

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2012-239430

(P2012-239430A)

(43) 公開日 平成24年12月10日(2012.12.10)

(51) Int.Cl.	F I	テーマコード (参考)
<b>C12Q 1/68 (2006.01)</b>	C12Q 1/68 ZNAA	4B024
<b>GO1N 33/53 (2006.01)</b>	GO1N 33/53 M	4B063
<b>GO1N 37/00 (2006.01)</b>	GO1N 37/00 I02	
<b>C12N 15/09 (2006.01)</b>	C12N 15/00 A	

審査請求 未請求 請求項の数 17 O L (全 46 頁)

(21) 出願番号 特願2011-112887 (P2011-112887)	(71) 出願人 301032942
(22) 出願日 平成23年5月19日 (2011.5.19)	独立行政法人放射線医学総合研究所 千葉県千葉市稲毛区穴川四丁目9番1号
特許法第30条第1項適用申請有り 平成22年11月19日「 <a href="http://www.aeplan.co.jp/bmb2010/abstract/index.html">http://www.aeplan.co.jp/bmb2010/abstract/index.html</a> 」を通じて発表	(71) 出願人 500569018 株式会社メイズ 東京都中野区中央3-13-11MGビル508
	(74) 代理人 100108855 弁理士 蔵田 昌俊
	(74) 代理人 100159651 弁理士 高倉 成男
	(74) 代理人 100091351 弁理士 河野 哲
	(74) 代理人 100088683 弁理士 中村 誠

最終頁に続く

(54) 【発明の名称】 網羅的フラグメント解析における遺伝子同定方法および発現解析方法

## (57) 【要約】

【課題】 簡便且つ高い信頼性を保持した遺伝子同定方法および発現解析方法、並びにここで使用される網羅的フラグメント解析のためのデータベース構築方法を提供する。

【解決手段】 試料に含まれるゲノムDNAまたは転写産物から得られたcDNAを断片化し、識別可能な指標配列を付与してフラグメントDNA混合液を得る段階、前記フラグメントDNA混合液の一部を高速DNAシーケンシングし、そこに含まれる全フラグメントDNAについてリード配列データを取得する段階、前記の全リード配列データについて前記指標配列部分の有無を検査し、前記指標配列を有するリード配列データを抽出する段階、前記抽出された全リード配列データについて、配列類似性と配列長のパラメータにより配列のクラスタリング・アッセンブリング処理を行い、複数のクラスタを形成し、前記クラスタについて、当該クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を取得する段階を具備するデータベース構築方法。

【選択図】 なし

**【特許請求の範囲】****【請求項 1】**

試料に含まれるゲノムDNAまたは転写産物から得られたcDNAを断片化し、且つ指標配列を付与することによってフラグメントDNA混合液を得る段階と、

前記フラグメントDNA混合液の第1の一部分を高速DNAシーケンシングすることによって、そこに含まれる全てのフラグメントDNAについてのリード配列データを取得する段階と、

前記リード配列データの全てについて、前記指標配列部分の有無を検査し、前記指標配列を有するリード配列データを抽出する段階と、

前記抽出されたリード配列データの全てについて、予め決定されたパラメータを用いて配列のクラスタリング処理とアッセンブリ処理を行うことにより、複数の配列クラスタを形成し、前記配列クラスタのそれぞれについて、当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を取得する段階と、

前記配列クラスタのそれぞれに対応付けられた当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を含むデータベースを構築する段階と、

を具備し、

前記パラメータが、配列の類似性と配列長と指標配列に関するパラメータであることを特徴とするゲノムまたは転写産物の網羅的フラグメント解析のためのデータベース構築方法。

**【請求項 2】**

前記フラグメントDNA混合液の第2の一部分を電気泳動し、得られた電気泳動結果から各フラグメントに由来するバンド群またはピーク群の強度と電気泳動配列長とをリファレンスプロファイリングの各データとして取得する段階と、

前記各配列クラスタの前記コンセンサス配列の配列情報、塩基数および当該配列クラスタを構成する配列数と、前記リファレンスプロファイリングのバンド群またはピーク群に対応付ける段階と、

を更に具備する請求項1に記載のデータベース構築方法。

**【請求項 3】**

前記対応付けが、前記コンセンサス配列の配列情報及び塩基数と前記リファレンスプロファイリングのバンドまたはピークの電気泳動で得られた分子量との関係を第1のパラメータとして用い、及び前記コンセンサス配列の配列クラスタを構成する配列数と前記リファレンスプロファイリングのバンドまたはピークの強度との関係を第2のパラメータとして使用することにより、当該コンセンサス配列と当該リファレンスプロファイリングとを対応付けることを特徴とする請求項2に記載のデータベース構築方法。

**【請求項 4】**

前記対応付けが、当該配列クラスタを構成する配列数および塩基数と、当該電気泳動で得られたバンドまたはピークの強度および電気泳動配列長とについての一致度をスコア化し、合計スコアが最大になる組み合わせを選択することにより、当該配列クラスタと当該バンドまたはピークとの対応付けが行われることを特徴とする請求項2または3に記載のデータベース構築方法。

**【請求項 5】**

前記対応付けにおけるずれと配列長および分子量との関係および/またはずれと内部塩基組成との関係に基づいて、前記対応付を校正することを更に具備する請求項1～4の何れか1項に記載のデータベース構築方法。

**【請求項 6】**

前記データベースを構築する段階に続いて、当該試料の由来する動物種と同じ動物種の既知遺伝子配列情報を検索し、当該コンセンサス配列と当該同種の既知遺伝子を比較することにより、当該方法において得られたコンセンサス配列の信頼性をデータ化する段階を更に具備する請求項1～5の何れか1項に記載のデータベース構築方法。

10

20

30

40

50

## 【請求項 7】

前記データベースを構築する段階に続いて、前記配列クラスタのコンセンサ配列と前記クラスタを構成するリード配列のアライメント情報に基づいて、各配列クラスタのコンセンサ配列についての信頼性をデータ化することを更に具備する請求項 1 ~ 6 の何れか 1 項に記載のデータベース構築方法。

## 【請求項 8】

前記データベースを構築する段階に続いて、既知遺伝子配列情報を検索し、当該コンセンサ配列と既知遺伝子配列情報とを比較することにより、当該方法において得られたコンセンサ配列に既知遺伝子配列情報を付与することを更に具備する請求項 1 ~ 7 の何れか 1 項に記載のデータベース構築方法。

10

## 【請求項 9】

対象試料に含まれるゲノムまたは転写産物から得られたDNAを断片化し、更に識別可能な指標配列を付与することによって対象フラグメントDNA混合液を得る段階と、

前記対象フラグメントDNA混合液を電気泳動し、得られた電気泳動結果からバンドまたはピークの強度および電気泳動配列長を遺伝子同定対象プロファイリングのデータとして取得する段階と、

前記遺伝子同定対象プロファイリングのデータと、当該対象試料の種類に依存して請求項 2 ~ 8 の何れか 1 項に記載の方法により予め構築されたデータベースの当該コンセンサ配列および当該リファレンスプロファイリングのデータとを対応付けることにより当該対象試料に含まれるゲノムまたは転写産物について遺伝子同定する方法。

20

## 【請求項 10】

前記対応付けが、遺伝子同定対象プロファイリングに含まれるバンドまたはピークの強度および電気泳動配列長のデータと、当該リファレンスプロファイリングのバンドまたはピークの強度および電気泳動配列長とを対応付けて、それにより当該対象試料に含まれるゲノムまたは転写産物の遺伝子情報を得ることを特徴とする請求項 9 に記載の方法。

## 【請求項 11】

前記対応付けが、遺伝子同定対象プロファイリングに含まれる電気泳動配列長のデータを、当該リファレンスプロファイリングのバンドまたはピークの電気泳動配列長および当該配列クラスタの塩基数とその配列数から作成された疑似プロファイリングとを対応付けて、それにより当該対象試料に含まれる転写産物の遺伝子情報を得ることを特徴とする請求項 9 に記載の方法。

30

## 【請求項 12】

第 1 ~ 第 n の対象試料にそれぞれ含まれるゲノムまたは転写産物から得られたDNAをそれぞれ断片化し、更に指標配列をそれぞれ付与することによって第 1 ~ 第 n のフラグメントDNA混合液をそれぞれ得る段階と（ここにおいて「n」は 2 以上の整数を示す）、

前記第 1 ~ 第 n のフラグメントDNA混合液を、それぞれ高速DNAシーケンシングすることによって、各フラグメントDNA混合液にそれぞれ含まれる全てのフラグメントDNAについての第 1 ~ 第 n のリード配列データをそれぞれ取得する段階と、

前記第 1 ~ 第 n のリード配列データそれぞれの全ての配列データについて、前記指標配列部分の有無をそれぞれ検査し、前記指標配列を有する第 1 ~ 第 n のリード配列データをそれぞれ抽出する段階と、

40

前記抽出された第 1 ~ 第 n のリード配列データそれぞれの全てについて、予め決定されたパラメータを用いて配列のクラスタリング処理とアッセンブリング処理をそれぞれ行うことにより、第 1 ~ 第 n のクラスタ群をそれぞれ形成し、前記第 1 ~ 第 n のクラスタ群のそれぞれについて、当該配列クラスタの構成配列数、コンセンサ配列およびコンセンサ配列長、並びにアライメント情報をそれぞれ取得する段階と、

前記第 1 ~ 第 n のそれぞれの配列クラスタ群にそれぞれ対応付けられた当該配列クラスタの構成配列数、コンセンサ配列およびコンセンサ配列長、並びにアライメント情報を含む、第 1 ~ 第 n の配列クラスタ群情報を含むデータベースをそれぞれ構築する段階と、

50

前記第1～第nの配列クラスタ群の間で、それぞれの当該コンセンサス配列の類似性によってそれぞれ互いに対応付け、当該それぞれに対応付けられた配列クラスタ間で、それらを構成する配列数をそれぞれ比較することにより、量の変化を伴う配列クラスタ群を検出する段階と、  
を具備する網羅的フラグメント解析方法。

【請求項13】

対象試料に含まれるゲノムまたは転写産物から得られたDNAを断片化し、更に指標配列を付与することによってフラグメントDNA混合液を得る段階と、

前記フラグメントDNA混合液を、高速DNAシーケンシングすることによって、そこに含まれる全てのフラグメントDNAについてのリード配列データを取得する段階と、

前記リード配列データの全ての配列データについて、前記指標配列部分の有無を検査し、前記指標配列を有するリード配列データを抽出する段階と、

当該対象試料の種類に依存して請求項2～6の何れか1項に記載の方法により予め構築されたデータベースの当該コンセンサス配列との配列類似性をパラメータとして使用して、当該リード配列データのそれぞれについて配列のクラスタリング処理を行い、前記対象試料の配列クラスタ群をそれぞれ得る段階と、

前記対象試料の配列クラスタ群と、前記データベースに含まれる当該配列クラスタ群とを比較して、前記対象試料の配列クラスタ群のみに存在するクラスタおよび/または前記データベースの配列クラスタ群のみに存在するクラスタを検出する段階と、  
を具備する網羅的フラグメント解析方法。

【請求項14】

第1～第nの対象試料にそれぞれ含まれるゲノムまたは転写産物から得られたそれぞれのDNAをそれぞれ断片化し、更に指標配列を付与することによって第1～第nのフラグメントDNA混合液をそれぞれ得る段階と（ここにおいて「n」は2以上の整数を示す）、

前記第1～第nのフラグメントDNA混合液を、それぞれ高速DNAシーケンシングすることによって、それぞれそこに含まれる全てのフラグメントDNAについての第1～第nのリード配列データをそれぞれ取得する段階と、

前記第1～第nのリード配列データのそれぞれの全ての配列データについて、前記指標配列部分の有無をそれぞれ検査し、前記指標配列を有する第1～第nのリード配列データをそれぞれ抽出する段階と、

当該第1～第nの対象試料それぞれの種類に依存して請求項2～6の何れか1項に記載の方法により予め構築されたデータベースの当該コンセンサス配列との配列類似性をパラメータとして使用して、当該第1～第nのリード配列データのそれぞれについて配列のクラスタリング処理をそれぞれ行い、第1～第nの配列クラスタ群をそれぞれ得る段階と、

前記第1～第nの配列クラスタ群にそれぞれ含まれる互いに同じコンセンサス配列毎に、当該クラスタをそれぞれ構成する配列数を比較して、前記第1～第nの配列クラスタ群間で異なる配列数を示す配列クラスタを特定する段階と、  
を具備する網羅的フラグメント解析方法。

【請求項15】

目的に応じて、請求項2～6の何れか1項に記載の方法により予めデータベースを構築する段階と、

構築されたデータベースに含まれるコンセンサス配列に基づいて設計して準備したプローブ群を基体に固定化することによりマイクロアレイを作成する段階と、

対象試料に含まれるゲノムまたは転写産物から得られたDNAを断片化し、更に指標配列を付与することによってフラグメントDNA混合液を得る段階と、

前記プローブ群に対して前記フラグメントDNA混合液を接触させ、ハイブリダイズ信号を得る段階と、

前記得られたハイブリダイズ信号に基づいて、当該対象試料に含まれる転写産物の存在を検出する段階と、  
を具備する網羅的フラグメント解析方法。

**【請求項 16】**

目的に応じて、請求項 2 ~ 6 の何れか 1 項に記載の方法により予めデータベースを構築する段階と、

構築されたデータベースに含まれるコンセンサス配列に基づいて設計して準備したプローブ群を 1 セットとして、 $n$  個の基体に 1 セットずつそれぞれ固定化することにより  $n$  個のマイクロアレイを作成する段階と（ここにおいて「 $n$ 」は 2 以上の整数を示す）、

第 1 ~ 第  $n$  の対象試料にそれぞれ含まれるゲノムまたは転写産物から得られた DNA をそれぞれ断片化し、更に指標配列をそれぞれ付与することによってフラグメント DNA 混合液をそれぞれ得る段階と、

それぞれ前記  $n$  個のマイクロアレイにそれぞれ固定された各プローブ群に対して前記第 1 ~ 第  $n$  のフラグメント DNA 混合液をそれぞれ接触させ、それぞれのハイブリダイズ信号をそれぞれ得る段階と、

前記それぞれ得られたハイブリダイズ信号に基づいて、当該第 1 ~ 第  $n$  の対象試料に含まれる転写産物の存在量を比較する段階と、

前記比較により、前記当該第 1 ~ 第  $n$  の対象試料の間で前記存在量に差がある転写産物を検出する段階と、

を具備する網羅的フラグメント解析方法。

**【請求項 17】**

断片化されて指標配列を付与された、試料に含まれるゲノムまたは転写産物からのフラグメント DNA 混合液が、高速 DNA シーケンシングされることによって取得されたリード配列データの全てについて、前記指標配列部分の有無を検査し、前記指標配列を有するリード配列データを抽出する手順と、

前記抽出されたリード配列データの全てについて、予め決定された配列の類似性と配列長と指標配列に関するパラメータを用いて配列のクラスタリング処理とアセンブリ処理を行うことにより、複数の配列クラスタを形成し、前記配列クラスタのそれぞれについて、当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を取得する手順と、

前記配列クラスタのそれぞれに対応付けられた当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を含むデータベースを構築する手段と、

を含む処理をコンピュータに実行させる、前記転写産物の網羅的フラグメント解析のためのデータベース構築用プログラム。

**【発明の詳細な説明】****【技術分野】****【0001】**

本発明は、網羅的フラグメント解析におけるフラグメント配列データベース構築方法、並びにそれを利用した遺伝子同定方法および発現解析方法に関する。

**【背景技術】****【0002】**

現在の種々の発現解析手法が存在し、また開発されている。例えば、マイクロアレイを使用する手法が広く使用されている。マイクロアレイは、基板に固定された検出しようとする配列を含むプローブと試料に含まれる核酸とのハイブリダイゼーションを検出する方法である。この方法では、プローブを準備するために、対象となる核酸の情報が必要である。また、マイクロアレイ技術では、配列情報のある生物種の場合であっても、定量的な発現量差を求めることは難しい。また、低発現遺伝子について検出される発現量の変化率の信頼性は低い。

**【0003】**

近年、ロシュ社製 454FLX、イルミナ社製 GAII シリーズ・HiSEQ シリーズ、LifeTechnology 社製 SOLiD シリーズ・イオントレント PGM シリーズ、ヘリコス社製、パシフィックバイオ社製などで代表される高速 DNA シーケンサを使用して、転写産物をシーケンシングし、遺

10

20

30

40

50

伝子ごとの配列数を集計して、発現解析を行う手法 (RNA-Seq) が報告されている。この方法では、配列情報が公知の生物種について、その配列情報に基づくリファレンス配列に対してシーケンシングされた配列をアライメントする必要がある。また、このような方法では、目的とする生物種のための配列情報が公知であっても、メジャーな転写産物由来の配列集団についてはその量 (即ち、リード配列数) を比較することができるが、量の少ないマイナーな転写産物由来の配列集団については、量の再現性が低く、比較結果の信頼性も低い。

