

(19) 日本国特許庁 (JP)

(12) 公表特許公報 (A)

(11) 特許出願公表番号

特表2017-500004

(P2017-500004A)

(43) 公表日 平成29年1月5日 (2017. 1. 5)

|                                |               |             |
|--------------------------------|---------------|-------------|
| (51) Int. Cl.                  | F I           | テーマコード (参考) |
| <b>C 1 2 Q 1/68 (2006.01)</b>  | C 1 2 Q 1/68  | 4 B 0 6 3   |
| <b>G 0 6 F 19/22 (2011.01)</b> | G 0 6 F 19/22 |             |

審査請求 未請求 予備審査請求 未請求 (全 40 頁)

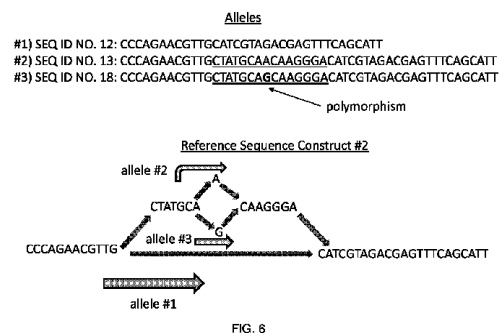
|                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (21) 出願番号 特願2016-523194 (P2016-523194)<br>(86) (22) 出願日 平成26年10月17日 (2014. 10. 17)<br>(85) 翻訳文提出日 平成28年6月3日 (2016. 6. 3)<br>(86) 国際出願番号 PCT/US2014/061156<br>(87) 国際公開番号 W02015/058093<br>(87) 国際公開日 平成27年4月23日 (2015. 4. 23)<br>(31) 優先権主張番号 61/892, 662<br>(32) 優先日 平成25年10月18日 (2013. 10. 18)<br>(33) 優先権主張国 米国 (US) | (71) 出願人 516050647<br>セブン ブリッジズ ジェノミクス イン<br>コーポレイテッド<br>アメリカ合衆国 マサチューセッツ O 2<br>1 4 2, ケンブリッジ, ブロードウェ<br>イ 1, 1 4 ティーエイチ フロア<br>(74) 代理人 100078282<br>弁理士 山本 秀策<br>(74) 代理人 100113413<br>弁理士 森下 夏樹<br>(74) 代理人 100181674<br>弁理士 飯田 貴敏<br>(74) 代理人 100181641<br>弁理士 石川 大輔 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

最終頁に続く

(54) 【発明の名称】 遺伝子試料について遺伝子型解析するための方法およびシステム

## (57) 【要約】

本発明は、参照配列構築物、例えば、ゲノムの各遺伝子座において公知のバリエーションを表す有向非巡回グラフ (DAG) を使用して、特異的な遺伝子座において特異的な塩基コールを行う方法およびシステムを提供する。配列リードは、アライメント時に DAG にアラインされるため、変異を、参照ゲノムと突き合わせて、公知の変異についての表と比較する後続のステップを廃することができる。開示される方法およびシステムは、ゲノム内の構造バリエーションまたは構造バリエーション内にある変異の扱いにおいて特に効率的である。



**【特許請求の範囲】****【請求項 1】**

遺伝子試料について遺伝子型解析する方法であって、  
遺伝子試料に対応する複数の配列リードを得るステップと；  
前記複数の配列リードを、参照配列構築物内の第 1 の位置において異なる遺伝子型を代替配列として表す前記配列構築物にアラインするステップと；  
前記試料について、前記参照配列構築物に対する前記複数の配列リードのアライメントに基づき遺伝子型解析するステップと  
を含む方法。

**【請求項 2】**

前記参照配列構築物が有向非巡回グラフ（DAG）である、請求項 1 に記載の方法。

**【請求項 3】**

前記配列リードをバリエーションコールフォーマット（VCF）ファイルまたは一塩基多型データベース（dbSNP）と比較するステップを含まない、請求項 1 に記載の方法。

**【請求項 4】**

代替配列の少なくとも一部が遺伝子構造バリエーションを含む、請求項 1 に記載の方法。

**【請求項 5】**

前記参照配列構築物が、前記構築物内の第 2 の位置において構造バリエーションを代替配列として表す、請求項 1 に記載の方法。

**【請求項 6】**

2 つまたはそれ超の関連するリードのアライメントを使用して、前記試料について遺伝子型解析する、請求項 1 に記載の方法。

**【請求項 7】**

前記参照配列構築物上の場所にアラインされた配列リードの数を使用して、前記試料について遺伝子型解析する、請求項 1 に記載の方法。

**【請求項 8】**

前記アライメントに基づき、前記試料について対立遺伝子を決定するステップをさらに含む、請求項 1 に記載の方法。

**【請求項 9】**

前記参照上の場所にアラインされた配列リードの数に基づき、前記試料について対立遺伝子頻度を決定するステップをさらに含む、請求項 8 に記載の方法。

**【請求項 10】**

前記試料について遺伝子型を決定するステップをさらに含む、請求項 1 に記載の方法。

**【請求項 11】**

前記配列リードと前記代替配列のうちの 1 つとの間で重複する塩基対の数に基づき前記遺伝子型解析の信頼度の値を決定するステップをさらに含む請求項 1 に記載の方法であって、重複が多いことが、信頼度が高いことと相関する、方法。

**【請求項 12】**

前記参照配列構築物が染色体を表す、請求項 1 に記載の方法。

**【請求項 13】**

前記参照配列構築物がゲノムを表す、請求項 1 に記載の方法。

**【請求項 14】**

前記代替配列が、塩基欠失、塩基挿入または多型によって互いと異なる、請求項 1 に記載の方法。

**【請求項 15】**

前記配列リードが、サンガーシーケンシング、パイロシーケンシング、イオン半導体シーケンシング、合成によるシーケンシング、ライゲーションによるシーケンシング、および単一分子リアルタイムシーケンシングから選択される方法を用いた、対象の遺伝子試料のシーケンシングによって得られる、請求項 1 に記載の方法。

10

20

30

40

50

## 【請求項 16】

前記参照配列構築物が、約 1,000 を超える塩基対を含む、請求項 1 に記載の方法。

## 【請求項 17】

前記参照配列構築物が、約 1,000,000 を超える記号を含む、請求項 16 に記載の方法。

## 【請求項 18】

前記複数の配列リードが、約 1000 を超える配列リードを含む、請求項 1 に記載の方法。

## 【請求項 19】

前記複数の配列リードの大半が、約 50 を超える塩基対の長さである、請求項 1 に記載の方法。

## 【請求項 20】

前記複数の配列リードの大半が、約 100 を超える塩基対の長さである、請求項 19 に記載の方法。

## 【請求項 21】

前記遺伝子試料が、血液、尿、唾液、痰、糞便、乳首吸引物、汗、毛包、口腔内スワブまたは組織に由来する、請求項 1 に記載の方法。

## 【請求項 22】

前記試料から複数の核酸を単離するステップをさらに含む、請求項 21 に記載の方法。

## 【請求項 23】

前記試料の遺伝子型に基づき疾患を診断するステップをさらに含む、請求項 1 に記載の方法。

## 【請求項 24】

遺伝子試料について遺伝子型解析するコンピュータ実行方法であって、

遺伝子試料に対応する複数の配列リードを提供するステップと；

コンピュータプロセッサを使用して、前記複数の配列リードを、参照配列構築物内の第 1 の位置において異なる遺伝子型を代替配列として表す参照配列構築物にアラインするステップと；

前記試料を、前記参照配列構築物に対する前記複数の配列リードのアライメントに基づき遺伝子型解析するステップとを含む方法。

## 【請求項 25】

前記複数の配列リードが、非一時的コンピュータ読取り型媒体において提供される、請求項 24 に記載の方法。

## 【請求項 26】

前記参照配列構築物が、非一時的コンピュータ読取り型媒体において提供される、請求項 24 に記載の方法。

## 【請求項 27】

前記参照配列構築物が有向非巡回グラフである、請求項 24 に記載の方法。

## 【請求項 28】

複数のリードをアラインするための、プロセッサおよびメモリを含むシステムであって、前記メモリが、実行されると、

遺伝子試料に対応する複数の配列リードを得る；

前記複数の配列リードを、参照配列構築物内の第 1 の位置において異なる遺伝子型を代替配列として表す参照配列構築物にアラインする；

前記試料について、前記参照配列構築物に対する前記複数の配列リードのアライメントに基づき遺伝子型解析することを前記プロセッサにさせる命令を含む、システム。

## 【請求項 29】

アラインするステップが、

10

20

30

40

50

配列リードに対応する記号の文字列を前記第 1 の位置における複数の異なる遺伝子型と同時に比較するステップ、

前記記号の文字列と前記複数の異なる遺伝子型の各々との間の重複をスコア付けするステップであって、ここで、高スコアは、重複の量が多いことに対応するステップ、

前記配列リードについて、最高のスコアに対応する前記遺伝子型の重複を同定するステップ

を含む、請求項 28 に記載のシステム。

【請求項 30】

同定された前記遺伝子型に対応してファイルをメモリに書き込むステップをさらに含む、請求項 29 に記載のシステム。

【請求項 31】

前記複数の配列リードが、非一時的コンピュータ読取り型媒体において提供される、請求項 28 に記載のシステム。

【請求項 32】

前記参照配列構築物が、非一時的コンピュータ読取り型媒体において提供される、請求項 28 に記載のシステム。

【請求項 33】

前記参照配列構築物が有向非巡回グラフである、請求項 28 に記載のシステム。

【請求項 34】

複数のプロセッサを含み、各プロセッサが、前記複数の配列リードの一部を前記参照配列構築物にアラインするように構成されている、請求項 28 に記載のシステム。

【請求項 35】

前記命令が、前記プロセッサに、前記試料についての遺伝子型を出力することをさらにさせる、請求項 28 に記載のシステム。

【発明の詳細な説明】

【技術分野】

【0001】

関連出願

本出願は、その全体が本明細書に参考として援用される、2013 年 10 月 18 日に出願された米国特許出願第 61/892,662 号に対する優先権を主張する。

【0002】

発明の分野

本発明は、遺伝子試料について遺伝子型解析するための方法およびシステムに関する。

【背景技術】

【0003】

シーケンシング技術の進歩により、個体のゲノムを 1 週間またはそれ未満でシーケンシングすることが可能になっている。典型的には、遺伝子試料は単離され、小片に切断され、増幅され、次いで、Illumina（商標）シーケンシング（Illumina, Inc.、San Diego, CA）などのハイスループットなシステムでシーケンシングされる。この工程により、配列を創出するために後でアSEMBルしなければならない、莫大な数の配列リードが生成される。しかし、配列それ自体によってもたらされる情報は、その診断および予測上の価値がゲノムにおけるその配列の相対的な位置に依存するので、有用なものは極めて少ない。すなわち、ゲノム内の配列の相対的な位置が公知である場合に、例えば、疾患についてのマーカーの存在を決定することが可能であるのみである。加えて、特定の位置における配列が公知であれば、例えば表現型、対立遺伝子同一性、遺伝子型などの、より高レベルの情報を決定することができる。

【0004】

各ゲノム内のサイズおよび変異性が原因で、各リードが属する場所の特定が、次世代シーケンシング（N. G. S.）における遺伝子型解析の実質的な障害である。リードのセットに基づき遺伝子型を決定することに関する問題は、証拠上の観点から自然に枠どられ

10

20

30

40

50

る。各リードにより、いくつかの遺伝子型または遺伝子型のセットの証拠がもたらされ、全てのリードからの証拠を組み合わせることにより、対象の遺伝子型に関する何かを結論づけることが可能になる。しかし、この工程の両方の部分——単一リードが対象のゲノムに関して何を明らかにするのかを解釈すること、および多くのリードの証拠を集合させること——が問題を提示する。さらに、実際の遺伝子データを扱う際には、極めて多数のリード、および遺伝子データ内のより大きな構造バリエーションの存在により、何百万のピースがあり、ピース間のバリエーションがわずかであるジグソーパズルと類似した難題が生じる。

#### 【0005】

現在の技術水準のアライメント法では、重複するリードを参照に対してアラインして、重要な遺伝情報または構造情報の探索を可能にするアセンブルされた配列（例えば、疾患に関するバイオマーカー）を生成するのに、膨大な計算能力を使用する。最終的に、配列アライメントの目標は、シーケンサーにより生成される核酸リードのセットを組み合わせ、被験体に由来する遺伝子試料に基づき、より長いリード（すなわち、コンティグ）、なおまたはその被験体の全ゲノムを達成することである。次世代シーケンサーからの配列データは、併せて標的配列の全体を表す、数百万もの短い配列を含むことが多いため、リードのアラインは、複雑で計算が高価である。加えて、ランダムシーケンシングエラー（すなわち、不正確なシーケンシングマシン出力）により引き起こされる配列の歪みを最小化するためには、プローブされた配列の各部分を、複数回にわたり（例えば、2～100回またはこれを超える回数にわたり）シーケンシングして、任意のランダムシーケンシングエラーの、作り出される最終アライメントおよび出力配列に対する影響を最小化する。最後に、核酸リードの全てに対応するデータの全てを収集したら、被験体の全ての配列（またはその一部）を決定するために、リードを、単一の参照配列（reference sequence）、例えばGRCh37にアラインする。多くの場合、個々のリードを実際に表すわけではなく、アラインされた配列を、サンプリングされた配列へとアセンブルし、サンプリングされた配列を、データファイルとして提示する。

#### 【0006】

典型的には、配列アライメントは、配列情報の2つの線形文字列（linear string）間のペアワイズアライメントを集約することにより構築される。アライメントの例として、2つの文字列である、S1（配列番号20：AGCTACGTACACTACC）およびS2（配列番号21：AGCTATCGTACTAGC）は、互いにアラインすることができる。典型的には、S1はリードに対応し、S2は参照配列の部分に対応する。互いに対して、S1およびS2は、置換、欠失、および挿入を含んでもよい。典型的には、用語は、文字列S1から文字列S2への変換に関して定義される：置換は、S2内の文字または配列が、S1内の同じ長さの異なる文字または配列で置き換えられる場合に生じ、欠失は、S2内の文字または配列が、S1の対応する区画（section）内で「スキップ」される場合に生じ、挿入は、文字または配列が、S1内の、S2内では隣接する2つの位置の間で生じる場合に生じる。例えば、2つの配列であるS1およびS2は、下記の通りにアラインすることができる。下記のアライメントは、13箇所のマッチ、長さ1の欠失、長さ2の挿入、および1箇所の置換：

(S1) AGCTA—CGTACACTACC（配列番号20）

(S2) AGCTATCGTAC—TAGC（配列番号21）

を表す。

#### 【0007】

当業者は、配列アライメントのための正確なアルゴリズムおよび近似的なアルゴリズムが存在することを十分に理解する。正確なアルゴリズムは、最高スコアのアライメントを見出すと予想されるが、計算が高価でありうる。2つの最も周知の正確なアルゴリズムは、Needleman—Wunsch（J Mol Biol、48巻（3号）：443～453頁、1970年）およびSmith—Waterman（J Mol Biol、147巻（1号）：195～197頁、1981年；Adv. in Math.、20巻（3号）：367～38

7頁、1976年)である。Gotoh (J Mol Biol、162巻(3号):705~708頁、1982年)による、Smith-Watermanに対するさらなる改善は、計算時間を、 $O(m^2n)$ から $O(mn)$  [ここで、 $m$ および $n$ は、比較される配列サイズであり、並列処理により適する]へと短縮する。バイオインフォマティクスの分野では、Gotohの改変アルゴリズムが、Smith-Watermanアルゴリズムと称されることが多い。並列計算リソースが、より広くかつ廉価に利用可能となりつつあるので、Smith-Waterman法は、より多くの配列セットをより多くの参照配列に対してアラインするのに使用されている。例えば、<http://aws.amazon.com>で入手可能な、Amazon.comのクラウドコンピューティングリソースを参照されたい。上記の雑誌論文の全ては、参照によりそれらの全体において本明細書に組み込まれる。

10

#### 【0008】

Smith-Waterman (SW) アルゴリズムでは、配列内の塩基間の重複に対して報酬を与え、配列間のギャップに対してペナルティーを課すことにより、直鎖状の配列をアラインする。Smith-Watermanはまた、SWは、短い配列が、長い配列を記載する文字の文字列にわたることを必要としないという点でも、Needleman-Wunschと異なる。すなわち、SWは、1つの配列が、他の配列の全体についてのリードであることを仮定しない。さらに、SWは、文字列の全長にわたり伸長するアライメントを見出さなくてもよいため、局所的アライメントは、2つの配列内のどこでも開始および終結させることが可能である。

