティグデータセットを得る工程；
隣接していない物理的に結合した配列情報に対応する少なくとも１つのリードペア配列リードを得る工程；および、
リードペア配列リードの少なくとも１つのリードがコンティグデータセットの２つの異なる座位にマッピングされる場合に、リードペア配列リードを除外する工程、
を含むことを特徴とする、方法である。
８０．　反復領域が、第１閾値を超えるあらかじめ与えられたショットガンリード深さを有する配列を含む、列挙された実施形態７９に記載の方法により実施される。
８１．　反復領域が、第２閾値を超えるリード深さを有する塩基位置を含む、列挙された実施形態８０に記載の方法により実施される。
８２．　第１閾値および第２閾値が、リード深さの全面的分布に関して固定される、列挙された実施形態８１に記載の方法により実施される。
８３．　第１閾値が、リード深さの全面的分布の３倍である、列挙された実施形態８２に記載の方法により実施される。
８４．　第２閾値が、リード深さの全面的分布の３．　５倍である、列挙された実施形態８２に記載の方法により実施される。
８５．　核酸サンプルがゲノムを含む、列挙された実施形態７９－８４のいずれか１つに記載の方法により実施される。
８６．　核酸サンプルが複数のゲノムを含む、列挙された実施形態７９－８４のいずれか

１つに記載の方法により実施される。
８７．　コンティグアセンブリ決定を誘導するための方法であって、当該方法が、
　第１コンティグと第２コンティグとの間におよぶリードペア配列の数および示唆される分離を観察する可能性を判定する工程であって、当該コンティグが、セット[＋＋、＋－、－＋、－－]内のｏの相対的配向を有し、間隙長さによって分離される、工程
を含むことを特徴とする、方法である。
８８．　分離距離にわたるリードペア配列の分布の確率を正規化する工程を含み、当該正規化は、リードペア配列と、核酸サンプルを独立してサンプリングするノイズペアとを比較することを含む、列挙された実施形態８７に記載の方法により実施される。
８９．　核酸サンプルがゲノムを含む、列挙された実施形態８８に記載の方法により実施される。
９０．　核酸サンプルが複数のゲノムを含む、列挙された実施形態８８に記載の方法により実施される。
９１．　ノイズペアの総数が、コンティグペアのサンプルに関する結合の密度を集計することにより判定される、列挙された実施形態８８に記載の方法により実施される。
９２．　密度の最も高い１％および最も低い１％が除外される、列挙された実施形態９１の方法により実施される。
９３．　コンティグ順序を判定する工程をさらに含む、列挙された実施形態８７の方法により実施される。
９４．　コンティグ配向を判定する工程をさらに含む、列挙された実施形態８７の方法により実施される。
９５．　コンティグ誤連結修正のための方法であって、当該方法が、
　初期の構成を有するコンティグ配列のセットを得る工程；
　ペアエンドリードのセットを得る工程；
　標準的なペアエンドリードの距離度数データを得る工程；
　少なくとも１つのペアエンドリードにおいて共存する配列を共有するコンティグペアをグループ分けする工程；
　コンティグのグループ分けのためのリードペア度数データを標準的なペアエンドリード距離度数データと比較する工程；
　グループ分けしたコンティグに切断を導入することにより、コンティグのグループ分けのためのリードペア度数データが、標準的なペアエンドリードの距離度数データにさらにより近く近似するかどうかを判定する工程；および、
　コンティグのグループ分けのためのリードペアの度数データが標準的なペアエンドリードの距離度数データとさらにより近く近似する場合には、コンティグに切断を導入する工程、
を含むことを特徴とする、方法である。
９６．　第１位置が、切断を導入する工程の前に前記尤度未満の前記対数閾値を有する少なくとも１つの隣接した第２位置と融合される、列挙された実施形態９５に記載の方法により実施される。
９７．　第２隣接位置が、第１位置から３００塩基対を超えない、列挙された実施形態９５に記載の方法により実施される。
９８．　第２位置が、第１位置から１０００塩基対を超える位置を含まない、列挙された実施形態９５に記載の方法により実施される。
９９．　対数尤度の変更を判定する工程は、コンティグのための平均のペアエンドマッピング密度を特定する工程、平均的なペアエンドマッピング密度の少なくとも３倍の数値のペアエンドマッピング密度を有するコンティグのセグメントを特定する工程、および平均的なペアエンドマッピング密度の少なくとも３倍の数値のペアエンドマッピング密度を有するコンティグのセグメントを除外する工程を含む、列挙された実施形態９５－９８のいずれか１つに記載の方法により実施される。
１００．　コンティグ配列のセットがゲノムに由来する、列挙された実施形態９５－９９

のいずれか１つに記載の方法により実施される。
１０１．　コンティグ配列のセットが複数のゲノムに由来する、列挙された実施形態９５－９９のいずれか１つに記載の方法により実施される。
１０２．　コンティグアセンブリのための方法であって、当該方法が：
開始アセンブリの切断されたコンティグを示す工程であって、当該切断されたコンティグがノードであり、切断されたコンティグのエッジが整数の順序付けられたペアのリストで標識され、切断のエッジがマッピングされたリードペア配列に対応する、工程、および、
マッピングされた結合部の閾値数よりも少数のエッジを除外する工程、
を含むことを特徴とする、方法である。
１０３．　閾値数が５％未満である、列挙された実施形態１０２に記載の方法により実施される。
１０４．　閾値数がｔＬ結合よりも少ない、列挙された実施形態１０２に記載の方法により実施される。
１０５．　コンティグ長さに対する対応ノードのグラフ中の程度の比が塩基対であるエッジを含むコンティグは、分布のすべての値の上限の約５％を超過する、列挙された実施形態１０２の方法により実施される。
１０６．　コンティグがゲノムに由来する、列挙された実施形態１０２－１０５のいずれか１つに記載の方法により実施される。
１０７．　コンティグが複数のゲノムに由来する、列挙された実施形態１０２－１０５のいずれか１つに記載の方法により実施される。
１０８．　少なくとも１つのスキャフォールドへコンティグ配列情報をアセンブルする方法であって、当該方法が、
複数のコンティグに対応する配列情報を得る工程、
複数のコンティグによって示される核酸サンプルからペアエンドリード情報を得る工程、および、
予測されるリードペア距離データセットからのリードペア距離パラメータの偏差が最小化されるように複数のコンティグを構成する工程であって、当該構成する工程が８時間未満で生じる、工程、
を含むことを特徴とする、方法である。
１０９．　予測されるリードペア距離データセットが、リードペア距離尤度曲線を含む、列挙された実施形態１０８に記載の方法により実施される。
１１０．　リードペア距離パラメータが、リードペア距離尤度曲線に対して最大距離尤度である、列挙された実施形態１０８または１０９に記載のいずれか１つに記載の方法により実施される。
１１１．　リードペア距離パラメータが、リードペア距離尤度曲線に対して最小限の変動である、列挙された実施形態１０８または１０９のいずれか１つに記載の方法により実施される。
１１２．　コンティグの局所的に隣接したセットが２つのコンティグを含む、列挙された実施形態１０８－１１１のうちいずれか１つに記載の方法により実施される。
１１３．　前期コンティグの局所的に隣接したセットが３つのコンティグを含む、列挙された実施形態１０８－１１１のうちいずれか１つに記載の方法により実施される。
１１４．　前期コンティグの局所的に隣接したセットが４つのコンティグを含む、列挙された実施形態１０８－１１１のうちいずれか１つに記載の方法により実施される。
１１５．　前期コンティグの局所的に隣接したセットが５つのコンティグを含む、列挙された実施形態１０８－１１１のうちいずれか１つに記載の方法により実施される。
１１６．　前期コンティグの局所的に隣接したセットが６つのコンティグを含む、列挙された実施形態１０８－１１１のうちいずれか１つに記載の方法。
１１７．　構成が７時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。

１１８．　構成が６時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。
１１９．　構成が５時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。
１２０．　構成が４時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。
１２１．　構成が３時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。
１２２．　構成が２時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。
１２３．　構成が１時間未満以内に生じる、列挙された実施形態１０８－１１６のいずれか１つの方法により実施される。
１２４．　コンティグ情報がゲノムに由来する、列挙された実施形態１０８－１２３のいずれか１つに記載の方法により実施される。
１２５．　コンティグ配列情報が複数のゲノムに由来する、列挙された実施形態１０８－１２３のいずれか１つに記載の方法により実施される。
１２６．　コンティグ配列のセットをスキャフォールドする方法であって、当該方法が、
核酸サンプルの代表的なコンティグ配列のセットを得る工程、
核酸サンプル用のリードペアデータを得る工程、および、
核酸サンプル用のリードペアが予想されるリードペア分布により近く近似するようにコンティグ配列のセットを順序付けし、配向する工程であって、当該コンティグ配列のセットの７０％が、わずか８時間で核酸サンプル中においてそれらの配列の相対的な順序と配向に一致するように順序付けられ、配向される、工程、
を含むことを特徴とする、方法である。
１２７．　スキャフォールドする工程が、コンティグのセットを順序付ける工程を含む、列挙された実施形態１２６に記載の方法により実施される。
１２８．　スキャフォールドする工程は、コンティグのセットを配向する工程を含む、列挙された実施形態１２６に記載の方法。
１２９．　スキャフォールドする工程は少なくとも２つのコンティグの端と端を融合する工程を含む、列挙された実施形態１２６に記載の方法。
１３０．　スキャフォールドする工程は１つのコンティグを第２のコンティグに挿入する工程を含む、列挙された実施形態１２６に記載の方法。
１３１．　スキャフォールドする工程はコンティグを少なくとも２つの構成コンティグへと切断する工程を含む、列挙された実施形態１２６に記載の方法。
１３２．　コンティグ配列のセットの８０％は、わずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように、順序付けされ、および配向される、列挙された実施形態１２６に記載の方法。
１３３．　コンティグ配列のセットの９０％は、わずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態１２６に記載の方法。
１３４．　コンティグ配列のセットの９５％は、わずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態１２６に記載の方法。
１３５．　コンティグ配列のセットの７０％は、わずか４時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態１２６に記載の方法。
１３６．　コンティグ配列のセットの７０％は、わずか２時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態１２６に記載の方法。
１３７．　コンティグ配列のセットの７０％は、わずか１時間で核酸サンプル中のそれら

の配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態１２６に記載の方法。
１３８．　コンティグ配列のセットはゲノムに由来する、列挙された実施形態１２６－１３７のいずれか１つに記載の方法。
１３９．　コンティグ配列のセットは複数のゲノムに由来する、列挙された実施形態１２６－１３７のいずれか１つに記載の方法。
１４０．　核酸配列データのセットを構成する方法であって、
複数のコンティグに対応する配列情報を得る工程、ペアエンドリード情報を得る工程、および、ペアエンドリード情報のためのペアエンドリード距離分布が、基準のペアエンドリード距離分布に近似するように全体的に最適化されるように、複数のコンティグを構成する工程を含み、
上記の構成はわずか８時間で生じる、方法。
１４１．　核酸配列データのセットはゲノムに由来する、列挙された実施形態１４０に記載の方法。
１４２．　核酸配列データのセットは複数のゲノムに由来する、列挙された実施形態１４０に記載の方法。
１４３．　上記の構成はわずか４時間で生じる、列挙された実施形態１４０に記載の方法。
１４４．　上記の構成はわずか２時間で生じる、列挙された実施形態１４０に記載の方法。
１４５．　スキャフォールドのアセンブリを改善する方法であって、
複数の連結したノードペアを含むスキャフォールドセットを得る工程であって、ノードペアの各ノードが少なくとも１つのコンティグ配列を含む、工程と、複数の連結したノードにマッピングされたペアエンドリード情報を得る工程と、連結されたノードペアによって共有されるリードペアの数を数える工程と、上記数を閾値と比較する工程と、上記数が閾値以下である場合にノードペアを連結していないノードへと切断する工程を含む、方法。
１４６．　固有のコンティグ配列にマッピングされたリードペアだけが数えられる、列挙された実施形態１４５に記載の方法。
１４７．　異なるペアエンド閾値数がマッピングされるコンティグ配列セグメントに対するリードペアのマッピングは廃棄される、列挙された実施形態１４５に記載の方法。
１４８．　閾値数は非反復配列について代表数の３倍である、列挙された実施形態１４５に記載の方法。
１４９．　スキャフォールドセットはゲノムを含む、列挙された実施形態１４５－１４８のいずれか１つに記載の方法。
１５０．　スキャフォールドセットは複数のゲノムを含む、列挙された実施形態１４５－１４８のいずれか１つに記載の方法。
１５１．　スキャフォールドのアセンブリを改善する方法であって、
複数の連結したノードペアを含むスキャフォールドセットを得る工程であって、ノードペアの各ノードが少なくとも１つのコンティグ配列を含む、工程と、連結されたノードペアによって共有されるリードペアの数を数える工程と、上記数を閾値と比較する工程と、標準的なペアエンドリード距離度数データを得る工程と、複数の連結されたノードにマッピングされたペアエンドリード情報に関するペアエンドリード度数データを、標準的なペアエンドリード距離度数データと比較する工程と、連結されたノードの切断が、標準的なペアエンドリード距離度数データにより近く近似する複数の連結されたノードにマッピングされたペアエンドリード情報のためのペアエンドリード度数データをもたらす場合に、少なくとも１つの連結したノードを切断する工程とを含む、方法。
１５２．　スキャフォールドセットはゲノムを含む、列挙された実施形態１５１に記載の方法。
１５３．　スキャフォールドセットは複数のゲノムを含む、列挙された実施形態１５１に記載の方法。

