(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2015 年 10 月 8 日(08.10.2015)

WIPO | PCT

(10) 国際公開番号
WO 2015/151758 A1

(51) 国際特許分類:
G06F 19/22 (2011.01)

(21) 国際出願番号: PCT/JP2015/057348

(22) 国際出願日: 2015 年 3 月 12 日(12.03.2015)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2014-077278 2014 年 4 月 3 日(03.04.2014) JP

(71) 出願人: 株式会社日立ハイテクノロジーズ(HITACHI HIGH-TECHNOLOGIES CORPORATION) [JP/JP]; 〒1058717 東京都港区西新橋一丁目２４番１４号 Tokyo (JP).

(72) 発明者: 木村　宏一(KIMURA Kouichi); 〒1008280 東京都千代田区丸の内一丁目６番６号　株式会社日立製作所内 Tokyo (JP).

(74) 代理人: 特許業務法人磯野国際特許商標事務所 (ISONO INTERNATIONAL PATENT OFFICE, P.C.); 〒1020093 東京都千代田区平河町２丁目７番４号　砂防会館別館内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): 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, 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,

[続葉有]

(54) **Title**: SEQUENCE DATA ANALYZER, DNA ANALYSIS SYSTEM AND SEQUENCE DATA ANALYSIS METHOD

(54) 発明の名称 : 配列データ解析装置、ＤＮＡ解析システムおよび配列データ解析方法

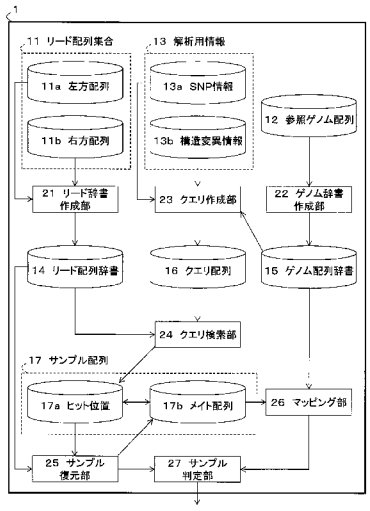


FIG. 2:
11 Lead sequence set
11a Left sequence
11b Right sequence
12 Referential genome sequence
13 Information for analysis
13a SNP information
13b Structural variation information
14 Lead sequence dictionary
15 Genome sequence dictionary
16 Query sequence
17 Sample sequence
17a Hit position
17b Mate sequence
21 Lead dictionary creating section
22 Genome dictionary creating section
23 Query creating section
24 Query searching section
25 Sample reconstructing section
26 Mapping section
27 Sample determining section

(57) **Abstract**: A sequence data analyzer (1) comprising: a lead dictionary creating section (21) creating a lead sequence dictionary (14) based on a connection string, said connection string consisting of a pair of a left sequence (11a) and a right sequence (11b), which are obtained by sequencing a sample DNA fragment respectively from the left and right ends, and connecting characters connecting these sequences (11a, 11b) together; and a sample reconstructing section (25) extracting, as a sample sequence (17), a string until the appearance of a final character positioned in the neighborhood of a hit position (17a) of a query sequence (16) in the lead sequence dictionary (14), and extracting, as a mate sequence (17b), the left sequence (11a) or right sequence (11b) until the appearance of a final character in the side where the hit position (17a) does not exist in the sample sequence (17).

(57) 要約:　配列データ解析装置（１）は、サンプルＤＮＡ断片の両端からそれぞれシーケンシングされたペアである左方配列（１１ａ）と右方配列（１１ｂ）との間を結合文字で連結した結合文字列をもとにリード配列辞書（１４）を作成するリード辞書作成部（２１）と、リード配列辞書（１４）内のクエリ配列（１６）のヒット位置（１７ａ）の周囲に位置する終端文字が出現するまでの文字列をサンプル配列（１７）として抽出し、サンプル配列（１７）内のヒット位置（１７ａ）が存在しない側の終端文字が出現するまでの左方配列（１１ａ）または右方配列（１１ｂ）をメイト配列（１７ｂ）として抽出するサンプル復元部（２５）と、を有する。

WO 2015/151758 A1

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, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (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), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ：

配列データ解析装置、ＤＮＡ解析システムおよび配列データ解析方法

参照による取り込み

[0001] 本出願は、２０１４年４月３日に出願された日本特許願第２０１４－７７２７８号の優先権を主張し、その内容を参照することにより本出願に取り込む。

技術分野

[0002] 本発明は、配列データ解析装置、ＤＮＡ解析システムおよび配列データ解析方法に関する。

背景技術

[0003] ゲノムＤＮＡは、その塩基配列がＤＮＡ全体ですでにシーケンシングされ（読み取られ）、その塩基文字列はインターネット上のサーバなどに公開されている。研究者は、そのゲノムＤＮＡをお手本（参照データ）として、シーケンサ装置が読み取った被験者のサンプルＤＮＡ断片との塩基位置の照合（ゲノムマッピング）を行うことにより、サンプルＤＮＡ内の変異を検出する。変異とは、例えば、一塩基多型（ＳＮＰ：Single Nucleotide Polymorphism）や、ＳＶ（構造変異）などのゲノムＤＮＡの塩基文字列とサンプルＤＮＡ断片の塩基文字列との相違箇所である。なお、サンプルＤＮＡ断片は、１本のサンプルＤＮＡがシーケンサ装置内のフラグメント処理により多数に断片化されたものである。

[0004] ここで、シーケンサ装置は限られた長さのＤＮＡ配列しか読取ることができないため、ほぼ一定の長さに揃えられたサンプルＤＮＡ断片それぞれの両端から限られた長さ（サンプルＤＮＡ断片の一部だけ）を読み取った２本の（ペアの）リード配列を扱うペアドエンド法が知られている。つまり、ペアドエンド法では、サンプルＤＮＡ断片の中央部に、どちらのペアのリード配列にも属さない、シーケンシングされない区間が存在することになる。

よって、ペアドエンド法で読み取られたペアのリード配列を、ゲノムＤＮＡ内の塩基位置にゲノムマッピングするときには、ペアのうちの片方のリード配列だけを位置照合するよりも、もう片方のリード配列も併せて位置照合することにより、１本のサンプルＤＮＡ断片を高精度にゲノムＤＮＡ内にマッピングできる。

[0005] ゲノムＤＮＡやサンプルＤＮＡ断片の塩基文字列のデータ量が膨大であるため、以上説明したような、ペアドエンド法による２本の（ペアの）リード配列をゲノムＤＮＡ内にマッピングする際には、処理の効率化やデータの圧縮化が求められる。そこで、以下に示すように、様々なゲノムマッピングの効率化技術が提案されている。

[0006] ゲノムワイド・トランスクリプトーム・プロファイリングでは、全長ｃＤＮＡ上での５’末端から３’末端に向かう順序が維持されたペアエンド・ダイタグのシーケンシングが行われ、指定されたスペーサ配列に隣接するダイタグ配列を効率的に検出する方法が知られている（特許文献１）。

[0007] 新型ＤＮＡシーケンサから得られる大量のショート・リードを参照ゲノム配列に高速かつ高精度にアラインメントするために、ゲノム配列をＢＷ（Burrows-Wheeler）変換したデータを作成し、これを利用して、ショート・リードの先頭の数十塩基のシード配列と一致する配列が参照ゲノム内に出現する位置を高速に同定する技術は、一般に広く用いられている（例えば、非特許文献１）。

[0008] また、このとき、アラインメントできる候補位置が複数あるという曖昧性を減らすために、ペアのショート・リードのマッピングも広く行われている（例えば、非特許文献１）。通常、２つのショート・リードがペアにあることを示すため、それらには同一の名前（例えば、非特許文献１では、ＱＮＡＭＥ）が与えられる。

[0009] 変異解析を行うために、大量のショート・リードを参照ゲノムにマッピングして得られる大量のデータを利用する方法が、一般に広く利用されている（例えば、非特許文献３）。また、大きな計算コストがかかるマッピングの

計算を行わずに変異解析を行うために、参照ゲノム配列のＢＷ変換とショート・リード配列のＢＷ変換を利用するアプローチが考えられる（特許文献２、非特許文献４）。

先行技術文献

特許文献

[0010] 特許文献1：特表２００８－５４７０８０号公報

特許文献2：特願２０１３－０３８９１９号公報

非特許文献

[0011] 非特許文献1：Li H., HandsakerB., Wysoker A., Fennell T., RuanJ., Homer N., Marth G., AbecasisG., Durbin R. and 1000 Genome Project Data Processing Subgroup: The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9 (2009).

非特許文献2：Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60 (2009).

非特許文献3：M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetetics 43(5): 491-498 (2011).

非特許文献4：K. Kimura and A. Koike: A new approach to DNA sequence variation analysis using Burrows-Wheeler transform of massive short-read data, in Proceedings of Advances in Genome Biology & Technology Conferenece (AGBT) 2013, Marco Island, p.202.