20

#### 【0009】

下記の等式(1)：

#### 【数1】

$$H_{k0} = H_{0l} = 0 \quad (0 \leq k \leq n \text{ および } 0 \leq l \leq m \text{ について}) \quad (1)$$

$$H_{ij} = \max\{H_{i-1,j-1} + s(a_i, b_j), H_{i-1,j} - W_{in}, H_{i,j-1} - W_{del}, 0\}$$

( $1 \leq i \leq n$  および  $1 \leq j \leq m$  について)

との関連で、SWアルゴリズムは、長さ $n$ および $m$ の2つの文字列を表す、 $n \times m$ 行列 $H$ で容易に表される。

#### 【0010】

30

上記の等式では、 $s(a_i, b_j)$ は、マッチボーナス( $a_i = b_j$ である場合)またはミスマッチペナルティー( $a_i \neq b_j$ である場合)を表し、挿入および欠失には、それぞれ、ペナルティー $W_{in}$ および $W_{del}$ が課される。大半の場合、結果として得られる行列は、ゼロである多くの成分を有する。この表示は、行列内の上行~下行、右列~左列のバックトレースを容易とし、これにより、アライメントの同定を容易とする。

#### 【0011】

行列にスコアを完全に追加したら、SWアルゴリズムにより、バックトラックを実施して、アライメントを決定する。アルゴリズムは、行列内の最大値から始めて、各セルの最終的な最大値を計算するのに3つの値( $H_{i-1, j-1}$ 、 $H_{i-1, j}$ 、または $H_{i, j-1}$ )のうちのいずれを使用したのかに基づき、バックトラックする。バックトラックは、ゼロに到達すると停止される。例えば、先行技術を表すものではなく、バックトラックの概念と、バックトラックが読み取られた場合の、対応する局所的アライメントとを説明するものである、図3(B)を参照されたい。したがって、アルゴリズムにより決定された「最良のアライメント」は、可能な最小数を超える挿入および欠失を含有しうるが、可能な最大数をはるかに下回る置換を含有する。

40

#### 【0012】

SWまたはSW-Gotohとして適用する場合、技法では、動的計画法アルゴリズムを使用して、それぞれ、サイズを $m$ および $n$ とする、2つの文字列 $S$ および $A$ の局所的配列アライメントを実施する。この動的計画法では、表または行列を援用して、マッチスコアを保存し、一連のセルについての再計算を回避する。文字列の各成分は、配列の文字に

50

関するインデックスが付されていてもよく、すなわち、Sが文字列A T C G A Aであれば、 $S[1] = A$ 、 $S[4] = G$ などである。最適のアライメントを $H_{i,j}$ （上記）と表す代わりに、最適のアライメントは、下記の等式（2）：

【数2】

$$B[j, k] = \max(p[j, k], i[j, k], d[j, k], 0) \quad (0 < j \leq m, 0 < k \leq n \text{ について}) \quad (2)$$

の $B[j, k]$ と表すことができる。

【0013】

最大値関数である $B[j, k]$ の引数を、下記の等式（3）～（5）[ここで、MISMATCH\_PENALTY、MATCH\_BONUS、INSERTION\_PENALTY、DELETION\_PENALTY、およびOPENING\_PENALTYは、全て定数であり、MATCH\_BONUSを除き、全て負である]に概括する。マッチの引数である $p[j, k]$ は、下記の等式（3）：

【数3】

$$\begin{aligned} p[j, k] &= \max(p[j-1, k-1], i[j-1, k-1], d[j-1, k-1]) + \text{MISMATCH\_PENALTY}, \quad S[j] \neq A[k] \text{ の場合} \quad (3) \\ &= \max(p[j-1, k-1], i[j-1, k-1], d[j-1, k-1]) + \text{MATCH\_BONUS}, \quad S[j] = A[k] \text{ の場合} \end{aligned}$$

で与えられ、挿入の引数である $i[j, k]$ は、下記の等式（4）：

【数4】

$$i[j, k] = \max(p[j-1, k] + \text{OPENING\_PENALTY}, i[j-1, k], d[j-1, k] + \text{OPENING\_PENALTY}) + \text{INSERTION\_PENALTY} \quad (4)$$

で与えられ、欠失の引数である $d[j, k]$ は、下記の等式（5）：

【数5】

$$d[j, k] = \max(p[j, k-1] + \text{OPENING\_PENALTY}, i[j, k-1] + \text{OPENING\_PENALTY}, d[j, k-1]) + \text{DELETION\_PENALTY} \quad (5)$$

で与えられる。

【0014】

3つの引数全てについて、 $[0, 0]$ 成分は、ゼロと置いて、バックトラックの完了を確認する、すなわち、 $p[0, 0] = i[0, 0] = d[0, 0] = 0$ とする。

【0015】

スコア付けパラメータは、ある程度任意のものであり、計算の挙動を達成するように調整することができる。DNAのためのスコア付けパラメータ設定の一例（Huang、3章：Bio-Sequence Comparison and Alignment、Curr Top Comp Mol Biolシリーズ、Cambridge、Mass.: The MIT Press、2002年）であれば、

MATCH\_BONUS : 10

MISMATCH\_PENALTY : -20

INSERTION\_PENALTY : -40

OPENING\_PENALTY : -10

DELETION\_PENALTY : -5

である。

【0016】

上記のギャップペナルティー（INSERTION\_PENALTY、OPENING\_PENALTY）の間の関係は、ギャップ挿入ペナルティーを、ギャップオープニングコストより大きく設定することにより、ギャップオープニングの数を制限する助けとなる、すなわち、ギャップをまとめてグループ化することを支援する。当然ながら、MISMATCH\_PENALTY、MATCH\_BONUS、INSERTION\_PENALTY、OPENING\_PENALTY、およびDELETION\_PENALTYの間の代替的な関係も可能である。

10

20

30

40

50

## 【0017】

アライメントが完了したら、アライメントされた配列を、参照（すなわち、遺伝子標準物質）と比較して、バリエーションを同定しうる配列を生成するように、アセンブルすることができる。アセンブルされたリードを参照と比較して初めて、試料の遺伝子型の決定を行うことが可能になる。アセンブルされた配列を参照配列と比較したら、差異を分類し、次いで、バリエーションコールフォーマット（VCF）ファイルまたは塩基多型データベース（dbSNP）などの参照変異ファイルと比較する。しかし、この遺伝子型解析の標準的な方法は時間のかかるものであり、シーケンシング／増幅エラーが真の変異と間違えられないことを保証するためには重複が大きいリードカバレッジが必要であることが多い。

## 【0018】

遺伝子型解析工程全体は、遺伝子試料に構造バリエーション、すなわち、「通常の」ゲノムに挿入されたまたはそこから欠失したより長い（250bpまたはそれ超、例えば、1000bpまたはそれ超）配列が存在することにより、さらに複雑である。重複、逆位または転座も存在しうる。多くの場合、試料は、そのような構造バリエーション、変異、逆位、転座などの存在にある程度起因して1つの遺伝子型に「コール」される。他の場合では、構造バリエーション内の変異により、試料がさらに異なる遺伝子型に「コール」される。他の場合には、構造バリエーションの近傍にあることに起因して、目的の配列が移動している（「正常な」位置と比較して）。

しかし、いずれかの特定のリードが由来しうる公知のバリエーションが多く存在するので、最新の遺伝子型解析方法に構造バリエーションを組み込むことは非常に難しい。各N構造バリエーションについて、配列について遺伝子型解析するためにアセンブルされた配列と比較しなければならない異なる参照は、およそ $2^N$ 存在する。言い換えれば、試料について遺伝子型解析するためには、1つの構造バリエーションを調整するために、アセンブルされた配列を少なくとも2つの別々の参照配列と比較しなければならないが、20の可能性がある構造バリエーションを調整するためには、配列をおよそ百万の異なる参照と比較することが必要とされる。並行計算を用いる最新の方法を使用したとしても、これは、非常に費用のかかる提案である。さらに、この組合せの急増により、数百の可能性がある構造バリエーションを含むより長い配列を適切に遺伝子型解析することが不可能になる。したがって、計算時間を減縮するという名目で近似が行われる。言い換えれば、現行の方法では、多くのゲノムにおいて共通して見出される構造バリエーションが完全には表されない。

## 【先行技術文献】

## 【非特許文献】

## 【0019】

【非特許文献1】Needleman-Wunsch (J Mol Biol、48巻(3号)：443～453頁、1970年)

【非特許文献2】Smith-Waterman (J Mol Biol、147巻(1号)：195～197頁、1981年)

【非特許文献3】Smith-Waterman (Adv. in Math.、20巻(3号)：367～387頁、1976年)

【非特許文献4】Gotoh (J Mol Biol、162巻(3号)：705～708頁、1982年)

【非特許文献5】Huang、3章：Bio-Sequence Comparison and Alignment、Curr Top Comp Mol Biolシリーズ、Cambridge、Mass.：The MIT Press、2002年

## 【発明の概要】

## 【課題を解決するための手段】

## 【0020】

## 要旨

本発明は、配列リードを、生物のゲノム内の複数の遺伝子座における複数の対立遺伝子を同時に構成する参照配列構築物に直接アラインすることによって配列リードについて効率的に遺伝子型解析するための方法およびシステムを提供する。加えて、本発明の方法お

10

20

30

40

50

よびシステムにより、構造バリエーションに効率的に対処し、次世代シーケンシング（N . G . S . ）を使用した遺伝子試料の遺伝子型解析に必要な計算力を著しく縮小させることが可能になる。加えて、参照配列構築物は、構築物内の種々の可能性のある対立遺伝子を構成するので、試料のリードを構築物にただ単にアラインすることによって、試料について直接遺伝子型解析することが可能である。特定の遺伝子型に対しては特定のパターンのアライメントのみが可能であり、したがって、アセンブルされた配列を参照配列と比較し、次いで、バリエーションをその参照に関連する変異ファイルと比較することは必要ない。

#### 【0021】

本発明の方法およびシステムは、例えば S m i t h - W a t e r m a n - G o t o h などの線形的な局所的配列アライメント工程を、並列化の増大、速度の増大、精度の増大、および全ゲノムを通してリードをアラインする能力をもたらし、多次元アライメントアルゴリズムに変換する。本発明のアルゴリズムは、配列情報についての「遡及」型の解析をもたらし（S m i t h - W a t e r m a n におけるように）が、公知の線形的方法とは対照的に、本発明の遡及は、全般的なミスマッチ率、欠失率、および挿入率の低下を達成しながら、複雑かつ長大な配列リードについてのより正確なアライメントをもたらしために、複数の経路および複数のノードを含む多次元空間を介して実行される。多くの場合、経路のいくつかは、生物についての特定の遺伝子型を表す。したがって、リードを、特定の経路を表す対立遺伝子のセットにアラインすることにより、遺伝子型は迅速に同定される。

#### 【0022】

実際には、本発明は、配列リードを、一連の有向非巡回配列であって、アライメント内の、可能な配列バリエーションの全てまたはほぼ全てであり、挿入、欠失、置換、および構造バリエーションを含む配列バリエーションを構成する分枝点間にわたる配列にアラインすることにより実行される。有向非巡回グラフ（D A G : d i r e c t e d a c y c l i c g r a p h）として表されることが多い、このような構築物は、「許容された」参照配列およびバリエーションフォーマット（V C F）のエントリーを含む、利用可能な配列データベースから容易にアセンブルすることができる。D A G または他の有向構築物と組み合わせると、開示されるアルゴリズムにより、アライメントの正確さを著しく改善し、従来のアルゴリズムでは不可能な配列分解能を可能にする、配列アライメントのための多次元手法がもたらされる。

#### 【0023】

本発明は、加えて、生物の配列内の位置における公知のバリエーションを表す有向非巡回グラフデータ構造（D A G）を構築するための方法も含む。D A G は、何千箇所もの位置において、複数の配列を含むことが可能であり、各位置において、欠失、挿入、翻訳、逆位、一塩基多型（S N P）、および構造バリエーションを含む、複数のバリエーションを含む。D A G 内の各バリエーションに、遺伝子型または他の相関する診断情報（例えば、「乳がん」）をタグ付けし、これにより、リード内の有益な診断情報を得るのに必要とされるステップを削減することも可能である。一部の実施形態では、バリエーションが相関しており、遺伝子型判定のより高い信頼度がもたらされる。一部の実施形態では、バリエーションは、スコア付けされるか、重み付けされるか、または他のバリエーションと相関させられて疾患についてのマーカーとしてのそのバリエーションの発生率を反映する。

#### 【0024】

本発明は、本発明の方法を実行するためのシステムもさらに含む。一実施形態では、システムは、複数の配列（すなわち、核酸配列、アミノ酸配列）を、ゲノム内またはゲノムの領域内で観察されるバリエーションを表す参照配列構築物（例えば、D A G）と比較することが可能な、プロセッサおよび記憶デバイスの分散型ネットワークを含む。システムは、加えて、効率的なアライメントアルゴリズムを使用して、連続的な配列を生成するように、核酸リードをアラインすることが可能である。参照配列構築物は、膨大な冗長情報を圧縮し、アライメントアルゴリズムは、極めて効率的であるため、市販のリソースを

使用して、リードにタグ付けし、全ゲノム上でアセンブルすることもできる。システムは、複数のリードと参照配列構築物との間の複数の比較を同時に実行する複数のプロセッサを含む。比較データは、蓄積し、医療提供者へと提示することができる。比較は、計算により扱いやすいため、配列リードの解析はもはや、NGSシーケンシングと患者の遺伝的危険性についての有意義な議論との間の障壁を表さない。

【図面の簡単な説明】

【0025】

【図1A】図1は、参照配列内の遺伝子バリエーションを表す有向非巡回グラフ(DAG)の構築について描示する図である。図1(A)は、出発参照配列および欠失の付加を示す図である。図1(B)は、挿入およびSNPの付加であり、これにより、アライメントに使用される最終的なDAGに到達することを示す図である。

10

【図1B】図1は、参照配列内の遺伝子バリエーションを表す有向非巡回グラフ(DAG)の構築について描示する図である。図1(A)は、出発参照配列および欠失の付加を示す図である。図1(B)は、挿入およびSNPの付加であり、これにより、アライメントに使用される最終的なDAGに到達することを示す図である。

【図2】図2は、有向非巡回グラフとして表される3つのバリエーションコールフォーマット(VCF)のエントリーについて描示する図である。

【図3A】図3(A)は、核酸配列リードを、挿入イベントならびに参照配列を構成する構築物にアラインすることについての、図解による表示である。

【図3B】図3(B)は、核酸配列リード「ATCGAA」の適正な場所を同定するために使用される行列およびバックトラックを示す図である。

20

【図4】図4は、生物のゲノム内のある位置における2つの代替対立遺伝子および両方の対立遺伝子が組み込まれた参照配列構築物を描示する図である。第2の対立遺伝子は、長い挿入、すなわち、構造バリエーションを含むという点で、第1の対立遺伝子とは異なる。参照配列構築物を通る経路の1つが第1の対立遺伝子を表し、構築物を通る第2の経路が第2の対立遺伝子を表す。

【図5】図5は、4つの別々のリードがどのように参照配列構築物にアラインされうるかを示す図である。アラインすると、いくつかのリードは対立遺伝子#1のみまたは対立遺伝子#2のみに対応するとして同定することができ、それによって、アセンブルされたリードと参照配列との間の差異を変異ファイルと比較して特定の対立遺伝子の存在を決定する必要性が除かれる。

30

【図6】図6は、生物のゲノム内の位置における3つの代替対立遺伝子および3つの対立遺伝子全てが組み込まれた参照配列構築物を描示する図である。第2の対立遺伝子および第3の対立遺伝子は、長い挿入、すなわち、構造バリエーションを含むという点で、第1の対立遺伝子とは異なる。第2の対立遺伝子と第3の対立遺伝子は一塩基多型で異なる。参照配列構築物を通る経路の1つが第1の対立遺伝子を表し、構築物を通る第2の経路が第2の対立遺伝子を表し、構築物を通る第3の経路が第3の対立遺伝子を表す。

【図7】図7は、3つの別々のリードがどのように参照配列構築物にアラインされうるかを示す図である。アラインすると、いくつかのリードが対立遺伝子#1のみ、対立遺伝子#2のみまたは対立遺伝子#3のみに対応するとして同定され得、それによって、アセンブルされたリードと参照配列との間の差異を変異ファイルと比較して特定の対立遺伝子の存在を決定する必要性が除かれる。