１５４．スキャフォールドのアセンブリの方法であって、
１セットのコンティグを得る工程と、ペアエンドリードを含む入力データを得る工程であって、ペアエンドリードの少なくとも１％が少なくとも１ｋｂのリードペア距離を含み、ペアエンドリードのセットが自然配向のペアエンドリードを含み、リードペアのための配列決定のエラー率が０．１％以下であり、および入力データのＲＮ５０が組み立てられたスキャフォールドの２０％以下である、工程と、
スキャフォールドを出力する工程であって、スキャフォールドのＲＮ５０が入力のＲＮ５０の少なくとも２倍である、工程を含む、方法。
１５５．　スキャフォールドのＲＮ５０が入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態１５４に記載の方法。
１５６．　スキャフォールドのアセンブリの方法であって、
Ｔ０コンティグ配列を含むコンティグ配列のセットを得る工程と、ペアエンドリードのセットを得る工程であって、ペアエンドリードの少なくとも１％が少なくとも１ｋｂのリードペア距離を含み、ペアエンドリードのセットが自然配向のペアエンドリードを含み、リードペアのための配列決定のエラー率が０．１％以下である、工程と、Ｔ１コンティグ配列を含むスキャフォールドを出力する工程であって、Ｔ１＜Ｔ０である、工程を含む、方法。
１５７．　Ｔ１は３未満である、列挙された実施形態１５６に記載の方法。
１５８．　Ｔ１はＴ０の１０％未満である、列挙された実施形態１５６に記載の方法。
１５９．　Ｔ１はＴ０の１％未満である、列挙された実施形態１５６に記載の方法。
１６０．　コンティグ配列のセットはゲノムを含む、列挙された実施形態１５６－１５９のいずれか１つに記載の方法。
１６１．　コンティグ配列のセットは複数のゲノムを含む、列挙された実施形態１５６－１５９のいずれか１つに記載の方法。
１６２．　核酸配列データ処理の方法であって、
リードペアを含む入力データを受け取る工程であって、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離した２つの核酸セグメントからの配列データを含み、入力データ用のＲＮ５０がアセンブルしたスキャフォールドの２０％以下であり、および、上記入力データのエラー率が０．１％以下である、工程と、
スキャフォールドを含む出力データを出力する工程であって、出力データのＲＮ５０が入力のＲＮ５０の少なくとも２倍である、工程を含む、方法。
１６３．　出力データのＲＮ５０が入力のＲＮ５０少なくとも１０倍である、列挙された実施形態１６２に記載の方法。
１６４．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態１６２に記載の方法。
１６５．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態１６２に記載の方法。
１６６．　コンティグ配列のセットはゲノムを含む、列挙された実施形態１６２－１６５のいずれか１つに記載の方法。
１６７．　コンティグ配列のセットは複数のゲノムを含む、列挙された実施形態１６２－１６５のいずれか１つに記載の方法。
１６８．　核酸配列データ処理の方法であって、
リードペアを含む前処理されたデータセットを出力する工程であって、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含み、前処理されたデータセットのＲＮ５０がアセンブルされたスキャフォールドの２０％以下であり、および上記出力データの誤り率が０．１％以下である、工程と、
スキャフォールドを含む処理されたデータセットを受け取る工程であって、出力データのＲＮ５０が入力のＲＮ５０の少なくとも２倍である、工程を含む、方法。
１６９．　出力データのＲＮ５０が入力のＲＮ５０の少なくとも１０倍である、列挙され

た実施形態１６８に記載の方法。
１７０．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態１６８に記載の方法。
１７１．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態１６８に記載の方法。
１７２．　上記リードペアの少なくとも１０％は少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含む、列挙された実施形態１６８に記載の方法。
１７３．　核酸配列データ処理を方法であって、
リードペアを含む入力データを受け取る工程であって、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離した２つの核酸セグメントからの配列データを含み、入力データのＲＮ５０がアセンブルしたスキャフォールドの２０％以下であり、および、上記出力データのエラー率が０．１％以下である、工程と、
および、スキャフォールドを含む出力データを出力する工程であって、出力データのＮ５０が入力のＮ５０少なくとも２倍である、工程を含む、方法。
１７４．　出力データのＮ５０は入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態１７３に記載の方法。
１７５．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態１７３に記載の方法。
１７６．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態１７３に記載の方法。
１７７．　上記リードペアの少なくとも１０％は少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含む、列挙された実施形態１７３に記載の方法。
１７８．　核酸配列データ処理を含む方法であって、
リードペアを含む前処理されたデータセットを出力する工程であって、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離した２つの核酸セグメントからの配列データを含み、入力データのＲＮ５０がアセンブルしたスキャフォールドの２０％以下であり、および、上記出力データのエラー率が０．１％以下である、工程と、
および、スキャフォールドを含む処理されたデータセットを受け取る工程であって、処理データのＮ５０が前処理されたデータセットのＮ５０の少なくとも２倍である、工程を含む、方法。
１７９．　出力データのＮ５０は入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態１７８に記載の方法。
１８０．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態１７８に記載の方法。
１８１．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態１７８に記載の方法。
１８２．　上記リードペアの少なくとも１０％は少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含む、列挙された実施形態１７８に記載の方法。
１８３．　核酸配列データはゲノムに由来する、列挙された実施形態１７８－１８２のいずれか１つに記載の方法。
１８４．　核酸配列データは複数のゲノムに由来する、列挙された実施形態１７８－１８２のいずれか１つに記載の方法。
１８５．　少なくとも１つのペアエンドリードを共有する２つの核酸コンティグを連結する尤度を評価する方法であって、
第１のコンティグに対するマッピングされたショットガンリードの密度を判定する工程と、第２のコンティグに対するマッピングされたショットガンリードの密度を判定する工程

と、第１のコンティグと第２のコンティグを連結するための尤度スコアを判定する工程と、第１のコンティグに対するマッピングされたショットガンリードの密度が第２のコンティグに対するマッピングされたショットガンリードの密度と大きく異なるときに、尤度スコアを減少させる工程とを含む、方法。
１８６．　尤度スコアは対数尤度スコアである、列挙された実施形態１８５に記載の方法。
１８７．　尤度スコアは本明細書で示される通りに減少する、列挙された実施形態１８５に記載の方法。
１８８．　尤度スコアは、第１のコンティグに対するマッピングされたショットガンリードの密度と第２のコンティグに対するマッピングされたショットガンリードの密度の小さい方対大きい方の比率として減少する、列挙された実施形態１８５に記載の方法。
１８９．　２つの核酸コンティグは異質的なサンプルに由来する、列挙された実施形態１８５に記載の方法。
１９０．　２つの核酸コンティグはメタゲノミクスサンプルに由来する、列挙された実施形態１８５に記載の方法。
１９１．　２つの核酸コンティグは別々の個々の生体に由来する、列挙された実施形態１８５に記載の方法。
１９２．　２つの核酸コンティグは別々の種に由来する、列挙された実施形態１８５に記載の方法。
１９３．　プロセッサを含む、コンピュータは、核酸配列情報のコンティグをスキャフォールドするためのコンピュータにより実施されるシステムであって、当該プロセッサが、
初期の構成を有するコンティグ配列のセットを受け取り、
ペアエンドリードのセットを受け取り、
標準的なペアエンドリード距離度数データを受け取り、
少なくとも１つのペアエンドリード中に共存する配列を共有するコンティグペアがグループ分けされるように、コンティグペアを処理し、
コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データが、初期の構成のコンティグ配列のリードペア度数データに対して標準的なペアエンドリード距離度数データにより近く近似するように、グループ化されたコンティグ配列をスキャフォールドし、および、
ネットワーク、スクリーン、またはサーバーへ処理されたコンティグスキャフォールドを出力する、ように構成される、システム。
１９４．　スキャフォールドすることはコンティグのセットを順序付けることを含む、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
１９５．　スキャフォールドすることはコンティグのセットを配向することを含む、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
１９６．　スキャフォールドすることは少なくとも２つのコンティグの端から端を融合することを含む、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
１９７．　スキャフォールドすることは第１のコンティグを第２のコンティグに挿入することを含む、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
１９８．　スキャフォールドすることはコンティグを少なくとも２つの構成コンティグへと切断することを含む、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
１９９．　標準的なペアエンドリード度数は　両方のリードが共通のコンティグにマッピングされるペアエンドリードから得られる、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
２００．　標準的なペアエンドリード度数は、事前に生成された曲線から得られる、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。

２０１．　初期の構成は無作為の構成である、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
２０２．　コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データは、リードペア距離尤度が増大するときにペアエンドリード距離度数データにより近く近似する、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
２０３．　リードペア距離尤度は最大化される、列挙された実施形態２０２に記載のコンピュータにより実施されるシステム。
２０４．　コンティグを分離するためにマッピングされるリードペアのリードペア距離度数データは、リードペアの距離度数データと標準的なペアエンドリード距離度数データとの間の差の統計的尺度が減少する際に、ペアエンドリード距離度数データにより近く近似する、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
２０５．　リードペア距離度数データと標準的なペアエンドリード距離度数データとの間の距離の統計的尺度は、ＡＮＯＶＡ、ｔ検定、およびＸ２乗検定の少なくとも１つを含む、列挙された実施形態２０４に記載のコンピュータにより実施されるシステム。
２０６．　コンティグを分離するためにマッピングされるリードペアのリードペア距離は、標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差が減少するときに、ペアエンドリード距離度数データと厳密に一致する、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
２０７．　標準的なペアエンドリード距離度数と比較して得られた順序付けされたコンティグ中のリードペア距離分布の偏差は最小限に抑えられる、列挙された実施形態２０６に記載のコンピュータにより実施されるシステム。
２０８．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグは、より多くの共有される末端リードを有するクラスターに割り当てられる、列挙された実施形態１９３に記載のコンピュータにより実施されるシステム。
２０９．　クラスター化は、生体中の染色体の数以上である多くのグループにコンティグをグループ分けすることを含む、列挙された実施形態１９３－２０８のいずれか１つに記載のコンピュータにより実施されるシステム。
２１０．　単一のペアエンドリードのみをクラスターの1つのコンティグと共有するコンティグは、そのクラスターには含まれない、列挙された実施形態１９３－２０９のいずれか１つに記載のコンピュータにより実施されるシステム。
２１１．　反復配列を含む少なくとも１つのペアエンドリードのみをクラスターと共有するコンティグは、そのクラスターには含まれない、列挙された実施形態１９３－２１０のいずれか１つに記載のコンピュータにより実施されるシステム。
２１２．　低品質の配列を含む少なくとも１つのペアエンドリードのみをクラスターと共有するコンティグは、そのクラスターには含まれない、列挙された実施形態１９３－２１１のいずれか１つに記載のコンピュータにより実施されるシステム。
２１３．　ペアエンドリードのセットは、核酸内で内部二本鎖切断を生じさせるためにサンプルＤＮＡを消化し、少なくとも１つの再連結接合部を形成するために二本鎖切断を再連結し、および少なくとも１つの再連結接合部にわたって配列決定することにより得られる、列挙された実施形態１９３－２１２のいずれか１つに記載のコンピュータにより実施されるシステム。
２１４．　ＤＮＡは少なくとも１つのＤＮＡ結合剤に架橋される、列挙された実施形態２１３に記載のコンピュータにより実施されるシステム。
２１５．　ＤＮＡは分離された裸のＤＮＡである、列挙された実施形態２１３に記載のコンピュータにより実施されるシステム。
２１６．　分離されたＤＮＡは再構成された染色質へ再度アセンブルされる、列挙された実施形態２１４に記載のコンピュータにより実施されるシステム。
２１７．　再構成された染色質は架橋される、列挙された実施形態２１６に記載のコンピュータにより実施されるシステム。