#### 【0004】

一方、ゲノムDNAの違いや転写産物の発現量の違いを検出する方法として、配列情報がない生物種にも適用できる方法も提案されている。そのような方法には、網羅的フラグメント解析手法とも称され、例えば、HiCEP、AFLP、T-RFLP、SAGE、CAGE、Differential Display などがある。これらの方法は、DNA配列を制限酵素で切断し、末端に特定の配列を付与した断片配列を調整して特定の配列を用いてPCRで増幅後電気泳動する、またはDNA配列を特定の配列を用いてPCRで増幅後電気泳動するものである。これらの方法では、更に、得られた断片DNA配列の電気泳動結果 (即ち、バンド群またはピーク群) について、異なるサンプル間で比較し、強度の異なるバンド群またはピーク群を検出する。このような網羅的フラグメント解析手法においては、発現解析を行うためには各バンド群またはピーク群を各々分取して、それらを1つ1つシーケンシングして塩基配列を決定する必要がある。そのような手法により遺伝子同定と発現解析を行うためには、膨大な時間と莫大な費用が必要である。

10

20

#### 【発明の概要】

#### 【発明が解決しようとする課題】

#### 【0005】

上記の状況に鑑み、本願発明の目的は、簡便且つ高い信頼性を保持した遺伝子同定方法および発現解析方法、並びにそこにおいて使用される網羅的フラグメント解析におけるフラグメント配列データベース構築方法を提供することである。

#### 【課題を解決するための手段】

#### 【0006】

本発明の1態様に従うと、

試料に含まれる転写産物を断片化し、更に指標配列を付与し、フラグメントDNA混合液を得る段階と、

30

前記フラグメントDNA混合液の第1の一部分を高速DNAシーケンシングすることによって、そこに含まれる全てのフラグメントDNAについてのリード配列データを取得する段階と

、  
前記リード配列データの全てについて、前記指標配列部分の有無を検査し、前記指標配列を有するリード配列データを抽出する段階と、

前記抽出されたリード配列データの全てについて、予め決定されたパラメータを用いて配列のクラスタリング処理とアッセンブリ処理を行うことにより、複数のクラスタを形成し、前記クラスタのそれぞれについて、当該クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を取得する段階と、

40

を具備し、

前記パラメータが、配列の類似性と配列長に関するパラメータであることを特徴とするデータベース構築方法

が提供される。

#### 【発明の効果】

#### 【0007】

本発明により、簡便且つ高い信頼性を保持した遺伝子同定方法および発現解析方法、並びにそこにおいて使用される網羅的フラグメント解析におけるフラグメント配列データベース構築方法が提供される。

50

#### 【図面の簡単な説明】

## 【 0 0 0 8 】

- 【図 1】データベースの構築方法の 1 例を示すフローチャート。
- 【図 2】データベースの構成の 1 例を示す図。
- 【図 3】遺伝子同定法の 1 例を示すフローチャート。
- 【図 4】データベースの構成の 1 例を示す図。
- 【図 5】図 3 の遺伝子同定法において使用できる更なる 1 例を示すフローチャート。
- 【図 6】高速DNAシーケンサを用いた解析方法の 1 例を示すフローチャート。
- 【図 7】高速DNAシーケンサを用いた解析方法の 1 例を示すフローチャート。
- 【図 8】マイクロアレイを用いた解析方法の 1 例を示すフローチャート。
- 【図 9 A】クラスタリング処理の 1 例を示すフローチャート。 10
- 【図 9 B】クラスタリング処理の 1 例を示すフローチャート。
- 【図 1 0】DNAフラグメント混合液の調製例を示すスキーム。
- 【図 1 1】選択的PCTを用いたDNAフラグメント混合液の調製例を示すスキーム。
- 【図 1 2】DNAフラグメント混合液の断片長による分離と検出との関係を示す模式図。
- 【図 1 3】アダプタ配列の評価の 1 例を示す図。
- 【図 1 4】エラークラスタの修正の 1 例について示す概念図。
- 【図 1 5】ヘテロSNPによるクラスタ分割の概念図。
- 【図 1 6】電気泳動長と配列長のズレの補正方法の 1 例を示す図。
- 【図 1 7 A】リード配列とピークとの対応付け方法の 1 例を示す図。
- 【図 1 7 B】リード配列とピークとの対応付け方法の 1 例を示す図。 20
- 【図 1 8】指標配列による検査および分類の 1 例を示すフローチャート。
- 【図 1 9】クラスタリング・アッセンブリング処理の 1 例示すフローチャート。
- 【図 2 0】対応付けの 1 例を示す模式図。
- 【図 2 1】高品質配列の 1 例について示す模式図。
- 【図 2 2】アライメントの出力例を示す図。
- 【図 2 3】指標配列の例を示す模式図。
- 【図 2 4】指標配列の例を示す模式図。
- 【図 2 5】フラグメント長による検索の入力画面の 1 例を示す図。
- 【図 2 6】遺伝子名による検索のための入力画面の 1 例を示す図。
- 【図 2 7】BLAST検索の入力画面の 1 例を示す図。 30
- 【図 2 8】較正の前後の例を示す図。
- 【図 2 9】ピークの 1 例を示す図。
- 【図 3 0】ピークの 1 例を示す図。
- 【図 3 1】スコア化の 1 例を示す図。
- 【図 3 2】正誤のアライメントを示す図。
- 【図 3 3】フレームと順序番号の概念図。
- 【図 3 4】高さから高さ順序番号への変換イメージを示す図。
- 【図 3 5】高さ順序番号のイメージを示す図。
- 【図 3 6】プロファイルピークの 1 例を示す図。
- 【図 3 7】プロファイルピークの 1 例を示す図。 40
- 【図 3 8】補正前後の対応付けの 1 例を示す図。
- 【図 3 9】対応付けの 1 例を示す図。
- 【図 4 0】配列長とズレの関係を示すグラフ。
- 【図 4 1】分子量とズレの関係を示すグラフ。
- 【図 4 2】含有アミノ酸とズレとの関係を示すグラフ。
- 【図 4 3】補正の計算方法を示す図。
- 【図 4 4】コンピュータの構成の 1 例を示すブロック図。
- 【発明を実施するための形態】

## 【 0 0 0 9 】

- ( 1 ) 高速DNAシーケンサを活用した網羅的フラグメント配列データベースの構築

以下、図1を用いて高速DNAシーケンサを活用した網羅的フラグメント配列データベースの構築の1例について説明する。

【0010】

まず、データベースを作成するためのフラグメントDNA混合液を調製する。フラグメントDNA混合液は、データベースを作成しようとする試料に含まれるゲノムまたは転写産物を断片化し、指標配列を付与し、調製すればよい。これを網羅的フラグメント解析法のための混合液とする。

【0011】

試料は、細胞、組織および臓器などからそれ自身公知の何れかの手段によりゲノムまたは転写産物を含む混合液に調製されればよい。ゲノムまたは転写産物の断片化に先駆けてそれ自身公知の何れかの手段により行ってもよい。好ましくは転写産物からcDNAを調製し、これを断片化して、標識配列を付与する。

10

【0012】

ゲノムまたは転写産物から得られたDNAの断片化は、それ自身公知の制限酵素を用いて行ってよい。断片化されたDNAへの識別可能な指標配列の付加は、例えば、アダプタ配列を当該断片に付与することにより行ってよい。アダプタの付与は、各断片の5'末端および/または3'末端であってもよい。また、例えば、メイトペア法において実施されるように、アダプタの付与は、各断片の5'末端および/または3'末端に付与された後に、アダプタを付与された1つの断片の5'末端と3'末端とを結合し、環状核酸を形成した後に当該アダプタに対応する配列以外の部位において切断することにより直鎖状核酸を調製してもよい。

20

【0013】

アダプタの塩基配列およびその長さは、識別可能な限りで任意に決定してよい。ここで、「指標配列」とは、指標となるべき配列が識別可能な数の塩基配列を含むことを示す。

【0014】

このようなcDNA断片への識別可能な指標配列の付与は、例えば、HiCEP法、AFLP法、T-RFLP法、CAGE法およびDifferential Display法などのフラグメント解析法における指標配列を付与する方法を利用してよく、より好ましくはAFLP法、T-RFLP法、CAGE法およびDifferential Display法、最も好ましくはHiCEP法を利用して行ってよい。上述のフラグメント解析法を利用して、cDNA断片への識別可能な指標配列の付与し、更に、それらのフラグメントの混合液についてゲル電気泳動および/またはキャピラリー電気泳動などの電気泳動によりバンドまたはピークおよび電気泳動配列長(ここでは「分子量」または「配列長」または「フラグメント長」ともいう)を得ることにより解析する方法を行うことにより、一般的には網羅的にフラグメントが解析されてもよい。

30

【0015】

このように調製されたフラグメント混合液を高速DNAシーケンサにかけてリード配列を得る。

【0016】

ここで「高速DNAシーケンサ」とは、長さの異なる複数種類の塩基配列について分離することなくシーケンシングできるシーケンサを示す。例えば、ロシュ社製454FLX、イルミナ社製GAIIシリーズ・HiSEQシリーズ、LifeTechnology社製SOLiDシリーズ・イオントレントPMGシリーズ、ヘリコス社製、パシフィックバイオ社製などにより提供されるシーケンサを使用することが可能であるが、これに限定するものではない。また、高速DNAシーケンサは、クローニング不要であってもよい。

40

【0017】

次に、リード配列の長さと同義性の2つの要素をパラメータとして利用して、コンピュータ処理により、リード配列をクラスタリング処理およびアセンブリ処理する。それにより、高精度な配列クラスタとコンセンサス配列を作成し、各々の配列クラスタを構成するリード配列数を集計する。

【0018】

50



コンピュータ処理によるリード配列のクラスタリング処理とアッセムプリング処理について図9Aおよび図9Bを用いて更に詳しく説明する。なお、図9Aと図9Bは、同じ一連の工程を示すものであるが、便宜上、図9Aでは工程1～工程3について詳細に記載し、図9Bでは工程4～6について詳細に記載する。またここで、配列のクラスタリング処理とアッセムプリング処理の両方の処理を行う場合、この処理を「クラスタリング・アッセムプリング」または「クラスタリング・アッセムプリング処理」とも記す。

【0019】

ここで「配列のクラスタリング」は、「クラスタリング」および「クラスタ化」と交換可能に使用される語であり、予め決定したパラメータ、好ましくは塩基配列の類似性および/または配列長に基づいてグループ分けすることを示す。クラスタリングにより生じたグループを「クラスタ」または「配列クラスタ」と呼ぶ。互いに同じ長さの複数の配列からなるクラスタを「整列クラスタ」と呼び、互いに異なる長さの複数の配列からなるクラスタを「非整列クラスタ」と呼ぶ。1つのみの配列からなるクラスタを「シングルトン」とも称するが、「シングルトン」もクラスタとして使用されてよい。

10

【0020】

ここで「アッセムプリング」は、「アッセムプリ」および「アッセムプル」と交換可能に使用される語であり、少なくとも部分的に共通する配列を有する複数の核酸配列から1つの代表的な配列であるコンセンサス配列を得ることをいい、また、アッセムプリングに供した配列のコンセンサス配列へのアライメント情報を得ることをいう。

20

【0021】

ここで「リード配列」とは、シーケンサから出力された配列をいう。

【0022】

ここで「コンセンサス配列」とはアッセムプリ処理により得られた人工的な配列をいう。

【0023】

工程1 配列の分類

網羅的フラグメント解析の検出対象となるフラグメントDNA配列の両端に特定の配列が出現する場合、その両端配列の両方または片方を評価し、クラスタリング・アッセムプリングに使用する配列を振り分ける。具体的には、即ち、リード配列が指標配列を含むかが判断され、両端または片方の末端に指標配列が含まれる場合には、データベース作成のためのリード配列として抽出され、以下の工程において使用される。

30

【0024】

指標配列を含むか否かの判断は、判断の対象となるリード配列における指標配列の存在を確認すればよい。確認のために使用される指標配列は、アダプタ配列としてDNA断片に対して付与された配列に対応する塩基配列であってよく、アダプタ配列の全体に対応する塩基配列であっても、アダプタ配列の一部に対応する塩基配列であっても、アダプタ配列に対応する配列に加えて更なる塩基を含む配列であってもよい。更なる配列を含ませる場合には、例えば、任意の数の任意の塩基N(アデニン、チミン、グアニンおよびシトシンから選択される塩基)を含ませてよい。また、任意の塩基Nを含ませる場合には、アダプタ配列に対応する配列の5'末端側または3'末端側に伸長するように含ませることが好ましい。任意の塩基Nを任意の数で含む場合、任意の塩基Nの数は、例えば、1以上、2以上、3以上、4以上、5以上であってよく、好ましくは、1アダプタに対して2つで、且つ1つの配列の5'末端と3'末端の両側に2つずつ含ませる。しかしながら、当該断片の両末端に付与する場合、5'末端側と3'末端側側とは互いに異なる数の任意の種類塩基を含んでもよい。

40

【0025】

なお、指標配列は、リード配列の内部に存在していてもよいが、両末端に存在するのが好ましい。

【0026】

工程2 クラスタリング・アッセムプリング

50

クラスタリング・アッセンブリングを行ない、配列クラスタとそのコンセンサス配列を得る。前記抽出されたリード配列データの全てについて、予め決定されたパラメータを用いて配列のクラスタリング処理とアッセンブリング処理を行う。それにより、複数の配列クラスタを形成し、前記配列クラスタのそれぞれについて、当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を取得する。予め決定されたパラメータとして、例えば、配列の類似性、配列長および/または指標配列に関するパラメータ、好ましくは、配列の類似性と配列長と指標配列に関するパラメータを用いてよい。

【0027】

工程3 クラスタリングエラーの修正

得られた配列クラスタについて、コンセンサス配列とクラスタを構成する配列のアライメント情報を使用して、配列クラスタを構成する配列どうしの配列類似性と配列長の同一性を評価し、更に、コンセンサス配列同士の配列類似性と配列長を評価し、クラスタリング・アッセンブリングの間違いや矛盾を検出し、作成されたクラスタを修正する。

【0028】

工程4 クラスタの信頼性のデータ化

工程3で得られた配列クラスタについて、コンセンサス配列とクラスタを構成するリード配列のアライメント情報から、各配列クラスタの代表配列としてのコンセンサス配列の信頼性をデータ化する。

【0029】

クラスタの信頼性を得るためには、例えば、クラスタのコンセンサス配列とそれを構成するリード配列について指標配列に隣接する塩基の評価を行なえばよい。その場合、指標配列に隣接する配列の数は、2以上、好ましくは2塩基であってよい。

【0030】

工程5 既知遺伝子情報を利用したコンセンサス配列の信頼性のデータ化

前記工程4で得られた配列クラスタのコンセンサス配列について、既知遺伝子情報（転写産物、ゲノム、EST情報など）が存在する生物種では、公知配列情報を検索し、コンセンサス配列の信頼性データを作成する。

【0031】

工程6 コンセンサス配列への遺伝子情報の付与

前記工程4で得られた配列クラスタのコンセンサス配列について、公知配列情報を検索し、配列に遺伝子情報を付与する。

【0032】

以上の工程により網羅的フラグメント解析データベース（以下、「DB」とも記す）を構築することが可能である。なお、工程4～工程6は、任意の工程であり、目的に応じて、例えば、より高い信頼性を担保したい場合や、具体的な遺伝子情報をデータベースに加えた場合に行えばよい。

【0033】

また、工程1～工程6により得られるデータベースに含まれる成分の例を図2に示す。上記の工程1～工程6によって、データベースは、「コンセンサス配列」、「配列クラスタの構成配列数」、「配列クラスタのコンセンサス配列長」、「アライメント情報」、「配列クラスタリングの信頼性データ」および「配列クラスタの遺伝子情報」を含み、これらの情報は関連して記憶部に格納されればよい。

【0034】

(2) 電気泳動で得られるバンドまたはピークと配列の対応付け

以下、電気泳動で得られるバンドまたはピークと配列の対応付けの手順の1例について図1を用いて説明する。

【0035】

上記(1)で得られた高精度なコンセンサス配列の配列情報と塩基数と配列クラスタを構成する配列数を利用して、シーケンス対象としたDNA混合液から得られる網羅的フラグ

10

20

30

40

50

メント解析の電気泳動のバンド群またはピーク群（これらのデータを総称して「リファレンスプロファイリング」と称す）に対応付ける。

【0036】

コンセンサ配列の配列情報及び塩基数とバンドまたはピークの電気泳動で得られた分子量（または電気泳動配列長）、及び、コンセンサ配列のクラスタを構成する配列数とバンドまたはピークの強度の2つの要素を使用して、コンセンサ配列とリファレンスプロファイリングとを対応付ける。

【0037】

コンセンサ配列の対応付けには、あらかじめ多量の配列長とその電気泳動で得られた塩基数との対応付け実験を行なったデータを基に得られた配列の分子量および塩基組成による塩基数の校正情報で、校正を行った値を用いてよい。

10

【0038】

これにより上記（1）のデータベースに含まれるコンセンサ配列が、リファレンスプロファイリングと対応付けられる。

【0039】

（3）（1）のデータベース、及び、（2）の対応付け情報を使用して、網羅的フラグメント解析で得られるバンドまたはピークの遺伝子同定法

以下、網羅的フラグメント解析で得られるバンドまたはピークの遺伝子同定法の手順の1例について図3を用いて説明する。

【0040】

20

[方法1]