40

【図8】図8は、並列処理のための連想計算モデルについて描示する図である。

【図9】図9は、並列計算のためのアーキテクチャーについて描示する図である。

【発明を実施するための形態】

【0026】

詳細な説明

本発明は、核酸配列を参照配列構築物にアラインするための方法、参照配列構築物を構築するための方法、およびアライメント法および構築物を使用して、アライメントおよびアセンブリーを生成するシステムを含む。参照配列構築物は、下記の、有向非巡回グラフ

50

(DAG)でありうるが、参照配列は、構築物が、アライメントのためにフォーマットされていることを条件として、種内の異なる生物の配列内の遺伝的変異性を反映する任意の表示でありうる。一般に、参照配列構築物は、異なる遺伝子型の間で同一な部分と、異なる遺伝子型の間で変化する部分とを含む。したがって、構築物は、様々な対立遺伝子を構成する、すなわち、異なる遺伝子型と相関する。本出願は、加えて、構築物内の種々の場所に対する核酸リードのアライメントに基づき、遺伝子型または疾患の危険性を同定するための方法を開示する。

#### 【0027】

本発明は、加えて、参照配列構築物、例えば、ゲノムの各遺伝子座における公知のバリエーションを表すDAGを使用して、特異的な遺伝子座にける特異的な塩基のコールを行う方法も提供する。配列リードは、アライメント時にDAGにアラインされるため、変異を、参照ゲノムと突き合わせて、公知の変異についての表と比較する後続のステップを廃することができる。開示される方法を使用すると、それは、核酸リードを、DAG上に表される公知の変異に位置するものとして同定し、その変異を判定することであるに過ぎない。あるいは、変異が公知でない（すなわち、参照配列構築物内に表されていない）場合、アライメントは見出され、バリエーションは、新たな変異または遺伝子型として同定されうる。この方法はまた、特異的な疾患の危険性または疾患の進行などのさらなる情報を、参照配列構築物に組み込まれている公知の変異と関連付けることも可能とする。加えて、参照配列構築物は、大量の計算資源を必要とせずに配列リードを潜在的な構造バリエーションにアラインすることを可能にする。

#### 【0028】

方法の効率が良いので、複数の配列リードを同じ参照配列構築物に対して迅速にアラインすることが可能である。リードは、典型的には少なくとも約20塩基対(bp)の長さ、例えば、少なくとも約50bpの長さ、例えば、少なくとも約80bpの長さ、例えば、少なくとも約100bpの長さ、例えば、少なくとも約150bpの長さ、例えば、少なくとも約200bpの長さである。一部の実施形態では、複数とは、約1000を超える配列リード、例えば、約10,000を超える配列リード、例えば、約100,000を超える配列リード、例えば、約1,000,000を超える配列リードを含む。一部の実施形態では、下記の通り、並列処理を使用して、複数の配列リードが参照配列構築物にアラインされる。一部の実施形態では、2つまたはそれ超の配列リードは、それらが、オリジナルの試料の同じ領域に起因することが公知であるという点で関連しうる。一部の実施形態では、配列リードは、リード間に様々な長さの挿入を有するペアメントでありうる。Illumina（商標）シーケンシングなどの種々の次世代シーケンシング技法で使用するためのペアメントを調製するための技法は公知である。

参照配列構築物

#### 【0029】

核酸リードをアラインして遺伝子型解析するのに単一の参照配列を使用する先行技術による配列アライメント法と異なり、本発明では、種内、集団内、なおまたは単一の生物体における異なる細胞間の遺伝子配列の変異性を構成しうる構築物を使用する。遺伝子バリエーションについての表示は、有向非巡回グラフ(DAG)(上記で論じた)または行一列によるアライメント行列として提示することができ、これらの構築物は、アライメントアルゴリズムのパラメータを適正に設定する(下記で論じる)ことを条件として、本発明のアライメント法と共に使用することができる。

#### 【0030】

本発明の好ましい実施形態では、構築物は、有向非巡回グラフ(DAG)である、すなわち、方向を有するが、巡回経路を有さない(すなわち、配列経路は、1回より多く参照構築物上の位置を通して進みえない)。DAGでは、配列内の遺伝子バリエーションを、代替的なノードとして表す。ノードは、保存的配列の区画の場合もあり、遺伝子の場合もあり、単に核酸の場合もある。構築物を通る、異なる可能な経路は、公知の遺伝子バリエーションを表す。DAGは、生物体の全ゲノムについて構築することもでき、DAGは、

ゲノムの部分、例えば、染色体、または遺伝情報のより小さなセグメントだけについて構築することもできる。一部の実施形態では、DAGは、1000を超える核酸、例えば、10,000を超える核酸、例えば、100,000を超える核酸、例えば、1,000,000を超える核酸を表す。DAGは、種（例えば、Homo sapiens）を表す場合もあり、選択された集団（例えば、乳がんを有する女性）を表す場合もあり、なおまたは同じ個体における異なる腫瘍細胞間の遺伝子バリエーションなど、より小さな部分集団を表す場合もある。

#### 【0031】

DAG構築の簡単な例を、図1に示す。図1（A）に示される通り、DAGは、図1（A）に配列番号1：CATAGTACCTAGGTC TTGGAGCTAGTCとして示される参照配列で始まる。実際的には、参照配列は、はるかに長いことが多く、全ゲノムでありうる。配列は、FASTAファイルまたはFASTQファイルとして保存される（FASTQは、次世代シーケンサーから生成された配列データのためのデフォルトフォーマットとなっている）ことが典型的である。一部の実施形態では、参照配列は、GRCh37などの標準的な参照でありうる。当業者により認識される通り、配列内の各文字（または記号）は、実際的には、ヌクレオチド（例えば、デオキシリボヌクレオチドまたはリボヌクレオチド）またはアミノ酸（例えば、ヒスチジン、ロイシン、リシンなど）に対応する。

#### 【0032】

次のステップでは、図1（A）の下図に示される通り、バリエントを、参照配列へと付加する。図1（A）に示されるとおり、バリエントは、図中の線の間での、参照からの配列「AG」の欠失、すなわち、配列番号2である。図上では、この欠失を、参照配列を、欠失の前後でノードへと切断し、ノードをエッジで接続し、また一方のノードから「AG」へ、次いで他方のノードへの経路を創出することにより表す。したがって、ノード間の1つの経路は参照配列を表し、他の経路は、欠失を表す。

#### 【0033】

実際的には、バリエントは、1000 Genomes Project ウェブサイトで見出されうるVCFファイルなどの、バリエントコールフォーマット（VCF）ファイル内のエントリを適用することにより、DAGに対してコールする。各VCFファイルは、特異的な場所における特異的な参照ゲノムに適合させてあるため、文字列がどこに位置するのかを同定することは、困難ではない。実際、VCFファイル内の各エントリは、図2に表される通り、参照と組み合わせて、別個のグラフを創製するエントリと考えることができる。図2中のVCFエントリは、図1のVCFエントリに対応しないことに注目されたい。

#### 【0034】

図1（B）に移ると、特異的な位置における挿入「GG」に対応する、第2のVCFエントリを付加して、伸長型DAG、すなわち、配列番号3および配列番号4を含むDAGを作製する。次に、第3のVCFエントリを、伸長型DAGに付加して、参照配列内の初期のSNP、すなわち、配列番号5～8を含むSNPを構成することができる。こうして、3つのステップで、核酸リードをそれに対してアラインさせることができるDAGが創製された（下記で論じられるとおり）。

#### 【0035】

実際的には、DAGは、コンピュータメモリ内（ハードディスク、フラッシュメモリ、クラウドメモリなどの中）に、ノードのセットSとして表され、各ノードは、文字の文字列、親ノードのセット、および位置により規定される。文字列とは、ノードの「内容物」、すなわち、配列であり、親ノードは、ノードの位置を、グラフ内の他のノードに照らして規定し、ノードの位置は、システム内のいくつかのカノニカル・オーダリング、例えば、参照ゲノムに対する位置である。グラフを、参照配列に照らして規定することが厳密に必要なわけではないが、これにより、出力データの操作が簡略となる。当然ながら、Sに対するさらなる制約は、それがループを含みえないことである。

10

20

30

40

50

## 【0036】

多くの実施形態では、ノードは、図1（A）および1（B）に示される通り、複数の文字を含むが、ノードは、図2に示される通り、例えば一塩基を表す単一の文字である可能性がある。ノードが文字の文字列を表す例では、従来のSmith-Waterman技法を用いて行われるような文字ごとの計算ではなく、ノード内の文字の全てを単一の比較ステップでアラインすることができる。結果として、計算負荷が最新の方法と比較して著しく低減される。計算負荷の低減により、アライメントを、より少ない資源を用いてより迅速に完了させることができる。数百万の小さなリードをアラインしアセンブルすることが必要とされる次世代シーケンシングにおいて使用する場合、この計算負荷の低減には、意味のある情報、すなわち、遺伝子型をより迅速に利用可能にしながら、アライメントの10  
コストを削減するという点で、明確な利点がある。患者の遺伝子型に対して処置を調整する例では、速度の増大により、患者が、最新の方法を使用した場合よりも数日早く処置を受け始めることが可能にし得る。

## 【0037】

このDAG法を、大型の構造へと外挿することにより、参照の所与の領域について、遺伝子配列内の公知の変異を表す、数千のVCFエントリを組み込むDAGを構築することが可能である。にも拘らず、DAGが高高くなると、計算も長くなるので、多くの適用では、配列の部分、例えば、染色体だけを表しうる、小型のDAGを使用する。他の実施形態では、DAGにより包含される集団のサイズを減じることにより、例えば、乳がんにおけるバリエーションを表すDAGから、トリプルネガティブ乳がんにおけるバリエーションを表すDAGへと移行することにより、DAGを小型とすることができる。あるいは、試料間で一致している、DAGの大部分を結果としてもたらしことが典型的な、容易に同定される遺伝子マーカーに基づきカスタマイズされた、長大なDAGも使用することができる。例えば、アフリカ系女性（African-ancestry female）に由来する核酸リードのセットを、アフリカ系女性（women of African ancestry）に由来するVCFエントリにより創製されたDAGに対してアラインすることの方が、同じ配列にわたりヒトにおいて公知の全てのバリエーションを構成するDAGと比較して速い。本発明のDAGは、それらが、時間の経過にわたって、新たに同定された変異を組み込むように改変されうるという点で、動的構築物であることを認識されたい。加えて、また、アライメント結果をDAGへと再帰的に付加するアルゴリズムも可能である。20  
30

## 【0038】

文字列対DAGアライメントの場合は、ギャップペナルティーを、ギャップ挿入のコストをなおより大きくし、これにより、全体的な配列内の新たなギャップのオープニングではなく、配列に対するアライメントを支援するように調整することができる。当然ながら、DAG内の改善（上記で論じた）により、変異は、DAG内で構成されるため、ギャップの発生は、なおさらに減少するはずである。

アライメントアルゴリズム

## 【0039】

一実施形態では、アルゴリズムを使用して、配列リードを、有向非巡回グラフ（DAG）に対してアラインする。「背景技術（Background）」で表されたアルゴリズムと異なり、アライメントアルゴリズムでは、DAG（例えば、参照配列構築物）上の位置において含有される各配列に対する最大スコアを同定することにより、 $C_{ij}$ の最大値を同定する。実際、先行する位置を「後ろ向きに（backwards）」見ることにより、複数の可能な経路にわたり最適のアライメントを同定することが可能である。40

## 【0040】

本発明のアルゴリズムは、上記で論じた通り、リード（別名「文字」）および有向非巡回グラフ（DAG）上で実行される。アルゴリズムを規定する目的で、Sを、アラインされる文字列とし、Dを、Sがアラインされる有向非巡回グラフとする。文字列Sの成分において、1で始まるインデックスがカッコ内に示される。したがって、Sが文字列ATCGAAであれば、 $S[1] = A$ 、 $S[4] = G$ などである。50

## 【0041】

DAGでは、ノードの配列の各文字は、別個の成分であるdとして表される。dの先行成分(predecessor)は、以下のように定義される。

(i) dが、そのノードの配列の第1の文字でなければ、そのノード内のdに先行する文字が、その(唯一の)先行成分であり、

(ii) dが、そのノードの配列の第1の文字であれば、任意のノードの配列の最後の文字であって、dのノードの親である文字が、dの先行成分である。

## 【0042】

全ての先行成分のセットは、 $P[d]$ として表す。

「最良の」アライメントを見出すために、アルゴリズムでは、Sの最初のj個の成分の、dに先行する(およびdを含む)DAGの部分による最適のアライメントについてのスコアである、 $M[j, d]$ の値を求める。このステップは、「背景技術」節中の等式1内の $H_{ij}$ を見出すステップと同様である。具体的に、 $M[j, d]$ を決定するステップは、下記：

## 【数6】

$$M[j, d] = \max\{a, i, e, 0\} \quad (6)$$

ここで、

$$P[d] \text{ 中の } p^* \text{ について、 } e = \max\{M[j, p^*] + \text{DELETE\_PENALTY}\}$$

$$i = M[j-1, d] + \text{INSERT\_PENALTY}$$

$$S[j] = d \text{ ならば } P[d] \text{ 中の } p^* \text{ について、 } a = \max\{M[j-1, p^*] + \text{MATCH\_SCORE}\};$$

$$S[j] \neq d \text{ ならば } P[d] \text{ 中の } p^* \text{ について、 } a = \max\{M[j-1, p^*] + \text{MISMATCH\_PENALTY}\}$$

で規定される通り、a、i、e、および0のうちの最大値を見出すことを伴う。

## 【0043】

上記で記載した通り、eとは、Sの最初のj個の文字の、DAGの部分であって、dまでであるが、dを含まない部分によるアライメントのうちの最高のアライメントに、追加のDELETE\_PENALTYを加えた値である。したがって、dが、ノードの配列の第1の文字でなければ、唯一の先行成分pが存在し、Sの最初のj個の文字の、DAG(pまでであり、pを含む)によるアライメントスコアは、 $M[j, p] + \text{DELETE\_PENALTY}$ と等しい。dが、そのノードの配列の第1の文字である場合、複数の可能な先行成分が存在することが可能であり、DELETE\_PENALTYは定数であるため、 $[M[j, p^*] + \text{DELETE\_PENALTY}]$ を最大化することは、先行成分を、Sの最初のj個の文字による最高のアライメントスコアと共に選択することと同じである。

## 【0044】

等式(6)では、iとは、文字列Sの最初のj-1個の文字の、dまでであり、dを含むDAGによるアライメントに、SWにおける挿入引数の定義(等式1を参照されたい)と同様のINSERT\_PENALTYを加えた値である。

## 【0045】

加えて、aとは、Sの最初のj個の文字の、DAGの部分であって、dまでであるが、dを含まない部分によるアライメントのうちの最高のアライメントに、MATCH\_SCORE(Sのj番目の文字が、文字dと同じである場合)またはMISMATCH\_PENALTY(Sのj番目の文字が、文字dと同じでない場合)を加えた値である。eと同様に、これは、dが、そのノードの配列の第1の文字でなければ、唯一の先行成分、すなわち、pが存在することを意味する。これは、aが、Sの最初のj-1個の文字の、DAG(pまでであり、pを含む)によるアライメントスコア、すなわち、dとSのj番目の文字とがマッチするのかどうかに応じて、MISMATCH\_PENALTYまたはMATCH\_SCOREを加えた $M[j-1, p]$ であることを意味する。dが、そのノードの配列の第1の文字である場合、複数の可能な先行成分が存在しうる。この場合、 $\{M[$

10

20

30

40

50

$j, p^*] + \text{MISMATCH\_PENALTY}$ または $\text{MATCH\_SCORE}$ を最大化することは、先行成分を、 $S$ の最初の $j-1$ 個の文字による最高のアライメントスコア（すなわち、 $M[j-1, p^*]$ の候補引数の最高値）と共に選択し、 $d$ と $S$ の $j$ 番目の文字とがマッチするのかどうかに応じて、 $\text{MISMATCH\_PENALTY}$ または $\text{MATCH\_SCORE}$ を加えることと同じである。

#### 【0046】

ここでもまた、「背景技術」で論じられたSWアルゴリズムの場合と同様に、ペナルティー、例えば、 $\text{DELETE\_PENALTY}$ 、 $\text{INSERT\_PENALTY}$ 、 $\text{MATCH\_SCORE}$ 、および $\text{MISMATCH\_PENALTY}$ は、少数のギャップを伴うアライメントを促すなどのように調整することができる。