２１８．　再構成された染色質はＤＮＡ結合タンパク質を含む、列挙された実施形態２１６に記載のコンピュータにより実施されるシステム。
２１９．　再構成された染色質はナノ粒子を含む、列挙された実施形態２１６に記載のコンピュータにより実施されるシステム。
２２０．　コンティグのクラスター化は生体に関して数または染色体とは無関係である、列挙された実施形態１９３－２１９のいずれか１つに記載のコンピュータにより実施されるシステム。
２２１．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグは、より多くの数の共有される末端リードを有するクラスターに割り当てられる、列挙された実施形態１９３－２２０のいずれか１つに記載のコンピュータにより実施されるシステム。
２２２．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグは、より大きなリードペア距離尤度値を有するクラスターに割り当てられる、列挙された実施形態１９３－２２０のいずれか１つに記載のコンピュータにより実施されるシステム。
２２３．　第１のクラスターと第２のクラスターに関連するペアエンドリード中の配列を共有するコンティグは、標準的なリードペア距離分布と比較してそのリードペア分布中で偏差が少ないクラスターに割り当てられる、列挙された実施形態１９３－２２０のいずれか１つに記載のコンピュータにより実施されるシステム。
２２４．　第１のクラスターと第２のクラスターに関連したペアエンドリードの中の配列を共有するコンティグは、各クラスターから除外される、列挙された実施形態１９３－２２１のいずれか１つに記載のコンピュータにより実施されるシステム。
２２５．　クラスター化は、生体中の染色体の数以上である多くのグループにコンティグを分類することを含む、列挙された実施形態１９３－２２４のいずれか１つに記載のコンピュータにより実施されるシステム。
２２６．　スキャフォールドすることは、クラスター化されたコンティグの推定上の隣接するコンティグの第１のセットを選択すること、上記リードペアに関してリードペアの距離の総計尺度を減少させる推定上の隣接するコンティグの第１のセットの最小限の距離順序を決定すること、および、リードペア距離の総計尺度を減少させるために、推定上の隣接するコンティグの第１のセットをスキャフォールドすることを含む、列挙された実施形態１９３－２２５のいずれか１つに記載のコンピュータにより実施されるシステム。
２２７．　推定上の隣接するコンティグの第１のセットが２つのコンティグからなる、列挙された実施形態２２６に記載のコンピュータにより実施されるシステム。
２２８．　推定上の隣接するコンティグの第１のセットが３つのコンティグからなる、列挙された実施形態２２６に記載のコンピュータにより実施されるシステム。
２２９．　推定上の隣接するコンティグの第１のセットが４つのコンティグからなる、列挙された実施形態２２６に記載のコンピュータにより実施されるシステム。
２３０．　推定上の隣接するコンティグの第１のセットが４つのコンティグを含む、列挙された実施形態２２６に記載のコンピュータにより実施されるシステム。
２３１．　スキャフォールドすることは、推定上の隣接するコンティグの第１のセットにおける各コンティグの順序と配向を決定することを含む、列挙された実施形態２２６に記載のコンピュータにより実施されるシステム。
２３２．　最小限の距離順序を決定することは、すべての可能なコンティグ構成についての上記の設定の２つのコンティグに対するリードマッピングを含む少なくとも１つのリードペアのための予想されるリードペア距離を比較することを含む、列挙された実施形態２２６－２２７のいずれか１つに記載のコンピュータにより実施されるシステム。
２３３．　上記のリードペアの最小限のリードペア距離に対応するコンティグ配向を選択することをさらに含む、列挙された実施形態２３２に記載のコンピュータにより実施されるシステム。
２３４．　最大尤度のリードペア距離分布に対応するコンティグ配向を選択することをさ

らに含む、列挙された実施形態２３２に記載のコンピュータにより実施されるシステム。
２３５．　上記のコンティグクラスターのリードペアの総計尺度のための最小限のリードペア距離に対応するコンティグ配向を選択することをさらに含む、列挙された実施形態２３２－２３３のいずれか１つに記載のコンピュータにより実施されるシステム。
２３６．　予想されるリードペア距離は、上記ペアエンドリード距離度数データと比較される、列挙された実施形態２３２－２３５のいずれか１つに記載のコンピュータにより実施されるシステム。
２３７．　上記ペアエンドリード距離度数データと比較することは、式１を用いることを含む、列挙された実施形態２３６に記載のコンピュータにより実施されるシステム。
２３８．　上記クラスター化されたコンティグの推定上の隣接するコンティグの第２のセットを選択することであって、上記第２のセットが上記第１のセットの１つのエンド末端のコンティグを除くすべてと、上記クラスター化されたコンティグの１つの追加のコンティグを含む、ことと、リードペアの距離の上記総計尺度を減少させるために、推定上の隣接するコンティグの第２のセットをスキャフォールドすることとをさらに含む、列挙された実施形態２２６－２３７のいずれか１つに記載のコンピュータにより実施されるシステム。
２３９．　上記クラスター化されたコンティグの推定上の隣接するコンティグの第３のセットを選択することであって、上記第３のセットが上記第２のセットの１つのエンド末端のコンティグを除くすべてと、第１のセットと第２のセットに含まれない上記クラスター化されたコンティグの１つの追加のコンティグを含む、ことと、リードペアの距離の上記総計尺度を減少させるために、推定上の隣接するコンティグの第３のセットをスキャフォールドすることとをさらに含む、列挙された実施形態２３８に記載のコンピュータにより実施されるシステム。
２４０．　上記クラスター化されたコンティグの大部分が順序付けされるまで、少なくとも１つの追加のセットを繰り返し選択することをさらに含む、列挙された実施形態２３９に記載のコンピュータにより実施されるシステム。
２４１．　上記クラスター化されたコンティグの各々が順序付けされるまで、少なくとも１つの追加のセットを繰り返し選択することをさらに含む、列挙された実施形態２４０に記載のコンピュータにより実施されるシステム。
２４２．　核酸配列はゲノムに由来する、列挙された実施形態１９３－２４１のいずれか１つに記載のコンピュータにより実施されるシステム。
２４３．　核酸配列は複数のゲノムに由来する、列挙された実施形態１９３－２４１のいずれか１つに記載のコンピュータにより実施されるシステム。
２４４．　クラスター中のコンティグをスキャフォールドするためのコンピュータにより実施されるシステムであって、該システムはプロセッサを含み、プロセッサはコンティグのセットを受け取り、
ａ）各ペアのコンティグについて対数尤度比スコアを割り当て、
ｂ）対数尤度比スコアによって接続を分類し、
および、ｃ）アセンブリの合計スコアを増加させるために、尤度比スコアの次数を減少させる際にコンティグ接続を受け入れるかまたは拒否することによって、上記コンティグを処理し、
ならびに、ネットワーク、スクリーン、またはサーバーへ処理されたコンティグスキャフォールドを出力するように構成される、コンピュータにより実施されるシステム。
２４５．　スキャフォールドすることはコンティグのセットを順序付けることを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。
２４６．　スキャフォールドすることはコンティグのセットを配向させることを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。
２４７．　スキャフォールドすることは少なくとも２つのコンティグの端から端を融合することを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。

２４８．　スキャフォールドすることは１つのコンティグを第２のコンティグに挿入することを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。
２４９．　スキャフォールドすることはコンティグを少なくとも２つの構成コンティグへと切断することを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。
２５０．　コンティグはゲノムを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。
２５１．　コンティグは複数のゲノムを含む、列挙された実施形態２４４に記載のコンピュータにより実施されるシステム。
２５２．　クラスター内の複数のコンティグの局所的に最適なコンティグ構成を決定するためのコンピュータにより実施されるシステムであって、コンピュータにより実施されるシステムはプロセッサを含み、プロセッサは、
コンティグのセットを受け取り；
ａ）コンティグのクラスターに沿った位置ｉから始まるサイズｗのコンティグの配列ウインドウを特定すること、
ｂ）ウインドウ中の各位置ｉにおける適合性のある順序と配向のスコアを調べることによりウインドウｗのコンティグ用のｗ！２^{w}順序付けと配向のオプションを考慮することと、
ｃ）最適なスコアを得るために、上記ウインドウ中の上記ｗコンティグを配向して順序付けすることと、
ｄ）ｉ＋１の位置を決めるためにウインドウを変えることと、および、
ｅ）最適なスコアを決定するために、上記ｗコンティグの配向と順序付けを用いて、位置ｉ＋１で上記ウインドウについて工程（ａ）、（ｂ）、および（ｃ）を繰り返すこと、によってコンティグの上記セットを処理し、
それにより、スコアに対する局所的に最適な構成中の上記複数のコンティグを配向および順序付けし；
ならびに、ネットワーク、スクリーン、またはサーバーに局所的に最適な構成を出力する、ように構成される、コンピュータにより実施されるシステム。
２５３．　クラスター中の複数のコンティグにマッピングされるリードペアデータが得られ、標準的なペアエンドリード度数データセットが得られ、および、上記のｗコンティグを配向させ、および順序付けるためのスコアは、クラスター中の複数のコンティグにマッピングされるリードペアデータのためのリードペア距離データセットが標準的なペアエンドリード度数データセットにどれだけより近く一致するかの尺度である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２５４．　クラスター中の複数のコンティグにマッピングされるリードペアデータが得られ、スコアは完全なリードペア距離であり、および、完全なリードペア距離が最小限に抑えられるとき、スコアは最適化される、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２５５．　ｗは３である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２５６．　ｗは４である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２５７．　ｗは５である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２５８．　ｗは６である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２５９．　ｗは第１のクラスターに対する第１の値を有し、ｗは第２のクラスターで第２の値を有する、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。

２６０．　ｗはセットのコンティグの１％を含むように選択される、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２６１．　ｗはセットのコンティグの５％を含むように選択される、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２６２．　ｗはセットのコンティグの１％を含むように選択される、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２６３．　スコアはリードペア距離尤度スコアであり、スコアは、所定のウィンドウサイズについて最大化されるときに最適である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２６４．　スコアは式１を使用して計算される、列挙された実施形態２６３に記載のコンピュータにより実施されるシステム。
２６５．　スコアは予想されるリードペア分布からの偏差であり、スコアは所定のウィンドウサイズについて最小限に抑えられるときに最適である、列挙された実施形態２５２に記載のコンピュータにより実施されるシステム。
２６６．　複数のコンティグはゲノムを含む、列挙された実施形態２５２－２６５に記載のいずれか１つに記載のコンピュータにより実施されるシステム。
２６７．　複数のコンティグは複数のゲノムを含む、列挙された実施形態２５２－２６５のいずれか１つに記載のコンピュータにより実施されるシステム。
２６８．　核酸配列アセンブリのための方法であって、
ａ）精製されたＤＮＡを得る工程と、
ｂ）ＤＮＡ／染色質複合体を形成するために精製されたＤＮＡをＤＮＡ結合剤に結合させる工程と、
ｃ）粘着末端を残すために制限酵素でＤＮＡ染色質複合体をインキュベートする工程と、
ｄ）ＤＮＡの末端を連結するために連結反応を行う工程と、；
ｅ）ペアエンドリードを生成するために連結されたＤＮＡ接合部全体を配列決定する工程と、および、
ｆ）ペアエンドリードを受け取るとともに処理し、ネットワーク、スクリーン、またはサーバーに、精製されたＤＮＡの配列を表すコンティグを含むスキャフォールド核酸データセットを出力するように構成されたプロセッサを含む、コンピュータにより実施されるシステムを使用する工程とを含む、方法。
２６９．　精製されたＤＮＡはゲノムに由来する、列挙された実施形態２６８に記載の方法。
２７０．　精製されたＤＮＡは複数のゲノムに由来する、列挙された実施形態２６８に記載の方法。
２７１．　反復のコンティグ領域に対するリードペアの配列リードマッピングを特定するためのコンピュータにより実施されるシステムであって、該システムはプロセッサを含み、該プロセッサは、核酸サンプル用のコンティグデータセットを受け取り、隣接していない物理的に連結した配列情報に対応する少なくとも１つのリードペアの配列リードを受け取り、および、リードペアの配列リードの少なくとも１つのリードがコンティグデータセットの２つの異なる座位にマッピングされる場合に、リードペアの配列リードを除外する、ように構成される、コンピュータにより実施されるシステム。
２７２．　反復領域は、第１の閾値を超えるショットガンリード深さを有する配列を含む、列挙された実施形態２７１に記載のコンピュータにより実施されるシステム。
２７３．　反復領域は、第２の閾値を越えるリード深さを有する塩基位置を含む、列挙された実施形態２７２に記載のコンピュータにより実施されるシステム。
２７４．　第１の閾値と第２の閾値はリード深さの全体的な分布に対して固定される、列挙された実施形態２７３に記載のコンピュータにより実施されるシステム。
２７５．　第１の閾値はリード深さの全体的な分布の３倍である、列挙された実施形態２７４に記載のコンピュータにより実施されるシステム。
２７６．　第２の閾値はリード深さの全体的な分布の３．５倍である、列挙された実施形