工程1 遺伝子同定対象の試料から得たプロファイリング結果と（2）で使用したリファレンスプロファイリングを対応付けたデータを作成する。

【0041】

工程2 遺伝子同定対象の試料から得た遺伝子同定対象バンド群またはピーク群から、上記工程1で作成した対応付けデータを利用して、リファレンスプロファイリングのバンドまたはピークを求め、さらに、（2）での対応付け情報から、上記（1）で作成したクラスタを求め、コンセンサ配列と遺伝子情報を求める。これにより、注目のバンド群またはピーク群と遺伝子情報との対応リストを作成する。

【0042】

30

工程3 加えて、上記（1）の工程6で作成した遺伝子情報により、注目するコンセンサ配列を決定し、上記（2）で対応付けられたリファレンスプロファイリングのバンドまたはピークを介して、遺伝子同定対象の試料から得たプロファイリングのバンドまたはピークを求める。

【0043】

[方法2]

遺伝子同定対象サンプルから得られた電気泳動結果のひとつもしくは複数のバンドまたはピークの電気泳動で得られた塩基数を上記（2）で使用したリファレンスプロファイリングのバンド、さらに、上記（1）で作成した配列クラスタとその配列数から作成した疑似プロファイリングおよびその遺伝子情報を並べて提示することで、遺伝子同定対象サンプルから得られた電気泳動結果の注目バンドまたはピークの遺伝子情報を得る。

40

【0044】

（4）網羅的フラグメント解析の高速DNAシーケンサによる検出

高速DNAシーケンサにより、網羅的にフラグメントを解析し、例えば、目的とする遺伝子を検出することも可能である。このような方法の1例について図6を用いて説明する。

【0045】

測定対象の複数サンプルについて、それぞれ網羅的フラグメント解析法

調整された混合液をサンプルごとに同じ種類の高速度DNAシーケンサにかけて配列を得る。

【0046】

50

測定対象のサンプルからそれぞれ得られたリード配列について、(1)のデータベースをそれぞれ作成し、コンセンサ配列の類似性により、配列クラスタどうしを対応付け、対応付けられた配列クラスタ間で、構成する配列数を比較し、量の変化を伴う配列群を検出し、第1の対象試料と第2の対象試料の間で発現解析を行う方法

配列数を比較する際は、全リード配列数もしくはクラスタリングに用いた配列数を使用して標準化を行ない比較してもよい。

【0047】

(5) データベースをリファレンスにした網羅的フラグメント解析の高速シーケンサによる検出

更なる高速シーケンサによる網羅的フラグメント解析の例について図7を用いて説明する。

【0048】

あらかじめ測定対象となるサンプルについて上記(1)の手順を実施し、データベースを作成しておく。

【0049】

測定対象の複数サンプルについて、それぞれ網羅的フラグメント解析法で調整された混合液をサンプルごとに同じ種類の高速度DNAシーケンサにかけて配列を得る。あらかじめ作成したデータベースで使用した高速度DNAシーケンサと同じである必要はない。

【0050】

この配列をあらかじめ作成したデータベースのコンセンサ配列をリファレンスとして、これにアライメント処理等を行なうことで測定対象のリード配列をクラスタリングする。

【0051】

同じコンセンサ配列にクラスタリングされた配列の数を、測定サンプル間で比較し、量の変化を伴う配列群を検出し 第1の対象試料と第2の対象試料の間で発現解析を行う方法

配列数を比較する際は、全リード配列数もしくはクラスタリングに用いた配列数を使用して標準化を行ない比較してもよい。

【0052】

(6) データベースからプローブを設計して作成したマイクロアレイによる網羅的フラグメント解析方法

更にマイクロアレイを利用する網羅的フラグメント解析方法の1例を図8を用いて説明する。

【0053】

あらかじめ測定対象となるサンプルについて(1)の手順を実施し、データベースを作成しておく。得られたコンセンサ配列をもとにプローブ設計を行い、マイクロアレイを作成する。

【0054】

測定対象の複数サンプルについて、網羅的フラグメント解析法で調整された混合液について、上記で作成したマイクロアレイを用いて量の変化を伴う配列群を検出し、第1の対象試料と第2の対象試料の間で発現解析を行う方法。

【0055】

(7) 指標配列による検査および分類工程

以下に、図18を用いて指標配列による検査および分類工程の更なる1例について更に説明する。

【0056】

シーケンスされた全リード配列を読み込み、それらのリード配列に必ず存在するべき既知指標配列との類似性データを算出する。その後、リード配列すべてに対して、一本ずつ、類似性データを参照し、既知の指標配列があるかどうかを確認する。既知の指標配列が確認できたリード配列は、クラスタリングに使用する配列として分類する。

10

20

30

40

50

## 【 0 0 5 7 】

## ( 8 ) クラスタリング・アッセンブリング処理

以下に、図 19 を用いてクラスタリング・アッセンブリング処理の更なる 1 例について更に説明する。

## 【 0 0 5 8 】

図 19 において各記号は次のことを意味する；

M：クラスタのシーズとなるリード配列の番号

N：シーズとなるM番目のリード配列の次のリード配列から最後の配列までを読み取るための番号

I：生成されたクラスタ番号。

10

## 【 0 0 5 9 】

クラスタリングに使用するリード配列をすべて読み込み、まず、クラスタのシーズとなるリード配列：M番目の配列を決定し、シーズ配列の次のリード配列から残り全部のリード配列について、順次シーズ配列とN番目の対象リード配列とで類似性と配列長を比較し、同じと判定されたならば、クラスタの記憶域のI番目のクラスタにリード配列を格納する。シーズ配列の検索がすべて終了した時点で、クラスタが確立する。その後、各クラスタでコンセンサス配列を得るために、各クラスタごとにアッセンブリングを行なう。

## 【 0 0 6 0 】

## ( 9 ) プログラム

本発明の何れの態様に従う方法を行うために、各方法に含まれる工程（ここでは「段階」とも記す）を各手順としてコンピュータに実行させるためのプログラムが提供されてもよい。例えば、上述の(1)、(2)、(3)、(4)、(5)、(6)、(7)および/または(8)に含まれる段階を各手順として実施するためのプログラムが提供される。

20

## 【 0 0 6 1 】

例えば、上記(1)の方法をコンピュータに実行させるためのプログラムが何れかの媒体に格納されて供されてもよい。

## 【 0 0 6 2 】

そのようなプログラムは、例えば、次のようなプログラムである：

断片化されて識別可能な指標配列を付与された、試料に含まれる転写産物からのフラグメントDNA混合液が、高速DNAシーケンシングされることによって取得されたリード配列データの全てについて、前記指標配列部分の有無を検査し、前記指標配列を有するリード配列データを抽出する手順と、

30

前記抽出されたリード配列データの全てについて、予め決定された配列の類似性と配列長に関するパラメータを用いて配列のクラスタリング処理とアッセンブリング処理を行うことにより、複数の配列クラスタを形成し、前記配列クラスタのそれぞれについて、当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を取得する手順と、

前記配列クラスタのそれぞれに対応付けられた当該配列クラスタの構成配列数、コンセンサス配列およびコンセンサス配列長、並びにアライメント情報を含むデータベースを構築する手段と、

40

を含む処理をコンピュータに実行させる、前記転写産物の網羅的フラグメント解析のためのデータベース構築用プログラム。

## 【 0 0 6 3 】

本発明の態様において使用されるコンピュータは、それ自体公知の何れかのコンピュータであればよい。本発明の態様に従い手続きを行うためのコンピュータの構成の1例を模式的に図44に示す。当該コンピュータは、処理管理部、記憶部、一時記録部、プログラム格納部、クラスタリング・アッセンブリング処理部、指標配列検査部、補正データ格納部、類似性判定部および配列長判定部を含む。少なくとも処理管理部に対して、他の全ての構成部、即ち、記憶部、一時記録部、プログラム格納部、クラスタリング・アッセンブリング処理部、指標配列検査部、補正データ格納部、類似性判定部および配列長判定部が

50

、信号の授受可能に接続される。また、所望に応じて処理を行うための更なる構成部が含まれてもよく、そのような構成部は、信号の授受可能に処理管理部に接続される。

【0064】

全てのプログラムはプログラム格納部に格納される。処理管理部は、プログラム格納部に格納されたプログラムに従って、全ての処理を管理し実行させる。本発明の態様により構成されたデータベースは、記憶部に格納される。リード配列は、記憶部または一時記録部に格納される。指標配列検査部は、プログラム格納部に格納されたプログラムに従う処理管理部の指示により、格納された構成部から出力され、入力されたリード配列に指標配列が含まれるか否かを検査する。クラスタリング・アッセンブリング処理部は、リード配列をクラスタリングおよびアッセンブリング処理する。補正データ格納部は、得られたデータについて補正を行うために使用されるデータを格納する。補正データ格納部に格納されたデータを出力させ、プログラム格納部から補正のためのプログラムを出力させ、それらに基づいて得られたデータについての補正を処理管理部が行う。類似性判定部は、プログラム格納部に格納されたプログラムに従う処理管理部の指示により、比較されるべき対象についての類似性に関する判定を行なう。配列長判定部は、プログラム格納部に格納されたプログラムに従う処理管理部の指示により、比較されるべき対象についての配列長に関する判定を行なう。

10

【0065】

更にコンピュータは、オペレータや高速DNAシーケンサなどからデータを入力するためにキーボードおよび/またはスキャナーなどの入力部を有してもよい。また更に、得られた結果を出力するためのモニターおよび/またはプリンターなどの出力部を有してもよい。尚、上記では、補正を処理管理部が行う例を示したが、コンピュータが更に補正部を有し、補正部が上述のようにデータの補正を行ってもよい。

20

【0066】

本発明者らは、従来技術においては、次のような問題があることを見出している。このような問題も本発明により解決される。

【0067】

HiCEP法に代表されるようなDNA配列を制限酵素で切断し、末端に特定の配列を付与した断片配列を調整して、特定の配列を用いてPCRで増幅後電気泳動する方法、または、DNA配列を特定の配列を用いてPCRで増幅後電気泳動する方法、などで得られた断片DNA配列の電気泳動結果（バンド群またはピーク群）を、異なるサンプル間で比較し、強度の異なるバンド群またはピーク群を検出する方法がある（以下、網羅的フラグメント解析手法と呼ぶ）。網羅的フラグメント解析手法の代表的なものは次のような手法である。例えば、そのような手法には、HiCEP、AFLP、T-RFLP、SAGE、CAGE、Differential Display と称される方法が含まれる。

30

【0068】

これらの網羅的なフラグメント解析法は、既存の配列データがなくとも、網羅的フラグメント解析が可能である。しかしながら、HiCEP法以外の方法は、網羅性が低い、または、フラグメントが短く遺伝子を特定できない、さらに、ひとつの配列から複数のバンドまたはピークが出現し解析が艱難であるという問題がある。

40

【0069】

HiCEP法は、他の網羅的なフラグメント解析とは異なり、解析対象となる1種類のmRNA配列またはゲノム配列断片（スタート配列）から、制限酵素で切断して1種類のフラグメントのみが生成されるように調整することを特徴とする方法で、さらに、検出のためのPCR工程で、アダプタ配列より内側に2塩基（セレクション配列）長いプライマーを使用し256通りのPCRと電気泳動を行うことで、約2万種類以上の断片配列を同時に独立した波形ピークとして得ることで、その1ピークが元となる配列1種類と対応付くという特徴を持った方法である。よって、HiCEPで得られた電気泳動結果を比較し、サンプル間で強度が変化したピークは、そのフラグメントの元となる配列も同様に量的な差があることを検出できる方法である。さらに、PCRを利用した手法であるため低発現量の転写産物も検出

50

可能であり、再現性も非常に良いため1.2倍以上の発現量差も検出できる。

【0070】

しかしながら、HiCEP法においても、その他の網羅的なフラグメント解析法同様、量に違いのあるバンドやピークを知ることができても、その配列を決定するには、分取という煩雑な工程を必要とする。

【0071】

これを解決するひとつの方法として、公知の配列情報を持つ生物種については、網羅的フラグメント解析をコンピュータで予測して配列を決定する方法が考えられたが、電気泳動長と配列長のズレや公知配列の情報過多による判別の困難さなどの問題で、バンドやピークの配列の予測ができて、その信頼性は低いというのが実状である。

10

【0072】

また、HiCEPにおいては、ES細胞を試料として、HiCEPで検出されたピークの約14000についてサンガー法のシーケンサを利用して配列を決定してデータベースを作成したが、このデータベースを作成するのに約3年の期間と大きなコストを必要として、この方法をHiCEP測定対象とするすべての試料で行なうのは現実的ではない。

【0073】

最近では、高速DNAシーケンサが登場し、これを利用してゲノムDNAやmRNAをシーケンスする研究がさかんに行なわれているが、読み取りの長さの制限やシーケンスする配列を作成する段階で配列を同じ長さにそろえる等、網羅的フラグメント解析の試料をシーケンスするには適当ではない方法が使われている。また、配列類似性のみでゲノム配列や転写産物にマッピングをして、遺伝子ごとにクラスタリングする方法を取るため、配列情報がない生物種に適用できないことはもちろんのこと、マッピングエラーなどのバイアスがかかり、メジャーな配列群以外は期待される再現性が得られない。

20

【0074】

また本発明は、次のような効果を奏することが可能である。

【0075】

(1) 注目バンドまたは注目ピークの配列決定方法の簡易化

網羅的フラグメント解析方法で得られた候補のバンド、または、ピークの配列を知るためには、それらのバンドやピークを分取してシーケンシングする必要がある。この注目しているバンド、または、ピークを検出した時点でその配列を知りえないということは、その後の解析を行う上で大きな障害となる。最も重要な欠点は、網羅的フラグメント解析の結果、多くの候補が得られた場合、それらのバンドまたはピークに情報が付与されていないため既知知見を利用して絞り込むことができず、科学的な根拠ではなく、その後の実験の容易さ等で分取対象を決定する必要があり、重要な遺伝子を候補から落としてしまうという問題である。もちろん、注目のバンド、または、ピークを全て分取して配列を決定するという方法も考えられるが、候補が多い場合には大きな費用と時間がかかる。もうひとつの問題点は、分取の手技そのものが煩雑であり、特に、バンドやピークが1ベース単位で蜜な状態では、クローニングを行なってシーケンスしなければならないなど、費用と時間のかかる工程になることである。

30

【0076】

この分取工程を省くためのひとつの方法として、ゲノム情報や転写産物情報のある生物に対して、HiCEP法をコンピュータによりシミュレーションし、既知の配列から得られる仮想的なフラグメント配列を、電気泳動のバンドもしくはピークの分子数とマッチングを行い、バンドあるいはピークの配列を予測する方法も構築した。しかしながら、電気泳動で得られる配列長（電気泳動長と呼ぶ）と対象のフラグメントの実際の配列長はかならずしも一致しないこと、また、バンドやピークの数に比べて既知の配列の種類が多くなり候補の配列が多くなってしまふこと、これらのことから配列長だけで正確に配列とバンドまたはピークと対応付けられないことがわかった。

40

【0077】

もうひとつの方法として、対象サンプルについて、すべてのバンドまたはピークについ

50

て、あらかじめ配列を決定してデータベースを作成しておくことで分取工程を省くことができる考えた。そこで、ES細胞について、HiCEPを実施し、サンガー法のシーケンサを仕様して、得られたピーク約14000について配列を決定し、データベースを作成した。その結果、ES細胞の解析には有用であったが、このデータベースを作成するには、膨大な時間と期間が必要で、HiCEPを適用する生物種や試料ごとにこの方法でデータベースを作成することは困難であることがわかった。

【0078】

網羅的フラグメント解析法に、高速DNAシーケンサを併用する本法で配列をデータベース化することで、これまでとは比べ物にならないほど、短い期間と低コストで、バンド群またはピーク群の配列を網羅的に同定することができるようになる。さらに、本法では、公知配列を必須としないため、公知配列がない生物種においても本法を適用できる。

10

【0079】

そのことは、本法を適用することで、網羅的フラグメント解析法のバンド群やピーク群の配列を同定するという効果だけではなく、ゲノムや転写産物の網羅的断片配列を手にいれることができるという効果もある。

【0080】

(2) 網羅的フラグメント解析の新たな検出方法

上記(1)の課題である網羅的フラグメント解析法で得られるバンドやピークの配列を決定することは重要な課題であるが、網羅的フラグメント解析法の差のあるフラグメントの検出方法として、高速DNAシーケンサを利用することも考えられる。

20

【0081】

しかしながら、高速DNAシーケンサを利用する場合、通常は、シーケンス対象の配列群をランダムに切断し、配列の長さをそろえてシーケンスしなければならない。

【0082】

また、シーケンスした配列をクラスタリングするには、リファレンスとなる既知配列が必要である。

【0083】

よって、網羅的フラグメント解析法で得られるcDNA調製液を高速DNAシーケンサでシーケンシングし解析することは難しいと考えられている。

【0084】

網羅的フラグメント解析法のcDNA調整液を高速DNAシーケンサでシーケンシングし、本法でデータベースを構築することにより、配列クラスタごとのコンセンサス配列と構成配列数を得ることができる。このコンセンサス配列を構成配列を使用して、量に差のある配列クラスタを求めることができる。これは、高速DNAシーケンサの問題はあるものの、バンド群やピーク群をPCRと電気泳動で求めるのではなく、直接、配列クラスタ間のリード配列数を比較することができ、バンド群やピーク群と配列クラスタとの対応付けを必要としないメリットがある。