10

#### 【0047】

上記の等式で記載されている通り、アルゴリズムでは、各リードについて、その成分についての挿入スコア、欠失スコア、およびマッチスコアを計算するだけでなく、DAG上の任意の先行ノードを後ろ向きに見て（DAGの方向と反対方向に）、最大のスコアを見出すことにより、最大値を見出す。こうして、アルゴリズムは、DAGを通る異なる経路であって、公知の変異を含有する経路を横断することが可能である。グラフは有向であるため、グラフの方向と反対方向に移動するバックトラックは、グラフの起点に向かって好ましいバリエーション配列に進み、最大値のアライメントスコアは、最も可能性の高いアライメントを、高い確実性で同定する。上記の等式は、「最大」値として表されるが、「最大」は、例えば、等式の全てにおいて記号を切り替え、最小値について解くことを含む、最適化の任意の形態を包含することを意図する。

20

#### 【0048】

開示されるアルゴリズムの実行について、図3で例示するが、ここで配列「ATCGA A」を、参照配列である配列番号10：TTGGATATATGGGと、公知の挿入イベントである配列番号11：

#### 【化1】

TTGGATCGAATTATGGG

「ここで、挿入には下線を付す」とを表すDAGに対してアラインする。図3（A）が、DAGと比較されるリードについての図解による表示を示すのに対し、図3（B）は、比較に対応する実際の行列を示す。「背景技術」で論じられたSmith-Waterman技法と同様に、本発明のアルゴリズムでは、最高のスコアを同定し、バックトラックを行って、リードの適正な場所を同定する。図3（A）および（B）はまた、本発明が、文字列について、構築物に対する実際のマッチをもたらすのに対し、公知の方法（例えば、SW）であったら、文字列を、参照の誤った部分にアラインする、または文字列を、アライメント内に含まれるのに十分に高いアライメントスコアを生成しないものとして棄却する可能性が高いことも強調する。配列リードが、DAG内に含まれていなかったバリエーションを含む場合、アラインされた配列は、ギャップ、挿入などを伴うと報告される。

30

参照配列構築物の適用

#### 【0049】

本発明の参照構築物およびアライメントアルゴリズムの1つの利点は、配列リードを参照配列構築物のある特定の位置において第1の配列または第2の配列のいずれかにアラインするその能力である。すなわち、本発明の参照配列構築物により、配列リードをある特定の位置において少なくとも2つの異なる配列経路、例えば、参照配列と同等である配列に進む経路、およびバリエーション（例えば、変異、多型、コピー数バリエーション、構造バリエーション）を含めた参照配列と同等である公知の配列に進む別の経路のうちの1つに対してアラインすることが可能になる。したがって、配列内の公知のバリエーションは、本発明の技法を使用して、公知のバリエーションを含有するリードを、変異を含む配列経路にアラインすることによって確実に構成し、同定することができる。

40

#### 【0050】

50

有向非巡回グラフ (DAG) を使用してリードについて遺伝子型解析する 2 つの例が図 4～7 に示されている。図 4 は、生物のゲノム内の位置における 2 つの潜在的な対立遺伝子：

配列番号 1 2 : C C C A G A A C G T T G C A T C G T A G A C G A G T T T C A G C A T T

配列番号 1 3 : C C C A G A A C G T T G C T A T G C A A C A A G G G A C A T C G T A G A C G A G T T T C A G C A T T

を示す。

#### 【0051】

本例では、2 つの対立遺伝子は、15 塩基の挿入、すなわち、構造バリエーションにより異なる。図 4 に示される通り、2 つの対立遺伝子は単一の参照配列構築物内に描示することができ、ここで、2 つの対立遺伝子は、構築物を通る異なる経路に対応する。

#### 【0052】

図 5 に示される通り、リードが参照配列構築物にアラインされると、リードは、対立遺伝子の一方または両方と即座に相関しうる。開示されるアライメントアルゴリズムを使用して、リードは、対立遺伝子 # 1 に対応する経路、対立遺伝子 # 2 に対応する経路のいずれかとアラインされる、または、経路は、リードとアラインされる領域において共通するので、リードがいずれかにアラインされうる。特定の経路にアラインされるリードの数に基づき、リードのセットを、アラインされたリードをアセンブルし、アセンブルされたリードを参照配列と比較するさらなるステップを伴わずに、対立遺伝子 # 1 に対応するとして、または対立遺伝子 # 2 に対応するとして即座にコールすることが可能である。加えて、図 5 に示される通り、方法により、リード # 4 について示される通り、大部分が共通する配列を含むリードが効率的にアラインされる。最新の方法、すなわち、線形アライメントを使用すると、リード # 4 は、考慮されないかまたは塩基の転位と仮定される尾部配列を有する第 1 の対立遺伝子にアラインされる可能性がある。しかし、開示される方法を使用すると、リード # 4 が実際に対立遺伝子 # 2 と相関することが明らかである。

#### 【0053】

図 5 の種々のリードのアライメントにより、DAG に基づく遺伝子型解析工程が、線形的な配列に基づく工程、特に単一の参照配列のみを使用し、ミスマッチを別々に取り扱う工程と比較してどのくらい明白かつ正確であるかが強調される。配列が参照 DAG にアラインされたらすぐに、リードがゲノム内のどこにあるかに関する良好な情報が得られる。そのような構築物から、特定の分枝配列を重み付けすることまたは相関させることによって追加の情報を得ることがさらに可能であり、したがって、対立遺伝子が稀であることまたはその対立遺伝子を有するという潜在的な結果を直接認識することが可能になる。

#### 【0054】

本発明の方法は、図 6 および 7 に示されるように構造バリエーション内に重要な遺伝的差異が存在する場合に、よりいっそうの汎用性を可能にする。図 6 において説明されているように、構造バリエーション内に SNP を有する第 3 の対立遺伝子というさらなる複雑さを組み込むことが可能である。したがって、図 7 に示される通り、リードを参照配列構築物 # 2 にアラインした場合、リードは、対立遺伝子 # 1 に対応する経路、対立遺伝子 # 2 に対応する経路、対立遺伝子 # 3 に対応する経路、または 3 つの対立遺伝子全てに共通する経路の一部にアラインされうる。特定の経路にアラインされるリードの数に基づき、アラインされたリードをアセンブルし、アセンブルされたリードを参照配列と比較するステップを伴わずに、リードのセットを、対立遺伝子 # 1 に対応するとして、対立遺伝子 # 2 に対応するとして、もしくは対立遺伝子 # 3 に対応するとして、またはいくつかのその組合せとして即座にコールすることが可能である。

#### 【0055】

図 4～7 に示されている状況は限定とみられるべきではなく、参照配列構築物は、多数の異なる経路を含んでよく、構築物は、遺伝的変異性の位置に対応する一連の異なる代替配列を含んでよい。現代のシーケンシングを用いた場合と同様に、数千（または数百万）

のリードを処理する場合には、統計解析および相互相関を使用して、遺伝子型コールにおける信頼レベルを評価することもできる。

#### 【0056】

重要なことに、開示される方法は、構造バリエーションの部分を含む配列リードをアラインし、結果として、遺伝子型解析することを有利に可能にする。対照的に、一次元参照配列アライメント（最新の）を使用すると、これらのリードは、アライメントスコアが低いことに起因して棄却され、構造バリエーションに対応するリードの部分内のいずれかのバリエーションは無視される可能性がある。いくつかの場合には、構造バリエーションは大きく、典型的には、1 Kbから3 Mbの間のサイズである。しかし、本出願の目的に関しては、構造バリエーションは、参照から3またはそれ超の連続した塩基対が逸脱した配列リード内の任意のバリエーションを含んでよい。ある特定の実施形態では、構造バリエーションの配列の長さは、約20 bp、50 bp、80 bp、100 bp、200 bp、300 bp、400 bp、500 bp、600 bp、700 bp、800 bp、1 Kb、1.1 Kb、1.2 Kb、1.3 Kb、1.4 Kb、1.5 Kb、1.6 Kb、1.7 Kb、1.8 Kb、1.9 Kb、2.0 Kb…2.0 Mb、2.1 Mb、2.2 Mb、2.3 Mb、2.4 Mb、2.5 Mb、2.6 Mb、2.7 Mb、2.8 Mb、2.9 Mb、3.0 Mbなどである。構造バリエーションは、遺伝的多様性および易罹患性に寄与するので、対象に対する重要な洞察をもたらす。

#### 【0057】

本発明とは異なり、従来のアライメント法（例えば、線形的な参照配列）では、構造バリエーションが同定される見込みはなく、さらには構造バリエーションの近くに位置する稀なバリエーションが同定される可能性も低い。稀なバリエーションは、所与の集団において低い確率で見出される任意の変異（インデルまたは多型など）を含む。例えば、稀なバリエーションは、例えば、25%またはそれ未満；20%またはそれ未満；15%またはそれ未満；10%またはそれ未満；または5%またはそれ未満にわたるマイナー対立遺伝子頻度を有する。（マイナー対立遺伝子頻度（MAF）とは、所与の集団において最小の共通対立遺伝子が存在する頻度を指す）。いくつかの場合には、稀なバリエーションは、まだ同定されていない、すなわち、リードをアラインする参照では表されないバリエーションを含む。いくつかの場合には、稀なバリエーションは、VCFファイル内に分類されていない。アライメント機構の展望から、そのようなバリエーションは、試料の集団におけるそれらの実際の頻度にかかわらず、事実上これまでに見られていないものである。構造バリエーションの近くに位置する稀なバリエーションは、構造バリエーションから、およそリードの長さ、すなわち、約100 bpまたはそれ未満離れていてよい。しかし、本発明は、この間隔に限定されない。いくつかの場合には、構造バリエーションの近くに位置する稀なバリエーションでは、稀なバリエーションと構造バリエーションとの間の間隔は、約1 bp～約1 Mb、例えば、約10 bp～約10,000 bp、例えば、約100 bp～約1000 bpにわたる。したがって、本発明は、試料について、大規模に、例えば、染色体または全ゲノムにおいて構造バリエーションの近くのマイナー対立遺伝子に基づき遺伝子型解析することをさらに可能にする。

#### 【0058】

いくつかの稀なバリエーションは、実質的な疾患の危険性を付与するので、配列アセンブルの間に稀なバリエーションを検出し、その後、そのような試料について遺伝子型解析する能力を最大にすることが極めて重要である。本発明の参照構築物は、多くの異なる公知の構造バリエーションを構成するので、本発明の参照構築物により、アライメント工程の間のアライメントされない構造バリエーションおよび稀なバリエーションが最小限になる。参照構築物におけるある特定の場所に少なくとも2つの構造バリエーションを含めることにより、本発明では、構造バリエーションの少なくとも1つの部分を含む配列リードを参照構築物にアラインすることが可能になる。すなわち、公知の構造バリエーションの部分を含む配列リードがアラインされ、構成されるが、線形的な参照構造では、同じ構造バリエーションはアラインすることに失敗する。本発明の結果は、構造バリエーションを含むリードは、リードがアライメント不可能なものではなくアライメント可能なものとして扱われるので、DAGに高い程度の信頼

10

20

30

40

50

度および正確性で適正にアラインすることができるというものである。

#### 【0059】

適正にアラインされた構造バリエーションでは、構造バリエーションを伴う配列リードの一部である他の配列データも同様に参照構築物にアラインされる。例えば、構造バリエーションに近接する（したがって、配列リードは構造バリエーションおよび稀なバリエーションの少なくとも一部を含む）稀なバリエーションは構造バリエーションと共に参照構築物にアラインされる。したがって、構造バリエーションに隣接する稀なバリエーションは、配列リード内の構造バリエーションがDAG参照構築物に適切にアラインされるので、多数の他のやり方で十分にアラインされた信頼できるリードに存在する。稀なバリエーションが一貫して存在することにより、バリエーションが参照構築物内で表されなくとも、それがシーケンシングエラーではなく、正当な遺伝的バリエーションであると認識されるようになる。

並列化の見込み

#### 【0060】

Smith-Waterman-Gotohアルゴリズムの逐次形は、大規模な並列化に適応し、大幅に改変されている。例えば、連想大規模並列処理 (Associative Massive Parallelism) を使用する Smith-Waterman 法 (SWAMP) と呼ばれる ASC モデルについては、参照によりその全体において本明細書に組み込まれる、米国特許公開第 2012/0239706 号において記載されている。SWAMP (および他の並列処理システム) のための並列化の一部は、任意の反対角成分 (anti-diagonal) に沿った値が、互いから独立であるという事実から来る。こうして、所与の反対角成分に沿ったセルの全ては、計算リソースを分散させるように、並列的に処理することができる。上記の再帰式で示されたデータの依存性により、達成可能な並列処理のレベルは制限されるが、ウェーブフロント法を使用することにより、この有用なアルゴリズムはさらに加速化される。Wozniak (Comput Appl in the Biosciences (CABIOS)、13 巻 (2 号) : 145~150 頁、1997 年) により、Sun Ultra SPARC 上で実行されるウェーブフロント法では、特化した SIMD 様のビデオ処理命令を使用する。Wozniak は、SIMD レジスターを使用して、副対角成分 (minor diagonal) に対応する値を保存したところ、同じマシン上の従来の実行に対して 2 倍の加速化を報告している。Wozniak の例に続く、コードを並列化する同様の方法は、ストリーミング SIMD 拡張 (SSE : Streaming SIMD Extension) セットを、x86 アーキテクチャーに使用することである。Intel により設計されたベクトル様演算では、少数の値 (通例、4 つ、8 つ、または 16 の値) に対する単一の演算 / 命令を、一度に完了させる。多くの AMD 製チップおよび Intel 製チップが、SSE の多様なバージョンを支援しており、Intel では、その最新チップセットのためのアドバンスドベクトルエクステンション (AVX) に関して、この技術の開発を継続している。

#### 【0061】

他の実行では、Rognes および Seeberg (Bioinformatics (Oxford, England)、16 巻 (8 号) : 699~706 頁、2000 年) は、Intel Pentium (登録商標) プロセッサを使用して、SSE の先行成分である、MMX SIMD 命令を、それらの実行のために使用している。Rognes および Seeberg (Bioinformatics、16 巻 (8 号) : 699~706 頁、2000 年) による、ParAlign のための作業から開発された手法では、ウェーブフロント法を使用しない (Rognes、Nuc Acids Res、29 巻 (7 号) : 1647~52 頁、2001 年 ; Saebo ら、Nuc Acids Res、33 巻 (増刊 2 号) : W535~W539 頁、2005 年)。代わりに、彼らは、クエリー配列と並行に SIMD レジスターをアラインさせ、あらかじめ計算されたクエリー特異的なスコア行列を使用して、8 つの値を一度に計算する。この方法のさらなる詳細は、参照により本明細書に組み込まれる、U. S. 7, 917, 302 において見出すことができる。Rognes および Seeberg が、SIMD レジスターを配置する (layout) 方式である、ノースネイバー依存方式によれば、SSE による並列「ベクトル」計算から得られる潜在的加速化のうちの最大 3 分の 1 が失われうる。これを克服するために、彼らは、SWAT 様最適化を組み込んで

いる。アフィングギャップペナルティを大きくすると、ノーザンネイバーは、大半の場合にゼロとなる。これが成り立つなら、プログラムは、ノースネイバーの値の計算をスキップすることが可能であり、これを、Farrar (Bioinformatics、23巻(2号)：156～161頁、2007年)は、「F遅延評価」と称している。RognesおよびSeebergの方法では、ノースネイバーの値がある特定の閾値を下回る場合には、それをスキップすることにより、等式1の計算回数を縮減して、それらのアルゴリズムを加速化することが可能である。RognesおよびSeeberg、Bioinformatics、16巻(8号)：699～706頁、2000年では、MMX/SSE命令およびSWAT様拡張を介する8元ベクトルを使用して、6倍の加速化が報告された。

#### 【0062】

Farrar (Bioinformatics、23巻(2号)：156～161頁、2007年)によりなされたSSE作業では、ストライプパターンまたはストライドパターンのアクセスを使用して、SIMDレジスターを、クエリーレジスターに沿って線形に並べる。このようにすることにより、いかなる依存性の重複も回避される。ここでもまた、SWAT様最適化(Farrar、Bioinformatics、23巻(2号)：156～161頁、2007年)を組み込むことにより、Wozniak (CABIOS、13巻(2号)：145～150頁、1997年)およびRognesおよびSeeberg (Bioinformatics (Oxford, England)、16巻(8号)：699～706頁、2000年)によるSIMD実装に対して、2～8倍の加速化が達成されている。ブロック置換行列、および効率的で巧妙な内部ループであって、ノーザン(F)条件により、その内部ループの外部へと移動させた内部ループは、重要な最適化である。16ビットエレメント、8ビットエレメントの処理のための、ストライドパターンによるメモリアクセス(strided memory pattern access)もまた、メモリアクセス時間を改善し、全体的な加速化に寄与する。