発明の概要

発明が解決しようとする課題

[0012] ペアエンド法でシーケンスされた大量のリード配列データに対して、ペア関係の情報を計算機で扱うためには、ペアの片方のリード配列にもう片方のリード配列を関連付けるための識別子（またはポインタ情報）が必要となる。これらの識別子は、ペアの相手を一意に特定できなければならないため、

多数のビット数を必要とする。そのため、識別子のデータサイズが大きくなり、計算効率の低下を招く要因となる。例えば、長さ100塩基のショート・リードが80億本ある場合は40億通りの識別子が必要となり、識別子当たり少なくとも4バイト必要で、全リードに対する識別子データのサイズは4GB（ギガバイト）にも達する。

[0013] そこで、本発明は、サンプルＤＮＡ断片の両端からシーケンシングされるペアの片方のリード配列から、もう片方のリード配列を効率的に求めることを、主な課題とする。

課題を解決するための手段

[0014] 前記課題を解決するために、本発明の配列データ解析装置は、

サンプルＤＮＡ断片の両端からそれぞれシーケンシングされたペアである左方配列と右方配列との間を結合文字で連結した結合文字列を前記サンプルＤＮＡ断片ごとに作成し、各前記サンプルＤＮＡ断片の前記結合文字列を終端文字で結合した文字列をもとにリード配列辞書を作成するリード辞書作成部と、

ゲノムＤＮＡの塩基文字列から生成されるクエリ配列が出現する前記リード配列辞書内の塩基文字座標であるヒット位置を検索するクエリ検索部と、

前記リード配列辞書内の前記ヒット位置を起点として、その周囲に位置する前記終端文字が出現するまでの文字列をサンプル配列として抽出し、

前記サンプル配列内の前記ヒット位置を起点として、その周囲に位置する前記結合文字を検査し、検査した前記結合文字から前記ヒット位置が存在しない側の前記終端文字が出現するまでの前記左方配列または前記右方配列をメイト配列として抽出するサンプル復元部と、

前記メイト配列が出現する前記ゲノムＤＮＡ内の塩基文字列の塩基文字座標を検索するマッピング部と、を有することを特徴とする。

その他の手段は、後記する。

発明の効果

[0015] 本発明によれば、サンプルＤＮＡ断片の両端からシーケンシングされるペ

アの片方のリード配列から、もう片方のリード配列を効率的に求めることができる。

図面の簡単な説明

[0016] [図1]本発明の一実施形態に関するＤＮＡ解析システムを示す構成図である。

[図2]本発明の一実施形態に関する配列データ解析装置を示す構成図である。

[図3]本発明の一実施形態に関する各辞書の作成処理を示すフロー説明図である。

[図4]本発明の一実施形態に関する辞書作成処理におけるＢＷ文字列の作成処理を示す説明図である。図４（ａ）は、リード配列から結合文字列を作成する処理を示す。図４（ｂ）は、結合文字列からＢＷ文字列を作成する処理を示す。

[図5]本発明の一実施形態に関する辞書作成処理におけるWavelet Tree形式への変換処理を示す説明図である。図５（ａ）は、Wavelet Tree形式に変換するために参照される二分木を示す。図５（ｂ）は、ＢＷ文字列をWavelet Tree形式に変換する処理を示す。

[図6]本発明の一実施形態に関するＳＮＰの解析処理を示すフロー説明図である。

[図7]本発明の一実施形態に関するＳＮＰの解析処理における各データを示す説明図である。図７（ａ）は、ＳＮＰ情報を示す。図７（ｂ）は、ＳＮＰ情報とゲノム配列辞書とから作成されるクエリ配列を示す。図７（ｃ）は、リード配列辞書からクエリ配列を検索する様子を示す。図７（ｄ）は、ＳＮＰの解析結果を示す。

[図8]本発明の一実施形態に関する構造変異の解析処理を示すフロー説明図である。

[図9]本発明の一実施形態に関する構造変異の解析処理における各データを示す説明図である。図９（ａ）は、構造変異情報を示す。図９（ｂ）は、構造変異情報とゲノム配列辞書とから作成されるクエリ配列を示す。図９（ｃ）は、リード配列辞書からクエリ配列を検索する様子を示す。図９（ｄ）は、

構造変異の解析結果を示す。

発明を実施するための形態

[0017] 以下、本発明の一実施形態を、図面を参照して詳細に説明する。

[0018] 図１は、ＤＮＡ解析システムを示す構成図である。

配列データ解析装置１は、通常の計算機の構成を有するサーバ等のコンピュータで実現される。

配列データ解析装置１は、中央処理部（ＣＰＵ：Central Processing Unit）２０１、プログラムなどが記憶される記憶部であるメモリ２０２、操作のためのＧＵＩ（Graphical User Interface）や、解析結果などを表示する表示部２０３、配列辞書（図２のリード配列辞書１４、ゲノム配列辞書１５）などを記憶する記憶部として機能するハードディスクドライブ（ＨＤＤ）２０４、ＳＮＰなどの変異情報やパラメータ入力などを行うキーボード等の入力部２０５、インターネット等に接続するためのネットワークインタフェース（ＮＩＦ）２０６がバス２０７に接続された構成を備えている。

ＨＤＤ２０４に記憶される配列辞書は、配列データ解析装置１の外部に設置された記憶装置に記憶してもよいし、ネットワークを介してデータセンタなどに記憶してもよい。

以下で説明される各種のフローチャートは、ＣＰＵ２０１のプログラム実行などで実現される。

[0019] 配列データ解析装置１のＮＩＦ２０６には、ゲノムサーバ８とシーケンサ９とがネットワークを介して接続されている。

シーケンサ９は、各サンプルＤＮＡ断片に対して、その両端（5’末側のリード配列と、3’末側のリード配列）のペアをシーケンシングし（読み取り）、その結果を配列データ解析装置１に提供する。

なお、リード配列（塩基配列）の表記法として、5’末側の塩基文字を左側に記載し、3’末側の塩基文字を右側に記載する方法が一般的であるため、以下では、5’末側を「左方」とし、3’末側を「右方」とする。

シーケンサ９は、超並列型（いわゆる次世代型）ＤＮＡシーケンサとして

構成され、多数の（例えば、１億本の）サンプルＤＮＡ断片を並列にシーケンシングすることができる。

同様に、ゲノムサーバ８は、ゲノムＤＮＡをシーケンシングした結果であるゲノム配列を、配列データ解析装置１に提供する。

[0020] なお、ゲノムＤＮＡをシーケンシングしたゲノム配列を参照配列として、サンプルＤＮＡ断片を解析対象とする代わりに、ｍＲＮＡ配列を参照配列として、ｃＤＮＡサンプルのペアド・エンド・シーケンシング・データを解析対象とすることにより、スプライス・バリアントの検出を行うことができる。スプライス・バリアントによりエクソンが失われることは、構造変異の「欠失」に相当し、新たなエクソンが取りこまれることは構造変異の「挿入」に相当するからである。

[0021] 図２は、配列データ解析装置１を示す構成図である。この図２の説明では、各構成要素の概要を説明し、各構成要素の詳細については、後記する説明で明らかにする。

配列データ解析装置１は、シーケンサ９からリード配列集合１１（左方配列１１ａと右方配列１１ｂ）の入力を受け付ける。

リード配列集合１１は、シーケンサ９にシーケンシングされたサンプルＤＮＡ断片ごとのリード配列（左方配列１１ａと右方配列１１ｂ）の集合である。

左方配列１１ａは、サンプルＤＮＡ断片のうちの５’末側の端点を基点として、３’末側に向かってシーケンシングされたリード配列である。

右方配列１１ｂは、サンプルＤＮＡ断片のうちの３’末側の端点を基点として、５’末側に向かってシーケンシングされたリード配列である。

ここで、左方配列１１ａおよび右方配列１１ｂの長さは、例えば１００塩基程度であり、サンプルＤＮＡ断片が３００塩基程度であるとすると、中央の１００塩基程度は左方配列１１ａにも右方配列１１ｂにも含まれないシーケンシング対象外の箇所である。または、２万塩基程度の長さのサンプルＤＮＡ断片の場合は、シーケンシング対象外の箇所は１９８００塩基程度とな

る。

[0022] ＤＮＡ解析処理の前準備として、リード辞書作成部２１は、リード配列集合１１からリード配列辞書１４を作成する。ゲノム辞書作成部２２は、参照ゲノム配列１２からゲノム配列辞書１５を作成する。

なお、参照ゲノム配列１２は、解析の対象とする生物種ごとに定められており、各染色体の全長の配列の集合である。

[0023] 配列データ解析装置１は、今回の解析対象となる変異を示す解析用情報１３として、ＳＮＰ情報１３ａまたは構造変異情報１３ｂの入力を受け付ける。ＳＮＰとは、特定箇所におけるサンプルＤＮＡ断片の塩基内容がゲノムＤＮＡの同じ箇所の塩基内容と異なることである。構造変異とは、連続した複数の塩基の並びの挿入や欠失である。