30

【0085】

さらに、本法で作成された配列クラスタのコンセンサス配列をリファレンス配列として利用する前提でし、測定対象の試料から網羅的フラグメント解析法で調整されるDNA混合液に対して、既存解析法(高速DNAシーケンサのRNA-seqやマイクロアレイ)を適用することで、既存解析法の欠点をおぎないながら網羅的フラグメント解析法の特徴を生かした解析が可能となり、さらに、信頼性データも格納された高精度な配列クラスタの情報を利用することで、既存解析法を適用するよりもより高精度な解析が可能となる。

40

【実施例】

【0086】

HiCEP法(High coverage expression profiling method)は、網羅的フラグメント解析の方法のひとつで、微量の試料から網羅的・高精度に遺伝子発現解析を行う方法である。HiCEP法の最大の特徴は、低発現転写物も再現性良く高精度に解析可能な点である。更に、本法はあらかじめ遺伝子配列情報を必要としないため、ゲノム情報が明らかではない生

50



物種にも適用可能である。しかしながら、プロファイリングピークとして得られる転写産物の塩基配列予測が困難であることも意味する。よって、HiCEP法で得られる網羅的フラグメント解析の発現プロファイルにおける電気泳動ピークの塩基配列同定を本法で実施した。

#### 【 0 0 8 7 】

本法の利用イメージは、図 2 0 に示す通り、あらかじめ測定対象となる試料をHiCEP法で調製し、本法でシーケンシング・クラスタリング・アッセンブリングし配列クラスタのデータベースを作成後、同じ調製試料から得たHiCEPのリファレンスプロファイリングのピークとクラスタの対応付けを行なったデータを保存しておく。その後、データベースを作成した試料と同様の生物種・組織ではあるが異なる試料について、HiCEPを実施し、解析対象プロファイリングを得、注目する電気泳動ピークをリストアップしたのちに、あらかじめ作成しておいた配列クラスタのデータベース及びクラスタとリファレンスプロファイリングのピークとの対応付けデータを使用して、注目ピークの配列を決定する方法である。

10

#### 【 0 0 8 8 】

HiCEPの具体的な手法は、図10で示すように、生物試料から抽出したRNA (Total RNA)、もしくは、精製したmRNAの試料をもとに、まず二重鎖のcDNA群を生成し、これを適切な2つの制限酵素によって切断し、それぞれの末端に特徴的なアダプタを付与したcDNA断片群のみの調製液を作成する方法である。このとき、両端に異なるアダプタ配列を付与されたcDNA断片は、スタートのmRNA 1種類から1種類しか生成されないのがHiCEPの特徴である。

20

#### 【 0 0 8 9 】

さらにHiCEP法では、図 1 1 で示すように、cDNA断片群の調整液を256分割し、両端のアダプタ配列(既知の配列)より2塩基長いプライマーを16種類作成し、256種類の異なるプライマーの組み合わせでPCRを行って、図 1 2 のようにそれぞれのPCR産物をサイズマーカーとともにキャピラリー電気泳動装置にかけて、電気泳動の波形パターンとピークの電気泳動配列長及び蛍光強度のデータを、プロファイリングデータとして得る手法である。

#### 【 0 0 9 0 】

このHiCEPで得られたプロファイリングピークの配列同定を、マウスES細胞(E14)を試料として本法により実施した。

30

#### 【 0 0 9 1 】

(1) 高速DNAシーケンサを活用した網羅的フラグメント配列データベースの構築

マウスES細胞(E14) total RNA 1 µgを用いてHiCEP法を実施した。

#### 【 0 0 9 2 】

次に、HiCEP法の工程の内、図 1 0 で示す工程で得られた「鋳型cDNAs」(両端に指標配列であるHiCEP法で用いるアダプターを有する配列の混合物。長さの分布は約60-baseから約800base)について、シーケンシングに必要なDNA量を得るため、アダプター上のプライマーにて増幅を行った。その後、プライマーダイマーおよびアダプターダイマー画分の除去を目的とし、アクリルアミドゲル電気泳動による精製を行い70baseから100base以下のフラグメントを除去した。その精製物をRoche社製高速DNAシーケンサであるGS 454 FLX Systemにてシーケンシングを行った。尚、シーケンシングライブラリー作製時、DNAの断片化は行わなかった。シーケンシングにより、1回目(2分の1プレート)は469,318配列、2回目(2プレート)は1,868,178配列を得た。これらの配列群について、配列の長さと同様性のふたつの要素を利用して、コンピュータ処理により、クラスタリング・アッセンブリングし、高精度な配列クラスタとコンセンサス配列を作成し、それを構成するリード配列数を集計してデータベース化する次のような工程を開発した。

40

#### 【 0 0 9 3 】

工程 1 : 指標配列(HiCEP法に用いるアダプタ配列)による検査と分類

HiCEP法の検出対象となる、図 1 0 のcDNA断片の両端には、必ず特定の指標配列で

50

あるアダプタ配列が付与される。すべてのリード配列について指標配列を評価し、クラスタリング・アッセムブリングに使用する配列を振り分ける。

【0094】

具体的には、図13で示すように、アダプター配列にセレクション塩基NNまでを加えたマスキング配列32種類で、cross\_match(ワシントン大学)プログラムにより、全リード配列を類似性検索し、一定の類似度以上でアダプター配列が両端または片側に確認できる配列をクラスタリング・アッセムブリングの対象とする。

【0095】

(A) cross\_matchのパラメータ

cross\_matchプログラムのパラメータは、次の通りである。

10

【0096】

A) ミスマッチ・ギャップのペナルティ値を最小にする

-penalty -1 -gap\_init -1 -gap\_ext -1

454のリードエラーの特性(モノポリマー(1種類の塩基)が連続している場合に、リード配列ごとにその連続している領域の塩基数のばらつきが大きくなる特性)を考慮して、ギャップのペナルティ値を最小にする。

【0097】

B) ワードサイズを小さめにとる

-minmatch 5

ペアワイズアライメントを出来るだけ多く検出・出力するようにする。

20

【0098】

C) 最低スコア値を小さくとる

-minscore 15

(B) 指標配列

MspI側, MseI側それぞれのアダプタ配列を指標配列としてcross\_matchへの入力マスク配列とする。実際に使用した配列は、図13のように、アダプタ配列だけではなくNNのセレクション塩基2塩基部分も加えた配列を指標配列とした。これによって、全パターンを網羅するためには、MspI, MseI各16種類ずつ計32種類のアダプタ配列を使用した。なお、NNのセレクション塩基を含まないアダプタ配列を指標配列ともできるが、32種類のアダプタを使用したほうが、確認できる指標配列をやや多く確認できるため、32種類のアダプタを採用した。

30

【0099】

(C) cross\_matchによって検出されたアダプタの分類

図27で示すようなcross\_matchの出力を使用して、正しいHiCEPフラグメントを得るために、cross\_matchによって検出されるアダプタを次の4種類に分けて考える。

【0100】

A) 高品質アダプタ: アライメントにNNを含む高スコアのアダプタ

B) 救済可能な低品質アダプタ: アライメントは短い置換・ギャップがなく、NNを含んでおり、内部配列は高品質であることが期待できるアダプタ

C) 低品質アダプタ: 低品質かつ救済できないアダプタ

40

D) 偽アダプタ: アダプタに似た内部配列をcross\_matchがアライメントしたと思われるもの(実際にはアダプタとして存在しないと思われる部分)。

【0101】

(D) 高品質アダプタの判定条件

アダプタ配列 + NN33bpの内、29bp以上が一致する(図13を参照)。

【0102】

(E) 救済可能な低品質アダプタの判定条件

HiCEP法の特徴を生かすためにセレクション塩基NNを含む19bpがすべて一致する(図13を参照)。

【0103】

50

## (F) 配列の分類

確認されたアダプタの種類が、高品質アダプタか救済可能な低品質アダプタである配列をクラスタリング・アッセンブリングに使用する配列とする。本実施例では、全リード配列(469,318)の内、300,635配列、(64.1%)が両端にアダプタが確認できた配列であった。また、112365配列(23.9%)が片側のアダプタのみが確認できた配列であった(図13を参照)。

【表1】

	4レーン	1レーン
リードレーン数		
リード数	1868178	469318
高品質配列数	1211522(64.8%)	300635(64.1%)
	修正後	修正前
クラスタ数	37239	37295
		修正前
クラスタ構成リード数	1144179	1148252
		修正後
シングルトン数	67343(5.6%)	63270
		修正前
		修正後
初期クラスタリング処理時間	492h(16core)	29h(16core)
クラスタリング	456h	25.5h
アッセンブリング	36h	3.5h
非整列クラスタ修正処理時間	38h	1.5h
非整列修正対象クラスタ数	3210	33
構成リード数	626049	7122
修正後クラスタ数	3148	45
構成リード数	621976	7017
修正によるシングルトン数	4073	105

10

20

30

40

## 【0104】

## 工程2：クラスタリング・アッセンブリング

クラスタリング・アッセンブリングに使用する配列の内、実施例では両端にアダプタが確認できた配列を対象として、クラスタリング・アッセンブリングして、HiCEPフラグメントのコンセンサ配列を生成する。これにより、個々のリード配列のエラーが除かれ、より正確なHiCEPフラグメント配列が得られる。加えて、コンセンサ配列を構成するリード配列の数をHiCEPフラグメントの転写量の参照データとすることができる。

## 【0105】

50

## (A) 前処理

前処理として、クラスタリング・アッセムブリングに使用するアダプタ配列が確認できたすべてのリード配列について、確認できたアダプタ配列を含むアダプタ配列の位置から外側を除去し（図 2 1 参照）、さらに、除去した配列の端に本来のアダプタの塩基配列を人工的に付与する。加えて、シーケンサーから出力された各リード配列のクオリティ値の情報についても、確認できたアダプタ配列の位置から外側を除去し、人工的に付与されたアダプタに対応する部分に、クオリティ値の最高点を付与する。

## 【 0 1 0 6 】

上記の前処理を行なった配列群を入力として、クラスタリング・アッセムブリングプログラムを実行し、配列クラスタ情報と各配列クラスタの配列アライメント情報、及び、各配列クラスタのコンセンサス配列を得る。

10

## 【 0 1 0 7 】

## (B) クラスタリング・アッセムブリングソフトウェア

クラスタリング・アッセムブリングには、配列類似性のみでクラスタリング・アッセムブリングを行なうTGICLプログラム（ハーバード大学のウェブサイト<http://compbio.dfci.harvard.edu/tgi/software/>で公開）を利用する。なお、配列のアッセムブリングはTGICLに付属したアッセムブリングプログラム CAP3を用いる。

## 【 0 1 0 8 】

## (C) TGICLのパラメータ

パラメータ“-v 2”はアッセムブリング時に許容するオーバーハング（配列の端を無効にして、アッセムブリング結果から除外する部分）の塩基数を最小にする設定である。アッセムブリングにおいては、入力配列は両端に共通のHiCEPアダプタ配列が付加された配列なので、オーバーハングのあるアライメントが出力された場合には、そのクラスタが正常にアッセムブリング出来てないことを示すため、エラークラスタが認識可能となる。

20

## 【 0 1 0 9 】

クラスタリング・アッセムブリングに関する以下のパラメータにはデフォルト値を使用した。

## 【 0 1 1 0 】

## ・クラスタリング

最小オーバーラップ長： 40bp (-l)

オーバーラップの一致率： 94% (-p)

## ・アッセムブリング

・最小一致率： 93% (-O -p)

30

## (D) シングルトン配列の収集

通常、ランダムに切断された試料をシーケンシングして得られる配列は、類似性で帰属するゲノム領域や遺伝子を判定し、クラスタを形成し、知見へと結び付けていく。このような場合、シングルトン配列はクラスタを形成していないとして知見を得るためのシグナルとしては扱われない場合もある。しかしながら、本法で得られるシングルトン配列は、指標配列が確認された配列（HiCEP法ではアダプタ配列が確認された配列）で、信頼性も高く、その配列が1本シーケンシングされたという事実も知見を得るために利用できる。よって、本法では、シングルトン配列についても有効利用できるように、処理対象とする（よって、本明細書で、クラスタという場合は、シングルトン配列のみのクラスタも指す場合がある）。

40

## 【 0 1 1 1 】

tgiclの処理において、シングルトン配列は、クラスタリング・アッセムブリングの2つの処理段階で別々に発生する。アッセムブリング時に生じたシングルトン配列は、アッセムブリングを実行したスレッド毎に作成されるsingletsファイルに出力されるが、クラスタリング時に除外されたシングルトン配列の情報は出力されない。このため、クラスタリング時に除外された配列を特定して、シングルトン配列として取り出さなければな

50

らない。tgiclはクラスタリング終了時にクラスタ化できた配列名のリストをファイルに出力するので、このファイルを利用して、クラスタリング時に除外されたシングルトン配列を取得する。

【0112】

以下にシングルトン配列の収集手順を示す。

【0113】

A) クラスタ化された配列名リストを編集して、1行-1配列名の形式でファイルを作成する

B) FASTA形式の入力ファイルから、全入力配列名の1行-1配列名の形式のファイルを作成する

C)上の2つのファイルを連結・ソートして配列名が1行だけの行をとりだす

D)入力配列の配列データベースを作成し、(3)で取得した配列をFASTA形式で抜き出す

E)アッセンブリング時のシングルトン配列(singlets)ファイルと(4)のファイルを連結する

工程3：配列長によるクラスタリングエラーの修正

tgiclのアッセンブリング結果のアライメント情報はace形式のファイルに出力される。HiCEP法で得られたcDNAから生産される配列クラスタの構成配列すべておよびそこから得られるコンセンサ配列は、配列が類似していることはもちろん、全長にわたり配列がアライメントされている整列クラスタでなければならない。

【0114】

(A)配列クラスタの検査

上記原則に従って、このaceファイルの内容をコンティグ毎に以下の点を検査して、アッセンブリングエラーの判定を行う。

【0115】

A)リードの全長が有効にアッセンブリングされている

コンティグを構成する全てのリード配列にクリップされた部分(アライメント中でリードの端が無効化され、コンセンサ配列の形成に寄与しない部分)が無いこと

B)リードがコンセンサ(コンティグ)配列の全長にアライメントされている

コンセンサ配列にアライメントされたリード配列の両端がコンセンサ配列の両端に一致し、コンセンサ配列の途中からアライメントされることがないこと。

【0116】

上のA),B)の条件を満たさない配列クラスタを非整列クラスタとし、これはエラークラスタであると判定する(図14参照)。

【0117】

(B)エラークラスタの修復

上記(A)でエラークラスタと判定されたクラスタ(コンティグ)の構成配列を取り出して、そのクラスタごとに個別に再度アッセンブリングしてコンセンサ配列を取得することで、エラークラスタを修正する(図14参照)。

【0118】

A)反復アッセンブリング

個別アッセンブリングにはCAP3を用い、始めに一致度93%でアッセンブリングを試みる。アッセンブリング結果がエラークラスタと判定されなくなるまで、一致度のパラメータを1%ずつ上げて反復してアッセンブリングを行っていく。

【0119】

起動パラメータには、オーバーハング長に0を指定する「-k 0」を設定する。

【0120】

B)アッセンブリング結果のマージ(aceファイルの修正)

エラークラスタ修復前のtgiclが出力したaceファイルに対して、修復後に生成されたアライメント情報で修正を行なう。

10

20

30

40

50

## 【 0 1 2 1 】

- i. 個別アッセンブリしたクラスタのリストを作成する
- ii. tgi-clが出力したaceファイル先頭から順に読み込み、個別アッセンブリしたクラスタがあれば、そのクラスタの情報を削除する
- iii. 削除した位置に個別アッセンブリしたクラスタ情報を挿入する。

## 【 0 1 2 2 】

このときに、コンセンサス（コンティグ）配列名は元ファイルの名前に枝番を付加した名前を付与する。枝番は '\_' に続く0埋めした4桁の数字を「\_0001」から順に付与する。

## 【 0 1 2 3 】

## D) シングルトン配列の収集

個別アッセンブリで発生したシングルトン配列は、個々の入力クラスタ毎のsingletsファイルに出力されている。これらのファイルを連結して、tgi-clの結果から作成したシングルトン配列のファイルに追加する。

## 【 0 1 2 4 】

本実施例では、表1で示すように、全リード配列469,318の内、両端アダプタが確認できたリード配列300,635配列にこれを実施し、15326のクラスタ（2本以上の配列が同じものと判定されたグループ）と284554配列のシングルトン（他に同じ配列がないと判定された配列）を得ることができた。全リード配列1,868,178の場合については表1を参照のこと。

## 【 0 1 2 5 】

## 工程4：クラスタの信頼性のデータ化

上記工程3で得られたクラスタ配列に対して、HiCEPのセレクション配列部分がどの程度確からしいかを評価する。具体的には、配列クラスタのコンセンサス配列と構成リード配列について、配列の類似性をスコア化した。これにより、(2)のプロファイリングの対応付け処理等、作成されたクラスタ情報を使用する際、ここで算出した配列クラスタの信頼性を閾値にした処理を行なうことができるようになる。