#### 【0063】

Farrar (Sequence Analysis、2008年)は、ソニー、東芝、およびIBMにより製造されたCell Processorのために、自身の作業を拡張した。このCell Processorは、1つの主コアおよび8つの副コアを有する。Cell Broadband Engineは、複数のさらなるSmith-Waterman実装であって、いずれもFarrarのストライピング法を使用する、Szalkowskiら(BMC Res Notes、1巻(107号)、2008年)によるSWPS3、およびWirawanら(BMC Bioinformatics、9巻(377号)、2008年)によるCBESWを含む実装のための、開発プラットフォームであった。Rudnickiら(Fund Inform.、96巻、181～194頁、2009年)は、PS3を使用して、複数のデータベース配列にわたる並列化を使用する方法を開発した。

#### 【0064】

Rognes (BMC Bioinformatics、12巻(221号)、2011年)はまた、SWIPEと呼ばれるマルチスレッド法であって、複数のデータベース配列を、並列的に処理するマルチスレッド法も開発している。焦点は、SIMD法を、「通常のCPU」上で使用することであった。粗視化並列処理を使用するこの探索は、複数のデータベース配列を並列的に使用する作業を分割するものであり、これは、Liuら(BMC Res Notes、2巻(73号)、2009年)ならびにLigowskiおよびRudnicki (Eight Annual International Workshop on High Performance Computational Biology、Rome、2009年)によるCUDASWに記載されているグラフィックプロセッサユニット(GPU: graphics processor unit)ベースのツールと同様である。GPU作業の他の実装は、Liuら(BMC Res Notes、3巻(93号)、2010年)およびLigowskiら(GPU Computing Gems, Emerald Edition、Morgan Kaufmann、155～157頁、2011年)によるCUDASW++2.0でなされている。

#### 【0065】

他の変形では、小スケールのベクトルによる並列化(8、16、または32元の並列処理)を、複数の配列を並列的にアラインするGPU実装を介して、計算をアクセス可能

10

20

30

40

50

とするのに使用することができる。計算の理論的なピーク加速化は、最適な加速化である  $m$  倍である。96 の処理エレメントを使用する、Clear Speed 実装について、96 倍の加速化がなされることから、理論的な加速化が確認される。

並列計算モデル

#### 【0066】

Smith-Waterman 配列アライメントを開発および拡張するのに使用される、主要な並列モデルは、連想計算 (ASC: Associative Computing) (Potterら、Computer、27 巻 (11 号): 19~25 頁、1994 年) である。本明細書では、Smith-Waterman アルゴリズムの効率的な並列バージョンが記載される。本節では、このモデルおよび他の 1 つのモデルが詳細に記載される。

10

#### 【0067】

ここでは、いくつかの関与性の語彙が定義される。フリンによるコンピュータアーキテクチャーの分類法からの 2 つの目的の用語は、並列計算の 2 つの異なるモデルである、MIMD および SIMD である。複数命令複数データ (MIMD: multiple-instruction, multiple-data) モデルと分類される、コンピュータクラスターを、超大スケールのアライメントにおけるメモリの限界を克服する概念実証として使用する。節 8 では、MIMD モデルの使用法について記載する。また、ASC として公知の、拡張型データ並列単一命令複数データ (SIMD: single-instruction multiple-data) モデルについても記載される。

複数命令複数データ (MIMD)

20

#### 【0068】

複数データ複数命令モデルまたは MIMD モデルは、現在利用可能な並列システムの大半について記載するものであり、流通している一般用コンピュータクラスターを含む。MIMD プロセッサは、各々がそれ固有のローカルメモリを伴う (Quinn、Parallel Computing: Theory and Practice、2 版、New York: McGraw-Hill、1994 年)、本格的中央処理装置 (CPU: central processing unit) を有する。SIMD モデルと異なり、MIMD プロセッサの各々は、それ固有のプログラムを、非同期的に保存および実行する。MIMD プロセッサは、それらが通信することを可能とするネットワークを介して接続されるが、使用されるネットワークは、マシン (クラスターノード) 間の Ethernet (登録商標) 接続、Myrinet 接続、および InfiniBand 接続にわたり、広く変化する。通信は、SIMD よりはるかに緩やかな通信構造を援用する傾向があり、単一のユニット内に収まらない。データは、ネットワークに沿って、個々のプロセッサにより、それらが実行している、それらの個々のプログラムの制御下で、非同期的に移送される。通信は、メッセージの送受信を支援する複数の異なる並列言語のうちの 1 つにより操作されることが典型的である。このための極めて一般的なライブラリーは、メッセージパッシングインターフェース (MPI) として公知である。「SIMD 様」方式の通信も可能であるが、データの移動は、非同期的である。MIMD による並列計算は通例、プロセッサにより実行される多様なタスクが、高度に独立 (すなわち、いわゆる「驚異的並列 (embarrassingly parallel)」問題または「あきれるほど並列 (pleasingly parallel)」問題) でない限りにおいて、広範な通信および頻繁な同期化を必要とする。節 8 で提示される作業では、InfiniBand を介して接続された、AMD Opteron クラスターを使用する。

30

40

#### 【0069】

SIMD と異なり、メッセージの送受信に必要とされる最悪の場合の時間は、予測するのが困難であるかまたは不可能である。MIMD ソフトウェアのためのメッセージの送受信の実行時間は、SIMD に典型的な、最悪の場合の理論的な評価によってではなく、試行により決定されることが多い、平均的な場合の推定値を使用して決定することが典型的である。MIMD ソフトウェアの最悪の場合、極めて悪いことが多く、生じるのはまれであるので、平均的な場合の推定値がはるかに有用である。結果として、特定の問題について MIMD に必要とされる通信時間は、SIMD の場合より長くなる可能性があり、通

50

例、有意に長い。これにより、MIMDのプログラミング（とりわけ、メッセージの送受信を使用する場合）における重要な目標であって、必要とされるプロセッサ間通信の数を最小化し、プロセッサ通信間の時間の量を最大化するという目標がもたらされる。これは、グラフィックプロセッサまたはGPUを使用する場合など、単一のカードによる加速化レベルでもなお成り立つ。

#### 【0070】

また、データ並列プログラミングも、MIMDのプログラミングで重要な技法であるが、この場合、全てのタスクは、異なるデータに対して同じ演算を実施し、多様な臨界点に限り同期化される。MIMDシステムのためのアルゴリズムの大半は、単一プログラム複数データ（SPMD: Single-Program, Multiple-Data）プログラミングパラダイムで書き込まれる。各プロセッサは、同じプログラムのそれ固有のコピーであって、そのプロセッサまたはコアに特異的なコードセクションを、そのローカルデータに対して実行するコピーを有する。SPMDパラダイムの一般性は、多数の異なるプログラムであって、異なるプロセッサにわたり同時に実行され、なおかつ、単一の問題を解くのに協同することが可能なプログラムを書き込むことは極めて困難であるという事実から来る。メモリ集約的ではあるが、計算集約的ではない問題に使用される別の手法は、節8で提示される作業を使用して、JumboMemによりなされる通り、バーチャルメモリサーバーを創出することである。ここでは、その基礎となる実行においてMPIが使用される。

単一命令複数データ（SIMD）

#### 【0071】

SIMDモデルは、PEと呼ばれる、複数の単純な演算処理エレメント（processing element）からなる。各PEは、それ固有のローカルメモリであって、PEがそこからフェッチおよび保存するメモリは有するが、プログラムをコンパイルまたは実行する能力は有さない。本明細書で使用される「並列メモリ」という用語は、計算システム内のローカルメモリを集合的に指す。例えば、並列メモリは、SIMDコンピュータシステム内のローカルメモリの集合体（例えば、PEのローカルメモリ）、MIMDコンピュータシステム内のプロセッサのローカルメモリの集合体（例えば、中央処理装置のローカルメモリ）などでありうる。プログラムの編集および実行は、制御装置（またはフロントエンド）と呼ばれるプロセッサにより操作される（Quinn, Parallel Computing: Theory and Practice、2版、New York: McGraw-Hill、1994年）。制御装置は、通例はバスにより、全てのPEへと接続される。

#### 【0072】

全てのアクティブなPEは、制御装置から受信されたプログラムの命令を、ロックステップで、同期的に実行する。「いかなる時間単位においても、単一の演算は、複数の処理装置であって、各々が異なるデータを操作する処理装置上で、同じ実行状態にある」（Quinn, Parallel Computing: Theory and Practice、2版、New York: McGraw-Hill、1994年、79頁）。全てのアクティブなPEは、同じ命令を、同時に並列的に実行するが、いくつかのPEは、任意の特定の命令をスキップすることを許容されうる（Baker, SIMD and MASC: Course notes from CS6/73301: Parallel and Distributed Computing--power point slides、（2004年）2004年）。これは通例、PEのうちの一部が、if命令を実行し、残りのPEが、else部分を実行する、「if-else」分枝構造を使用して達成される。このモデルは、「データ並列的」な性質の問題であって、たかだか少数のif-else分枝構造であり、図像処理および行列演算など、同時に生じる分枝構造を有する問題に理想的である。

#### 【0073】

制御装置は、データを、全てのアクティブなPEへとブロードキャストすることができ、制御装置はまた、制御装置とPEとの接続（通例、バス）を使用して、データ値を、特定のPEから得ることもできる。加えて、PEのセットは、線形アレイ、2Dメッシュ、またはハイパーキューブなどの相互接続ネットワークであって、PE間の並列データの移

動をもたらす相互接続ネットワークによっても接続される。データは、このネットワークを通して、同期的並列方式で、PEにより移送され、PEは、データの移動を含む命令を、ロックステップで実行する。命令を、PEへとブロードキャストするのは、制御装置である。特に、SIMDネットワークは、今日大半の並列コンピュータにより使用される、メッセージ送受信パラダイムを使用しない。このことの重要な利点は、SIMDネットワークによる通信は、極めて効率的であり、通信に必要とされる最大の時間を、その特定の通信を制御するアルゴリズムの最悪の場合の時間により決定しうることである。

#### 【0074】

本節の残りは、拡張型SIMD ASCモデルについて記載することに充てる。ASCは、本論のためのアルゴリズムの設計および開発の中心にある。

連想計算モデル

#### 【0075】

連想計算(ASC)モデルとは、Goodyear AerospaceのKenneth Batcher博士により設計されたSIMD式連想コンピュータであるSTARAN、および米国海軍で縦横に活用されているその後継モデルであるASPROに基づく拡張型SIMDである。

#### 【0076】

ケント州立大学コンピュータ科学科で開発された、ASCとは、連想計算のためのアルゴリズムモデルである(Potterら、Computer、27巻(11号):19~25頁、1994年)(Potter、Associative Computing: A Programming Paradigm for Massively Parallel Computers、Plenum Publishing、1992年)。ASCモデルは、Goodyear Aerospaceにより組み立てられた連想プロセッサであるSTARAN上およびMPP上の作業から成長した。現在ハードウェアではサポートされていないが、現在の研究努力は、このモデルを効率的にシミュレートし、かつ、このモデルのためにコンピュータを設計しようとしてなされている。

#### 【0077】

拡張型SIMDモデルとして、ASCでは、マルチタスク処理および非同期的ポイント・ツー・ポイント通信経路決定(asynchronous point-to-point communication routing)の両方を回避する、同期的データ並列プログラミングを使用する。いかなる時点においても、1つのタスクだけが実行され、このタスクの複数のインスタンスは、全てのアクティブな処理エレメント(PE)上で、ロックステップで実行されるので、マルチタスク処理は、不要である。SIMDプログラマーと同様、ASCも、ロードバランシング、同期化、および動的タスクスケジューリングを伴う課題、MPIパラダイムおよび他のMIMDクラスターパラダイムでは明示的に取り組まなくてはならない問題を回避する。

#### 【0078】

図8は、ASCコンピュータの概念モデルを示す。命令列(IS)としてもまた公知の、単一の制御装置と、各々がそれ固有のローカルメモリを伴う、複数の処理エレメント(PE)とがある。制御装置とPEアレイとは、ブロードキャスト/縮約ネットワーク(reduction network)を介して接続され、PEは、PEデータ相互接続ネットワークを介して一体に接続される。

#### 【0079】

図8で見られる通り、PEは、それ固有のローカルメモリ内に置かれたデータへのアクセスを有する。データは、その場にとどまり、応答する(アクティブな)PEが、それらのローカルデータを並列的に処理する。連想という語に対する言及は、データを、メモリアドレスではなく、内容により位置決定するための検索の使用に関する。それは、連想メモリを援用せず、その代わりに、ASCモデルとは、一般的なサイクルが、検索する~処理する~読み出す(retrieve)である、連想プロセッサである。ASCモデルについての概観は、(Potterら、Computer、27巻(11号):19~25頁、1994年)において入手可能である。

#### 【0080】

10

20

30

40

50

アルゴリズムの表形式の特徴は、それ自体、A S C データ構造本来の表形式の構造に起因して、A S C を使用する計算をもたらす。S W A M P では、ロックステップによるノースネイバーおよびノースウェストネイバーのデータシフトのための、P E 相互接続ネットワークにわたる、高度に効率的な通信、ならびに検索および並列計算にわたる最大値のための、高速定数時間による (fast constant time) 連想機能を十分に活用する。

#### 【0081】

連想演算は、A S C モデルにより必要とされる、追加のハードウェアに起因して、定数時間で実行される (Jinら、15th International Parallel and Distributed Processing Symposium (IPDPS '01) Workshops, San Francisco、193頁、2001年)。これらの演算は、任意のS I M D 様マシンにより、効率的に (それほど速くはないが) 実施することができ、複数のS I M D ハードウェアプラットフォーム上で、効率的になされるように適応させることに成功している (Yuanら、Parallel and Distributed Computing Systems (PDCS)、Cambridge, MA、2009年; Trahanら、J. of Parallel and Distributed Computing (JPDC)、2009年)。したがって、S W A M P アルゴリズムおよび他のA S C アルゴリズムは、S I M D と近縁の他のシステムであって、ベクトルマシンを含むシステム上でも効率的に実行することができ、このために、モデルは、パラダイムとして使用されている。

#### 【0082】

制御装置は、プログラムの命令を、フェッチおよび解読し、制御信号を、P E へとブロードキャストする。P E は、制御装置の指示下で、それらの固有のローカルデータを使用して、これらの命令を実行する。全てのP E は、命令を、命令間の暗黙の同期化を伴って、ロックステップ方式で実行する。A S C は、複数の関与性の高速大域演算：連想検索、最大値／最小値検索、およびレスポnderの選択／検出を有する。これらについては、以下の節において記載される。

連想機能

#### 【0083】

S W A M P アルゴリズムに関与性の機能については、下記で論じる。

連想検索

#### 【0084】

A S C アルゴリズムにおける基礎的演算は、連想検索である。連想検索では、そのローカルデータが、所与の検索キーにマッチするP E を、同時に位置決定する。マッチするデータを有するP E は、レスポnderと呼ばれ、非マッチしないデータを伴うP E は、非レスポnderと呼ばれる。検索を実施した後、次いで、アルゴリズムは、非レスポnderを無効化することにより、さらなる処理を、レスポnderに影響を及ぼす処理だけに制限することができる (またはこの逆も成り立つ)。さらなる検索を実施することにより、レスポnderのセットをさらに精緻化することができる。連想検索は、どのP E が、対角成分内の並列動作中でアクティブなのかを選択するとき、S W A M P + により縦横に活用される。

最大値／最小値検索

#### 【0085】

各P E が、標準的な比較演算子 (等しい、未満など) を使用して、そのローカルデータを、検索キーに照らして比較する、単純検索に加えて、連想コンピュータはまた、全P E アレイからのデータを一体に組み合わせて、レスポnderのセットを決定する、大域検索も実施しうる。大域検索の最も一般的な種類は、レスポnderを、それらのデータが、全P E アレイにわたる最大値または最小値であるP E とする、最大値／最小値検索である。S W A M P + は、それが処理するあらゆる対角成分内で最大値を使用して、それまでに計算された最高値を探知ける。最大値検索の使用は、高頻度で、論理的並列動作において1回ずつ、アライメント当たり  $m + n$  回生じる。

レスポnderの選択／検出

#### 【0086】

連想検索は、複数のレスポnderを結果としてもたらしることが可能であり、連想アルゴリズムは、3つの異なるモード：並列選択、逐次選択、または単独選択のうちの1つにおいて、これらのレスポnderを処理しうる。並列レスポnder処理では、同じ演算セットを、各レスポnderに対して、同時に実施する。逐次レスポnder処理では、各レスポnderを、個別に選択し、各レスポnderについて、異なる演算セットを許容する。単独レスポnder選択（pick Oneとしてもまた公知の）では、1つの任意選択されたレスポnderを選択して、処理にかける。複数のレスポnderに加えてまた、連想検索は、レスポnderを結果としてもたらしえない可能性もある。この場合を取り扱うために、ASCモデルでは、その場合に、別個のアクションのセットを検索および実施するのに任意のレスポnder（any Respondersとして公知の）が存在するのかどうかを検出することが可能である。SWAMPでは、アラインされた文字を含有する複数のレスポnderを、上述の連想検索に基づき、並列的に選択および処理する。単独レスポnder選択は、最大値／最小値検索を使用する場合に、正確な同じ最大値を有する複数の値が存在する場合、または存在するときに、生じる。