[0024] クエリ作成部２３は、ゲノム配列辞書１５を参照して、解析用情報１３が示す変異を含むクエリ配列１６を作成する。

クエリ検索部２４は、リード配列辞書１４内から、クエリ配列１６にマッピングする（クエリ配列１６が出現する）塩基位置座標であるヒット位置１７ａを検索する。

サンプル復元部２５は、ヒット位置１７ａを含むサンプルＤＮＡ断片のリード配列（サンプル配列１７）を復元する。ここで、ヒット位置１７ａは、左方配列１１ａか、右方配列１１ｂかのいずれかに含まれるので、ヒット位置１７ａが含まれない片方のリード配列を、メイト配列１７ｂとする。

[0025] マッピング部２６は、ゲノム配列辞書１５を参照して、サンプルＤＮＡ断片のメイト配列１７ｂが由来するゲノムＤＮＡ内の位置を特定（ゲノム・マッピング）する。

サンプル判定部２７は、メイト配列１７ｂのマッピングの成否により、メイト配列１７ａのサンプルＤＮＡ断片が解析用情報１３が示す変異を含むか否かを判定する。そして、サンプル判定部２７は、解析用情報１３の判定結果を出力する。

[0026] 図３は、各辞書の作成処理を示すフロー説明図である。まず、リード辞書

作成部２１が、リード配列集合１１からリード配列辞書１４を作成するリード辞書作成処理（Ｓ１０１～Ｓ１０５）を説明する。

Ｓ１０１において、配列データ解析装置１は、シーケンサ９からリード配列集合１１（左方配列１１ａと右方配列１１ｂ）の入力を受け付ける。

図４（ａ）では、説明をわかりやすくするために、サンプルＤＮＡ断片３０１，３０５の２本を例示し、サンプルＤＮＡ断片３０１の左方配列１１ａ「ＧＡ」２塩基分、右方配列１１ｂ「Ｔ」１塩基分とし、サンプルＤＮＡ断片３０５の左方配列１１ａ「Ｃ」２塩基分、右方配列１１ｂ「ＴＡ」２塩基分とする。

[0027] シーケンサ９は、読み取ったリード配列集合１１を、左方配列１１ａを示す符号３６１と、右方配列１１ｂを示す符号３６２のように、ＦＡＳＴＱ形式で配列データ解析装置１に通知する。符号３６１，符号３６２では、２つのサンプルＤＮＡ断片３０１，３０５のリード配列を列挙しており、１つのリード配列に対して４行の情報が記載される。

ＦＡＳＴＱ形式の１行目「＠ｓｅｑ１，＠ｓｅｑ２」は、サンプルＤＮＡ断片の識別子（ＩＤ）であり、２行目「ＧＡ，Ｔ，Ｃ，ＴＡ」は、リード配列である。例えば、符号３６１の１行目「＠ｓｅｑ１」と、符号３６２の１行目「＠ｓｅｑ１」とが一致するので、同じサンプルＤＮＡ断片３０１から読み取られたペアであることがわかる。

[0028] 図３のＳ１０２において、リード辞書作成部２１は、ペアをなす左方配列１１ａの塩基文字と、右方配列１１ｂの塩基文字とを結合文字「＆」で結合することで、１本のサンプルＤＮＡ断片を示す結合文字列を作成する。なお、塩基文字とは、４種類の塩基それぞれを示す「Ａ、Ｃ、Ｇ、Ｔ」と、不明の塩基を示す「Ｎ」である。

図４（ａ）では、同じサンプルＤＮＡ断片３０１のペアである左方配列１１ａ「ＧＡ」と右方配列１１ｂ「Ｔ」とを結合文字「＆」３０２で結合して、末尾に終端文字「＄」３０３を付加することで、結合文字列３０４「ＧＡ＆Ｔ＄」を得る。同様に、サンプルＤＮＡ断片３０５のペアからは、結合文

字列３０６「Ｃ＆ＴＡ＄」を得る。

なお、サンプルＤＮＡ断片の長さに応じて、複数種類の結合文字を使い分けてもよい。例えば、約３００塩基のサンプルＤＮＡ断片から生成する結合文字列内の結合文字には「＆」を用い、約２００００塩基のサンプルＤＮＡ断片から生成する結合文字列内の結合文字には「＃」を用いてもよい。これにより、結合文字からサンプルＤＮＡ断片のペア相手だけではなく、サンプルＤＮＡ断片の長さを取得することができる。

[0029] 図３のＳ１０３において、リード辞書作成部２１は、結合文字列３０４，３０６をＢＷ変換してＢＷ文字列３１１を作成する。

図４（ｂ）では、例えば、以下の手順により、ＢＷ文字列３１１が作成される。この計算過程で、先頭から１文字ずつ比較して文字列比較を行う際、＄どうしが比較されたら比較を終了し、＆どうしが比較されたら比較を継続する。

（手順１）結合文字列３０４を巡回シフト（cyclic shift）して文字列のリスト３０７を得るとともに、結合文字列３０６も巡回シフトして文字列のリスト３０８を得る。

（手順２）２つのリスト３０７、３０８をマージすることで、マージ済みリスト３０９を得る。

（手順３）マージ済みリスト３０９をアルファベット順にソートすると、ソート済みリスト３１０を得る。そのとき、文字のソート順位は、例えば、「＄＜＃＜＆＜Ａ＜Ｃ＜Ｇ＜Ｔ＜Ｎ」とする。

（手順４）ソート済みリスト３１０の各行の末端の文字を連結して、ＢＷ文字列３１１を得る。

このようにして得られたＢＷ文字列３１１は、ソート済みなので、同じ文字が連続する頻度が高い。よって、ＢＷ文字列３１１をランレングス圧縮することで、データ量を圧縮することができる。

[0030] 図３のＳ１０５において、リード辞書作成部２１は、ＢＷ文字列３１１をWavelet Tree形式に変換することにより、検索を効率的に行うためのリード配

列辞書１４を作成する。

図５（ａ）は、Wavelet Tree形式に変換するために参照される二分木３２０を示す。この二分木３２０では、文字列に使用される全文字（＄、＆、Ａ、Ｃ、Ｇ、Ｔ、Ｎ）３２１が根となる。

二分木３２０は、文字列に使用される全文字（＄、＆、Ａ、Ｃ、Ｇ、Ｔ、Ｎ）３２１を、再帰的に二分類し、分類の末端で高々２種類の文字しか含まないように分類する方法を示した二分木である。

二分木３２０の根では、全文字３２１を、Ａと、Ｔ（Ｗ）３２４と、それ以外（Ｓ）３２５と、に分類する。Ｓに分類された文字（＄、＆、Ｃ、Ｇ、Ｎ）３２５も、同様にＭとＫに二分類する。以下、これを再帰的に繰り返し、分類の末端で高々２種類の文字しか含まないようにする。但し、ペアを表す符号＆（３０２）と終端記号＄（３０３）は、分類の末端３３１に一緒に現れるように分類する。

[0031] 図５（ｂ）のWavelet Tree３４０は、ＢＷ文字列３１１を示す二分木である。Wavelet Tree３４０の根は、二分木３２０のＷとＳの二分類に従って、ＢＷ文字列３１１をバイナリ文字列３４１に変換したものである。リード辞書作成部２１は、Ｗに分類される文字を抜き出した部分文字列３４２、および、Ｓに分類される文字を抜き出した部分文字列３４３を作成する。

リード辞書作成部２１は、（ＡとＴの２種類の文字しか含まない）部分文字列３４２を、０と１のバイナリ文字列３４４に変換する。

リード辞書作成部２１は、部分文字列３４３も同様に、Ｓに分類された文字のＭとＫへの二分類（二分木３２０で示される）に従ってバイナリ文字列３４５を作成する。

リード辞書作成部２１は、Ｍに分類される文字を抜き出した部分文字列３４６、および、Ｋに分類される文字を抜き出した部分文字列３４７を作成する。

[0032] 以上説明した各部分文字列を作成する処理を繰り返すことにより、図示した６本のバイナリ文字列が得られる。分解の末端３５１では、ペアを表す符

号＆（３０２）と終端記号＄（３０３）からなる文字列３５０がバイナリ文字列３５１で表される。バイナリ文字列３５１の長さ（ビット数）はリードの総数（ペアの総数の２倍）に等しい。

よって、Wavelet Tree３４０は、ＢＷ文字列３１１を可逆変換したものであり、Wavelet Tree３４０からＢＷ文字列３１１を復元することができる。

[0033] 図３のＳ１０６において、リード辞書作成部２１は、Wavelet Tree３４０から作成されたリード配列辞書１４を記憶手段に出力する。ここで、リード辞書作成部２１は、Wavelet Tree３４０に加えて、そのWavelet Tree３４０のrank関数やselect関数を計算するための、わずかな（ＢＷ文字列３１１に対して３．５％程度のデータ量の）補助データを、リード配列辞書１４に付加してもよい。

rank(p,c)とは、配列要素0～pのうちの文字「c」の出現回数を返す関数である。

select(i,c)とは、(i+1)番目の文字「c」が出現する配列位置を返す関数である。

補助データは、例えば、参考文献「Kouichi Kimura, Yutaka Suzuki, Sumio Sugano, and Asako Koike. Journal of Computational Biology. November 2009, 16(11): 1601-1613.」に記載の「hierarchical binary string」である。この補助データは、ＢＷ文字列３１１から、任意に与えられた塩基配列と一致するリード配列断片を全て求める、などの検索を効率的に行うためのデータである。