## 【 0 1 2 6 】

## (A) 前処理

クラスタリング・アッセンブリ結果から、セレクション評価を行うために下記の前処理を行う。

## 【 0 1 2 7 】

- ・ アダプタ配列から配列の向きを確認し、順鎖方向(CGG-TTA)に変換する
- ・ アダプタ配列をCCGG/TTAAに置換する。

## 【 0 1 2 8 】

## (B) コンセンサス構成配列による評価

コンセンサス構成配列による評価では、下記の二種類のスコアを5'端と3'端についてそれぞれ計算する。

## 【 0 1 2 9 】

- ・ セレクション塩基の構成配列割合(第3候補まで)

コンセンサス構成配列のセレクション一致率を評価し、ピーク対応付けがうまくいかなかった場合の修正候補となる。理想な場合のスコアは第一候補が100%で、第二および第三候補は0%である。ただし、ヘテロのSNPがセレクション部分に入っていた場合、第一候補と第二候補が50%ずつとなる(図2.3参照)。

## 【 0 1 3 0 】

- ・ 制限酵素サイトから内側5塩基のコンセンサス配列と構成配列の編集距離平均(図2.4参照)

理想な場合のスコアは0である。ただし、ヘテロのSNPがセレクション部分に入っていた場合、0.5となる。セレクション塩基の構成配列割合を計算する際に、セレクション塩基として認識される塩基を図2.3に示す。

## 【 0 1 3 1 】

10

20

30

40

50

## (C) クラスタの分割

256通りのPCR及び電気泳動を行なうHiCEP法の場合、セレクション塩基（アダプタ配列から内側の2塩基）は、(2)のプロファイリングの対応付け処理を行なう際に重要なデータとなる。そこで、本実施例では、すべての配列クラスタについて、アダプタ配列の内側2塩基のそれぞれの位置について、コンセンサス配列の塩基と配列クラスタを構成する各リード配列の塩基をデータ化した。これにより、本実施例のHiCEP法の場合、(2)のプロファイリングの対応付けの処理において、ひとつの配列クラスタの構成リード配列のセレクション塩基位置の塩基が2種類に分かれる場合は、そのセレクション部分にヘテロなSNPが存在すると判定し、クラスタをふたつに分けることができるようになる(図15)。

10

## 【0132】

工程5：既知遺伝子情報を利用したコンセンサス配列の信頼性のデータ化

上記工程4で得られた配列クラスタのコンセンサス配列について、既知遺伝子情報（転写産物、ゲノム、EST情報など）が存在する生物種では、公知配列情報を検索し、コンセンサス配列の信頼性データを作成する。

## 【0133】

## (A) 類似性検索の実行

すべての配列クラスタとシングルトンについて、コンセンサス配列またはシングルトンのリード配列を既知公共データベースに類似性検索をかけ、その出力をデータ化する。

## 【0134】

- ・mRNA：blastn -dust no -task megablast
- ・ゲノム：blat パラメータなし（デフォルトのまま）。

20

## 【0135】

## (B) カテゴリ分類

公共データベースへの類似性検索結果を、下記の4つのカテゴリに分ける。ただし、95-95は塩基の一致率95%以上・アライメント長がクエリ長の95%以上を表し、95-20baseは塩基の一致率が95%以上・アライメント長が20塩基以上を表す。

## 【0136】

1. 95-95かつCCGG-TTAAが存在する
2. 95-95でCCGG/TTAAの両方または片方が1塩基違いで存在する
3. 95-20baseで末端部分にヒット(クエリ配列の開始/終了位置が端から4塩基以内)し、CCGG/TTAAが存在する
4. 95-20baseで末端部分にヒットし、CCGG/TTAAが1塩基違いで存在する。

30

## 【0137】

また、制限酵素サイトを探す場合は、下記のようにアライメント位置から前後2塩基を加えた中で探す。

## 【0138】

- ・クエリ配列：ccGG XX XXXX...
- ・サブジェクト配列：...yy zzZZ YY XXXX...

40

## 【0139】

大文字がアライメントされた配列で、この例ではクエリ配列の3塩基目からアライメントが始まっている。サブジェクト配列でクエリ配列の制限酵素サイトと同じ座標となる配列(zzZZ)に前後2塩基を加えたyyzzZZYYの中でCCGGを探すこととなる。

## 【0140】

1塩基違いで制限酵素サイトを探した場合、yyzzZZYYの中に複数個所の候補が存在する場合がある。たとえば下記の配列の場合；

...CCNGGX...

CCNG GXと考えるとGXをセレクションととるか、CNGG XXと考えるとXXをセレクションととるかの二通りがある。これらは、配列のみでは判断できないので、候補が複数ある場合は、両方ともデータとして残すこととする。

50

## 【 0 1 4 1 】

すべての配列クラスタとシングルトンについて、コンセンサ配列またはシングルトンのリード配列と類似性のあった既知転写産物のIDと類似性のあった領域及びスコア、また、類似性のあったゲノム配列の染色体番号と類似性のあった領域及びスコアをデータと格納した。これにより、(2)のプロファイリングの対応付け処理等、作成されたクラスタ情報を使用する際、ここで算出した配列クラスタの信頼性を閾値にした処理を行なうことができるようになる。

## 【 0 1 4 2 】

加えて、すべての配列クラスタのコンセンサ配列について、アダプタ配列の内側2塩基のそれぞれの位置について、コンセンサ配列の塩基と類似性のあった公知配列の塩基をデータ化した。これにより、本実施例のHiCEP法の場合、(3)の配列同定処理において、異なる個体の試料の場合、対応する配列クラスタが存在しない場合も、SNPが存在すると仮定して検索処理を行なうことができる。

10

## 【 0 1 4 3 】

工程6：コンセンサ配列への遺伝子情報の付与

上記工程4で得られた配列クラスタのコンセンサ配列について、公知配列情報を検索し、配列に遺伝子情報を付与する。

## 【 0 1 4 4 】

対象生物が既知遺伝子情報(転写産物、ゲノム、EST情報など)が存在する生物種であれば、工程5において、その情報は付与されている。しかしながら、網羅的フラグメント解析では、未知の転写産物を多く検出することもある。よって、工程6においては、配列クラスタのコンセンサ配列について、すべての生物種、または、特定の複数の生物種の公知配列情報を類似性検索し、それぞれのコンセンサ配列に類似性の高い公知の配列を対応付ける。

20

## 【 0 1 4 5 】

(2)電気泳動で得られるバンドまたはピークと配列の対応付け

個々の配列クラスタとその配列のシーケンス対象のHiCEP法で得られた「鋳型cDNA液」からPCRと電気泳動を行って得られたプロファイリングである電気泳動のピーク群(ES細胞のリファレンスプロファイリング)の対応付け方法を開発した。対応付けには、(1)で得られた各配列クラスタのコンセンサ配列の配列と配列長、配列クラスタを構成するリード配列数を使用する。

30

## 【 0 1 4 6 】

工程1：配列長の補正

電気泳動長と電気泳動対象となった配列の配列長は、かならずしも一致しないことが知られている(図39参照)。本法では、ピークと配列を一致させるためには、電気泳動長と配列長のズレはひとつの課題である。この課題を解決するために、既存のマウスES細胞にHiCEP法を適用したデータベースにおいて、対応付け済みのピーク37,675とその配列のデータを利用して、塩基組成や分子量と電気泳動長と配列長のズレとの関係を検討した。その結果、塩基組成や分子量による補正が可能で、ピーク対応付け精度を向上させることができることがわかった。ひとつは、ズレが配列の塩基組成はTG(またはAC)含量と相関していることがわかった(図42参照)。また、分子量によっても、ズレに傾向があることがわかった(図41参照)。

40

## 【 0 1 4 7 】

その効果としては、図16のように、既存のマウスES細胞で補正しない場合、ズレが $\pm 2\text{bp}$ 以内の配列が89%であるのに対して、塩基組成と分子量で補正することで、96%に増加することがわかった。また、工程2についても、補正を行わずに工程2を実施した場合の正解率は66%であったが、補正を行なうことによって77%に増加することがわかった。

## 【 0 1 4 8 】

(A)ズレと配列長、分子量との関係を用いた較正

50



既知のシーケンスデータより、明らかに不適切なデータを除去した後、既知のシーケンスデータの電気泳動長と対応する配列長とのずれと配列長との関係を検討したところ、図40のように、配列長によりずれに偏りがあることがあることが分かった。また、ずれと分子量との関係を散布図で表すと図41のようになった。分子量は配列長と比べるとより単位が細かいために、細かい較正を行うには分子量を利用した較正が良いので、分子量での較正表を採用する。較正表の作成には、局所回帰平滑化関数であるloess関数を用いた。

【0149】

(B) ずれとシーケンスの内部塩基組成との関係を用いた較正

ずれと配列の内部塩基組成との関係を調べるために、(A)で計算した「ずれと分子量との関係を用いた較正」後のずれ(以下、(A)で較正した後のずれ)と配列の内部塩基組成との関係を検討した結果、A、C含量割合と(A)で較正した後のずれとの間に負の相関関係があること、T、G含量割合と(A)で較正した後のずれとの間に正の相関関係があることがわかった。較正の結果を大きくするために(1)で較正した後のずれとAC含量割合との関係、及びTG含量割合との関係を散布図で表すと図42のようになった。A、C、T、G単体よりもはっきりとした相関があるように見えた。上記(1)で較正した後のずれとAC含量割合との関係、及びTG含量割合との関係より較正表を作成し、較正を行う。較正表の作成には、局所回帰平滑化関数であるloess関数を使用する。

10

【0150】

(C) 較正表を用いた電気泳動長の予測

20

補間には線形補間を用いた。求めたい点 $X(x_x, y_x)$ の前後に較正表に記載されている点 $A(x_A, y_A)$ 、 $B(x_B, y_B)$ が存在する場合の求めたい点 $X$ の較正值 $y_x$ は下式の通りになる(図43参照)。

【数1】

$$\text{較正值 } y_x = \frac{(y_B - y_A)x_x - y_B x_A + y_A x_B}{x_B - x_A}$$

30

【0151】

工程2：配列クラスタとピークとの対応付け処理

各配列クラスタのコンセンサス配列の工程1で補正された配列長とHiCEP法で得られたプロファイリングである電気泳動のピークの電気泳動長、及び、配列クラスタを構成するリード配列数と電気泳動で得られたピークの強度のふたつの値を使用して、配列クラスタとHiCEPのリファレンスプロファイリングのピークとの対応付けを行なった。

【0152】

具体的には、次の手順で行なう。

【0153】

- (A) クラスタリング・アッセンブリ処理結果から " 擬似ピーク " を生成
- (B) 擬似ピークへピーク長較正を適用
- (C) 同じピーク長の擬似ピークを1つの擬似ピークにまとめる
- (D) ピーク対応付けアルゴリズムによる対応付け。

40

【0154】

この結果、ES細胞のHiCEP法で得られたプロファイリングピーク21,778に対応付けられたピークの数12,551ピーク(57.6%)で、その内77%をコンピュータ処理によって同定することができた。

【0155】

(A) クラスタリング・アッセンブリ処理結果から " 擬似ピーク " を生成  
配列クラスタのコンセンサス配列またはシングルトン配列に対し、ピーク長・高さを次の

50

ように割り当て、擬似ピークを生成する。

【0156】

- ・ピーク長： コンセンサス配列のセレクション塩基を含む塩基数 + 34
- ・ピーク高さ： 当該配列クラスタのリード数。

【0157】

配列長から電気泳動長への補正值 + 34 は次のように決定される。

【0158】

PCR時に使用されるプライマー配列の長さ40塩基とPCRで人工的にチミンが末端に結合された分41塩基が、HiCEPのセレクション塩基を含みフラグメントDNA配列の外側に付加された配列がPCR産物となる。配列長は、アダプタ配列を除去した塩基数を用いるため、41塩基から、配列長に含まれるセレクション塩基2塩基の両端分の4塩基を引いた37塩基が、配列長を電気泳動長にするための補正值である。しかしながら、アプライドバイオシステムズ社製のキャピラリー電気泳動装置（特に3100）では、この理論的な補正值より3塩基少ない電気泳動位置に現われることがわかっており、よって、理論的な補正值から3塩基引いた値の34塩基を補正值とした。

10

【0159】

ピークの高さについては、リード数をそのままピークの高さに適用すると、プロファイルピークよりもかなり低い値になる。本システムのピーク対応付けアルゴリズムでは、高さの絶対値に影響を受けないためこれは問題にならない。一方、擬似ピークを視覚化する場合は、リード数 = 高さでは高さの関係が見えにくい。そこで、本仕様書での擬似ピークの描画では、リード数に一定の係数をかけて、プロファイルピークと同レベルの高さまで引き上げている（図30を参照）。

20

【0160】

(B) 擬似ピークへピーク長較正を適用

擬似ピークのピーク長を、上記(2)工程1で生成した較正表により較正する（図28参照）。

【0161】

(C) 同じピーク長の擬似ピークを1つの擬似ピークにまとめる

内部配列は異なるが配列長が同じフラグメントが存在した場合、HiCEPの電気泳動結果では、ひとつのピークとして現われ、一方、上記(1)で作成した配列クラスタは異なる配列クラスタとなる。よって、ピークと配列クラスタの対応付けを行なう場合、擬似ピークの較正後の配列長が同一である場合は、これらの擬似ピークをまとめてひとつの擬似ピークとし、高さは合計してひとつの擬似ピークの高さとする（図29参照）。

30

【0162】

なお、配列クラスタのコンセンサス配列の配列長さが同じであっても、配列長較正を施した場合、擬似ピークの電気泳動長が大きく異なる可能性がある。よって、較正後の配列長が+/-0.25ベース以内のピーク同士を1つのピークにまとめることとする。

【0163】

(D) ピーク対応付けアルゴリズムによる対応付け

シーケンス・ピーク対応付けアルゴリズムは、DPマッチングを基本的な枠組みとし、各エッジのスコアリングを独自に行うようにしたものである。

40

【0164】

A) 擬似ピークの特徴

以下に擬似ピークの特徴をまとめる；

(I) リード数が発現量を反映していると考えられるが、ばらつきが大きく、同じ電気泳動同士で比較する場合のように一定係数をかけることで全体の高さをそろえることが難しい；

(II) リード数が多いコンセンサス配列ほど信頼性が高いと考えられる。リード数が少ないコンセンサス配列ほどセレクション塩基や配列長に誤りが発生する可能性が高くなると考えられる；

50

(III) 較正対象となる配列長と電気泳動長のずれは、領域単位で一定のずれがある場合よりも、単独のピークごとにずれ幅が異なる。

【0165】

これらの特徴を前提にピーク対応付けアルゴリズムで。

【0166】

B) DPマッチングによるピークアライメント

ペアのスコア値を得るための範囲をフレームと呼ぶ。一定のフレーム領域を設定して、そのフレーム内のリファレンスプロファイリングピークと擬似ピークのすべての組み合わせをペア候補としてスコアを付け、スコアの合計が最も高くなるペアの組み合わせをDPマッチング法で求める(図17A、図31、図32参照)。各ペア候補のスコアの最高値は1.0とし、0以上の場合に最終的なペアとなる可能性が生じる。スコアがマイナス値の場合、そのペア候補が最終的なペアになる可能性はない。

10

【0167】

DPマッチングの具体的な方法としては、「Needleman & Wunsch, 1970の方法を修正したもの」を利用した。

【0168】

C) ペア候補のスコア

ペア候補のスコアは、ピークの高さスコアとサイズスコアそれぞれに重みを掛けて合計したものとする。高さスコア、サイズスコアそれぞれの最高値は1.0であり、それぞれに重み係数を掛けることでペア候補のスコアの最高値も1.0となるようにする；

20

ペア候補のスコア = (高さスコア × 高さの重み) + (サイズスコア × サイズの重み)

実施例では、

高さの重み = 0.5, サイズの重み = 0.5を使用した(図17B参照)。

【0169】

I) 高さのスコア

高さスコアはペア候補ごとに次のように計算する。ピークの高さスコアを計算するにあたって、高さ値の代わりにフレーム内での高さの順序番号を使う(図17B参照)。

【0170】

高さスコア =  $(\text{error} - \text{abs}(\text{p.order} - \text{r.order})) / \text{error}$

30

error = 高さ順序番号の許容差(現状は10)

p.order = プロファイルピークの高さ順序番号

r.order = 擬似ピークの高さ順序番号

abs(n) = nの絶対値。

【0171】

II) 高さ順序番号とフレーム

高さ順序番号はフレーム内で最も高いピークから順に1, 2, 3...nと割り振る(図33参照)。プロファイルピークと擬似ピークそれぞれで別々に割り振る。フレーム及び高さ順序番号は、着目するプロファイルピークごとに計算する(1プライマーセットあたりプロファイルピーク数と同じ数のフレームが生成される)。

40

【0172】

なお、高さを順序番号に置き換える(図34参照)ことにより、高さの関係を考慮しつつ、データの特徴(I)「擬似ピークの高さのばらつき」の影響を受けないようにできる。また、低いピークほど高さ順序番号の一致精度が悪くなるが、これはデータの特徴(II)による。