PE相互接続ネットワーク

#### 【0087】

大半の連想プロセッサは、アレイ内の並列データの移動を可能とする、一部の種類のPE相互接続ネットワークを含む。ASCモデルそれ自体は、任意の特定の相互接続ネットワークを指定せず、実際、多くの有用な連想アルゴリズムは、相互接続ネットワークを必要としない。連想プロセッサは、1D線形アレイまたは2Dメッシュなど、単純なネットワークを実装することが典型的である。これらのネットワークは、実装が簡単であり、データを、迅速に、同期方式で転送することを可能とする。例えば、1D線形アレイは、SWAMPアルゴリズムにおける、PE間の明示的通信に十分である。

並列計算システム

#### 【0088】

一般化された並列処理アーキテクチャーを、図9に示す。各コンポーネントは、直接的な接続を有するものとして示されるが、多様なエレメントは、地理的に隔てられうるが、ネットワーク、例えば、インターネットを介して、接続されうることを理解されたい。ハイブリッドコンフィギュレーションも可能であるが、並列コンピュータ内のメインメモリは、単一のアドレス空間内の全ての処理エレメント間で共有されているか、または分散されている、すなわち、各処理エレメントが、それ固有のローカルアドレス空間を有することが典型的である。（分散型メモリとは、メモリが論理的に分散されているという事実を指すがまた、それが、物理的に分散されていることもしばしば示唆する）。処理エレメントが、それ固有のローカルメモリおよび非ローカルプロセッサ上のメモリへのアクセスを有する場合、分散共有メモリおよびメモリの視覚化は、2つの手法を組み合わせる。ローカルメモリへのアクセスは、非ローカルメモリへのアクセスより速いことが典型的である。

#### 【0089】

メインメモリの各エレメントに、等しい待ち時間およびバンド幅でアクセスしうる、コンピュータアーキテクチャーは、ユニフォームメモリアクセス（UMA：Uniform Memory Access）システムとして公知である。UMAは、メモリが物理的に分散されていない、共有メモリシステムだけにより達成しうるものが典型的である。この特性を有さないシステムは、非ユニフォームメモリアクセス（NUMA：Non-Uniform Memory Access）アーキテクチャーとして公知である。分散型メモリシステムは、非ユニフォームメモリアクセスを有する。

#### 【0090】

プロセッサ間通信およびプロセッサメモリ間通信は、共有（マルチポート型またはマルチプレックス型）メモリ、クロスバースイッチ、共有バス、またはスター、リング、ツリー、ハイパーキューブ、ファットハイパーキューブ（ノードにおいて複数のプロセッサを伴うハイパーキューブ）、またはn次元メッシュを含む無数のトポロジーを有す

る相互接続ネットワークを介する方式を含む、複数の方式で、ハードウェア内に実装することができる。

【0091】

相互接続されたネットワークに基づく並列コンピュータは、直接的に接続されていないノード間のメッセージの送受信を可能とする経路決定を組み込まなければならない。プロセッサ間の通信に使用される媒体は、大型のマルチプロセッサマシン内で階層的である可能性が高い。このようなリソースは、市販されていて購入して専用で使用するか、または「クラウド」、例えば、アマゾンクラウドコンピューティングを介して、これらのリソースにアクセスすることができる。

【0092】

コンピュータは一般に、バスを介してメモリへと連結されたプロセッサを含む。メモリは、RAMまたはROMを含むことが可能であり、少なくとも1つの有形の非一時的媒体であって、システムが、本明細書に記載される機能を果たすようにさせる実行可能な命令を保存する媒体を含むことが好ましい。当業者であれば、本発明の方法の実施に必要であるかまたは最適であると認識する通り、本発明のシステムは、バスを介して互いに通信する、1または複数のプロセッサ（例えば、中央処理装置（CPU）、グラフィックプロセッサユニット（GPU）など）、コンピュータ読取り型記憶デバイス（例えば、メインメモリ、スタティックメモリなど）、またはこれらの組合せを含む。

【0093】

プロセッサは、当技術分野で公知の、任意の適切なプロセッサであって、Intel（Santa Clara、CA）により、Xeon E7という商標で販売されているプロセッサ、またはAMD（Sunnyvale、CA）により、Opteron 6200という商標で販売されているプロセッサなどのプロセッサでありうる。

【0094】

メモリは、コンピュータ読取り型記憶デバイスを指す場合があり、命令（例えば、本明細書で見出される任意の方法または機能を統合するソフトウェア）、データ（例えば、患者の染色体内で見出される遺伝子配列など、任意の有形の物理オブジェクトを統合すること）、またはこれらの両方の1または複数のセットが保存された、任意のマシン読取り型媒体を含むうる。例示的な実施形態では、コンピュータ読取り型記憶デバイスは、単一の媒体でありうるが、「コンピュータ読取り型記憶デバイス」という用語は、命令またはデータの1または複数のセットを保存する、単一の媒体または複数の媒体（例えば、集中型データベースもしくは分散型データベース、ならびに／または関連するキャッシュおよびサーバー）を含むものと理解されたい。したがって、「コンピュータ読取り型記憶デバイス」という用語は、限定なしに、ソリッドステートメモリ（例えば、加入者識別モジュール（SIM）カード、セキュアデジタルカード（SDカード）、マイクロSDカード、またはソリッドステートドライブ（SSD））、光学媒体および磁気媒体、ならびに他の任意の有形記憶媒体を含むものと理解されたい。好ましくは、コンピュータ読取り型記憶デバイスは、有形の非一時的媒体を含む。このような非一時的媒体は、例えば、一過性の波動および信号を除外する。「非一時的メモリ」は、信号それ自体など、コンピュータ読取り型伝送媒体を除外すると解釈されたい。

【0095】

本発明に従う入力／出力デバイスは、ビデオディスプレイユニット（例えば、液晶ディスプレイ（LCD）またはブラウン管（CRT）モニター）、英数字入力デバイス（例えば、キーボード）、カーソル制御デバイス（例えば、マウスまたはトラックパッド）、ディスクドライブユニット、信号発生器（例えば、スピーカー）、タッチスクリーン、加速度計、マイクロフォン、セルラー式ラジオ波アンテナ、および、例えば、ネットワークインターフェースカード（NIC）、Wi-Fiカード、またはセルラー式モデムでありうる、ネットワークインターフェースデバイスを含むうる。

試料の収集および調製

【0096】

本発明は、生物学的試料から回収された核酸に対応する配列（例えば、核酸配列、アミノ酸配列）を生成するための方法を含む。一部の実施形態では、結果として得られる情報を使用して、被験体から得られた核酸材料中に存在する変異を同定することができる。一部の実施形態では、試料、すなわち、核酸（例えば、DNAまたはRNA）を被験体から得、核酸を処理し（溶解させ、増幅し、かつ／または精製し）、下記に記載される方法を使用して、核酸をシーケンシングする。多くの実施形態では、シーケンシングの結果は、直鎖状の核酸配列ではなく、数千または数百万もの個々の短い核酸リードであって、被験体についての配列へと再アセンブルしなければならない核酸リードのコレクションである。リードをアラインして配列を生成したら、アラインされた配列を、参照配列と比較して、例えば、疾患を指し示し得る変異を同定することができる。他の実施形態では、リードの、参照配列構築物、すなわち、上記で記載した、有向非巡回グラフ（「DAG」）に対するアライメントに基づき、特定の変異を有する被験体を同定することができる。

10

#### 【0097】

上記の目的のうちのいずれのためにも、方法を生物学的試料へと適用することができる。生物学的試料は、例えば、血液試料、全血、血漿、涙液、乳首吸引物、血清、糞便、尿、唾液、循環細胞、組織、生検試料、毛包、または患者の生物学的材料を含有する他の試料を含みうる。このような試料に基づき検査を行うときの1つの問題は、大半の場合において、目的の変異を含有するDNAまたはRNAであって、試料中に存在しうるDNAまたはRNAは、ごく微量でありうることである。これは、とりわけ、口腔内スワブ試料または血液試料などの非侵襲的試料であって、バリエーション核酸が、極めて少量で存在する非侵襲的試料に当てはまる。一部の実施形態では、核酸断片は、天然の短鎖でありうる、すなわち、試料中の関与性の核酸のランダムなせん断により、短い断片が作り出されうる。他の実施形態では、処理を容易とするため、またはシーケンシング技法では、1000塩基未満、例えば、500塩基未満、例えば、200塩基未満、例えば、100塩基未満、例えば、50塩基未満のリードだけをシーケンシングしうるため、核酸を意図的に断片化する。本明細書に記載される方法を使用して、様々な長さの配列をアラインしうるが、一部の実施形態では、複数の核酸リードの大部分は、シーケンシング法から得られ、1000塩基未満、例えば、500塩基未満、例えば、200塩基未満、例えば、100塩基未満、例えば、50塩基未満を含む。

20

#### 【0098】

核酸は、当技術分野で公知の方法により得ることができる。一般に、核酸は、その内容が、参照によりその全体において本明細書に組み込まれる、Maniatisら、Molecular Cloning: A Laboratory Manual、Cold Spring Harbor、N.Y.、280～281頁（1982年）により記載されている技法など、様々な技法により生物学的試料から抽出することができる。

30

#### 【0099】

十分に純粋な核酸調製物を得るためには、まず、試料の抽出物を調製し、次いで、さらなるステップ（すなわち、分別沈殿、カラムクロマトグラフィー、有機溶媒による抽出など）を実施することが必要でありうる。抽出物は、当技術分野における標準的な技法を使用して、例えば、細胞の化学的溶解または機械的溶解により調製することができる。次いで、抽出物は、例えば、濾過および／もしくは遠心分離により、かつ／あるいはイソチオシアン酸グアニジニウムもしくは尿素などのカオトロピック塩、またはフェノールおよび／もしくはHCCl<sub>3</sub>などの有機溶媒によりさらに処理して、任意の夾雑するタンパク質および潜在的に干渉するタンパク質を変性させることができる。一部の実施形態では、試料は、対象試料、例えば、血液試料から収集されたRNA、例えば、mRNAを含みうる。当技術分野では、RNA抽出のための一般的な方法が周知であり、Ausubelら、Current Protocols of Molecular Biology、John Wiley and Sons（1997年）を含む、分子生物学の標準的な教科書において開示されている。パラフィン包埋組織からのRNA抽出のための方法は、例えば、RuppおよびLocker、Lab Invest.、56巻：A67頁（1987年）、およびDe Andresら、BioTechniques、18巻：42044頁（1995

40

50

年)において開示されている。これらの参考文献の各々の内容は、参照によりそれらの全体において本明細書に組み込まれる。特に、RNAの単離は、Qiagenなど、商業的製造元からの精製キット、緩衝液セット、およびプロテアーゼを、製造元の指示に従い使用して、実施することができる。例えば、培養物中の細胞に由来する全RNAは、Qiagen RNeasy miniカラムを使用して単離することができる。他の市販のRNA単離キットは、MASTERPURE Complete DNA and RNA Purification Kit (EPICENTRE、Madison、Wis.)、およびParaffin Block RNA Isolation Kit (Ambion、Inc.)を含む。組織試料に由来する全RNAは、RNA Stat-60 (Tel-Test)を使用して単離することができる。腫瘍から調製されたRNAは、例えば、塩化セシウム密度勾配遠心分離により単離することができる。

10

解析的シーケンシング

#### 【0100】

シーケンシングは、当技術分野で公知の任意の方法によることができる。DNAシーケンシング技法は、標識されたターミネーターまたはプライマーおよびスラブ内またはキャピラリー内のゲル分離を使用する、古典的なジデオキシシーケンシング反応(サンガー法)、可逆的終結型標識ヌクレオチドを使用する、合成によるシーケンシング、パイロシーケンシング、454シーケンシング、標識されたオリゴヌクレオチドプローブのライブラリーとの、対立遺伝子特異的ハイブリダイゼーション、標識されたクロンのライブラリーとの対立遺伝子特異的ハイブリダイゼーションに続いてライゲーションを使用する、合成によるシーケンシング、重合化ステップの間における、標識されたヌクレオチドの組込みについての、リアルタイムモニタリング、ポロニーシーケンシング、およびSOLiDシーケンシングを含む。分離された分子のシーケンシングは、より近年になって、ポリメラーゼまたはリガーゼを使用する、逐次的伸長反応または単一の伸長反応によるほか、プローブのライブラリーとの単一のディファレンシャルハイブリダイゼーションまたは逐次的なディファレンシャルハイブリダイゼーションによっても裏付けられている。シーケンシングの前に、試料中の核酸の一部または全部を増幅することは、さらに有益でありうる。一部の実施形態では、核酸を、当技術分野で公知のポリメラーゼ連鎖反応(PCR)技法を使用して増幅する。

20

#### 【0101】

本発明の方法で使用されうるシーケンシング技術の一例は、DNAまたはRNAを増幅するのに活用されうる、合成によるポリメラーゼベースの配列(polymerase-based sequence-by-synthesis)である、Illuminaシーケンシング(例えば、MiSeq(商標)プラットフォーム)である。DNAのためのIlluminaシーケンシングは、固体表面上のDNAの増幅であって、フォールドバックPCRおよびアンカリングされたプライマーを使用する増幅に基づく。ゲノムDNAを、断片化し、アダプターを、断片の5'末端および3'末端へと付加する。フローセルチャネルの表面へと結合させたDNA断片を伸長させ、ブリッジ増幅する。断片は二本鎖となり、二本鎖分子を変性させる。複数サイクルにわたる固相増幅に続く変性により、フローセルの各チャネル内に、同じ鋳型の約1,000コピーの一本鎖DNA分子による数百万のクラスターを創製することができる。プライマー、DNAポリメラーゼ、および4つのフルオロフォアで標識された可逆的終結型ヌクレオチドを使用して、逐次シーケンシングを実施する。ヌクレオチド組込みの後、レーザーを使用して、フルオロフォアを励起し、画像を捕捉し、第1の塩基の同定を記録する。3'側ターミネーターおよび組み込まれた各塩基からフルオロフォアを除去し、組込みステップ、検出ステップ、および同定ステップを繰り返す。Illuminaシーケンシングを使用して、RNAを検出する場合、試料のRNA発現を決定するために、RNA断片を単離および増幅することを除き、同じ方法が適用される。配列は、シーケンサーで直接情報を取った後、生物学的配列および品質スコアを保存するための、テキストベースのフォーマットである、FASTQファイルなどのデータファイルに出力することができる(上記の議論を参照されたい)。

30

40

50

## 【0102】

本発明の方法で使用されうるDNAシーケンシング技法の別の例は、Life Technologies製のIon Torrent（商標）シーケンシングである。それらの各々の内容が、参照によりその全体において本明細書に組み込まれる、米国特許出願第2009/0026082号、同第2009/0127589号、同第2010/0035252号、同第2010/0137143号、同第2010/0188073号、同第2010/0197507号、同第2010/0282617号、同第2010/0300559号、同第2010/0300895号、同第2010/0301398号、および同第2010/0304982号を参照されたい。Ion Torrent（商標）シーケンシングでは、DNAを、約300～800塩基対の断片へとせん断すると、断片は、平滑末端となる。次いで、オリゴヌクレオチドアダプターを、断片の末端へとライゲーションする。アダプターは、断片の増幅およびシーケンシングのためのプライマーとして働く。断片を、表面へと結合させ、断片が個別に分解可能となるような分解能で結合する（attached）。1または複数のヌクレオチドの付加により、プロトン（H<sup>+</sup>）が放出され、このシグナルは、シーケンシング計器により検出および記録される。シグナル強度は、組み込まれたヌクレオチドの数に比例する。Ion Torrent データはまた、FASTQファイルとしても出力される。