[0034] ＢＷ変換された文字列上でrank関数とselect関数を効率的に計算するための補助データが与えられているため、＄と＆の２種類の区切り文字を用いたＢＷ変換の特徴を利用して、任意に与えられた文字列ｓに対して、ｓに一致する全てのリード断片配列を効率的に求めることができ、また、さらに、各リード断片配列に対して、その左方と右方にある配列を（＄が現れるまで延長して）復元する計算を効率的に行える。

[0035] 以上、図３～図５を参照して、リード辞書作成処理（Ｓ１０１～Ｓ１０５

）を説明した。一方のゲノム辞書作成処理（図３のＳ１０１ｂ～Ｓ１０５ｂ）もリード辞書作成処理（Ｓ１０１～Ｓ１０５）と同様に作成できる。

Ｓ１０１ｂでは、ゲノム辞書作成部２２は、Ｓ１０１のリード配列集合１１の代わりに参照ゲノム配列１２の入力を受け付ける。

Ｓ１０２ｂでは、ゲノム辞書作成部２２は、参照ゲノム配列１２が示す複数のゲノムＤＮＡの染色体配列（塩基文字列）を、そのまま終端文字「＄」で連結して、１本の文字列を作成する。ここで、参照ゲノム配列１２はペア形式ではないので、Ｓ１０２のような結合文字「＆」によるペア連結処理は不要である。

Ｓ１０５ｂでは、ゲノム辞書作成部２２は、Ｓ１０５のリード配列辞書１４の代わりにゲノム配列辞書１５を出力する。

[0036] 図６は、ＳＮＰの解析処理を示すフロー説明図である。

Ｓ１２１において、配列データ解析装置１は、今回の解析対象となる変異を示す解析用情報１３として、ＳＮＰ情報１３ａの入力を受け付ける。

図７（ａ）では、ＳＮＰ情報１３ａを示すテーブル４００を例示する。テーブル４００の各行に示すように、ＳＮＰごとに、染色体名、染色体上の塩基位置座標、参照ゲノム配列内での塩基の種類（標準塩基）、ＳＮＰとして現れる塩基の種類（変異塩基）の情報を含む。

テーブル４００の１行目は、ＳＮＰが染色体７番上の１２３４５６塩基目の位置にあり、参照ゲノムの塩基「Ａ」が塩基「Ｇ」に変異することを示す。

[0037] 図６のＳ１２２において、クエリ作成部２３は、ゲノム配列辞書１５を参照して、ＳＮＰ情報１３ａが示すＳＮＰを含むクエリ配列１６を作成する。

図７（ｂ）の説明欄４２０は、Ｓ１２２の例示であり、横軸４２１は染色体上の塩基位置座標である。まず、クエリ作成部２３は、ゲノム配列辞書１５を参照して、テーブル４００の１行目で示されたＳＮＰの位置４２４の周辺（例えば、左右に１０塩基程度）の塩基配列４２２を求める。クエリ作成部２３は、ＳＮＰの位置４２４で塩基配列４２２の塩基を変異させた配列４

２３を作成し、これをクエリ配列１６とする。または、クエリ作成部２３は、変異を含む塩基配列４２３の代わりに、変異を含まない塩基配列４２２をクエリ配列１６とすることで、ＳＮＰの位置４２４に現れる標準塩基を検出することができる。

[0038] 図６のＳ１２３において、クエリ検索部２４は、リード配列辞書１４内から、クエリ配列１６にマッピングする（塩基文字列の並びが一致する）塩基位置座標であるヒット位置１７ａを検索する。

なお、ＢＷ変換されたリード配列辞書１４をもとに計算する過程で、２本の文字列（リード配列辞書１４、クエリ配列１６）を先頭から１文字ずつ比較する際には、終端文字「＄」どうしが比較される場合は比較を終了し、結合文字列「＆」どうしが比較される場合は比較を継続する。

[0039] Ｓ１３１～Ｓ１３９において、配列データ解析装置１は、Ｓ１２３で求めたヒット位置１７ａごとのループ処理を実行する。

Ｓ１３２において、サンプル復元部２５は、Ｓ１３１からのループで現在選択しているヒット位置１７ａを含むサンプルＤＮＡ断片のリード配列（サンプル配列１７）を復元する。この復元処理では、ＢＷ変換により作成されているリード配列辞書１４から、rank関数とselect関数を用いることにより、ヒット位置１７ａを起点として、終端文字「＄」が出現するまでリード配列辞書１４内を延長して走査することにより、左右の終端文字「＄」に挟まれたサンプル配列１７が取得される。

このようなrank関数とselect関数を用いる方法は、例えば、文献「Ferragina, P. and Manzini, G著,"Opportunistic data structures with applications",In 41st IEEE Symposium on Foundations of Computer Science,FOCS,pages390-398」に記載されている。

サンプル配列１７は、結合文字列３０４，３０６のように、結合文字「＆」および終端文字「＄」を含む。このサンプル配列１７を結合文字列「＆」で分離することで、サンプルＤＮＡ断片ごとの識別子を使わずに、ペアをなす２本のリード配列（左方配列１１ａ、右方配列１１ｂ）が得られる。

[0040] Ｓ１３３において、サンプル復元部２５は、以下に示すように、サンプル配列１７から結合文字「＆」が見つかるまで走査（検査）することで、メイト配列１７ｂを取得する（Ｓ１３４）。

図７（ｃ）の［１］の場合のように、ヒット位置１７ａの配列４２３から右方に延長した先に結合文字「＆」が現れるので、取得されるメイト配列１７ｂは、結合文字「＆」より右側の右方配列１１ｂである。

図７（ｃ）の［２］の場合のように、ヒット位置１７ａの配列４２３から左方に延長した先に結合文字「＆」が現れるので、取得されるメイト配列１７ｂは、結合文字「＆」より左側の左方配列１１ａである。

図７（ｃ）の［３］の場合のように、ヒット位置１７ａの配列４２３から右方に延長しても、左方に延長しても結合文字「＆」が現れない場合は、サンプル配列１７はペアを構成する相手がいない単独のリード配列である。このような単独のリード配列は信頼性が低いとみなして無視してよい。

あるいは、変異を導入する前の配列４２３をゲノム配列辞書１５に問い合せて、ゲノム内で配列４２３が出現する箇所が一つしかないと確認できた場合に限り、（信頼性は低いが）ＳＮＰを検出できたと判定してもよい。

[0041] Ｓ１３５において、マッピング部２６は、ゲノム配列辞書１５を参照して、Ｓ１３４で取得したメイト配列１７ｂが由来するゲノムＤＮＡ内の位置を特定（ゲノム・マッピング）する。マッピング部２６は、例えば、メイト配列１７ｂの中から短い（例えば、２０塩基程度の）部分配列を切り出して、その部分配列がゲノム内に１か所だけ現れるか否かをゲノム配列辞書１５に問い合せる。

もし、１か所も現れない場合は、マッピング部２６は、その部分配列中にシーケンシング・エラーや多型が含まれていると考えられるため、別の部分配列で再度、問い合わせを行う。

また、もし、複数箇所に現れる場合は、マッピング部２６は、部分配列の長さを増やすか、または、別の部分配列を用いて、再度、問い合わせを行う。こうして、短い部分配列のゲノム内の位置を特定できた場合、それを含む

メイト配列１７ｂのゲノム上に位置も特定できる（換言すると、マッピングに成功する）。一方、位置を特定できない場合は、マッピングに失敗する。

Ｓ１３５でＹｅｓならＳ１３６に進み、Ｎｏなら今回のループを終了して（Ｓ１３９）、次のヒット位置を選択するためにＳ１３１に戻る。

[0042] Ｓ１３６において、サンプル判定部２７は、Ｓ１３２の配列１７ａのサンプルＤＮＡ断片が解析用情報１３で示されるＳＮＰの変異を含むか否かを、以下で説明するように距離が正常（整合）かによって判定する。例えば、サンプル判定部２７は、マッピングが成功したメイト配列１７ｂのマッピング位置と、ＳＮＰ情報１３ａが示すＳＮＰの位置（ヒット位置１７ａ）との距離が、サンプルＤＮＡ断片の長さにほぼ等しければ、サンプル配列１７を構成する左方配列１１ａと右方配列１１ｂとが整合したとして、「ＳＮＰ検出」（つまり、ＳＮＰ情報１３ａに整合）と判定する。

Ｓ１３６でＹｅｓならＳ１３７に進み、ＮｏならＳ１３８に進む。

[0043] Ｓ１３７において、サンプル判定部２７は、「ＳＮＰ検出」と判定されたＳＮＰ情報１３ａの検出数カウンタ値を１つ増やす。

Ｓ１３８において、サンプル判定部２７は、「ＳＮＰ非検出」と判定されたことにより、検出数カウンタ値の増加を行わない。

[0044] Ｓ１３１～Ｓ１３９のループを実行した後、Ｓ１４１において、サンプル判定部２７は、ＳＮＰ情報１３ａと、その検出数カウンタ値（ＳＮＰが検出されたサンプルＤＮＡ断片の個数）とを対応づけた解析用情報１３の判定結果を出力（ユーザに報告）する。