【0173】

III) 高さが同じピークの扱い

同じ高さのピークは、擬似ピークの高さが配列数であるため高い頻度で発生する可能性がある。フレーム内で高さの同じピークには同じ順序番号を割り振る(図35参照)。このときの順序番号は、同じ高さのピーク数を加算したものとする。同じ高さのピーク群

50

は、リード数が少ないかシングルトンである可能性が高い。そのようなピークは、より離れた順序番号をつけておくほうが一致精度が良くなる（ノイズの影響を少なくできる）。

【0174】

IV) フレーム幅の決定方法

実施例においては、着目プロファイルピーク前後でプロファイルピーク数27以内、かつ80塩基以内の範囲をフレームとする（擬似ピークは考慮しない）（図36）。

【0175】

V) 高さ順序番号の許容差

ペア候補の高さ順序差が「高さ順序番号の許容差」以内であればペナルティにならない。

10

【0176】

VI) サイズスコア

サイズスコアはペア候補ごとに次のように計算する（図17B参照）。

【0177】

サイズスコア =  $(\text{error} - \text{abs}(\text{p.size} - \text{r.size})) / \text{error}$

error = サイズ許容差(現状は2から4:後述)

p.size = プロファイルピークのサイズ

r.size = 擬似ピークのサイズ

abs(n) = nの絶対値。

【0178】

VII) サイズ許容差

ペア候補のサイズ差が「サイズ許容差」以内であればペナルティにならない。サイズ許容差は、2ベースから4ベースの間で可変とする。

20

【0179】

以下にサイズ許容差を求める手順を示す。

【0180】

- ・着目プロファイルピーク前後両方向それぞれに、隣のピークまでの距離を求める；
- ・前後2つの距離のうち、短いほうの距離の1/2(二分の一)を候補値とする；
- ・候補値が2～4ベース以内なら、候補値をそのままサイズ許容差とする；
- ・候補値が2ベースより小さければ2、4ベースより大きければ4をサイズ許容差とする（図37）。

30

【0181】

サイズが大きくなるにつれて、許容差を大きくしていく方法も考えられるが、本法では、上記方法を採用した。

【0182】

VIII) 近傍ピークの判定の矯正

対応すべきプロファイリングピークと擬似ピークが近傍にある場合、強度の大きい擬似ピークとサイズに近いが強度が低いプロファイリングピークのペアは高さ順序番号の差が大きくなることでスコアが低くなり、最終的なペアにはならないことが多い。しかし、低いプロファイリングピークに対応する擬似ピークがなく、本来対応すべき擬似ピークとプロファイリングピークが一定のサイズ以上はなれているとペアになってしまうことがある。

40

【0183】

このような場合への対処として、強度の強いピークとその近傍のピークについてペアを矯正する（図38）。矯正方法は以下の通り。

【0184】

- ・対応付け対象となるピークの前後0.75ベース以内に、強度が対象ピークの30%以下のピークがあればそれを“近傍の低いピーク”とする（左右は区別する）；
- ・スコア計算において一方が裾野ピークの場合、同じ近傍の強度が低いピーク同士の場合のみスコアを計算する。それ以外の場合、スコアを-1（ペナルティ）とする。

50

## 【0185】

この矯正は、配列長較正を適用した場合にのみ効果がある（配列長較正を行わない場合、擬似ピークは均等に1ベース刻みになる）。

## 【0186】

(3) 上記(1)データベース、及び上記(2)の対応付け情報を使用して、HiCEPで得られるピークの遺伝子同定法

上記(2)によって、HiCEP法で調整した同じ鋳型cDNAを使用して作成した配列クラスタとリファレンスプロファイリングのピークとを対応付け、リファレンスプロファイリングのピークについて配列を同定できるようになった。次は、同じ細胞で、別の試料にHiCEP法を適用して得られたプロファイリングのピークの配列を上記(1)のデータベースと  
10  
上記(2)の対応付け情報を使用して、同定する方法である（図25、図26、図27参照）。

## 【0187】

## 方法1

工程1：遺伝子同定対象のサンプルから得たHiCEPのプロファイリング結果と(2)で使用したリファレンスプロファイリングを電気泳動長とピークの強度で対応付けたデータを作成する。

## 【0188】

工程2：遺伝子同定対象のサンプルから得た遺伝子同定対象ピーク群から、上記工程1で作成した対応付けデータを利用して、リファレンスプロファイリングのピークを求め、さらに、(2)での対応付け情報から、(1)で作成したクラスタを求め、コンセンサス配列と遺伝子情報を求める。これにより、注目のピーク群と遺伝子情報との対応リスト  
20  
が作成される。

## 【0189】

工程3：(1)の工程6で作成した遺伝子情報により、注目するコンセンサス配列を決定し、(2)で対応付けられたリファレンスプロファイリングのピークを介して、遺伝子同定対象のサンプルから得た電気泳動のピークを求める。

## 【0190】

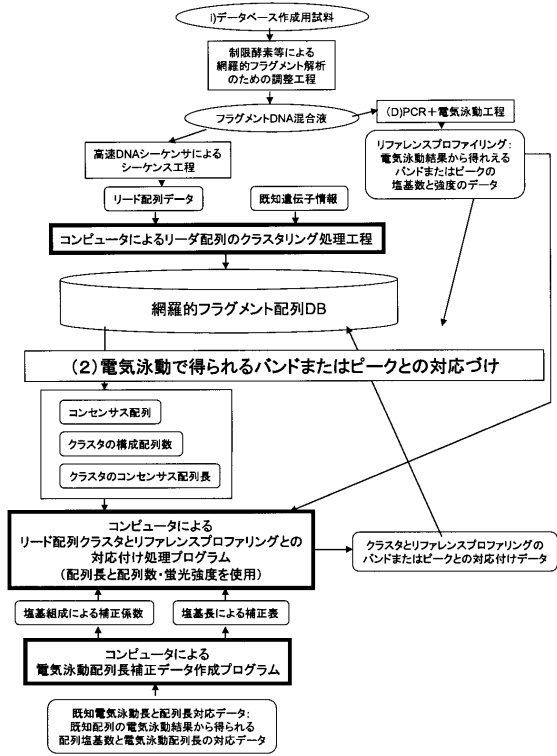
## 方法2

遺伝子同定対象サンプルから得られた電気泳動結果のひとつもしくは複数のピークの電気泳動で得られた塩基数を(2)で使用したリファレンスプロファイリングのバンド、さらに、(1)で作成した配列クラスタとその配列数から作成した擬似プロファイリングおよびその遺伝子情報を並べて提示することで、遺伝子同定対象サンプルから得られた電気泳動結果の注目ピークの遺伝子情報を得る。  
30

【 図 1 】

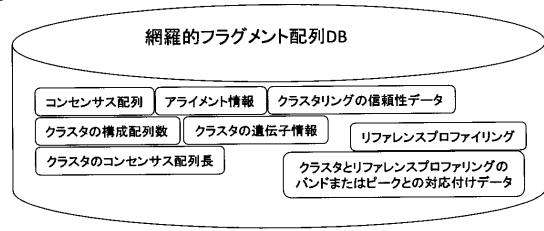
図 1

(1) 高速DNAシーケンサを活用した網羅的フラグメント配列DBの構築



【 図 2 】

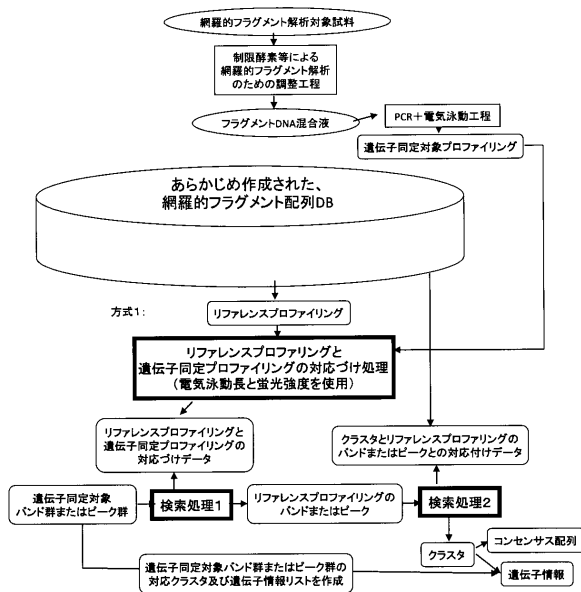
図 2



【 図 3 】

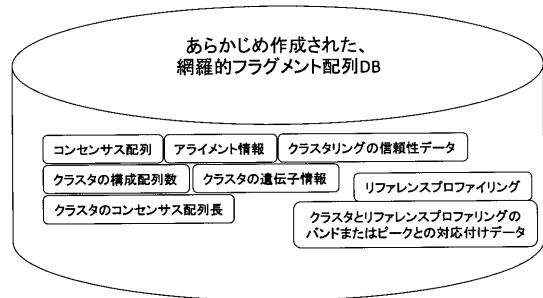
図 3

(3) 対応づけ情報を使用した網羅的フラグメント解析で得られるバンドまたはピークの遺伝子同定法



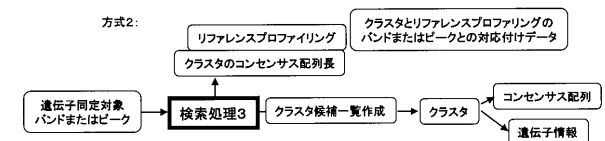
【 図 4 】

図 4



【 図 5 】

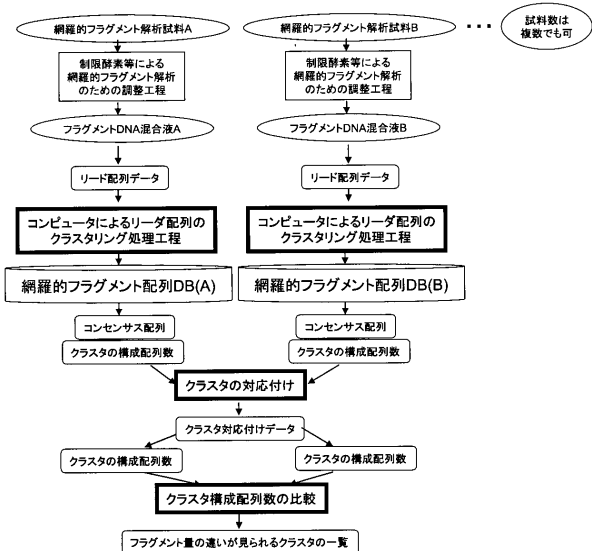
図 5



【 図 6 】

図 6

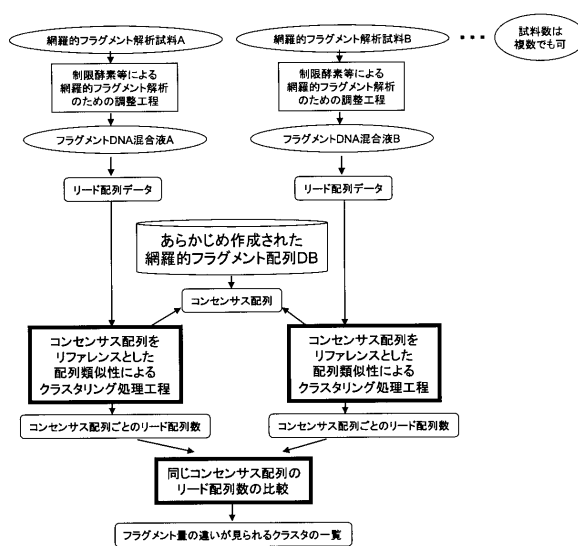
(4) 網羅的フラグメント解析の高速DNAシーケンサによる検出



【 図 7 】

図 7

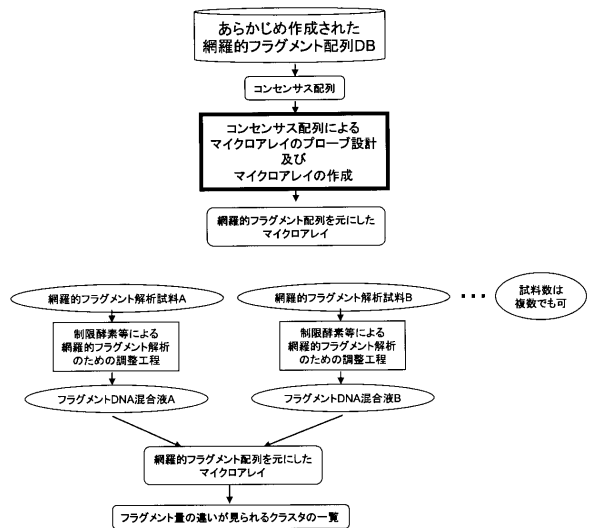
(5) データベースをリファレンスにした網羅的フラグメント解析の高速シーケンサによる検出



【 図 8 】

図 8

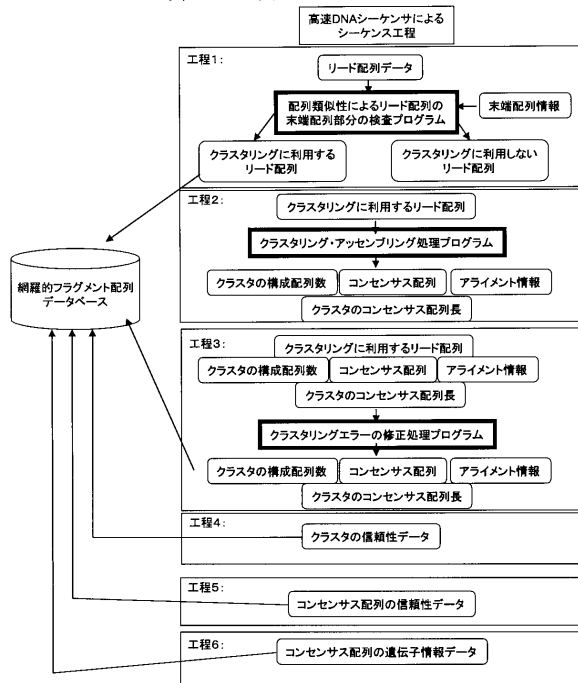
(6) データベースからプローブを設計して作成したマイクロアレイによる網羅的フラグメント解析の検出