態２７４に記載のコンピュータにより実施されるシステム。
２７７．　核酸サンプルはゲノムを含む、列挙された実施形態２７１－２７６のいずれか１つに記載のコンピュータにより実施されるシステム。
２７８．　核酸サンプルは複数のゲノムを含む、列挙された実施形態２７１－２７６のいずれか１つに記載のコンピュータにより実施されるシステム。
２７９．　プロセッサを含む、コンティグアセンブリ決定をガイドするためのコンピュータにより実施されるシステムであって、該プロセッサは、
コンティグデータセットを受け取ること、第１のコンティグと第２のコンティグとの間にまたがるリードペア配列の数と黙示的な分離を観察する確率を決定するためにデータセットを処理することであって、コンティグがセット［＋＋、＋－、－＋、－－］内のｏの相対的な配向を有し、ギャップ長により分離される、こと、および、ネットワーク、スクリーン、またはサーバーへデータセットと決定された確率とを出力することを行うように構成される、システム。
２８０．　分離距離にわたるリードペア配列の分布の確率を正規化する工程を含み、正規化は、リードペア配列を、核酸サンプルを独立してサンプリングするノイズペアと比較することを含む、列挙された実施形態２７９に記載のコンピュータにより実施されるシステム。
２８１．　核酸サンプルはゲノムを含む、列挙された実施形態２８０に記載のコンピュータにより実施されるシステム。
２８２．　核酸サンプルは複数のゲノムを含む、列挙された実施形態２８０に記載のコンピュータにより実施されるシステム。
２８３．　ノイズペアの総数はコンティグペアのサンプルに関する連結の密度を集計することにより決定される、列挙された実施形態２８０に記載のコンピュータにより実施されるシステム。
２８４．　密度の最大と最少の１％が除外される、列挙された実施形態２８３に記載のコンピュータにより実施されるシステム。
２８５．　コンティグ順序を決定することをさらに含む、列挙された実施形態２７９に記載のコンピュータにより実施されるシステム。
２８６．　コンティグ配向を決定することをさらに含む、列挙された実施形態２７９に記載のコンピュータにより実施されるシステム。
２８７．　プロセッサを含むコンティグ誤連結部修正のためのコンピュータにより実施されるシステムであって、該プロセッサは、
初期の構成を有するコンティグ配列のセットを受け取り、
ペアエンドリードのセットを受け取り、
標準的なペアエンドリード距離度数データを受け取り、
少なくとも１つのペアエンドリード中に共存する配列を共有するグループへとコンティグペアを処理し、
標準的なペアエンドリード距離度数データと比較することと、および、グループ分けのコンティグ中に切断を導入することにより、コンティグのグループ分けのためのリードペア度数データが標準的なペアエンドリード距離度数データにより近く近似するかどうかを判定することとによって、コンティグのグループ分けに関してリードペア度数データを処理し、および、
もし近似する場合に、上記切断を導入し、ならびに、
ネットワーク、スクリーン、またはサーバーへ処理されたコンティグデータセットを出力するように構成される、システム。
２８８．　第１の位置は、切断を導入する前に、上記閾値以下の対数尤度を有する少なくとも１つの隣接する第２の位置と融合される、列挙された実施形態２８７に記載のコンピュータにより実施されるシステム。
２８９．　第２の隣接する位置は、第１の位置からの３００以下の塩基対である、列挙された実施形態２８７に記載のコンピュータにより実施されるシステム。

２９０．　第２の位置は、第１の位置からの１０００を超える塩基ペアの位置を含まない、列挙された実施形態２８７に記載のコンピュータにより実施されるシステム。
２９１．　対数尤度の変化の判定は、コンティグの平均的なペアエンドマッピング密度を特定すること、コンティグのセグメントを特定すること、平均的なペアエンドマッピング密度の少なくとも３倍のペアエンドマッピング密度を有するコンティグのセグメンを特定すること、および、平均的なペアエンドマッピング密度の少なくとも３倍のペアエンドマッピング密度を有するコンティグのセグメンを除外することを含む、列挙された実施形態２８７－２９０のいずれか１つに記載のコンピュータにより実施されるシステム。
２９２．　コンティグ配列のセットはゲノムに由来する、列挙された実施形態２８７－２９１のいずれか１つに記載のコンピュータにより実施されるシステム。
２９３．　コンティグ配列のセットは複数のゲノムに由来する、列挙された実施形態２８７－２９１のいずれか１つに記載のコンピュータにより実施されるシステム。
２９４．　プロセッサを含む、コンティグアセンブリのためのコンピュータにより実施されるシステムであって、該プロセッサは、コンティグのセットを受け取り、かつ、以下の工程：
開始アセンブリの切断したコンティグを示す工程であって、切断されたコンティグがノードであり、切断されたコンティグのエッジが整数の順序付けられたペアのリストで標識され、切断のエッジはマッピングされたリードペア配列に対応する、工程、
および、マッピングされた接続の閾値数よりも少ないエッジを除外する工程、によってコンティグの上記セットを処理するように構成される、システム。
２９５．　閾値数は５％未満である、列挙された実施形態２９４に記載のコンピュータにより実施されるシステム。
２９６．　閾値数はｔＬリンクよりも少ない、列挙された実施形態２９４に記載のコンピュータにより実施されるシステム。
２９７．　コンティグ長さに対する対応ノードのグラフ中の程度の比が塩基対であるエッジを含むコンティグは、分布のすべての値の上限の約５％を超過する、列挙された実施形態２９４に記載のコンピュータにより実施されるシステム。
２９８．　コンティグはゲノムに由来する、列挙された実施形態２９４－２９７のいずれか１つに記載のコンピュータにより実施されるシステム。
２９９．　コンティグは複数のゲノムに由来する、列挙された実施形態２９４－２９７のいずれか１つに記載のコンピュータにより実施されるシステム。
３００．　プロセッサを含む、少なくとも１つのスキャフォールドへコンティグ配列情報をアセンブルするコンピュータにより実施されるシステムであって、該プロセッサは、
複数のコンティグに対応する配列情報を受け取ること、複数のコンティグによって表される核酸サンプルからペアエンドリード情報を受け取ること、予測されるリードペア距離情報データセットからのリードペア距離パラメータの偏差が最小限に抑えられるように複数のコンティグを構成することによって複数のコンティグを処理することであって、この構成が８時間未満で生じること、および、ネットワーク、スクリーン、またはサーバーへ最小限に抑えられた偏差を含む上記構成されたコンティグを出力することを行うように構成される、システム。
３０１．　予測されたリードペア距離情報データセットはリードペア距離尤度曲線を含む、列挙された実施形態３００に記載のコンピュータにより実施されるシステム。
３０２．　リードペア距離パラメータはリードペア距離尤度曲線に対して最大距離尤度である、列挙された実施形態３００－３０１に記載のいずれか１つに記載のコンピュータにより実施されるシステム。
３０３．　リードペア距離パラメータはリードペア距離尤度曲線に対して最小限の変動である、列挙された実施形態３００－３０１のいずれか１つに記載のコンピュータにより実施されるシステム。
３０４．　コンティグの局所的に隣接するセットは２つのコンティグを含む、列挙された実施形態３００－３０３のいずれか１つに記載のコンピュータにより実施されるシステム

。
３０５．　コンティグの局所的に隣接するセットは３つのコンティグを含む、列挙された実施形態３００－３０３のいずれか１つに記載のコンピュータにより実施されるシステム。
３０６．　コンティグの局所的に隣接するセットは４つのコンティグを含む、列挙された実施形態３００－３０３のいずれか１つに記載のコンピュータにより実施されるシステム。
３０７．　コンティグの局所的に隣接するセットは５つのコンティグを含む、列挙された実施形態３００－３０３のいずれか１つに記載のコンピュータにより実施されるシステム。
３０８．　コンティグの局所的に隣接するセットは６つのコンティグを含む、列挙された実施形態３００－３０３のいずれか１つに記載のコンピュータにより実施されるシステム。
３０９．　構成は７時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１０．　構成は６時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１１．　構成は５時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１２．　構成は４時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１３．　構成は３時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１４．　構成は２時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１５．　構成は１時間未満で生じる、列挙された実施形態３００－３０８のいずれか１つに記載のコンピュータにより実施されるシステム。
３１６．　コンティグ情報はゲノムに由来する、列挙された実施形態３００－３１５のいずれか１つに記載のコンピュータにより実施されるシステム。
３１７．　コンティグ配列情報は複数のゲノムに由来する、列挙された実施形態３００－３１５のいずれか１つに記載のコンピュータにより実施されるシステム。
３１８．　プロセッサを含むコンティグ配列のセットに足場を設けるコンピュータにより実施されるシステムであって、該プロセッサは、
核酸サンプルの代表的なコンティグ配列のセットを受け取ること、核酸サンプル用のリードペアデータを受け取ること、および、核酸サンプル用のリードペアデータが予想されるリードペア分布によりより近く近似するようにコンティグのセットを順序付けて配向することにより受け取ったデータを処理することであって、コンティグ配列のセットの７０％がわずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けられ配向される、こと、
ならびに、ネットワーク、スクリーン、またはサーバーに順序付けされて配向されたデータを出力すること、を行うように構成される、システム。
３１９．　スキャフォールドすることは、コンティグのセットを順序付けることを含む、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２０．　スキャフォールドすることはコンティグのセットを配向することを含む、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２１．　スキャフォールドすることは少なくとも２つのコンティグの端から端を融合することを含む、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２２．　スキャフォールドすることは１つのコンティグを第２のコンティグに挿入することを含む、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム

。
３２３．　スキャフォールドすることはコンティグを少なくとも２つの構成コンティグへと切断することを含む、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２４．　コンティグ配列のセットの８０％は、わずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように、順序付けされ、および配向される、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２５．　コンティグ配列のセットの９０％は、わずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２６．　コンティグ配列のセットの９５％は、わずか８時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２７．　コンティグ配列のセットの７０％は、わずか４時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２８．　コンティグ配列のセットの７０％は、わずか２時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３２９．　コンティグ配列のセットの７０％は、わずか１時間で核酸サンプル中のそれらの配列の相対的な順序と配向に一致するように順序付けされ、および配向される、列挙された実施形態３１８に記載のコンピュータにより実施されるシステム。
３３０．　コンティグ配列のセットはゲノムに由来する、列挙された実施形態３１８－３２９のいずれか１つに記載のコンピュータにより実施されるシステム。
３３１．　コンティグ配列のセットは複数のゲノムに由来する、列挙された実施形態３１８－３２９のいずれか１つに記載のコンピュータにより実施されるシステム。
３３２．　プロセッサを含む核酸配列データのセットを構成するコンピュータにより実施されるシステムであって、該プロセッサは、
複数のコンティグに対応する配列情報を受け取ること、ペアエンドリード情報を受け取ること、ペアエンドリード情報のためのペアエンドリード距離分布が基準のペアエンドリード距離分布に全体的に最適化されるように、複数のコンティグを構成することによって受信されたデータを処理することであって、構成がわずか８時間で生じる、こと、および、ネットワーク、スクリーン、またはサーバーに上記構成された複数のコンティグを出力することを行うように構成される、システム。
３３３．　核酸配列データのセットはゲノムに由来する、列挙された実施形態３３２に記載のコンピュータにより実施されるシステム。
３３４．　核酸配列データのセットは複数のゲノムに由来する、列挙された実施形態３３２に記載のコンピュータにより実施されるシステム。
３３５．　構成はわずか４時間で生じる、列挙された実施形態３３２に記載のコンピュータにより実施されるシステム。
３３６．　構成はわずか２時間で生じる、列挙された実施形態３３２に記載のコンピュータにより実施されるシステム。
３３７．　プロセッサを含むスキャフォールドアセンブリを改善するコンピュータにより実施されるシステムであって、該プロセッサは、
複数の連結したノードペアを含むスキャフォールドセットを受け取ることであって、ノードペアの各ノードが少なくとも１つのコンティグ配列を含む、こと、複数の連結したノードにマッピングされたペアエンドリード情報を受け取ること、連結したノードペアによって共有されるリードペアの数を数え、上記数を閾値と比較し、および、上記の数が閾値以下である場合にノードペアを連結していないノードに切断することとることにより受信データを処理すること、ならびに、ネットワーク、スクリーン、またはサーバーに上記処理さ