10

## 【0103】

本発明の方法で使用されうるDNAシーケンシング技法およびRNAシーケンシング技法の別の例は、454（商標）シーケンシング（Roche）（Margulies, Mら、2005年、Nature、437巻、376～380頁）である。454（商標）シーケンシングは、合成によるシーケンシング技術であって、パイロシーケンシングもまた活用する技術である。DNAの454（商標）シーケンシングは、2つのステップを伴う。第1のステップでは、DNAを、約300～800塩基対の断片へとせん断し、断片は、平滑末端となる。次いで、オリゴヌクレオチドアダプターを、断片の末端へとライゲーションする。アダプターは、断片の増幅およびシーケンシングのためのプライマーとして働く。断片は、例えば、5'-ビオチンタグを含有するAdaptor Bを使用して、DNA捕捉ビーズ、例えば、ストレプトアビジンでコーティングされたビーズへと結合させることができる。ビーズへと結合させた断片は、油-水エマルジョンの液滴内でPCR増幅する。結果は、各ビーズ上でクローン増幅されたDNA断片の複数のコピーである。第2のステップでは、ビーズを、ウェル（ピコリットルサイズの）内で捕捉する。パイロシーケンシングは、各DNA断片に対して並行的に実施する。1または複数のヌクレオチドの付加により、光シグナルが発生し、この光を、シーケンシング計器内のCCDカメラで記録する。シグナル強度は、組み込まれたヌクレオチドの数に比例する。パイロシーケンシングでは、ヌクレオチドが付加されると放出される、ピロリン酸（PPi）を使用する。PPiは、アデノシン5'-ホスホ硫酸の存在下で、ATPスルフィラーゼにより、ATPへと転換される。ルシフェラーゼは、ATPを使用して、ルシフェリンを、オキシルシフェリンへと転換し、この反応が、光を発生させ、これが検出および解析される。別の実施形態では、パイロシーケンシングを使用して、遺伝子発現を測定する。RNAについてのパイロシーケンシングも、DNAについてのパイロシーケンシングと同様に適用され、部分rRNA遺伝子配列（partial rRNA gene sequencings）を微小ビーズへと結合させ、次いで、結合物を個々のウェルに入れることにより達成する。次いで、遺伝子発現プロファイルを決定するために、結合させた部分rRNA配列を増幅する。Sharon Marsh、Pyrosequencing(登録商標) Protocols, Methods in Molecular Biology、373巻、15～23頁（2007年）。

20

30

40

## 【0104】

本発明の方法で使用されうるDNA検出技法およびRNA検出技法の別の例は、SOLID（商標）技術（Applied Biosystems）である。SOLID（商標）技術システムとは、ライゲーションベースのシーケンシング技術であって、DNAおよびRNAのいずれについての超並列次世代シーケンシングを行うのにも活用されうる技術

50

である。DNA SOLiD (商標) シーケンシングでは、ゲノムDNAを、断片へとせん断し、アダプターを、断片の5'末端および3'末端へと結合させて、断片ライブラリーを生成する。あるいは、内部アダプターは、アダプターを、断片の5'末端および3'末端へとライゲーションし、断片を環状化し、環状化させた断片を消化させて、内部アダプターを生成し、アダプターを、結果として得られる断片の5'末端および3'末端へと結合させて、メートペア (MP: m a t e - p a i r e d) ライブラリーを生成することにより導入することができる。次に、クローンビーズ集団を、ビーズ、プライマー、鋳型、およびPCR成分を含有するマイクロリアクター内で調製する。PCR後、鋳型を変性させ、ビーズを富化して、伸長した鋳型を伴うビーズを分離する。選択されたビーズ上の鋳型を、スライドガラスへの結合を可能とする3'修飾にかける。配列は、逐次ハイブリダイゼーションと、中央部の決定された塩基 (または塩基対) であって、特異的なフルオロフォアにより同定される塩基を伴う、部分的にランダムなオリゴヌクレオチドのライゲーションとにより決定することができる。色を記録した後で、ライゲーションされたオリゴヌクレオチドを切断および除去し、次いで、プロセスを繰り返す。

#### 【0105】

他の実施形態では、SOLiD (商標) 遺伝子発現連鎖解析 (SAGE: S e r i a l A n a l y s i s o f G e n e E x p r e s s i o n) を使用して、遺伝子発現を測定する。遺伝子発現連鎖解析 (SAGE) とは、各転写物についての個別のハイブリダイゼーションプローブを準備する必要なしに、多数の遺伝子転写物についての同時的で定量的な解析を可能とする方法である。まず、タグが、各転写物内の固有の位置から得られることを条件として、短い配列タグ (約10~14bp) であって、転写物を固有に同定するのに十分な情報を含むタグを生成する。次いで、多くの転写物を併せて連結して、長い連鎖分子であって、シーケンシングすることが可能であり、複数のタグの識別を同時に明らかにする分子を形成する。転写物の任意の集団の発現パターンは、個々のタグの存在度を決定し、各タグに対応する遺伝子を同定することにより、定量的に評価することができる。さらなる詳細については、例えば、それらの各々の内容が、参照によりそれらの全体において本明細書に組み込まれる、Velculescuら、Science、270巻: 484~487頁 (1995年) ; およびVelculescuら、Cell、88巻: 243~51頁 (1997年) を参照されたい。

#### 【0106】

本発明の方法で使用されうる別のシーケンシング技法は、例えば、Helicosの真の1分子のシーケンシング (tSMS: T r u e S i n g l e M o l e c u l e S e q u e n c i n g) (Harris T. D.ら (2008年)、Science、320巻: 106~109頁) を含む。tSMS技法では、DNA試料を、約100~200ヌクレオチドの鎖へと切断し、polyA配列を、各DNA鎖の3'末端へと付加する。各鎖を、蛍光標識されたアデノシンヌクレオチドの付加により標識する。次いで、DNA鎖を、フローセル表面へと固定化された、数百万ものオリゴ-T捕捉部位を含有するフローセルとハイブリダイズさせる。鋳型は、1cm<sup>2</sup>あたりの鋳型約1億個の密度でありうる。次いで、フローセルを、計器、例えば、Heliscope (商標) シーケンサーへとローディングし、レーザーでフローセルの表面を照射し、各鋳型の位置を明らかにする。CCDカメラにより、フローセル表面上の鋳型の位置をマッピングすることができる。次いで、鋳型の蛍光標識を、切断し、洗い落とす。DNAポリメラーゼと、蛍光標識されたヌクレオチドとを導入することにより、シーケンシング反応を開始する。オリゴ-T核酸は、プライマーとして働く。ポリメラーゼにより、標識されたヌクレオチドを、プライマーへと、鋳型指向的な様式で組み込む。ポリメラーゼおよび組み込まれなかったヌクレオチドは、除去する。蛍光標識されたヌクレオチドの組み込みを方向付けた鋳型は、フローセル表面をイメージングすることにより検出する。イメージングの後、切断ステップにより、蛍光標識を除去し、所望のリード長が達成されるまで、他の蛍光標識されたヌクレオチドについても、プロセスを繰り返す。配列情報は、各ヌクレオチドの付加ステップにより収集する。tSMSについてのさらなる記載は、例えば、Lapidusら (米国特許第7, 169, 56

10

20

30

40

50

0号)、Lapidusら(米国特許出願第2009/0191565号)、Quakeら(米国特許第6,818,395号)、Harris(米国特許第7,282,337号)、Quakeら(米国特許出願第2002/0164629号)、およびBraslavskyら、PNAS(USA)、100巻:3960~3964頁(2003年)において示されており、これらの参考文献の各々の内容は、参照によりその全体において本明細書に組み込まれている。

#### 【0107】

本発明の方法で使用されうるシーケンシング技術の別の例は、DNAおよびRNAのいずれもシーケンシングする、Pacific Biosciencesによる単一分子リアルタイム(SMRT:single molecule, real-time)技術を含む。SMRTでは、4つのDNA塩基の各々を、4つの異なる蛍光色素のうちの1つへと結合させる。これらの色素は、リン酸連結されている(phospholinked)。単一のDNAポリメラーゼを、鋳型である一本鎖DNAの単一の分子と共に、ゼロモード導波管(ZMW:zero-mode waveguide)の底部に固定化する。ZMWとは、単一のヌクレオチドの、DNAポリメラーゼによる組み込みの、ZMWの内外へと急速に(数マイクロ秒間で)拡散する蛍光ヌクレオチドのバックグラウンドに対する観察を可能とする閉じ込め構造である。ヌクレオチドを成長しつつある鎖へと組み込むには、数ミリ秒間かかる。この時間中に、蛍光標識が励起され、蛍光シグナルをもたらし、蛍光タグが切断される。色素の対応する蛍光の検出により、どの塩基が組み込まれたのかが指し示される。プロセスを繰り返す。RNAをシーケンシングするためには、ZMWでは、DNAポリメラーゼを、逆転写酵素で置き換え、相応のプロセスに従う。

#### 【0108】

本発明の方法で使用されうるシーケンシング技法の別の例は、ナノ細孔シーケンシング(Soni G VおよびMeller, A, Clin Chem, 53巻:1996~2001頁、2007年)である。ナノ細孔とは、直径が1ナノメートルのオーダーの小孔である。ナノ細孔を、導電性流体中に浸漬し、ナノ細孔にわたり電位を印加する結果として、ナノ細孔を通るイオンの伝導に起因する微弱な電流がもたらされる。流れる電流の量は、ナノ細孔のサイズに対して感受性である。DNA分子が、ナノ細孔を通して通過するとき、DNA分子上の各ヌクレオチドは、ナノ細孔を、異なる程度で閉塞させる。こうして、DNA分子が、ナノ細孔を通して通過するときに、ナノ細孔を通して通過する電流の変化は、DNA配列の読取りを表す。

#### 【0109】

本発明の方法で使用されうるシーケンシング技法の別の例は、化学感受性電界効果トランジスター(chemFET:chemical-sensitive field effect transistor)アレイを使用して、DNAをシーケンシングするステップ(例えば、米国特許出願公開第20090026082号において記載されている)を伴う。技法の一例では、DNA分子を、反応チャンバー内に入れることができ、鋳型分子を、ポリメラーゼに結合したシーケンシングプライマーへとハイブリダイズさせることができる。シーケンシングプライマーの3'末端における、1または複数の三リン酸の、新たな核酸鎖への組み込みは、電流の変化によって、chemFETにより検出することができる。アレイは、複数のchemFETセンサーを有しうる。別の例では、単一の核酸を、ビーズへと結合させることができ、核酸を、ビーズ上で増幅することができ、個々のビーズを、chemFETアレイ上の個々の反応チャンバーであって、各チャンバーがchemFETセンサーを有するチャンバーへと移送することができ、核酸をシーケンシングすることができる。

#### 【0110】

本発明の方法で使用されうるシーケンシング技法の別の例は、電子顕微鏡(Moudrianakis E. N.およびBeer M., Proc Natl Acad Sci USA., 1965年3月、53巻:564~71頁)を使用するステップを伴う。技法の一例では、電子顕微鏡を使用して識別可能な金属標識を使用して、個々のDNA分子を標識する。次いで、これらの分子を、平面上で伸長させ、配列を測定するのに電子顕微鏡を使用してイメージングする。

## 【0111】

さらなる検出法では、マイクロアレイへの結合を、後続の蛍光検出または非蛍光検出、質量分析的方法を使用する、バーコードによる質量検出、発せられたラジオ波の検出、アラインされたバーコードからの散乱光の検出、定量的PCR法またはデジタルPCR法を使用する蛍光の検出のために活用することができる。比較核酸ハイブリダイゼーションアレイとは、患者の試料DNA中のコピー数バリエーションを検出するための技法である。試料DNAと、参照DNAとを、例えば、顕著に異なるフルオロフォアを使用して、異なる様式で標識し、次いで、多数のプローブとハイブリダイズさせる。次いで、試料および参照の蛍光強度を測定し、次いで、蛍光強度比を使用して、コピー数バリエーションを計算する。比較ゲノムハイブリダイゼーションアレイの方法については、Shinawi M、Chung SW、The array CGH and its clinical applications、Drug Discovery Today、13巻（17～18号）：760～70頁においてより詳細に論じられている。マイクロアレイによる検出から、FASTQファイルを直接生成することはできないが、マイクロアレイシーケンサーにより作成されたデータを、FASTQまたは同様のフォーマットへと転換するプログラムが利用可能である。

## 【0112】

DNA分子、RNA分子、およびコピー数を検出する別の方法は、蛍光in situ ハイブリダイゼーション（FISH：fluorescent in situ hybridization）である。In Situ Hybridization Protocols（Ian Darby編、2000年）。FISHとは、DNA配列内の変異およびコピー数変動など、特異的な染色体再配列を検出する、分子細胞遺伝学技法である。DNA分子を化学的に変性させ、2つの鎖へと分離する。次いで、一本鎖プローブを、変性させたDNA鎖と共にインキュベートする。一本鎖プローブ（signals stranded probe）は、標的配列部分に応じて選択され、相補的配列部分に対する高アフィニティを有する。プローブは、反復配列プローブ、全染色体プローブ、および遺伝子座特異的プローブを含みうる。インキュベート中に、組み合わせられたプローブとDNA鎖とをハイブリダイズさせる。次いで、任意のバリエーションを評価するために、結果を、顕微鏡下で視覚化および定量する。

## 【0113】

別の実施形態では、Mass ARRAY（商標）ベースの遺伝子発現プロファイリング法を使用して、遺伝子発現を測定する。Sequenom, Inc.（San Diego, Calif.）により開発されたMass ARRAY（商標）ベースの遺伝子発現プロファイリング法では、RNAの単離および逆転写の後、得られたcDNAを、単一の塩基を除く全て位置において、ターゲティングされるcDNA領域にマッチし、内部標準として働く、合成DNA分子（コンペティター）とスパイクする。cDNA／コンペティター混合物を、PCR増幅し、PCR後、小エビアルカリホスファターゼ（SAP）酵素処理にかけ、その結果として、残りのヌクレオチドの脱リン酸化をもたらす。アルカリホスファターゼを不活化させた後、コンペティターおよびcDNAに由来するPCR産物を、プライマー伸長にかけ、これにより、コンペティターに由来するPCR産物およびcDNAに由来するPCR産物について、顕著に異なる質量シグナルを発生させる。精製後、これらの産物を、マトリックス支援レーザー脱着イオン化飛行時間質量分析（MALDI-TOF MS：matrix-assisted laser desorption ionization time-of-flight mass spectrometry）による解析に必要とされる成分をあらかじめローディングされたチップアレイ上に分注する。次いで、反応物中に存在するcDNAを、作成された質量スペクトル内のピーク面積の比を解析することにより定量する。さらなる詳細については、例えば、DingおよびCantor、Proc. Natl. Acad. Sci. USA、100巻：3059～3064頁（2003年）を参照されたい。

## 【0114】

さらなるPCRベースの技法は、例えば、ディファレンシャルディスプレイ（LiangおよびPardee、Science、257巻：967～971頁（1992年））；増幅フラグメン

ト長多型 (iAFLP) (Kawamotoら、Genome Res.、12巻：1305～1312頁 (1999年)) ; Bead Array (商標) 技術 (Illumina、San Diego、Calif. ; Oliphantら、Discovery of Markers for Disease (Biotechniquesへの付録)、2002年6月 ; Fergusonら、Analytical Chemistry、72巻：5618頁 (2000年)) ; 市販のLuminex 100 LabMAPシステムおよび複数色でコードされたマイクロスフェア (Luminex Corp.、Austin、Tex.) を、遺伝子発現のための迅速アッセイで使用される、遺伝子発現の検出のためのビーズアレイ (BADGE) (Yangら、Genome Res.、11巻：1888～1898頁 (2001年)) ; ならびに高カバレッジ発現プロファイリング (HiCEP) 解析 (Fukumuraら、Nucl. Acids. Res.、31巻 (16号) e94頁 (2003年)) を含む。それらの各々の内容が、参照によりそれらの全体において本明細書に組み込まれる。

10

#### 【0115】

ある特定の実施形態ではまた、遺伝子発現の変動も、例えば、Affymetrix (Santa Clara、CA) から市販されているアレイなど、ナイロン膜アレイ、マイクロチップアレイ、およびスライドガラスアレイを含む、マイクロアレイ技法を使用して、同定または確認することができる。一般に、RNA試料は、単離され、逆転写を介して、標識されたcDNAへと転換される。次いで、標識されたcDNAを、ナイロン膜、マイクロチップ、またはスライドガラス上で、目的の細胞または組織に由来する、特異的なDNAプローブとハイブリダイズさせる。次いで、ハイブリダイズさせたcDNAを検出および定量し、結果として得られる遺伝子発現データを、解析のために対照と比較することができる。標識化法、ハイブリダイゼーション法、および検出法は、マイクロアレイの支持体が、ナイロン膜であるのか、マイクロチップであるのか、スライドガラスであるのかに応じて変化する。ナイロン膜アレイは、P-dNTPで標識されたプローブとハイブリダイズさせることが典型的である。スライドガラスアレイは、2つの顕著に異なる、蛍光標識されたヌクレオチドによる標識化を伴うことが典型的である。マイクロアレイを作製し、遺伝子産物の発現 (例えば、RNAまたはタンパク質) を決定するための方法は、その内容が参照によりその全体において本明細書に組み込まれる、Yeatmanら (米国特許出願第2006/0195269号) に示されている。