【 図 9 A 】

図 9A

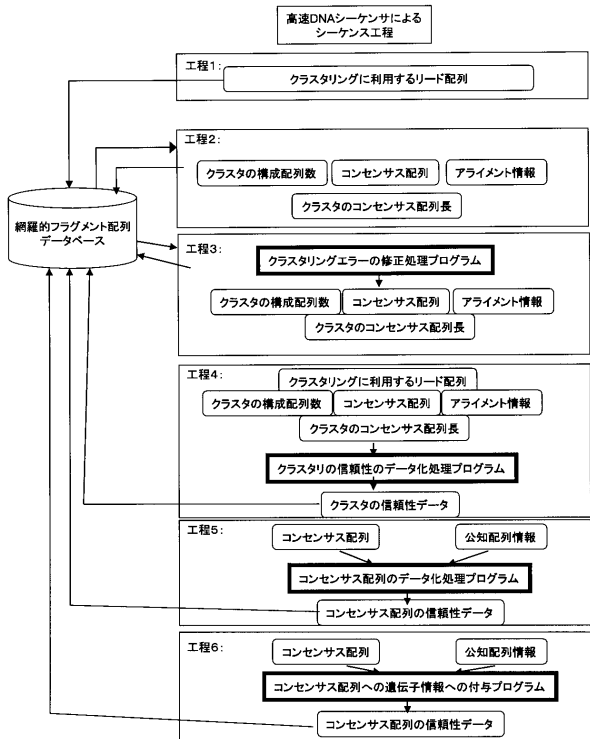
コンピュータ処理によりリード配列のクラスタリング処理工程



【 図 9 B 】

図 9B

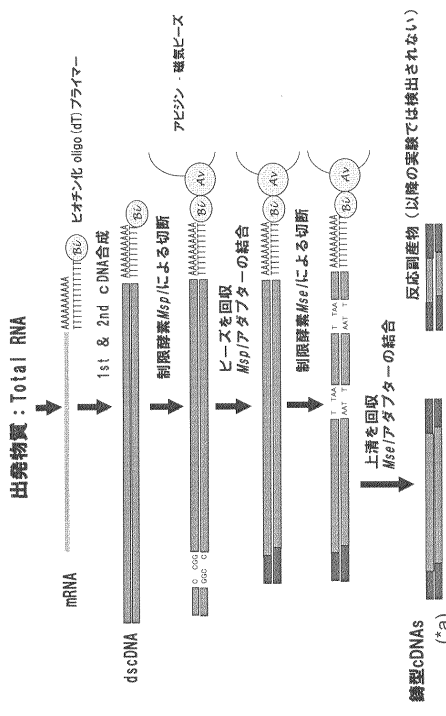
コンピュータ処理によりリード配列のクラスタリング処理工程



【 図 1 0 】

図 10

鋳型cDNA作成

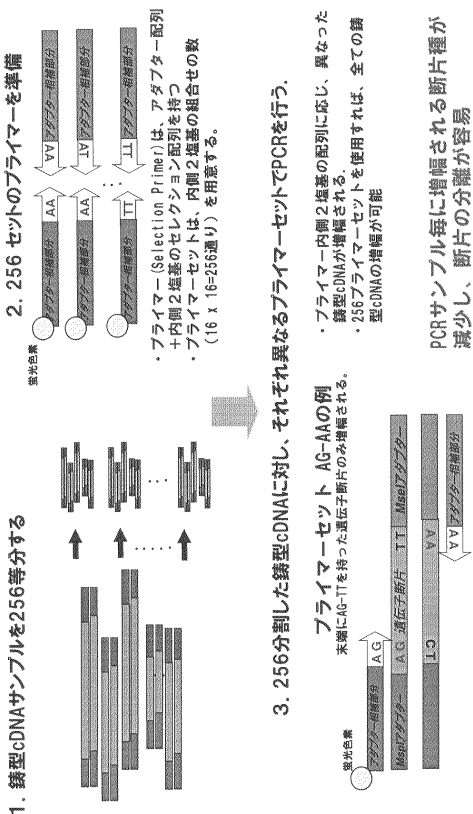


1種類の転写物から検出対象フラグメントは1種類しか得られない

【 図 1 1 】

図 11

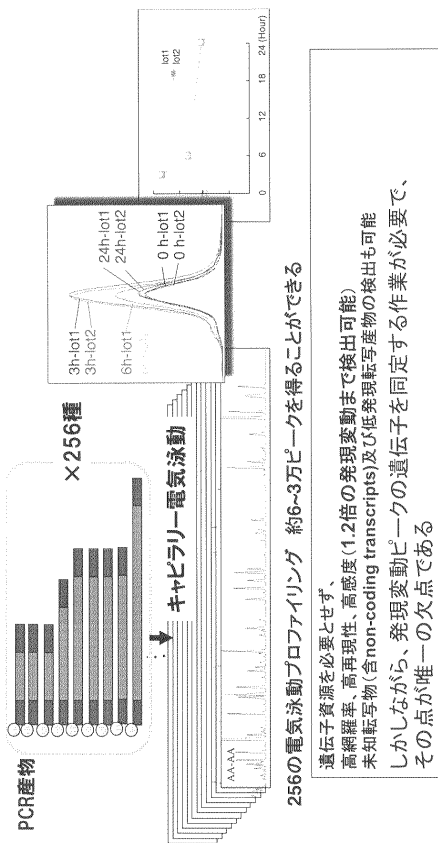
Selective PCR



【 図 1 2 】

図 12

断片長による分離と検出



256の電気泳動プロファイルが約6~3万ピークを得ることができる

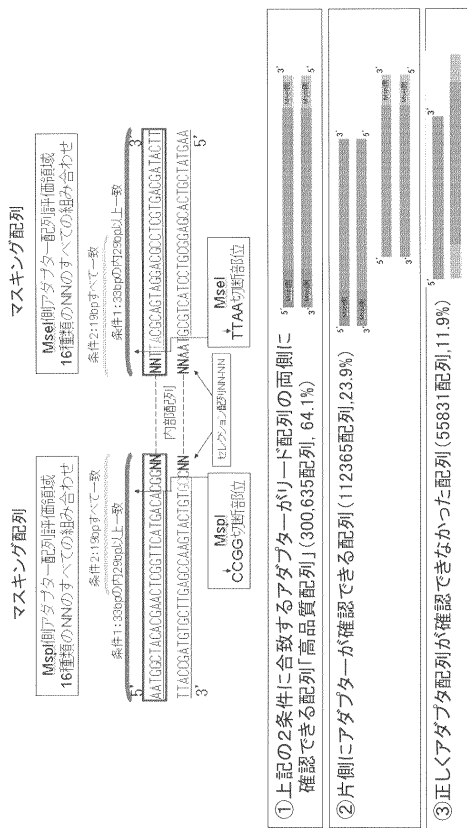
遺伝子資源を必要とせず、  
高感度、高再現性、高感度(1.2倍の発現変動まで検出可能)  
未知転写物(含non-coding transcripts)及び低発現転写産物の検出も可能  
しかしながら、発現変動ピークの遺伝子を同定する作業が必要で、  
その点が唯一の欠点である



【 図 1 3 】

図 13

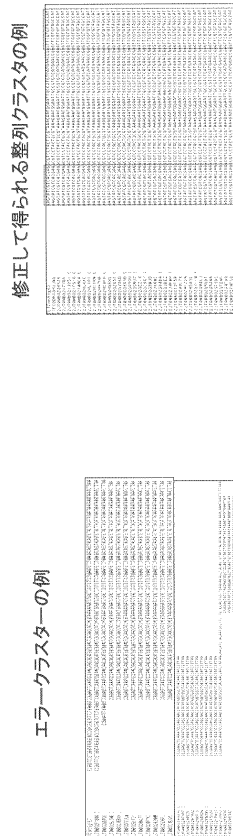
### アダプタ配列の評価



【 図 1 4 】

図 14

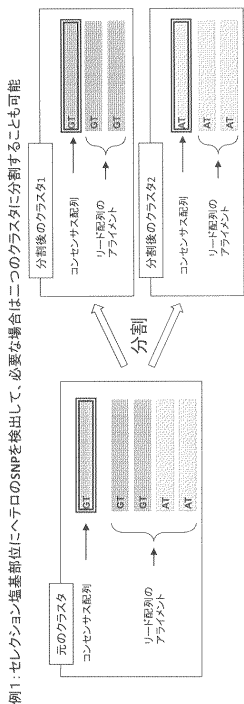
### エラークラスタの修正



【 図 1 5 】

図 15

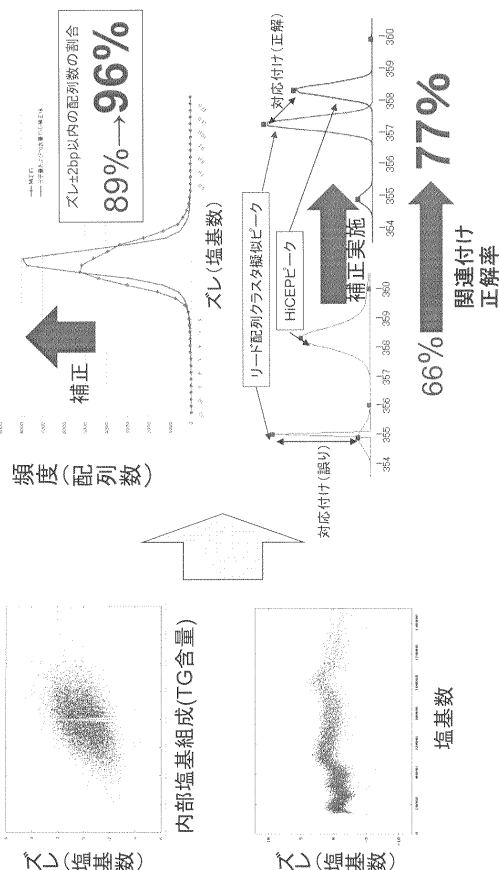
### セレクション塩基評価とヘテロSNPによるクラスタ分割



【 図 1 6 】

図 16

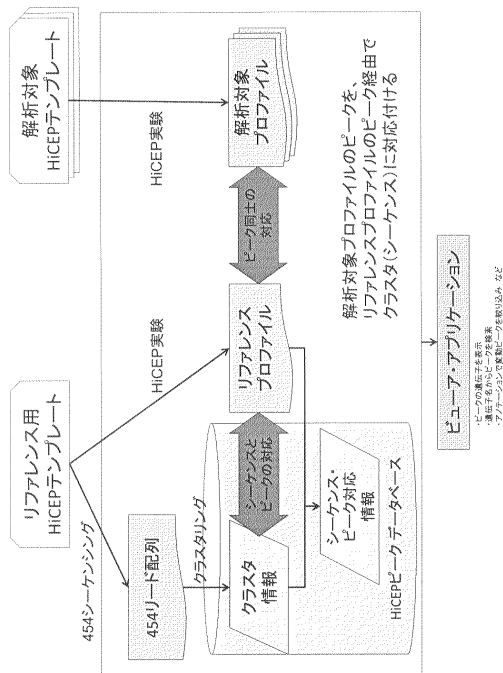
### 電気泳動長と配列長のズレ補正





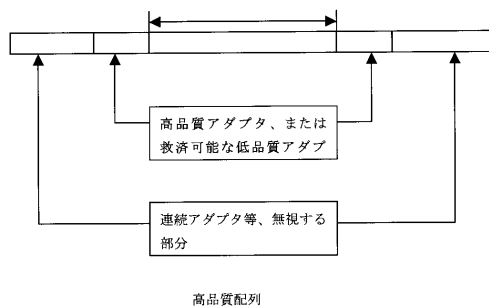
【 図 2 0 】

図 20



【 図 2 1 】

図 21



【 図 2 2 】

図 22

cross\_match のアライメント出力例(下線したアライメントが偽アダプタ):

```

33 0.00 0.00 0.00 F2JTNB02GRSRN 1 33 (203) MspI_TG 1 33 (0)
F2JTNB02GRSRN 1 AATGGCTACACGAACCGGTCATGACACGGTG 33
MspI_TG 1 AATGGCTACACGAACCGGTCATGACACGGTG 33
Transitions / transversions = 1.00 (0 / 0)
Gap_init rate = 0.00 (0 / 33), avg. gap size = 0.00 (0 / 0)

16 4.35 4.35 8.70 F2JTNB02GRSRN 171 193 (43) MseI_tt 8 29 (4)
F2JTNB02GRSRN 171 CAGATAGGACCCCTCGTG-CTGAT 193
MseI_tt 8 CAG-TAGGACGCCTCGTGAC-GAT 29
Transitions / transversions = 0.00 (0 / 1)
Gap_init rate = 0.14 (3 / 22), avg. gap size = 1.00 (3 / 3)

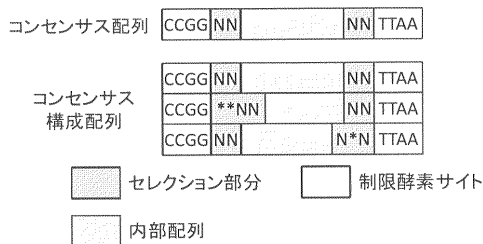
33 0.00 0.00 0.00 F2JTNB02GRSRN 204 236 (0) MseI_ca 1 3 (0)
F2JTNB02GRSRN 204 CATTACGCAGTAGGACGCCTCGTGACGATACTT 236
MseI_ca 1 CATTACGCAGTAGGACGCCTCGTGACGATACTT 33
Transitions / transversions = 1.00 (0 / 0)
Gap_init rate = 0.00 (0 / 33), avg. gap size = 0.00 (0 / 0)

```

【 図 2 3 】

図 23

セレクション塩基の構成配列割合を計算する際に、セレクション塩基として認識される塩基を下記に示す。



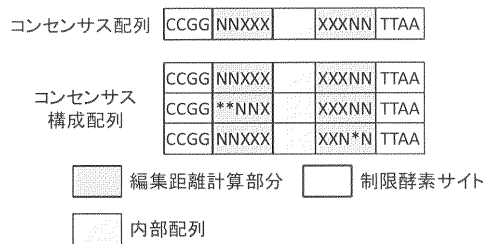
セレクション割合

「NN」の部分がそれぞれの配列のセレクションとなる。アセンブリのアライメントにギャップ「\*」が入っていた場合、それを除いたうえでセレクション塩基を決める。

【 図 2 4 】

図 24

コンセンサス配列と構成配列の編集距離平均を計算する際に、計算の対象となる塩基を下記に示す。



セレクション近傍 5 塩基

【 図 2 5 】

図 25

用語注: 公開 WEB 目 では、Fragment Length=ピーク長、Sequence Length=配列長、の ようである。

Primer Set: 5' AA 3' AA

Fragment Length:  bp

Kind of Fragment Search:
 Fragment Length & Sequence Length
 Profiling Fragment Length Only
 Includes Simulation Data

Range Parameter:
Profiling Fragment Length Range:  ±  bp
Sequence Length Range:  ±  bp

【 図 2 7 】

図 27

Subject DB:  Experimental Data  Simulated Data

Query Sequence:

E-Value:  10.0

【 図 2 6 】

図 26

Search condition of Cluster

1: for cDNA Cluster
UniGene ID: 
UniGene Title: 
cDNA Accession:

2: for cDNA and Genome Cluster
Gene Symbol: 
Locus ID:

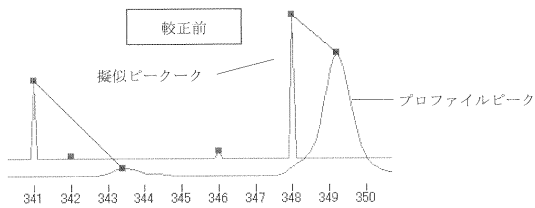
3: for Genome Cluster
Genome Position:

【 図 2 8 】

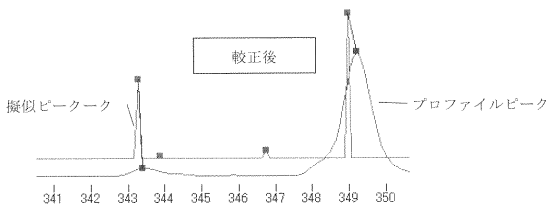
図 28

較正前と較正後の例

(a)

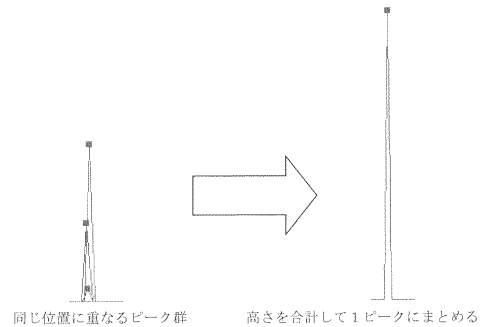


(b)