図７（ｄ）のテーブル４６０は、Ｓ１４１で出力される情報の一例である。テーブル４６０には、図７（ａ）のテーブル４００で示される各ＳＮＰごとに、ＳＮＰ位置４２４で変異塩基（ＳＮＰ）が検出されたリード断片の数（変異塩基検出数）、および、標準塩基が検出されたリード断片の数（標準塩基検出数）が、検出数カウンタ値から読み取られて、書き込まれている。

[0045] 図８は、構造変異の解析処理を示すフロー説明図である。図６のＳＮＰの解析処理との差分に着目して、以下で図８を説明する。

Ｓ１２１ｂでは、配列データ解析装置１は、今回の解析対象となる変異を示す解析用情報１３として、ＳＮＰ情報１３ａの代わりに、構造変異情報１３ｂの入力を受け付ける。

図９（ａ）の構造変異情報１３ｂは、テーブル６００の各行に示すように、各変異ごとに、染色体名、染色体上の塩基位置座標、変異のタイプ（挿入か欠失か）、変異長の情報を含む。

テーブル６００の１行目は、構造変異が染色体３番上の６５４３２１塩基目の位置にあり、標準ゲノムから５００個の連続した塩基の並びが失われる欠失が生ずることを示す。

[0046] 図８のＳ１２２ｂでは、クエリ作成部２３は、ゲノム配列辞書１５を参照して、構造変異情報１３ｂが示す構造変異周辺のクエリ配列１６を作成する。

図９（ｂ）の説明欄６２０は、Ｓ１２２ｂのクエリ配列１６の作成方法を示す。横軸４２１は染色体上の塩基位置座標である。まず、ゲノム配列辞書１５を参照して、構造変異が生ずる位置６２４の周辺（例えば、左方と右方に数十塩基程度離れた位置）の短い（例えば２０塩基程度の）塩基配列６２２と６２３を求め、これらをそれぞれクエリ配列１６とする。つまり、クエリ配列１６は構造変異が生ずる位置６２４の（左方または右方の）近くにある短い配列である。

[0047] 図８のＳ１３３ｂでは、Ｓ１３３の処理（図８では図示省略）の実行後に、構造変異解析の対象外（変異情報が得られない）か否かを判定する。Ｓ１３３ｂでＹｅｓなら今回のループを終了し（Ｓ１３９）、ＮｏならＳ１３４へ進む。以下、Ｓ１３３ｂの判定処理について、図９（ｃ）の説明欄６４０を参照して説明する。

[0048] 図９（ｃ）の［１］の場合のように、ヒット位置１７ａである位置６２４の左方にあるクエリ配列６２２から右方に延長して復元されたサンプル配列１７内に結合文字「＆」が現れる場合は、クエリ配列６２２は左方配列１１ａに含まれる。

よって、結合文字「＆」の右方の配列６４１（右方配列１１ｂ）がメイト配列１７ｂとして復元される。さらに、クエリ配列６２２と配列６４１との間には構造変異が生じる位置６２４が含まれる可能性がある。

一方、図示は省略するが、クエリ配列６２２の左方に延長して復元された文字列内に結合文字「＆」が現れる場合は、構造変異の可能性は無いので、対象外（変異情報が得られない）と判定する。

[0049] 図９（ｃ）の［２］の場合のように、ヒット位置１７ａである位置６２４の右方にあるクエリ配列６２３に対しても、図９（ｃ）の［１］に対して左右を反転して同様な判定を行う。クエリ配列６２３から左方に延長して復元されたサンプル配列１７内に結合文字「＆」が現れる場合は、クエリ配列６２３は右方配列１１ｂに含まれる。

よって、結合文字「＆」の左方の配列６４２（左方配列１１ａ）がメイト配列１７ｂとして復元される。さらに、クエリ配列６２３と配列６４２との間には構造変異が生じる位置６２４が含まれる可能性がある。

一方、図示は省略するが、クエリ配列６２３の右方に延長して復元された文字列内に結合文字「＆」が現れる場合は、構造変異の可能性は無いので、対象外（変異情報が得られない）と判定する。

以上、Ｓ１３３ｂの判定処理について、説明した。

[0050] Ｓ１３６の代わりに行われるＳ１３６ｂでは、サンプル判定部２７は、メイト配列１７ｂと、その抽出元であるサンプル配列１７におけるペアのリード配列とのペア間距離が、サンプルＤＮＡ断片の長さとほぼ等しい場合は、距離が正常値であると評価する（Ｓ１３６ｂ，Ｙｅｓ）。

Ｓ１３６ｂでＮｏなら、構造変異を検出したので、サンプル判定部２７は、該当する構造変異情報１３ｂの「変異有り検出数」のカウンタ値を１つ増やす（Ｓ１３７ｂ）。

ここで、サンプル判定部２７は、「欠失」の変異タイプと、「挿入」の変異タイプとで別々に検出数を集計してもよい。そのため、ペア間の距離がサンプルＤＮＡ断片の長さから予想されるよりも長い場合は（差分の長さに対

応する）欠失が生じていると判定し、逆に、ペア間の距離がサンプルＤＮＡ断片の長さから予想されるよりも短い場合は（差分の長さに対応する）挿入が生じていると判定してもよい。

Ｓ１３６ｂでＹｅｓなら、構造変異は非検出なので、サンプル判定部２７は、該当する構造変異情報１３ｂの「変異無し検出数」のカウンタ値を１つ増やす（Ｓ１３８ｂ）。

[0051] Ｓ１４１ｂでは、配列データ解析装置１は、Ｓ１３７ｂおよびＳ１３８ｂのカウンタ値を構造変異の検出結果として報告するため、図９（ｄ）のテーブル６６０で示すような情報を出力する。

このテーブル６６０には、テーブル６００で示される各構造変異ごとに、変異有りと判定されたリード断片の数、および、変異無し（即ち、標準ゲノムと一致する）と判定されたリード断片の数を、変異有り検出数、および、変異無し検出数として報告する。

[0052] 以上説明した本実施形態では、リード辞書作成部２１は、サンプルＤＮＡ断片のペアである左方配列１１ａと右方配列１１ｂとの間を結合文字で連結し、サンプルＤＮＡ断片のペア間を終端文字で結合することで、結合文字列を作成する。そして、リード辞書作成部２１は、結合文字列をＢＷ変換してからWavelet Tree化したリード配列辞書１４を作成する。

サンプル復元部２５は、リード配列辞書１４内のクエリ配列１６のヒット位置１７ａから、そのヒット位置１７ａを含むサンプルＤＮＡ断片のペア（メイト配列１７ｂ）を復元するときに、リード配列辞書１４内に埋め込まれた結合文字を手がかりに、メイト配列１７ｂを復元することができる。よって、サンプルＤＮＡ断片ごとの識別子は不要となるため、任意のリード配列に対してペアをなすメイト配列１７ｂを効率良く計算できる。

[0053] なお、サンプルＤＮＡ断片ごとの識別子を用いないことにより、以下に示すように、配列データ解析装置１が負担するデータ量および処理量を削減することができる。

比較例として、図４の符号３６１の「＠ｓｅｑ１」のようにサンプルＤＮ

Ａ断片ごとの識別子を用いる場合、サンプルＤＮＡ断片から読み取られるリード配列集合１１が左方配列１１ａと右方配列１１ｂとを合わせて１０億本存在する場合、５億種類の識別子が必要となる。１つの識別子当たり４バイトのデータサイズとすると、全識別子データのサイズは４ギガバイトを要する。さらに、サンプルＤＮＡ断片ごとの識別子を用いる場合には、ヒットしたリード配列の識別子を検索キーとして、そのペアをなすメイト配列１７ｂを検索するための負荷がかかってしまう。

[0054] 一方、本実施形態では、リード配列辞書１４内に埋め込む制御用文字（結合文字、終端文字）のデータサイズは、１つのサンプルＤＮＡ断片あたり１ビットで済むため、全サンプルＤＮＡ断片の合計で１０億ビット（＝０．１２５ギガバイト）となり、識別子を用いた場合のデータサイズの３２分の１程度で済む。さらに、ヒットしたリード配列と、そのペアをなすメイト配列１７ｂとは結合文字をはさんでリード配列辞書１４内に隣接して配置されているため、メイト配列１７ｂの検索負荷を低く抑えることができる。

[0055] なお、本発明は前記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、前記した実施例は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。

また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。

また、各実施例の構成の一部について、他の構成の追加・削除・置換をすることが可能である。また、上記の各構成、機能、処理部、処理手段などは、それらの一部または全部を、例えば集積回路で設計するなどによりハードウェアで実現してもよい。

また、前記の各構成、機能などは、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。

[0056] 各機能を実現するプログラム、テーブル、ファイルなどの情報は、メモリや、ハードディスク、ＳＳＤ（Solid State Drive）などの記録装置、または、ＩＣ（Integrated Circuit）カード、ＳＤカード、ＤＶＤ（Digital Versatile Disc）などの記録媒体に置くことができる。