20

#### 【0116】

一部の実施形態では、質量分析 (MS) による解析は、生物学的試料中の、本明細書で開示される、1または複数のバイオマーカーの存在および/または量を決定するのに、単独で使用することもでき、他の方法 (例えば、イムノアッセイまたはRNA測定アッセイ) と組み合わせることもできる。一部の実施形態では、MS解析は、例えば、ダイレクトスポットMALDI-TOFまたは液体クロマトグラフィーMALDI-TOF質量分析による解析など、マトリックス支援レーザー脱着イオン化 (MALDI) 飛行時間 (TOF) MS解析を含む。一部の実施形態では、MS解析は、例えば、液体クロマトグラフィー (LC) ESI-MSなどのエレクトロスプレーイオン化 (ESI) MSを含む。質量分析は、市販の分光光度計を使用して達成することができる。当技術分野では、MALDI-TOF MSおよびESI-MSを含むMS解析を活用して、生物学的試料中のバイオマーカーペプチドの存在および量を検出するための方法が公知である。さらなる指針については、例えば、それらの各々が参照によりその全体において本明細書に組み込まれる、米国特許第6,925,389号 ; 同第6,989,100号 ; および同第6,890,763号を参照されたい。

30

40

#### 【0117】

本発明の方法、配列構築物、およびシステムを伴う使用のためのタンパク質配列は、当業者に公知の多数の技法を使用して決定することができる。例えば、アミノ酸配列およびアミノ酸配列リードは、質量分析により、またはエドマン分解を使用して、タンパク質またはタンパク質の部分を解析することにより作製することができる。質量分析は、例えば、ダイレクトスポットMALDI-TOFまたは液体クロマトグラフィーMALDI-TOF質量分析による解析などの、マトリックス支援レーザー脱着イオン化 (MALDI)

50

飛行時間 (TOF) MS 解析、例えば、液体クロマトグラフィー (LC) ESI-MS などのエレクトロスプレーイオン化 (ESI) MS、または MS-MS など、他の技法を含みうる。エドマン分解による解析は、Model 49X Procise タンパク質／ペプチドシーケンサー (Applied Biosystems / Life Technologies) など、市販の計器を使用して実施することができる。シーケンシングされたアミノ酸配列、すなわち、ポリペプチド、すなわち、タンパク質は、少なくとも 10 アミノ酸の長さ、例えば、少なくとも 20 アミノ酸の長さ、例えば、少なくとも 50 アミノ酸の長さでありうる。

## 参照による組込み

【 0 1 1 8 】

本開示を通して特許、特許出願、特許公開、雑誌、書籍、論文、ウェブコンテンツなど、他の文献に対する言及および引用を行ってきた。全てのこのような文献は、参照によりそれらの全体において全て目的で本明細書に組み込まれる。

同等物

【 0 1 1 9 】

当業者には、本明細書で示され、記載される実施形態に加えて、本発明の多様な改変およびその多くのさらなる実施形態も、本明細書で引用される研究文献および特許文献への言及を含む、本明細書の全内容から明らかとなろう。本明細書における対象物は、その多様な実施形態における本発明およびその同等物の実施に適応させうる、重要な情報、例示、および指針を含有する。

【図 1 A】

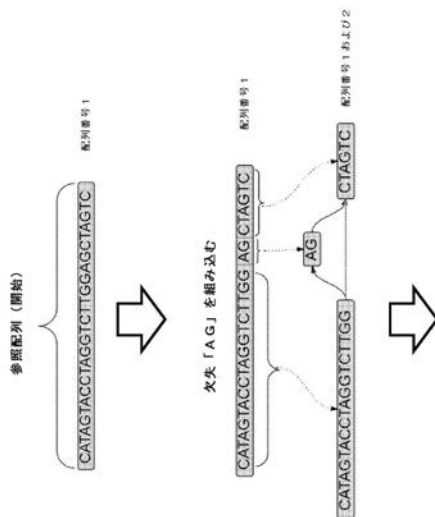


FIG. 1(A)

## 【图 1 B】

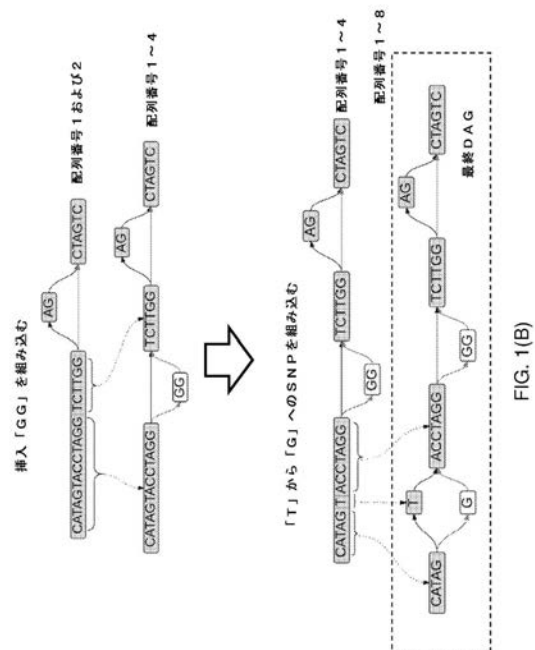


FIG. 1(B)

【図 2】

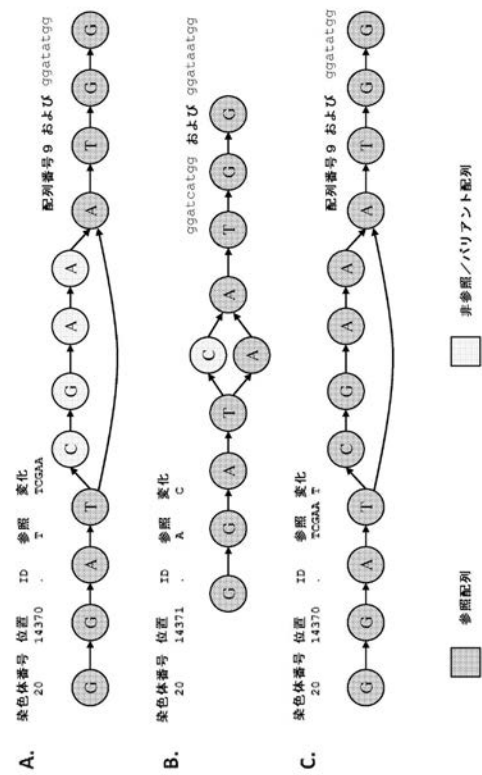
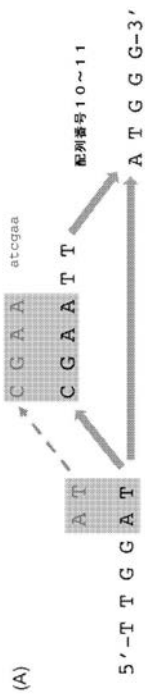


FIG. 2

【図 3 A】



【図 3 B】

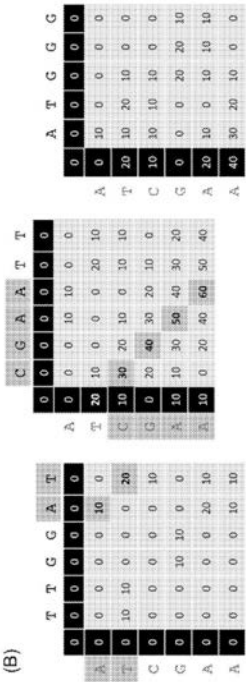


FIG. 3

【図 4】

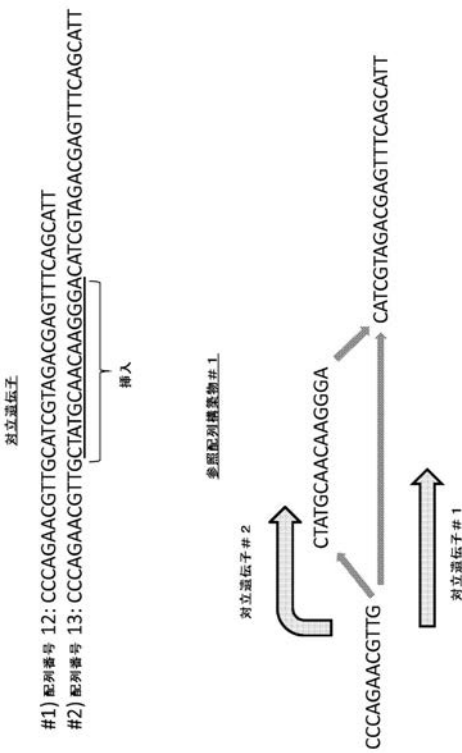


FIG. 4



【図 9】

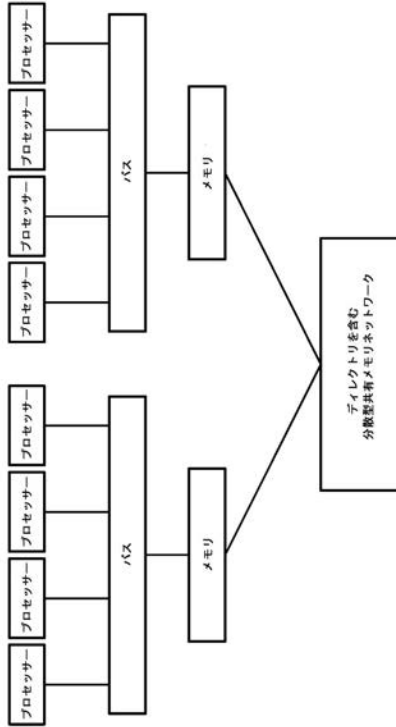


FIG. 9

【配列表】

2017500004000001.app

## 【国際調査報告】

| INTERNATIONAL SEARCH REPORT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                               | International application No.<br><b>PCT/US2014/061156</b>                                                                                                    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>A. CLASSIFICATION OF SUBJECT MATTER</b><br><b>G01N 1/00(2006.01)i</b><br><br>According to International Patent Classification (IPC) or to both national classification and IPC                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                               |                                                                                                                                                              |
| <b>B. FIELDS SEARCHED</b><br>Minimum documentation searched (classification system followed by classification symbols)<br>G01N 1/00; C12P 19/34; G06F 19/10; C12Q 1/70; C12Q 1/68<br><br>Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched<br>Korean utility models and applications for utility models<br>Japanese utility models and applications for utility models<br><br>Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)<br>cKOMPASS(KIPO internal) & Keywords: genotyping, genetic sample, sequence read, align, reference sequence and allele                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                               |                                                                                                                                                              |
| <b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                               |                                                                                                                                                              |
| Category*                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Citation of document, with indication, where appropriate, of the relevant passages                            | Relevant to claim No.                                                                                                                                        |
| X                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | US 2013-0073214 A1 (HYLAND et al.) 21 March 2013<br>See paragraphs [0006]-[0085], claims 1-10 and figure 1.   | 1-35                                                                                                                                                         |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | WO 2013-106737 A1 (DATA2BIO et al.) 18 July 2013<br>See page 2, line 16 - page 5, line 25 and figures 1-4.    | 1-35                                                                                                                                                         |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | US 2008-0003571 A1 (MCKERNAN et al.) 03 January 2008<br>See abstract, paragraphs [0010]-[0031] and figure 1A. | 1-35                                                                                                                                                         |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | US 2009-0325145 A1 (SABLON et al.) 31 December 2009<br>See abstract and paragraphs [0010]-[0062].             | 1-35                                                                                                                                                         |
| A                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | WO 2012-096579 A2 (KEYGENE N.V.) 19 July 2012<br>See page 1, line 22 - page 2, line 6 and figure 1.           | 1-35                                                                                                                                                         |
| <input type="checkbox"/> Further documents are listed in the continuation of Box C. <input checked="" type="checkbox"/> See patent family annex.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                               |                                                                                                                                                              |
| * Special categories of cited documents:<br>"A" document defining the general state of the art which is not considered to be of particular relevance<br>"E" earlier application or patent but published on or after the international filing date<br>"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)<br>"O" document referring to an oral disclosure, use, exhibition or other means<br>"P" document published prior to the international filing date but later than the priority date claimed<br>"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention<br>"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone<br>"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art<br>"&" document member of the same patent family |                                                                                                               |                                                                                                                                                              |
| Date of the actual completion of the international search<br>16 February 2015 (16.02.2015)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                               | Date of mailing of the international search report<br><b>17 February 2015 (17.02.2015)</b>                                                                   |
| Name and mailing address of the ISA/KR<br> International Application Division<br>Korean Intellectual Property Office<br>189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,<br>Republic of Korea<br>Facsimile No. +82 42 472 3473                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                               | Authorized officer<br>LEE, Hun Gil<br>Telephone No. +82-42-481-8525<br> |

**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/US2014/061156**

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Publication<br>date                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|-------------------------------------------|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| US 2013-0073214 A1                        | 21/03/2013          | EP 2758908 A1<br>WO 2013-043909 A1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 30/07/2014<br>28/03/2013                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| WO 2013-106737 A1                         | 18/07/2013          | EP 2802666 A1<br>US 2013-0184165 A1                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 19/11/2014<br>18/07/2013                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| US 2008-0003571 A1                        | 03/01/2008          | AU 2006-210553 A1<br>CA 2596496 A1<br>EP 1844162 A2<br>EP 1844162 B1<br>EP 2003214 A2<br>EP 2003214 A3<br>EP 2003214 B1<br>EP 2230315 A1<br>EP 2230316 A1<br>EP 2233581 A1<br>EP 2233582 A1<br>EP 2233583 A1<br>EP 2236628 A2<br>EP 2236628 A3<br>EP 2239342 A2<br>EP 2239342 A3<br>EP 2241637 A1<br>EP 2272983 A1<br>EP 2316977 A1<br>JP 2008-528040 A<br>KR 10-2007-0112785 A<br>US 2009-0181385 A1<br>US 2009-0181860 A1<br>US 2010-0297626 A1<br>US 2011-0077169 A1<br>US 2012-0191363 A1<br>US 2014-0248610 A1<br>US 2014-0342353 A1<br>US 8329404 B2<br>US 8431691 B2<br>WO 2006-084132 A2<br>WO 2006-084132 A3 | 10/08/2006<br>10/08/2006<br>17/10/2007<br>15/10/2014<br>17/12/2008<br>17/06/2009<br>10/04/2013<br>22/09/2010<br>22/09/2010<br>29/09/2010<br>29/09/2010<br>29/09/2010<br>06/10/2010<br>13/10/2010<br>13/10/2010<br>03/11/2010<br>20/10/2010<br>12/01/2011<br>04/05/2011<br>31/07/2008<br>27/11/2007<br>16/07/2009<br>16/07/2009<br>25/11/2010<br>31/03/2011<br>26/07/2012<br>04/09/2014<br>20/11/2014<br>11/12/2012<br>30/04/2013<br>10/08/2006<br>05/07/2007 |
| US 2009-0325145 A1                        | 31/12/2009          | EP 2121990 A2<br>EP 2463389 A1<br>EP 2463390 A1<br>WO 2008-046922 A2<br>WO 2008-046922 A3                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 25/11/2009<br>13/06/2012<br>13/06/2012<br>24/04/2008<br>12/06/2008                                                                                                                                                                                                                                                                                                                                                                                           |
| WO 2012-096579 A2                         | 19/07/2012          | AU 2012-205884 A1<br>CA 2823815 A1<br>CN 103476946 A<br>EP 2663655 A2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 29/08/2013<br>19/07/2012<br>25/12/2013<br>20/11/2013                                                                                                                                                                                                                                                                                                                                                                                                         |

**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/US2014/061156**

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s) | Publication<br>date |
|-------------------------------------------|---------------------|----------------------------|---------------------|
|                                           |                     | JP 2014-502513 A           | 03/02/2014          |
|                                           |                     | KR 10-2014-0040697 A       | 03/04/2014          |
|                                           |                     | US 2013-0331277 A1         | 12/12/2013          |
|                                           |                     | WO 2012-096579 A3          | 13/09/2012          |

フロントページの続き

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

(74)代理人 230113332

弁護士 山本 健策

(72)発明者 クラル, デニス

アメリカ合衆国 マサチューセッツ 02142, ケンブリッジ, ブロードウェイ 1, 14  
フォーエイチ フロア

Fターム(参考) 4B063 QA11 QA19 QQ03 QQ42 QQ52 QR08 QR20 QR42 QR58 QR62  
QS25 QX01