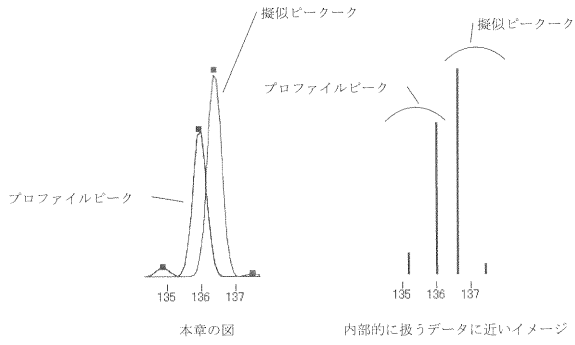
【 図 2 9 】

図 29



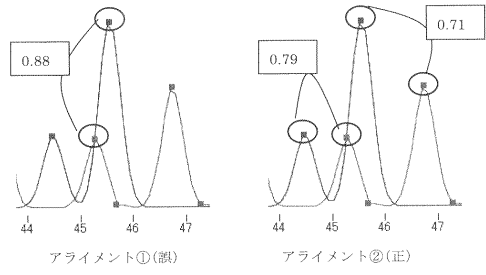
【 図 3 0 】

図 30



【 図 3 2 】

図 32



【 図 3 1 】

図 31

DP マッチング表

	r1	r2	r3	r4
p1	0.9	0.5	0.4	0.3
p2	0.7	0.8	-0.2	-0.3
p3	-0.2	0.3	0.4	0.7

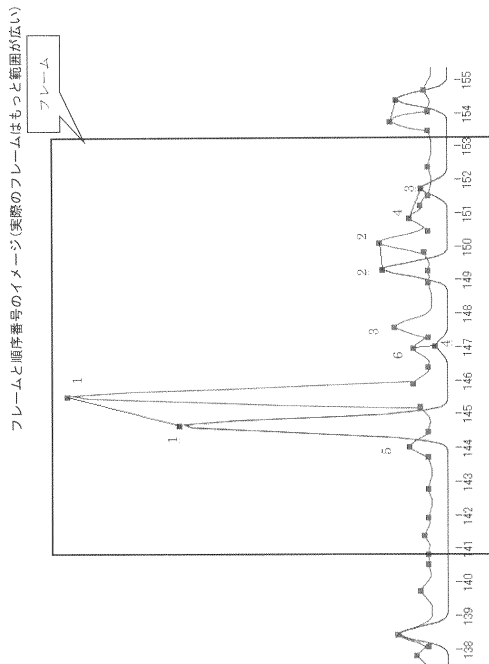
アライメント

p1	p2	p3	
⋮	⋮	⋮	
r1	r2	r3	r4

プロフィールピーク: p1, p2, p3  
 擬似ピーク: r1, r2, r3,

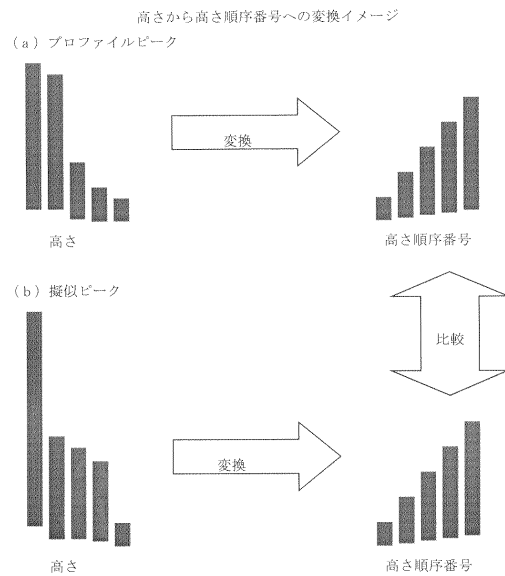
【 図 3 3 】

図 33



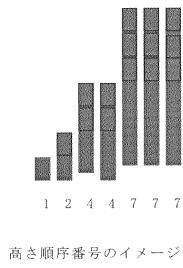
【 図 3 4 】

図 34



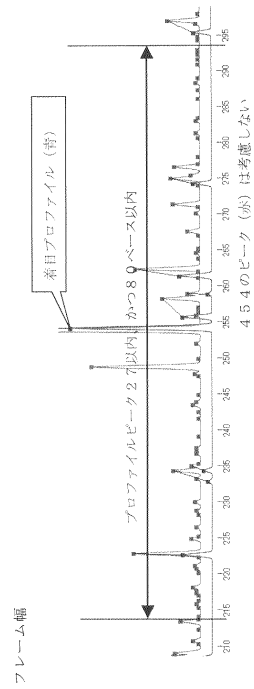
【 図 3 5 】

図 35



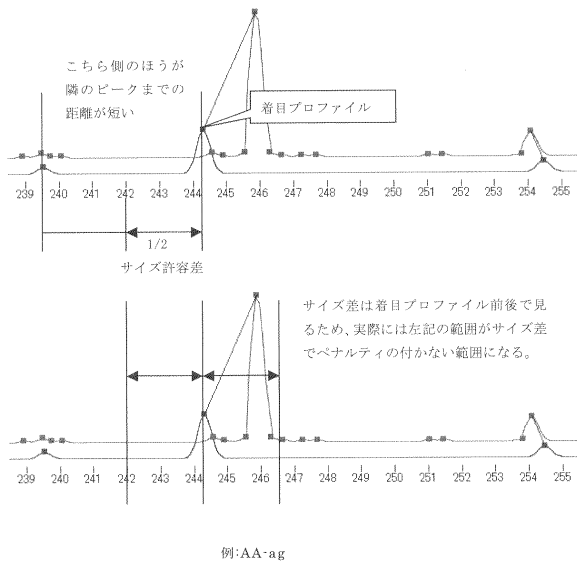
【 図 3 6 】

図 36



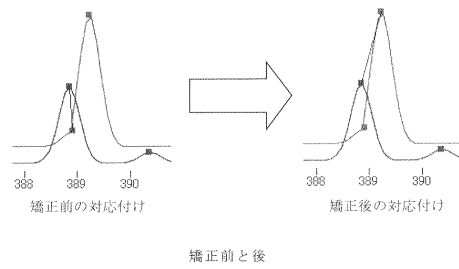
【 図 3 7 】

図 37



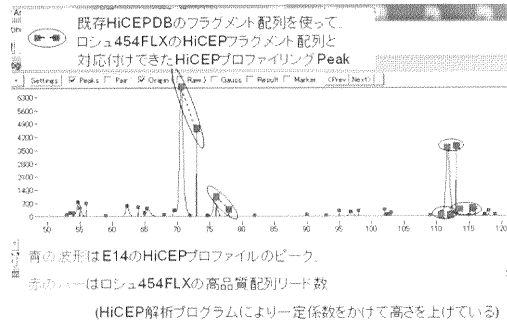
【 図 3 8 】

図 38



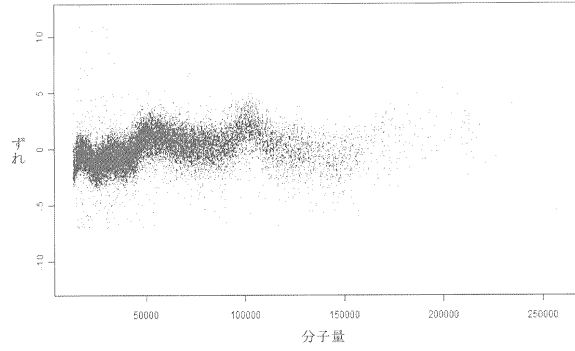
【 図 3 9 】

図 39



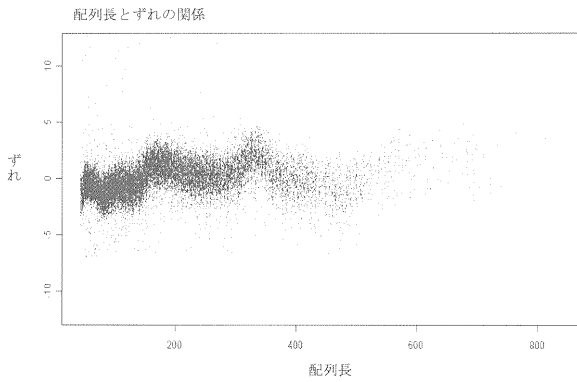
【 図 4 1 】

図 41



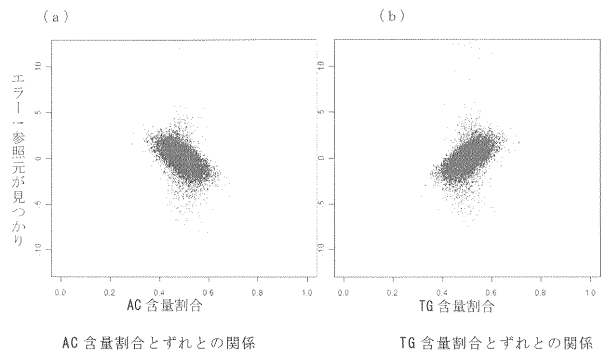
【 図 4 0 】

図 40



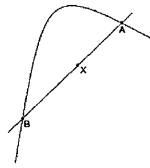
【 図 4 2 】

図 42



【 図 4 3 】

図 43



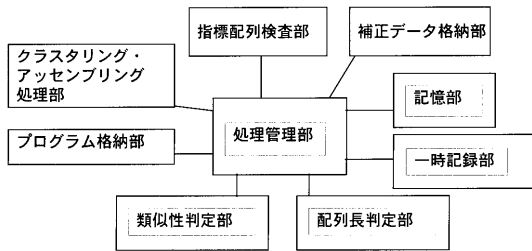
線形補間による計算方法

・校正値の計算方法  
補間には線形補間を用いた。求めたい点 X(x<sub>x</sub>, y<sub>x</sub>)の前後に校正表に記載されている点 A(x<sub>A</sub>, y<sub>A</sub>)、B(x<sub>B</sub>, y<sub>B</sub>)が存在する場合の求めたい点 X の校正値 y<sub>x</sub> は下式の通りになる。

$$\text{校正値 } y_x = \frac{(y_B - y_A)x_x - y_B x_A + y_A x_B}{x_B - x_A}$$

【 図 4 4 】

図 44



## 【手続補正書】

【提出日】平成24年3月1日(2012.3.1)

## 【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】0138

【補正方法】変更

【補正の内容】

【0138】

・クエリ配列:ccGG XX XXXX...

(配列番号13)

・サブジェクト配列:...yy zzZZ YY XXXX... (配列番号14)。

## 【手続補正2】

【補正対象書類名】図面

【補正対象項目名】図10

【補正方法】変更

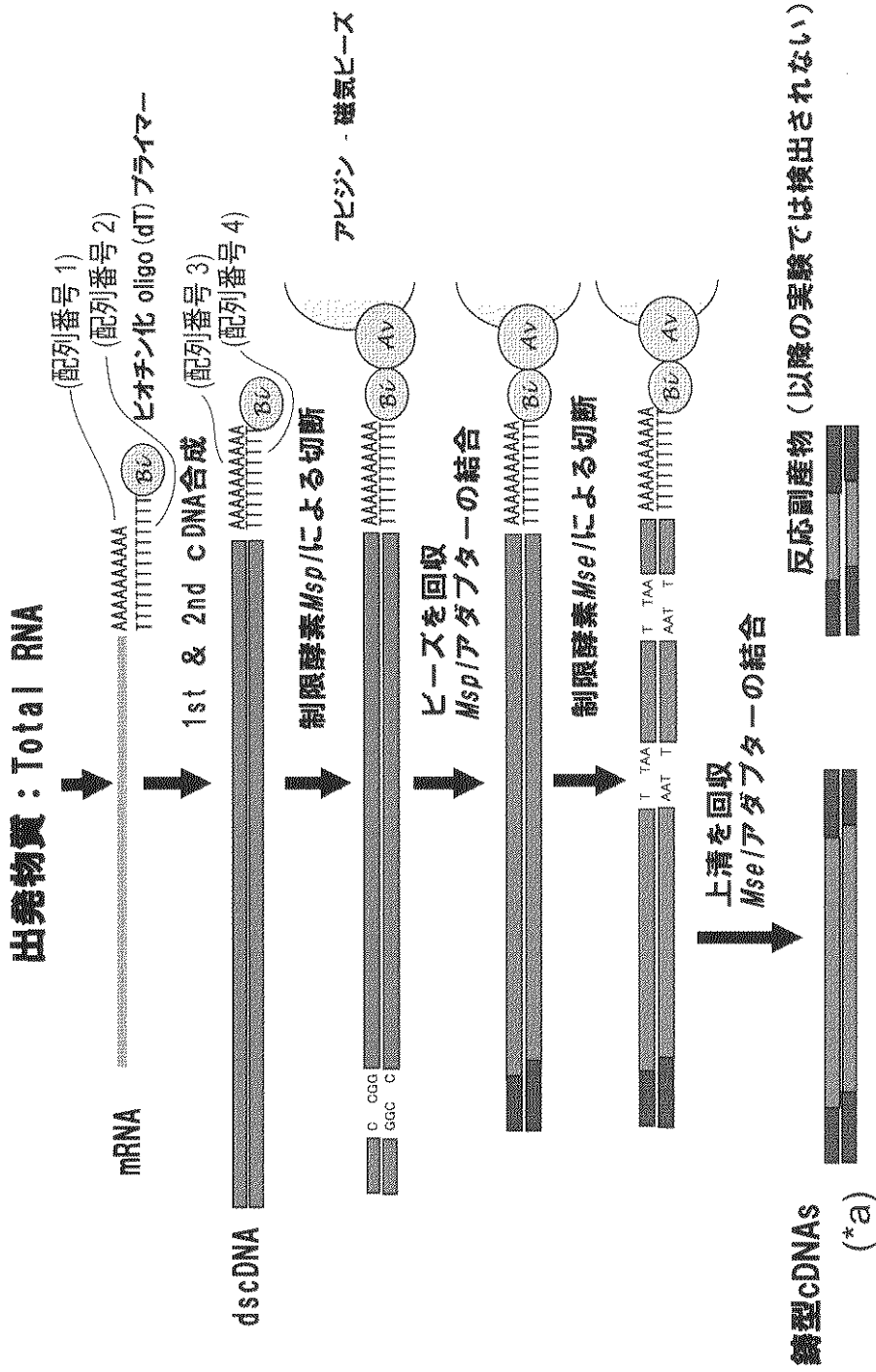
【補正の内容】



【 図 1 0 】

図 10

# 鋳型cDNA作成



1種類の転写物から検出対象フラグメントは1種類しか得られない

【手続補正3】

【補正対象書類名】図面

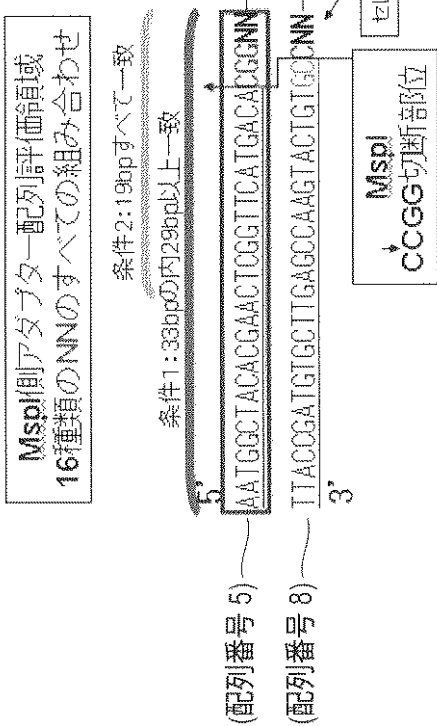
【補正対象項目名】図13

【補正方法】変更

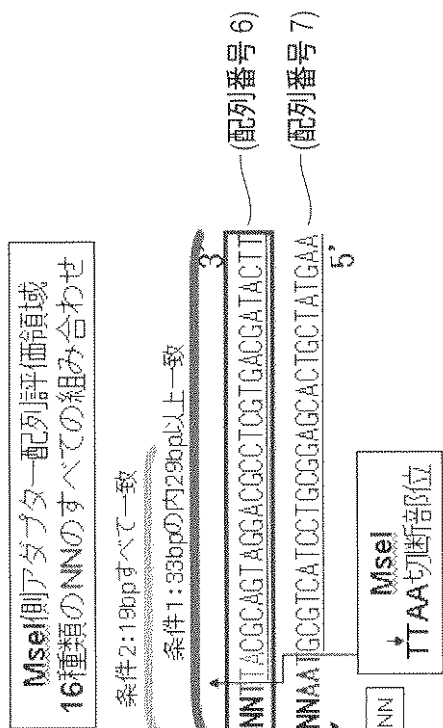
【補正の内容】

# アダプタ配列の評価

## マスキング配列



## マスキング配列



- ①上記の2条件に合致するアダプターがリード配列の両側に確認できる配列「高品質配列」(300,635配列, 64.1%)
  - ②片側にアダプターが確認できる配列(112365配列,23.9%)
  - ③正しくアダプタ配列が確認できなかった配列(55831配列,11.9%)
-

【手続補正4】

【補正対象書類名】図面

【補正対象項目名】図22

【補正方法】変更

【補正の内容】

【 図 2 2 】

図 22

cross\_match のアライメント出力例(下線したアライメントが偽アダプタ):

33	0.00	0.00	0.00	F2JTQNB02GRSRN	1	33 (203)	MspI_TG	1	33 (0)
F2JTQNB02GRSRN	1	AATGGCTACACGAACTCGGTTTCATGACACGGTG	33	(配列番号 9)					
MspI_TG	1	AATGGCTACACGAACTCGGTTTCATGACACGGTG	33	(配列番号 10)					
Transitions / transversions = 1.00 (0 / 0)									
Gap_init rate = 0.00 (0 / 33), avg. gap size = 0.00 (0 / 0)									
16	4.35	4.35	8.70	F2JTQNB02GRSRN	171	193 (43)	MseI_tt	8	29 (4)
F2JTQNB02GRSRN	171	CAGATAGGACCCCTCGTG-CTGAT	193	(配列番号 11)					
		- v - -							
MseI_tt	8	CAG-TAGGACGCCTCGTGAC-GAT	29	(配列番号 12)					
Transitions / transversions = 0.00 (0 / 1)									
Gap_init rate = 0.14 (3 / 22), avg. gap size = 1.00 (3 / 3)									
33	0.00	0.00	0.00	F2JTQNB02GRSRN	204	236 (0)	MseI_ca	1	3 (0)
F2JTQNB02GRSRN	204	CATTACGCAGTAGGACGCCTCGTGACGATACTT	236	(配列番号 13)					
MseI_ca	1	CATTACGCAGTAGGACGCCTCGTGACGATACTT	33	(配列番号 14)					
Transitions / transversions = 1.00 (0 / 0)									
Gap_init rate = 0.00 (0 / 33), avg. gap size = 0.00 (0 / 0)									

## フロントページの続き

- (74)代理人 100109830  
弁理士 福原 淑弘
- (74)代理人 100075672  
弁理士 峰 隆司
- (74)代理人 100095441  
弁理士 白根 俊郎
- (74)代理人 100084618  
弁理士 村松 貞男
- (74)代理人 100103034  
弁理士 野河 信久
- (74)代理人 100119976  
弁理士 幸長 保次郎
- (74)代理人 100153051  
弁理士 河野 直樹
- (74)代理人 100140176  
弁理士 砂川 克
- (74)代理人 100158805  
弁理士 井関 守三
- (74)代理人 100124394  
弁理士 佐藤 立志
- (74)代理人 100112807  
弁理士 岡田 貴志
- (74)代理人 100111073  
弁理士 堀内 美保子
- (74)代理人 100134290  
弁理士 竹内 将訓
- (72)発明者 安倍 真澄  
千葉県千葉市稲毛区穴川4丁目9番1号 独立行政法人放射線医学総合研究所内
- (72)発明者 湯野川 春信  
東京都中野区中央3-13-11 MGビル508 株式会社メイズ内
- (72)発明者 佐藤 伸司  
千葉県長生郡長生村岩沼918番2 株式会社メイズ内
- (72)発明者 近藤 一弘  
東京都中野区中央3-13-11 MGビル508 株式会社メイズ内
- (72)発明者 日永田 隆志  
東京都中野区中央3-13-11 MGビル508 株式会社メイズ内
- Fターム(参考) 4B024 AA11 CA04 CA09 CA20 HA14  
4B063 QA01 QQ42 QR08 QR55 QR62 QS25 QS34