また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線や情報線を示しているとは限らない。実際にはほとんど全ての構成が相互に接続されていると考えてもよい。

符号の説明

[0057]

１	配列データ解析装置
８	ゲノムサーバ
９	シーケンサ
１１	リード配列集合
１１ａ	左方配列
１１ｂ	右方配列
１２	参照ゲノム配列
１３	解析用情報
１３ａ	ＳＮＰ情報
１３ｂ	構造変異情報
１４	リード配列辞書
１５	ゲノム配列辞書
１６	クエリ配列
１７	サンプル配列
１７ａ	ヒット位置
１７ｂ	メイト配列
２１	リード辞書作成部
２２	ゲノム辞書作成部
２３	クエリ作成部
２４	クエリ検索部

２５　　サンプル復元部

２６　　マッピング部

２７　　サンプル判定部

請求の範囲

[請求項1] サンプルＤＮＡ断片の両端からそれぞれシーケンシングされたペアである左方配列と右方配列との間を結合文字で連結した結合文字列を前記サンプルＤＮＡ断片ごとに作成し、各前記サンプルＤＮＡ断片の前記結合文字列を終端文字で結合した文字列をもとにリード配列辞書を作成するリード辞書作成部と、

ゲノムＤＮＡの塩基文字列から生成されるクエリ配列が出現する前記リード配列辞書内の塩基文字座標であるヒット位置を検索するクエリ検索部と、

前記リード配列辞書内の前記ヒット位置を起点として、その周囲に位置する前記終端文字が出現するまでの文字列をサンプル配列として抽出し、

前記サンプル配列内の前記ヒット位置を起点として、その周囲に位置する前記結合文字を検査し、検査した前記結合文字から前記ヒット位置が存在しない側の前記終端文字が出現するまでの前記左方配列または前記右方配列をメイト配列として抽出するサンプル復元部と、

前記メイト配列が出現する前記ゲノムＤＮＡ内の塩基文字列の塩基文字座標を検索するマッピング部と、を有することを特徴とする

配列データ解析装置。

[請求項2] 前記配列データ解析装置は、さらに、クエリ作成部と、サンプル判定部とを備えており、

前記クエリ作成部は、前記ゲノムＤＮＡの塩基文字列のうちの所定位置の塩基文字を別の塩基文字に変異させ、その変異させた塩基文字とその周囲に位置する前記ゲノムＤＮＡの塩基文字列とを含めた前記クエリ配列を作成し、

前記サンプル判定部は、前記マッピング部により前記メイト配列の塩基文字座標が特定できたときに、前記ヒット位置と、前記メイト配列の塩基文字座標との文字間隔から特定される前記サンプルＤＮＡ断

片の長さをもとに、前記サンプルＤＮＡ断片内のＳＮＰ（Single Nucleotide Polymorphism）を検出することを特徴とする

請求項１に記載の配列データ解析装置。

[請求項3] 前記配列データ解析装置は、さらに、クエリ作成部と、サンプル判定部とを備えており、

前記クエリ作成部は、前記ゲノムＤＮＡの塩基文字列のうちの所定位置周囲に位置する前記ゲノムＤＮＡの塩基文字列から前記クエリ配列を作成し、

前記サンプル判定部は、前記マッピング部により前記メイト配列の塩基文字座標が特定できたときに、前記ヒット位置と、前記メイト配列の塩基文字座標との文字間隔から特定される前記サンプルＤＮＡ断片の長さをもとに、前記サンプルＤＮＡ断片内の構造変異を検出することを特徴とする

請求項１に記載の配列データ解析装置。

[請求項4] 前記リード辞書作成部は、各前記サンプルＤＮＡ断片の前記結合文字列を前記終端文字で結合した文字列に対してＢＷ（Burrows-Wheeler）変換したＢＷ文字列を作成し、その前記ＢＷ文字列をWavelet Tree形式に変換することで、前記リード配列辞書を作成することを特徴とする

請求項１ないし請求項３のいずれか１項に記載の配列データ解析装置。

[請求項5] 前記リード辞書作成部は、各前記サンプルＤＮＡ断片の長さに対応した複数種類の前記結合文字を用いて、前記結合文字列を作成することを特徴とする

請求項１ないし請求項３のいずれか１項に記載の配列データ解析装置。

[請求項6] 前記サンプル復元部は、前記ヒット位置の周囲に前記結合文字が存在しない場合、その前記ヒット位置について前記マッピング部の処理

対象から除外することを特徴とする

請求項１ないし請求項３のいずれか１項に記載の配列データ解析装置。

[請求項7] 請求項１ないし請求項６のいずれか１項に記載の配列データ解析装置と、前記サンプルＤＮＡ断片を両端からそれぞれシーケンシングし、その結果である前記左方配列と前記右方配列とのペアのリード配列を、前記配列データ解析装置に送信するシーケンサとを含めて構成される

ＤＮＡ解析システム。

[請求項8] 配列データ解析装置は、リード辞書作成部と、クエリ検索部と、サンプル復元部と、マッピング部とを有しており、

前記リード辞書作成部は、サンプルＤＮＡ断片の両端からそれぞれシーケンシングされたペアである左方配列と右方配列との間を結合文字で連結した結合文字列を前記サンプルＤＮＡ断片ごとに作成し、各前記サンプルＤＮＡ断片の前記結合文字列を終端文字で結合した文字列をもとにリード配列辞書を作成し、

前記クエリ検索部は、ゲノムＤＮＡの塩基文字列から生成されるクエリ配列が出現する前記リード配列辞書内の塩基文字座標であるヒット位置を検索し、

前記サンプル復元部は、

前記リード配列辞書内の前記ヒット位置を起点として、その周囲に位置する前記終端文字が出現するまでの文字列をサンプル配列として抽出し、

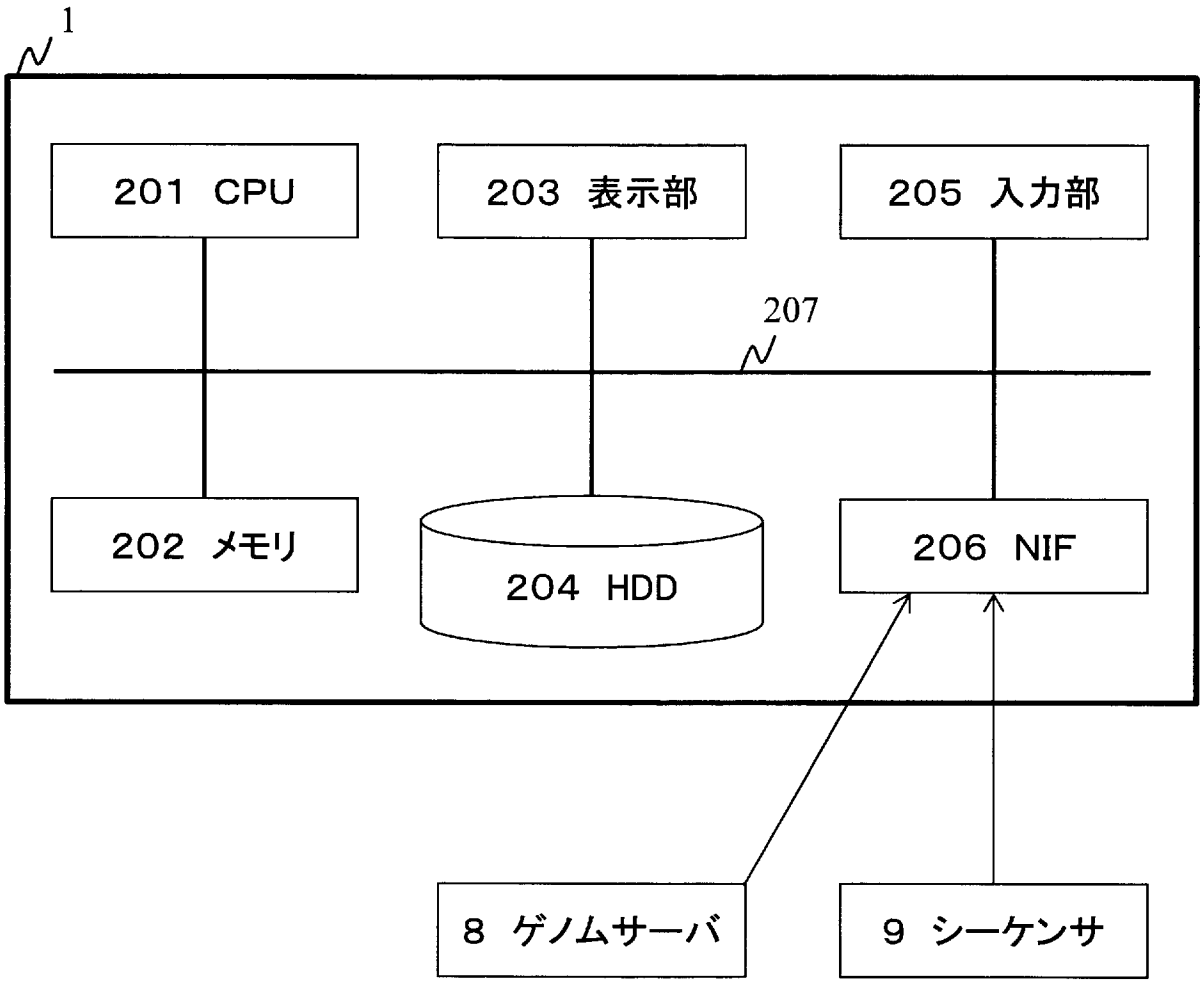
前記サンプル配列内の前記ヒット位置を起点として、その周囲に位置する前記結合文字を検査し、検査した前記結合文字から前記ヒット位置が存在しない側の前記終端文字が出現するまでの前記左方配列または前記右方配列をメイト配列として抽出し、