れたデータを出力することを行うように構成される、システム。
３３８．　固有のコンティグ配列にマッピングされたリードペアだけが数えられる、列挙された実施形態３３７に記載のコンピュータにより実施されるシステム。
３３９．　異なるペアエンド閾値数がマッピングされるコンティグ配列セグメントに対するリードペアのマッピングは廃棄される、列挙された実施形態３３７に記載のコンピュータにより実施されるシステム。
３４０．　閾値数は非反復配列について平均数の３倍である、列挙された実施形態３３７に記載のコンピュータにより実施されるシステム。
３４１．　スキャフォールドセットはゲノムを含む、列挙された実施形態３３７－３４０のいずれか１つに記載のコンピュータにより実施されるシステム。
３４２．　スキャフォールドセットは複数のゲノムを含む、列挙された実施形態３３７－３４０のいずれか１つに記載のコンピュータにより実施されるシステム。
３４３．　プロセッサを含むスキャフォールドアセンブリを改善するコンピュータにより実施されるシステムであって、該プロセッサは、
複数の連結したノードペアを含むスキャフォールドセットを受け取ることであって、ノードペアの各ノードが少なくとも１つのコンティグ配列を含む、こと、複数の連結したノードにマッピングされたペアエンドリード情報を受け取ること、標準的なペアエンドリード距離度数データを受け取ること、複数の連結されたノードにマッピングされたペアエンドリード情報用のペアエンドリード度数データを、標準的なペアエンドリード距離度数データと比較し、および、連結されたノードの切断が、標準的なペアエンドリード距離度数データにより近く近似する複数の連結されたノードにマッピングされたペアエンドリード情報のためのペアエンドリード度数データをもたらす場合に、少なくとも１つの連結したノードを切断することによって、受信されたデータを処理すること、ならびに、ネットワーク、スクリーン、またはサーバーに処理されたデータを出力することを行うように構成される、システム。
３４４．　スキャフォールドセットはゲノムを含む、列挙された実施形態３４３に記載のコンピュータにより実施されるシステム。
３４５．　スキャフォールドセットは複数のゲノムを含む、列挙された実施形態３４３に記載のコンピュータにより実施されるシステム。
３４６．　プロセッサを含むスキャフォールドアセンブリのコンピュータにより実施されるシステムであって、該プロセッサは、
コンティグ配列のセットを受け取ること、ペアエンドリードのセットを含む入力されたデータを受け取ることであって、ペアエンドリードの少なくとも１％が少なくとも１ｋｂのリードペア距離を含み、ペアエンドリードのセットが自然配向のペアエンドリードを含み、リードペアのための配列決定のエラー率が０．１％以下であり、および入力データのＲＮ５０が組み立てられたスキャフォールドの２０％以下である、こと、ならびに、スキャフォールドを出力することであって、スキャフォールドのＲＮ５０が入力のＲＮ５０の少なくとも２倍である、ことを行うように構成される、システム。
３４７．　スキャフォールドのＲＮ５０が入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態３４６に記載のコンピュータにより実施されるシステム。
３４８．　プロセッサを含むスキャフォールドアセンブリのコンピュータにより実施されるシステムであって、該プロセッサは、
Ｔ０コンティグ配列を含むコンティグ配列のセットを受け取ること、ペアエンドリードのセットを受け取ることであって、ペアエンドリードの少なくとも１％が少なくとも１ｋｂのリードペア距離を含み、ペアエンドリードのセットが自然配向のペアエンドリードを含み、リードペアのための配列決定のエラー率が０．１％以下である、こと、およびＴ１コンティグ配列を含むスキャフォールドを出力することであって、Ｔ１＜Ｔ０である、ことを行うように構成される、システム。
３４９．　Ｔ１は３未満である、列挙された実施形態３４８に記載のコンピュータにより実施されるシステム。

３５０．　Ｔ１はＴ０の１０％未満である、列挙された実施形態３４８に記載のコンピュータにより実施されるシステム。
３５１．　Ｔ１はＴ０の１％未満である、列挙された実施形態３４８に記載のコンピュータにより実施されるシステム。
３５２．　コンティグ配列のセットはゲノムを含む、列挙された実施形態３４８－３５１のいずれか１つに記載のコンピュータにより実施されるシステム。
３５３．　コンティグ配列のセットは複数のゲノムを含む、列挙された実施形態３４８－３５１のいずれか１つに記載のコンピュータにより実施されるシステム。
３５４．　プロセッサを含む、核酸配列データ処理のコンピュータにより実施されるシステムであって、該プロセッサは、
リードペアを含む入力データを受け取ること、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離した２つの核酸セグメントからの配列データを含み、入力データ用のＲＮ５０がアセンブルしたスキャフォールドの２０％以下であり、および、上記入力データのエラー率が０．１％以下である、こと、
および、スキャフォールドを含む出力データを出力することであって、出力データ用のＲＮ５０が入力のＲＮ５０の少なくとも２倍である、ことを行うように構成される、システム。
３５５．　出力データのＲＮ５０が入力のＲＮ５０少なくとも１０倍である、列挙された実施形態３５４に記載のコンピュータにより実施されるシステム。
３５６．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態３５４に記載のコンピュータにより実施されるシステム。
３５７．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態３５４に記載のコンピュータにより実施されるシステム。
３５８．　コンティグ配列のセットはゲノムを含む、列挙された実施形態３５４－３５７のいずれか１つに記載のコンピュータにより実施されるシステム。
３５９．　コンティグ配列のセットは複数のゲノムを含む、列挙された実施形態３５４－３５７のいずれか１つに記載のコンピュータにより実施されるシステム。
３６０．　プロセッサを含む核酸配列データ処理のコンピュータにより実施されるシステムであって、該プロセッサは、
リードペアを含む前処理されたデータセットを出力することであって、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含み、前処理されたデータセットのＲＮ５０がアセンブルされたスキャフォールドの２０％以下であり、および上記出力データの誤り率が０．１％以下である、ことと、
および、スキャフォールドを含む処理されたデータセットを受け取ることであって、出力データ用のＲＮ５０が入力のＲＮ５０の少なくとも２倍である、ことを行うように構成される、システム。
３６１．　出力データのＲＮ５０が入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態３６０に記載のコンピュータにより実施されるシステム。
３６２．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態３６０に記載のコンピュータにより実施されるシステム。
３６３．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態３６０に記載のコンピュータにより実施されるシステム。
３６４．　上記リードペアの少なくとも１０％は少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含む、列挙された実施形態３６０に記載のコンピュータにより実施されるシステム。

３６５．　プロセッサを含む、核酸配列データ処理のコンピュータにより実施されるシステムであって、該プロセッサは、
リードペアを含む入力データを受け取ること、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離した２つの核酸セグメントからの配列データを含み、入力データのＲＮ５０がアセンブルしたスキャフォールドの２０％以下であり、および、上記出力データのエラー率が０．１％以下である、ことと、
および、スキャフォールドを含む出力データを出力することであって、出力データのＮ５０が入力のＮ５０少なくとも２倍である、ことを行うように構成される、システム。
３６６．　出力データのＮ５０は入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態３６５に記載のコンピュータにより実施されるシステム。
３６７．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態３６５に記載のコンピュータにより実施されるシステム。
３６８．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態３６５に記載のコンピュータにより実施されるシステム。
３６９．　上記リードペアの少なくとも１０％は少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含む、列挙された実施形態３６５に記載のコンピュータにより実施されるシステム。
３７０．　プロセッサを含む核酸配列データ処理のコンピュータにより実施されるシステムであって、該プロセッサは、
リードペアを含む前処理されたデータセットを出力することであって、上記リードペアの少なくとも１％が少なくとも１ｋｂによっておよび自然配向に分離した２つの核酸セグメントからの配列データを含み、入力データのＲＮ５０がアセンブルしたスキャフォールドの２０％以下であり、および、上記出力データのエラー率が０．１％以下である、ことと、
スキャフォールドを含む、処理されたデータセットを受け取ることであって、処理データのＮ５０が前処理されたデータセットのＮ５０の少なくとも２倍である、ことを行うように構成される、システム。
３７１．　出力データのＮ５０は入力のＲＮ５０の少なくとも１０倍である、列挙された実施形態３７０に記載のコンピュータにより実施されるシステム。
３７２．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９０％を含む、列挙された実施形態３７０に記載のコンピュータにより実施されるシステム。
３７３．　スキャフォールドは正確な順序と配向の標的ゲノムのサンプル配列の少なくとも９９％を含む、列挙された実施形態３７０に記載のコンピュータにより実施されるシステム。
３７４．　上記リードペアの少なくとも１０％は少なくとも１ｋｂによっておよび自然配向に分離された２つの核酸セグメントからの配列データを含む、列挙された実施形態３７０に記載のコンピュータにより実施されるシステム。
３７５．　核酸配列データはゲノムに由来する、列挙された実施形態３７０－３７４のいずれか１つに記載のコンピュータにより実施されるシステム。
３７６．　核酸配列データは複数のゲノムに由来する、列挙された実施形態３７０－３７４のいずれか１つに記載のコンピュータにより実施されるシステム。
３７７．　少なくとも１つのペアエンドリードを共有する２つの核酸コンティグを連結する尤度を評価するコンピュータにより実施されるシステムであって、該プロセッサは、
コンティグのセットを受け取ることと、
第１のコンティグに対するマッピングされたショットガンリードの密度を判定し、第２のコンティグに対するマッピングされたショットガンリードの密度を判定し、第１のコンティグと第２のコンティグを連結するために尤度スコアを決定し、および、第１のコンティ

グに対するマッピングされたショットガンリードの密度が第２のコンティグに対するマッピングされたショットガンリードの密度と大きく異なるときに、尤度スコアを減少させることにより、コンティグの上記セットを処理することと、ならびに、ネットワーク、スクリーン、またはサーバーへのコンティグの処理されたセットを出力することを行うように構成される、システム。
３７８．　尤度スコアは対数尤度スコアである、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
３７９．　尤度スコアは本明細書で示される通りに減少する、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
３８０．　尤度スコアは、第１のコンティグに対するマッピングされたショットガンリードの密度と第２のコンティグに対するマッピングされたショットガンリードの密度の小さい方対大きい方の比率として減少する、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
３８１．　２つの核酸コンティグは異質的なサンプルに由来する、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
３８２．　２つの核酸コンティグはメタゲノミクスサンプルに由来する、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
３８３．　２つの核酸コンティグは別々の個々の生体に由来する、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
３８４．　２つの核酸コンティグは別々の種に由来する、列挙された実施形態３７７に記載のコンピュータにより実施されるシステム。
【０２０３】
実施例
実施例１：再構築された染色質から生成されたペアリードを用いるゲノム解読
５．５μｇの高分子量ＤＮＡが、ヒト細胞株ＧＭ１２８７８と野生の捕獲されたアメリカアリゲーターの血液から抽出された。高分子量ＤＮＡは約１５０Ｋｂｐの断片中で抽出された。染色質は、精製されたヒストンおよび染色質アセンブリ因子をＤＮＡと組み合わせることにより、再構成された。次いで、再構成された染色質をホルムアルデヒドで固定し、配列データライブラリを作製した。図１のＡからＦは、これらの工程の概略図を示す。
【０２０４】
ヒトＧＭ１２８７８サンプルについては、４ｂｐ５’オーバーハングを生成する制限酵素ＭｂｏＩおよびＭｌｕＣＩ、これらを用いて２つのＤＮＡライブラリが生成された。これらのバーコードが付けられたライブラリはプールされ、ペア１００ｂｐリード中のシングルＩｌｌｕｍｉｎａ　Ｈｉｓｅｑ２５００レーン上で配列決定され、４６Ｍ　ＭｂｏＩおよび５２Ｍ　ＭｌｕＣＩライブラリリードペアを生成した。比較のために、図２に示すように、名目上は４０ＫｂｐのＤＮＡについて第３のライブラリを調製した。
【０２０５】
アメリカアリゲーターゲノム（Ａｌｌｉｇａｔｏｒ　ｍｉｓｓｉｓｓｉｐｐｉｅｎｓｉｓ）については、我々は単一のＭｂｏＩライブラリを構築し、シングルレーン上でそれを配列決定し、１３２Ｍリードペアを得た。スキャフォールドアセンブリとハプロタイプ位相合わせ用のこれらのデータの有用性を判定するために、我々は、ＧＭ１２８７８ライブラリデータを基準のヒト・アセンブリ（ｈｇ１９）にアライメントした（図２）。生成されたライブラリは、関連していない５００Ｋｂｐゲノムウィンドウ間におおよそ１つのスプリアスリンクのバックグラウンドノイズ率を有するゲノム上の最大１５０Ｋｂｐの分離の有用な結合情報を提供した（このような結合の平均値は０．９７）。ＧＭ１２８７８ライブラリからの配列のシングルレーンは、それぞれ０－１、１－５、５－１０、１０－２５、２５－５０、および５０－２００Ｋｂｐライブラリにおいて３．８倍、８．４倍、８．６倍、１８．６倍、１３．５倍、６．５倍の物理的カバレッジに等しい結合情報を提供し、一方でアリゲーターについては、図３に示すように、比較可能なカバレッジの推定値