前記マッピング部は、前記メイト配列が出現する前記ゲノムＤＮＡ

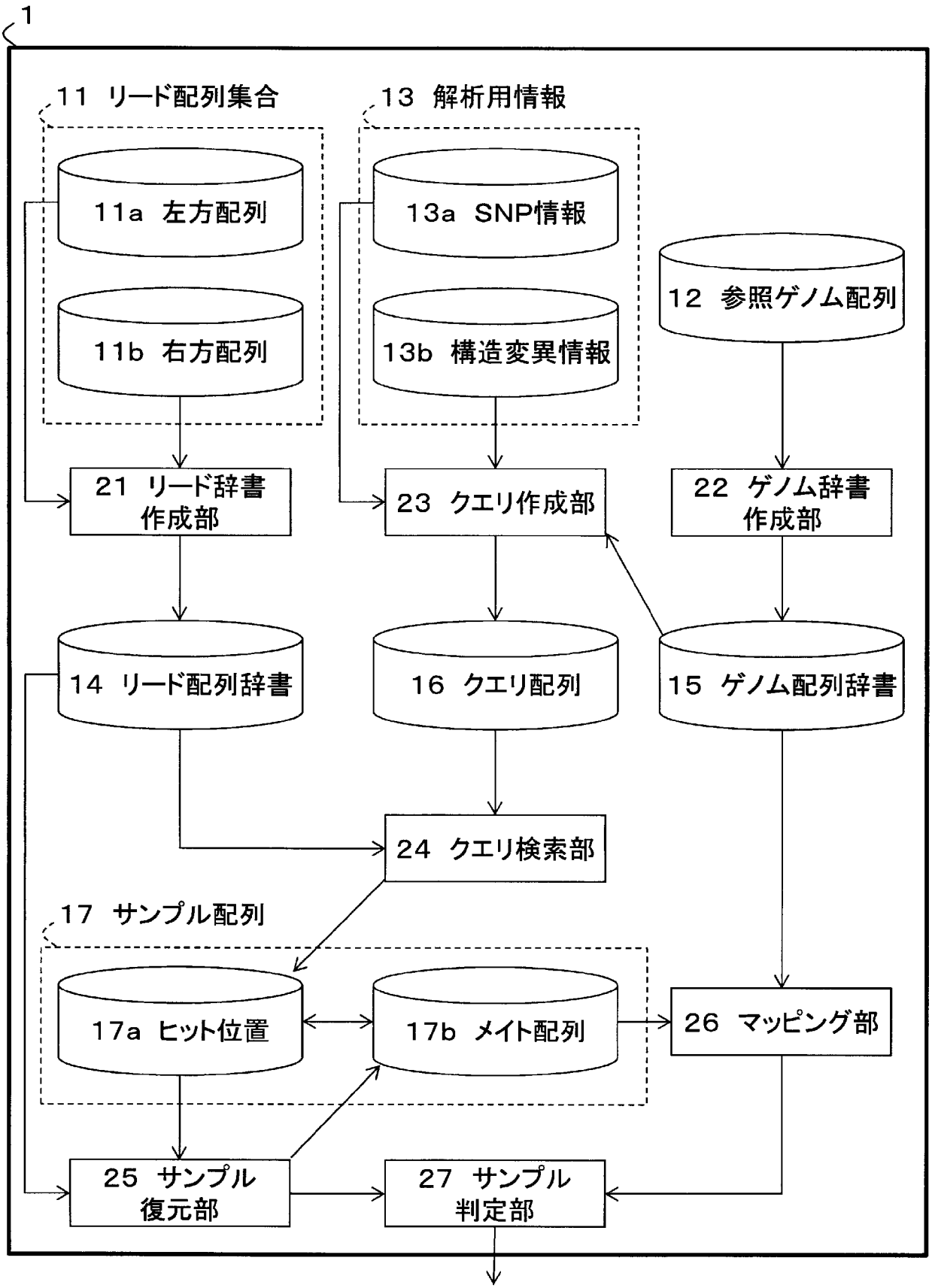
内の塩基文字列の塩基文字座標を検索することを特徴とする

配列データ解析方法。

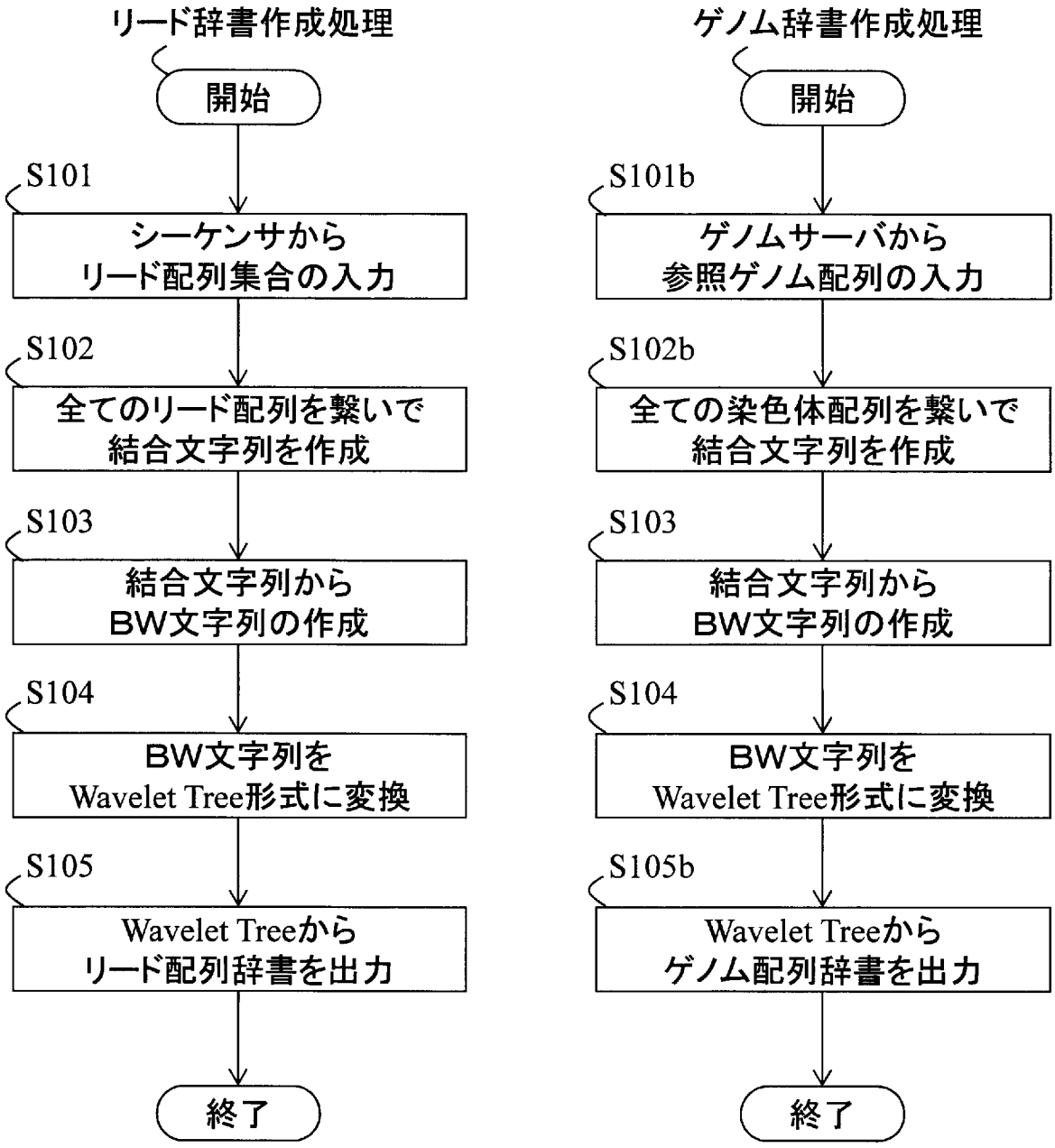
[図1]


1
201 CPU
203 表示部
205 入力部
207
202 メモリ
204 HDD
206 NIF
8 ゲノムサーバ
9 シーケンサ

[図2]