は、それぞれ５．４倍、１６．７倍、１６．７倍、４２．２倍、３６．１倍、および１６．５倍であった。

【０２０６】

実施例２：リードペアデータに基づいた核酸スキャフォールディング

ライブラリから抽出されたデータのパワーおよび有用性を判定するために、コンティグアセンブリおよびスキャフォールディングは、一般的な３００－５００ｂｐインサートＩｌｌｕｍｉｎａショットガンライブラリおよび上記のライブラリのみを用いて行った。ＭＥＲＡＣＵＬＯＵＳ［ｐｍｉｄ２１８７６５４］を３３Ｋｂｐという典型的なサイズ（Ｎ５０）のスキャフォールドに用いてＧＭ１２８７８（Ｃｈａｐｍａｎら、２０１１）から得られた８４倍１０１ｂｐペアエンドＩｌｌｕｍｉｎａショットガンデータセットが最初にアセンブルされた。生成されたライブラリから得られたリードペアは、本明細書に記載されている通り、この初期アセンブリに対してマッピングされた。リードペアの６８．９％は順方向および逆方向リードの両方が２０以上のマップ品質を有するようにマッピングされ、それゆえアセンブリ内で一意にマッピングされているとみなされ、複製物ではなかった。これらのリードペアの２６．８％は、異なるコンティグにマッピングされた順方向および逆方向リードを有し、ゆえに、アセンブリをさらにスキャフォールドするために情報を与えられる可能性があった。同じライブラリデータが、ペア２５０ｂｐリード中の５０倍カバレッジのＤｉｓｃｏｖｅｒアセンブリをスキャフォールドするためにも使用された（Ｓｈａｒｐｅら、２０１５）。

【０２０７】

尤度モデルは、生成されたライブラリがどのようにゲノムＤＮＡをサンプルするかを記述するために開発され、リードペアに基づきコンティグを切断し再スキャフォールドする“ＨｉＲｉｓｅ”と呼ばれるソフトウェアパイプラインは、コンティグと結合する。モデリングには、結果として得られたアセンブリの局所的および全体的スケールでの完全性、近接性および正確性を、２つの主要なＷＧＳアセンブラ：ＭＥＲＡＣＵＬＯＵＳ（Ｃｈａｐｍａｎら、２０１１）およびＡＬＬＰＡＴＨＳ－ＬＧ（ＡＰＬＧ）（Ｇｎｅｒｒｅら、２０１１）、により作成されたフォスミドエンドペアでの広範なカバレッジを含む、豊富なＷＧＳデータセットのアセンブリに対して比較することが含まれた（表１）。アライメントに基づくアセンブリ品質の比較の構築に関わる任意の選択を回避するために、比較は、二倍体１２８７８ＮＡアセンブリの各ハプロタイプにおいて正確に１度生じるすべての異なる１０１ｂｐ配列の、ランダムに選択されたサブセットである２５４０万の１０１ｂｐ配列のアセンブルされた位置に基づいていた（Ｒｏｚｏｗｓｋｙら、２０１１）。

【０２０８】

【表１】

表１

	誤連結部 5 Kbp	誤連結部 10 Kbp	誤連結部 50 Kbp	N50 (Mbp)	95% CI 50 Kbp Δ	%C	% 誤配向
MERACULOUS	0.032	0.030	0.022	9.1	1.3 Kbp	94.8	0.09
APLG	0.245	0.187	0.130	12.1	6.4 Kbp	92.2	0.4
MERAC (33Kbp N50) + HiRise:	0.028	0.011	0.009	12.6	7.7 Kbp	94.1	1.3
Discovar (178Kbp N50) + HiRise:	0.102	0.097	0.076	29.9	3.2 Kbp	97.9	1.4

表１：スキャフォールディング結果。誤連結部を特定するための３つの異なる閾値での誤連結部を含むスキャフォールドにおける各アセンブリの画分。スキャフォールドＮ５０５０Ｋｂｐの分離不一致（ｓｅｐａｒａｔｉｏｎ ｄｉｓｃｒｅｐａｎｃｙ） ９５％の信頼区間（Ｃｏｎｆｉｄｅｎｃｅ Ｉｎｔｅｒｖａｌ）（９５％ＣＩ＝ｘ平均：アセンブリ内に固有の１０１－ｍｅｒタグのペアがあると仮定した場合、それらの９５％は基準における互いの５０Ｋｂｐプラスマイナス以内である。）完全性（％Ｃ）；誤配向１０１ｍｅｒの画分。

【０２０９】
実施例３：長距離スキャフォールディング精度の判定
ＨｉＲｉｓｅパイプラインが生産したスキャフォールドは、公開されているＭＥＲＡＣＵＬＯＵＳおよびＡＰＬＧアセンブリよりも長く、かつ全体的な誤アセンブリの率が低く、どちらもペアフォスミドエンドリード中の深いカバレッジに依存する。表１は、誤連結を含むスキャフォールド中に見られる全アセンブリの画分を示し、誤連結は、二倍体基準における１以上の染色体からの少なくとも５Ｋｂｐ、１０Ｋｂｐまたは５０Ｋｂｐの範囲に及ぶ一続きの一意の１０１－ｍｅｒを有するものとして定義される。また、表１は、ＮＡ１２８７８の他のアセンブリと比較された、ＨｉＲｉｓｅアセンブリの４回の連続的なラウンドの完全性と近接性の尺度も示す。
【０２１０】
本明細書で提供される方法によって作製されたリードペアを作製するＤＮＡ連結事象は、定義された相対的な鎖のリードペアを生成するように強制されていないので、スキャフォールド中のコンティグの相対的な配向はリード密度情報から推測せねばならない。その結果、ＨｉＲｉｓｅ計算を用いて到達したスキャフォールドは、他のアセンブリに見られるよりも誤配向１０１－ｍｅｒの割合が高く（１．３％）、そのほとんどが小さなコンティグで発生した。誤配向１０１－ｍｅｒを含むコンティグの中央サイズは２．１Ｋｂｐであった。
【０２１１】
実施例４：改良されたアリゲーターアセンブリ
本明細書に記載されている新規染色質リモデリング方法を用いてアメリカアリゲーター（Ａｌｌｉｇａｔｏｒ ｍｉｓｓｉｓｓｉｐｐｉｅｎｓｉｓ）について構築されたシングルＤＮＡ断片ライブラリが生成され、Ｉｌｌｕｍｉｎａ Ｈｉｓｅｑ ２５００上で２億１０７０万のリードが配列決定された。リードペアは、公開されているデータ（Ｇｒｅｅｎら、２０１４）を使用して作製された新規アセンブリ（Ｎ５０ ８１Ｋｂｐ）に対しマッピングされ、ＨｉＲｉｓｅスキャフォールディングパイプラインを適用された。結果として生じるアセンブリは１０．３ＭｂｐのスキャフォールドＮ５０を有する。これらのスキャフォールドの精度を評価するために、前もって生成された１，４８５個の細菌人工染色体（ＢＡＣ）末端配列（Ｓｈｅｄｌｏｃｋら、２００７）のあつまりをアセンブリにアライメントさせた。それらのうち１，２９８のペアが、コンティグアセンブリおよびＨｉＲｉｓｅにスキャフォールドされたバージョンに対し９０％のカバレッジと９５％の同一性を有するＧＭＡＰ（ＷｕおよびＷａｔａｎａｂｅ、２００５）によって一意にアライメントされた。入力アセンブリにおいて、１２．５％のＢＡＣエンドペアが予期された配向および分離を伴い同スキャフォールドにおいて捕捉された。ＨｉＲｉｓｅアセンブリにおいては、９６．５％のＢＡＣエンドペアは同スキャフォールドにおいてアライメントされ、９８．１％のＢＡＣエンドペアが同じスキャフォールドにあり正しく相対的に配向されている。５つの（０．３９％）ＢＡＣエンドペアは、同じスキャフォールド上に置かれているが、インサートサイズよりも著しく長い距離をとって置かれ、そして、１４の（１．０８％）ＢＡＣエンドペアは別々のスキャフォールド上に置かれているが、インサートサイズよりも長く、スキャフォールドのエッジから充分離れた距離をとって置かれたが、これは誤連結部の全体的な密度が８．３６Ｍｂｐのアセンブリにつき１未満であると示唆している。
【０２１２】
実施例５：位相精度の評価
順方向および逆方向リードがヘテロ接合部位をカバーしているリードペアが、ハプロタイプ位相を直接リードするために用いられた。本明細書に記載されている新規染色質リモデリング方法および断片化方法によって生成されたリードペア中でカバーされている距離は、インプットＤＮＡのサイズと同程度である可能性があるため、ＧＭ１２８７８サンプルにおけるハプロタイプの位相を判定するための位相情報およびその有用性が評価された。ＧＭ１２８７８はトリオシーケンシングされた個体由来であるため、信頼できるハプロ

タイプ位相情報を用いて位相合わせ情報の精度を評価した。ハプロタイプ情報を与えた、１０Ｋｂｐと１５０Ｋｂｐの間の範囲内のリードペアは、ＧＭ１２８７８についての既知のハプロタイプ位相と９９．８３％一致した。

【０２１３】

実施例６：構造変異体の特定

基準に対し１つの個体からペア配列リードをマッピングすることは、連続した核酸またはゲノム構造における相違、例えば、逆位、欠失および重複を特定するための最も一般的に使用される配列に基づく方法である（Ｔｕｚｕｎら、２００５）。図４Ａおよび４Ｂは、ヒト基準ゲノムＧＲＣｈ３８にマッピングされたＧＭ１２８７８由来の再アセンブリされた染色質から得られたＤＮＡの近接連結によって生成されたリードペアが、そのような２つの構造差をどのようにして明らかにするかを示す。構造差を特定するためのリードペアデータの感度および特異性を算定するために、ヘテロ接合性逆位の効果をシミュレートするように構築された模擬データセット上の最大尤度識別器がテストされた。テストデータは、ＧＲＣｈ３８基準配列に生成されたＮＡ１２８７８リードのマッピングから定義された長さＬの区間をランダムに選択し、生成された各リードペアを独立してランダムに逆位または基準ハプロタイプに割り当て、それに応じてマッピングされた座標を編集することによって構築された。非対立遺伝子相同組換えが、ヒトゲノムで観察される構造変異体の大部分の要因であり、結果として、多くの変異体切断点が繰返された配列の長いブロックにおいて生じる（Ｋｉｄｄら、２００８）。逆位切断点を囲む繰返しの配列の長さの変化の影響は、それらの距離Ｗの範囲内でマッピングされたすべてのリードを除去することによってシミュレートされた。逆位切断点に繰返しの配列が存在しない場合、それぞれ１Ｋｂｐ、２Ｋｂｐおよび５Ｋｂｐの逆位について、感度（特異性）はそれぞれ０．７６（０．８８）、０．８９（０．８９）および０．９７（０．９４）であった。逆位切断点での繰返しの（マッピングできない）配列の１Ｋｂｐの領域をシミュレーションに使用した時、５Ｋｂｐ逆位に対する感度（特異性）は０．８１（０．７６）であった。

【０２１４】

実施例７：ＤＮＡ調製

ＤＮＡは、製造業者のインストラクションに従いＱｉａｇｅｎ　ＢｌｏｏｄおよびＣｅｌｌ　Ｍｉｄｉ　ｋｉｔｓにより抽出された。要するに、細胞を溶解し、遠心分離して核を単離した。核は、さらに、プロテイナーゼＫとＲＮＡｓｅ　Ａを組み合わせたものにより消化された。ＤＮＡはＱｉａｇｅｎ　ｇｅｎｏｍｉｃ　ｃｏｌｕｍｎに結合され、洗浄され、溶出され、イソプロパノール中で沈殿され、遠心分離によってペレット化された。乾燥の後、ペレットは、２００μＬのＴＥ（Ｑｉａｇｅｎ）中に再懸濁された。

【０２１５】

実施例８：染色質アセンブリ

染色質は、Ａｃｔｉｖｅ　Ｍｏｔｉｆ　ｉｎ　ｖｉｔｒｏ　Ｃｈｒｏｍａｔｉｎ　Ａｓｓｅｍｂｌｙ　ｋｉｔを用いて、一晩中２７度でゲノムＤＮＡからアセンブルされた。インキュベーション後、サンプルの１０％をＭＮａｓｅ消化に用いて、染色質アセンブリの成功を確認した。

【０２１６】

実施例９：ビオチン化および制限消化

染色質をヨードアセチル－ＰＥＧ－２－ビオチン（ＩＰＢ）でビオチン化した。ビオチン化に続いて、染色質を１％ホルムアルデヒド中において室温（ＲＴ）で１５分間固定し、続いて２．５Ｍグリシンの２倍モル過剰量でクエンチした。Ｓｌｉｄｅ－Ａ－Ｌｙｚｅｒ　２０ＫＤａ　ＭＷＣＯ　ｄｉａｌｙｓｉｓ　ｃａｓｓｅｔｔｅｅ（Ｐｉｅｒｃｅ）中で、１Ｌの透析バッファー（１０ｍｍ　Ｔｒｉｓ－Ｃｌ、ｐＨ８．０、１ｍＭ　ＥＤＴＡ）に対して、染色質を、４度で最低３時間透析することにより、過剰ＩＰＢおよび架橋したグリシンを除去した。続いて、染色質を３７度で４時間、１倍のＣｕｔＳｍａｒｔ中のＭｂｏＩまたはＭｌｕＣＩのいずれかで消化した。染色質を、５０　ＫＤａ　ＭＷＣＯｄｉａｌｙｓｉｓ　Ｆｌｅｘ　ｔｕｂｅ（ＩＢＩ　Ｓｃｉｅｎｔｉｆｉｃ＃ＩＢ４８２６２

）中で、４度で２時間、再度透析した後、新鮮なバッファーで一晩中再度透析して、酵素および短い遊離ＤＮＡ断片を除去した。

【０２１７】

ＰＢＳ＋０．１％　Ｔｗｅｅｎ－２０中で洗浄および再懸濁した後、染色質に添加し、室温で1時間インキュベートすることによって、Ｄｙｎａｂｅａｄ　ＭｙＯｎｅ　Ｃ１ストレプトアビジンビーズが調製された。次いで、ビーズを磁気濃縮器ラック上で濃縮し、その後、洗浄し、再濃縮し、１００μＬの１倍ＮＥＢｕｆｆｅｒ　２の中で再懸濁した。

【０２１８】

実施例１０：ｄＮＴＰ充填

充填反応中に標識化されたｄＮＴＰ（図１のＡからＦ）を捕捉させないために、遊離ビオチンの存在下にて１５分間室温でビーズをインキュベートすることにより、非結合ストレプトアビジン部位が占められた。続いて、ビーズを２回洗浄した後、１００μＬの１ＸＮＥＢｕｆｆｅｒ　の中で再懸濁した。全容量１６５μｌ中２５ＵのＫｌｅｎｏｗ（＃Ｍ０２１０Ｍ、ＮＥＢ）とともにａ－Ｓ－ｄＧＴＰおよびビオチン化ｄＣＴＰを含むｄＮＴＰによって２５度で４０分間インキュベートすることにより、粘着末端が充填された。充填反応は７μＬの０．５Ｍ　ＥＤＴＡを加えることにより止まった。次いで、ビーズを連結前ウォッシュバッファー（ＰＬＷＢ：５０ｍＭ　Ｔｒｉｓ７．４；　０．４％　Ｔｒｉｔｏｎ　Ｘ－１００；　０．１ｍＭ　ＥＤＴＡ）の中で２回洗浄した後、１００μＬのＰＬＷＢの中で再懸濁した。

【０２１９】

実施例１１：連結

連結は、少なくとも１ｍＬのＴ４連結バッファー中、１６℃で少なくとも４時間行われた。異なる染色質凝集体間の交差連結を最小にするために、大量の連結が使用された。連結反応は４０μＬの０．５Ｍ　ＥＤＴＡを加えることにより止められた。ビーズは濃縮され、１００μＬの抽出バッファー（５０ｍＭ　Ｔｒｉｓ－Ｃｌ　ｐＨ８．０、　１ｍＭ　ＥＤＴＡ、０．２％ＳＤＳ）の中で再懸濁された。４００ｕｇのプロテイナーゼＫ（＃Ｐ８１０２Ｓ、ＮＥＢ）を添加した後、ビーズを５５度で一晩インキュベートし、続いて５５度で追加の２００μｇのプロテイナーゼＫで２時間消化した。ＤＮＡは、２：１の比のＳＰＲＩビーズ、カラム精製キット、またはフェノール：クロロホルム抽出のいずれかを用いて、再び回収された。ＤＮＡは、低ＴＥ（１０ｍＭ　Ｔｒｉｓ－Ｃｌ　ｐＨ８．０、　０．５ｍＭ　ＥＤＴＡ）へ溶出された。

【０２２０】

実施例１２：エキソヌクレアーゼ消化

次に、ＤＮＡを、４０分間３７度で１００Ｕのエキソヌクレアーゼ ＩＩＩ（＃Ｍ０２０６Ｓ、ＮＥＢ）を用いて消化し、ビオチン化した遊離末端を除去して、続いてＳＰＲＩクリーンアップおよび１０１μＬの低ＴＥへの溶出を行った。

【０２２１】

実施例１３：せん断およびライブラリ調製

ＤＮＡは、３０秒間オン／３０秒間オフの６０回のサイクルの間、「低」に設定されたＤｉａｇｅｎｏｄｅ　Ｂｉｏｒｕｐｔｏｒを用いてせん断された。せん断の後、ＤＮＡはＫｌｅｎｏｗポリメラーゼとＴ４　ＰＮＫ（＃ＥＫ００３２　Ｔｈｅｒｍｏ　Ｓｃｉｅｎｔｉｆｉｃ）で２０度で３０分間充填された。充填反応後、ＤＮＡは、２００μＬの2倍ＮＴＢ（2Ｍ　ＮａＣｌ、１０ｍＭ　Ｔｒｉｓ　ｐＨ８．０、０．１ｍＭ　ＥＤＴＡｐＨ８．０、０．２％　Ｔｒｉｔｏｎ　Ｘ－１００）の中で再懸濁させる前にＴｗｅｅｎ　ウォッシュバッファーで２回洗浄することによって調製されたＣ１ビーズ上にプルダウンされた。一旦サンプルを加えたら、ビーズはロッキングしながら室温で２０分間インキュベートされた。続いて、ビオチン化されていないＤＮＡ断片を、低ＴＥ中に再懸濁させる前にビーズを３回洗浄することにより除去した。配列決定ライブラリは確立されたプロトコルを用いて生成された。（ＭｅｙｅｒおよびＫｉｒｃｈｅｒ、２０１０）

【０２２２】

実施例１４：リードマッピング

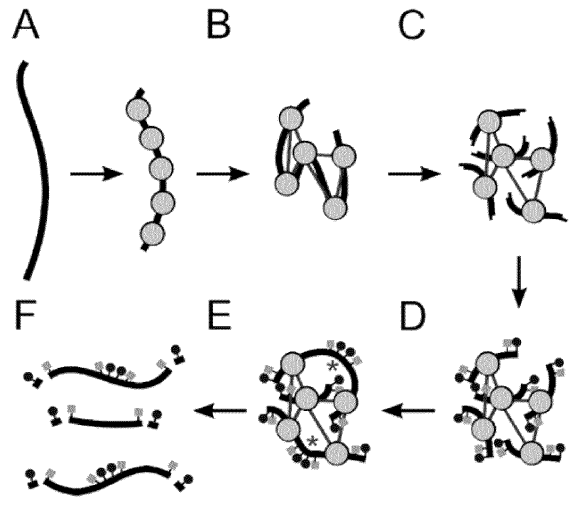
接合部が存在する場合は常に配列リードは切り詰められた。（ＭｂｏＩについては、ＳＥＱ　ＩＤ　ＮＯ．１：ＧＡＴＣＧＡＴＣ、ＭｌｕＣＩについては、ＳＥＱ　ＩＤ　ＮＯ．２：ＡＡＴＴＡＡＴＴ）その後、リードは、独立して順方向および逆方向リードをアライメントするための－ｘオプションを含むＳＭＡＬＴ［ｈｔｔｐ：／／ｗｗｗ．ｓａｎｇｅｒ．ａｃ．ｕｋ／ｒｅｓｏｕｒｃｅｓ／ｓｏｆｔｗａｒｅ／ｓｍａｌｔ／］を用いて、アライメントされた。ＰＣＲ複製物は、Ｐｉｃａｒｄ－ｔｏｏｌｓ　ＭａｒｋＤｕｐｌｉｃａｔｅｓ［ｈｔｔｐ：／／ｂｒｏａｄｉｎｓｔｉｔｕｔｅ．ｇｉｔｈｕｂ．ｉｏ／ｐｉｃａｒｄ／］を用いてマークされた。両方のリードがマッピングされ、マッピング品質が１０より大きい場合は、非重複リードペアを分析に使用した。

【０２２３】

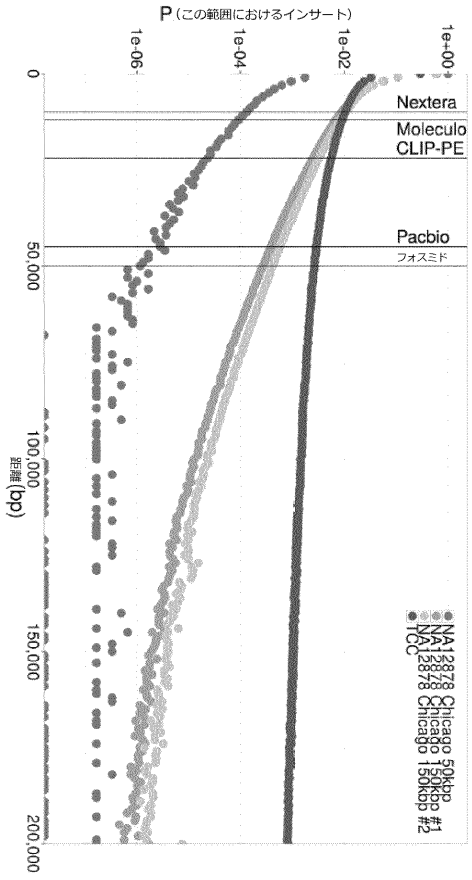
実施例１５：新規アセンブリ

ヒトおよびアリゲーターの新規ショットガンアセンブリは、公開されているショートインサートおよびメイトペアリード（ＳｉｍｐｓｏｎおよびＤｕｂｒｉｎ、２０１２；　Ｇｒｅｅｎら、２０１４）を用いて、Ｍｅｒａｃｕｌｏｕｓ２．０．３（Ｃｈａｐｍａｎら、２０１１）により生成された。アリゲーターメイトペアリードはＴｒｉｍｍｏｍａｔｉｃ（Ｂｏｌｇｅｒら、２０１４）でアダプタートリミングされた。いくつかの重複するアリゲーターショートインサートリードは“融合”された。これらは、順方向および逆方向リードへと戻された（ｕｎｍｅｒｇｅｄ　ｂａｃｋ）。

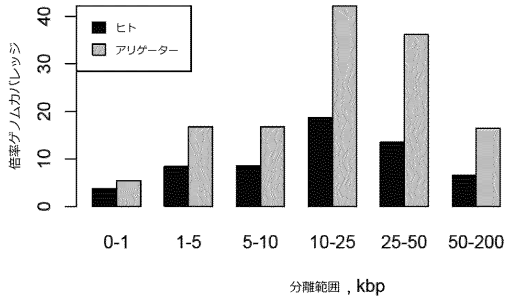
【図１】



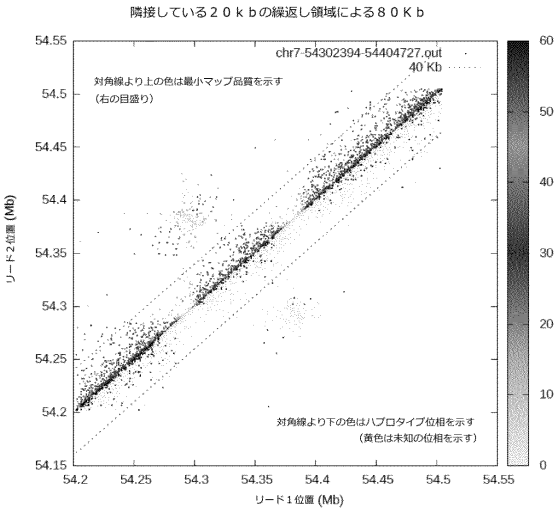
【図２】



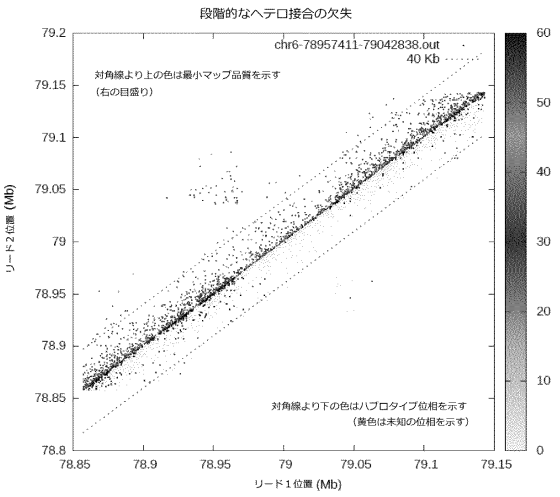
【図３】


ヒト
アリゲーター
倍率ゲノムカバレッジ
0
10
20
30
40
0-1
1-5
5-10
10-25
25-50
50-200
分離範囲 , kbp

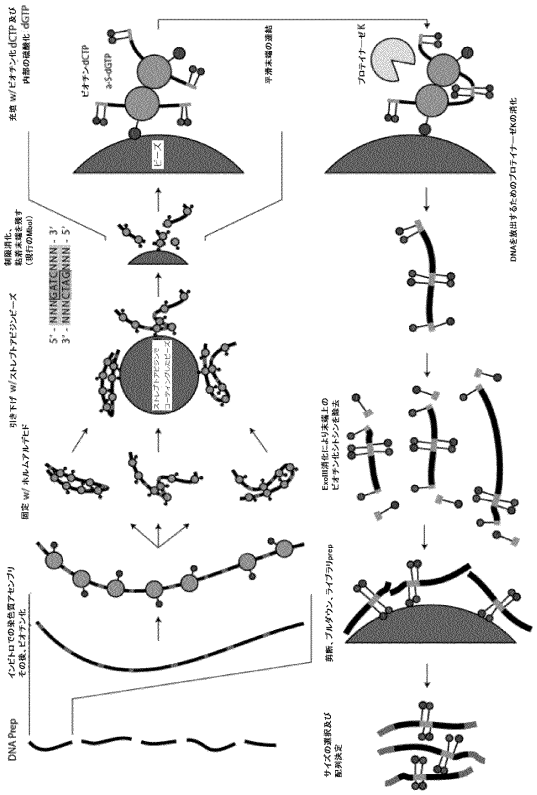
【図４Ａ】


隣接している２０ｋｂの繰返し領域による８０Ｋｂ
chr7-54302394-54404727.out
40 Kb
対角線より上の色は最小マップ品質を示す
（右の目盛り）
対角線より下の色はハプロタイプ位相を示す
（黄色は未知の位相を示す）
リード２位置 (Mb)
リード１位置 (Mb)
54.15
54.2
54.25
54.3
54.35
54.4
54.45
54.5
54.55
0
10
20
30
40
50
60

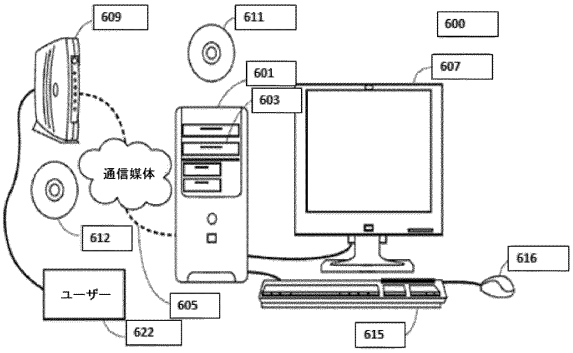
【図４Ｂ】


段階的なヘテロ接合の欠失
chr6-78957411-79042838.out
40 Kb
対角線より上の色は最小マップ品質を示す
（右の目盛り）
対角線より下の色はハプロタイプ位相を示す
（黄色は未知の位相を示す）
リード２位置 (Mb)
リード１位置 (Mb)
78.8
78.85
78.9
78.95
79
79.05
79.1
79.15
79.2
0
10
20
30
40
50
60

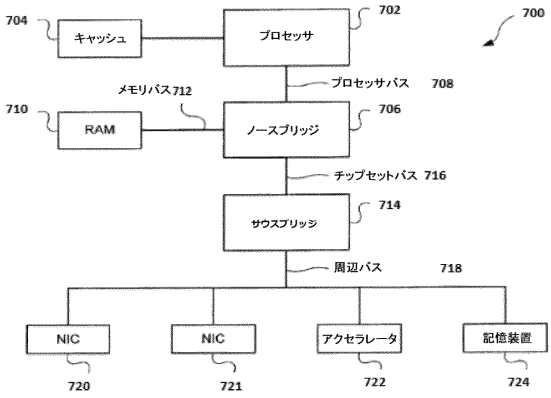
【図５】


DNA Prep
インビトロでの染色質アセンブリ
その後、ビオチン化
固定 w/ ホルムアルデヒド
引き下げ w/ ストレプトアビジンビーズ
ストレプトアビジン
コーティングしたビーズ
5' - NNNGATCNNN - 3'
3' - NNNCTAGNNN - 5'
解離消化、
粘着末端を残す
（現行のMbsu）
ビーズ
ビオチン-dCTP
ｄ-S-dGTP
平滑末端の連結
プロテイナーゼK
DNAを放出するためのプロテイナーゼKの消化
ExoIII消化により末端上の
ビオチン化ントシンを除去
剪断、プルダウン、ライブラリprep
サイズの選択及び
配列決定

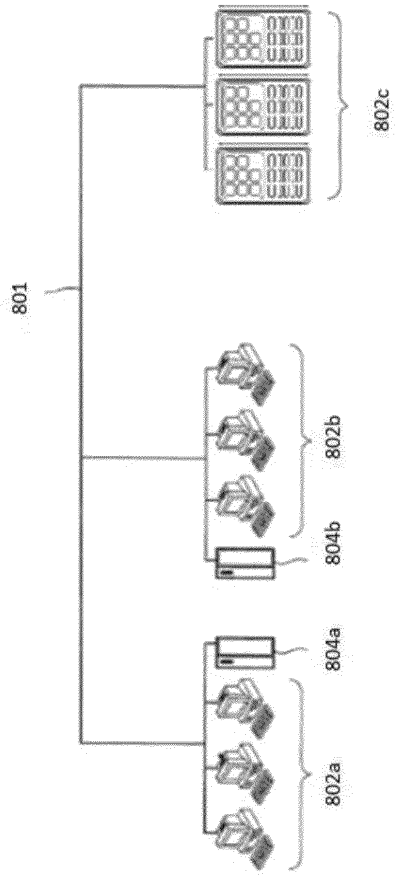
【図６】



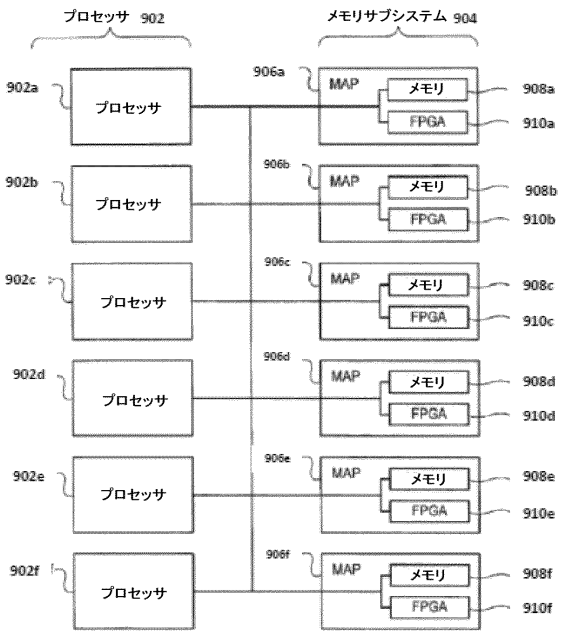
【図７】



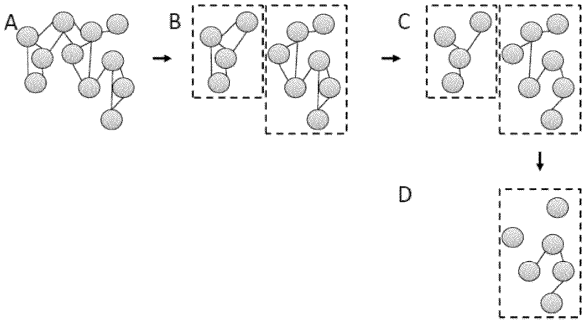
【図８】



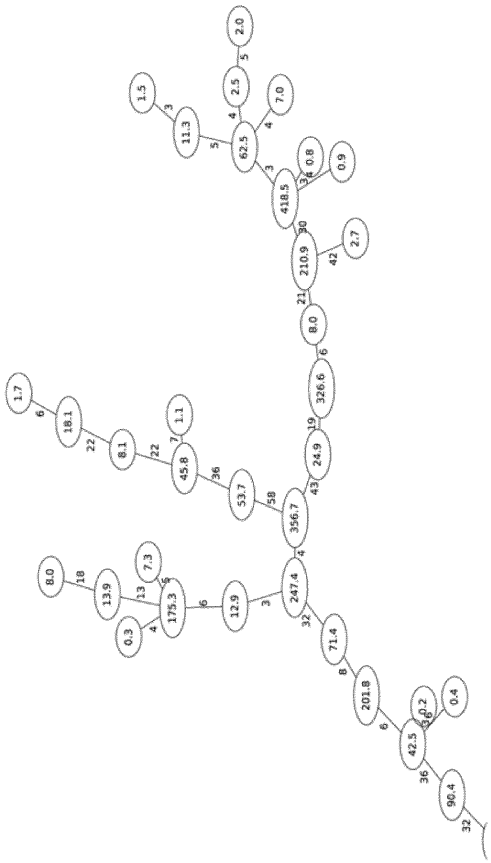
【図９】



【図１０】



【図１１】


2.0
5
2.5
4
7.0
4
1.5
3
11.3
5
62.5
3
0.8
3
0.9
418.5
30
2.7
42
210.9
21
8.0
6
326.6
19
24.9
43
1.7
6
18.1
22
8.1
22
1.1
7
45.8
36
53.7
58
356.7
4
247.4
32
71.4
8
201.8
6
0.2
36
0.4
42.5
36
90.4
32
8.0
18
13.9
13
7.3
5
175.3
6
12.9
3
0.3
4

フロントページの続き

(72)発明者　スタイツ，ジョナサン　シー.
アメリカ合衆国　９５０６０　カリフォルニア州　サンタ・クルーズ　シルバー・ストリート　１２１　ナンバー３
(72)発明者　ライス，ブランドン　ジェイ.
アメリカ合衆国　９５０６０　カリフォルニア州　サンタ・クルーズ　エルム・ストリート　２１４　アパートメント　１

審査官　渡邉　加寿磨

(56)参考文献　特表２０１２－５１１７５３（ＪＰ，Ａ）
特表２０１２－５１４９７７（ＪＰ，Ａ）
米国特許出願公開第２０１１／０２８８８４５（ＵＳ，Ａ１）
米国特許出願公開第２０１４／０１３６１２１（ＵＳ，Ａ１）
米国特許出願公開第２０１４／０２２０５８７（ＵＳ，Ａ１）
米国特許出願公開第２０１２／０１９７５３３（ＵＳ，Ａ１）
吉川　舜亮，”Ｓｐａｒｓｅ　ｋ－ｍｅｒ　ｇｒａｐｈアルゴリズムの評価とＶｅｌｖｅｔへの実装”，電子情報通信学会技術研究報告，日本，一般社団法人電子情報通信学会，２０１３年６月２０日，第113巻第111号，p.1-7，ISSN:0913-5685

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１６Ｂ　　５／００　－　９９／００
Ｇ０６Ｑ　１０／００　－　９０／００
Ｇ１６Ｈ　１０／００　－　８０／００
Ｇ１６Ｚ　９９／００