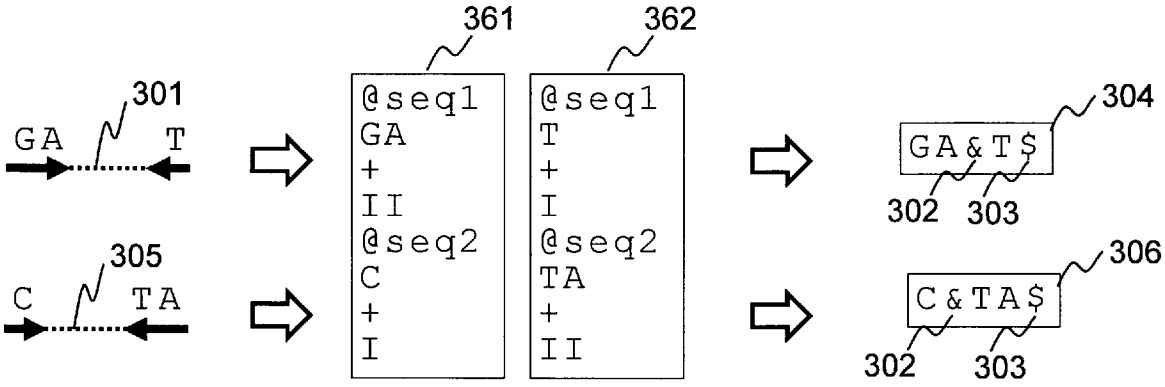

1
11 リード配列集合
11a 左方配列
11b 右方配列
13 解析用情報
13a SNP情報
13b 構造変異情報
12 参照ゲノム配列
21 リード辞書作成部
23 クエリ作成部
22 ゲノム辞書作成部
14 リード配列辞書
16 クエリ配列
15 ゲノム配列辞書
24 クエリ検索部
17 サンプル配列
17a ヒット位置
17b メイト配列
26 マッピング部
25 サンプル復元部
27 サンプル判定部

[図3]

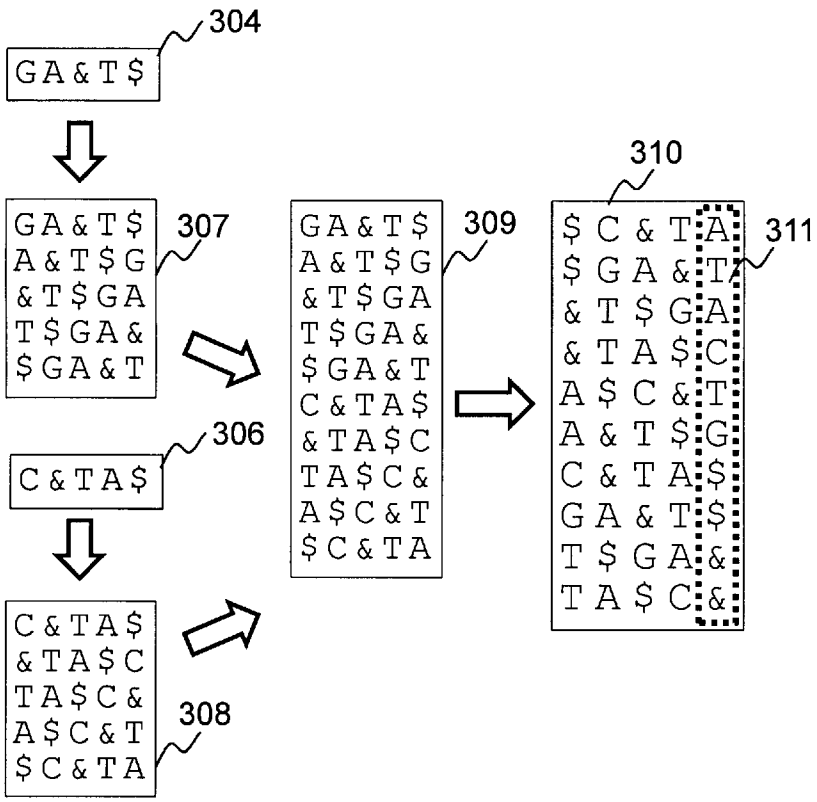

リード辞書作成処理
開始
S101
シーケンサから
リード配列集合の入力
S102
全てのリード配列を繋いで
結合文字列を作成
S103
結合文字列から
BW文字列の作成
S104
BW文字列を
Wavelet Tree形式に変換
S105
Wavelet Treeから
リード配列辞書を出力
終了
ゲノム辞書作成処理
開始
S101b
ゲノムサーバから
参照ゲノム配列の入力
S102b
全ての染色体配列を繋いで
結合文字列を作成
S103
結合文字列から
BW文字列の作成
S104
BW文字列を
Wavelet Tree形式に変換
S105b
Wavelet Treeから
ゲノム配列辞書を出力
終了

[図4]

(a)

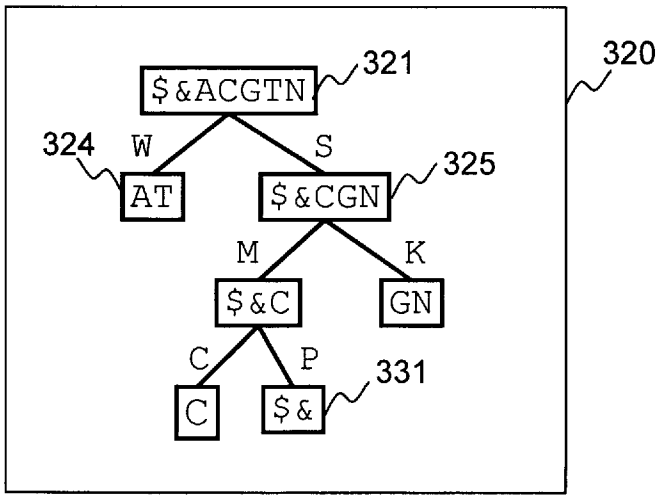

361
362
301
GA
T
@seq1
GA
+
II
@seq2
C
+
I
@seq1
T
+
I
@seq2
TA
+
II
304
GA&T$
302
303
305
C
TA
306
C&TA$
302
303

(b)

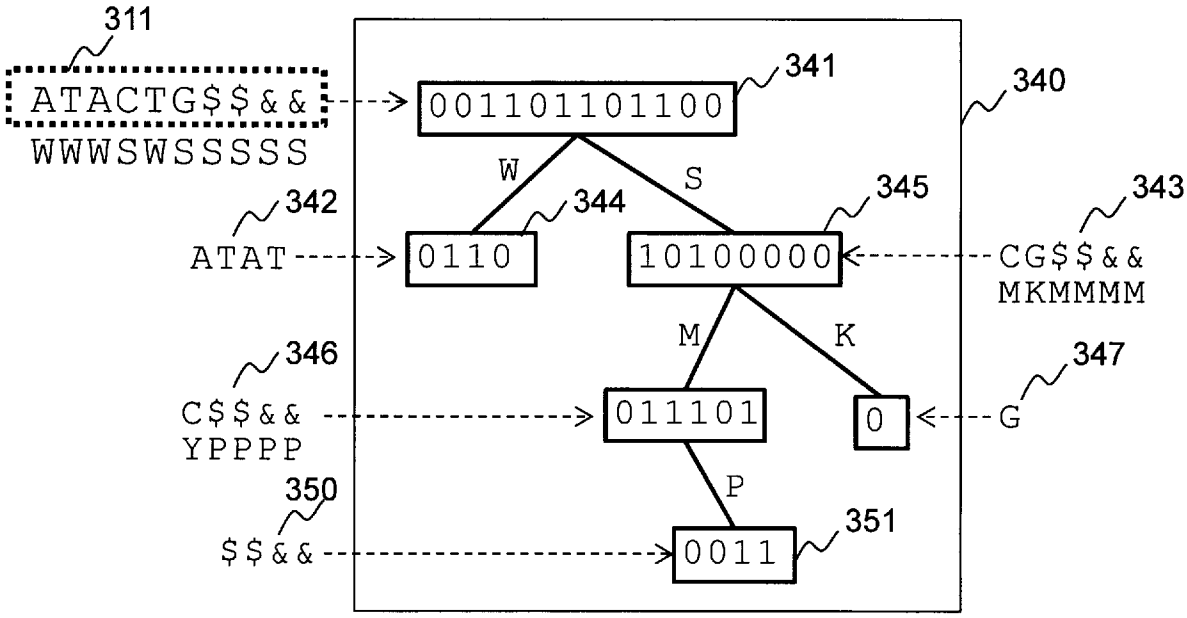

304
GA&T$
GA&T$
A&T$G
&T$GA
T$GA&
$GA&T
307
C&TA$
306
C&TA$
&TA$C
TA$C&
A$C&T
$C&TA
308
GA&T$
A&T$G
&T$GA
T$GA&
$GA&T
C&TA$
&TA$C
TA$C&
A$C&T
$C&TA
309
310
$C&TA
$GA&T
&T$GA
&TA$C
A$C&T
A&T$G
C&TA$
GA&T$
T$GA&
TA$C&
311

[図5]

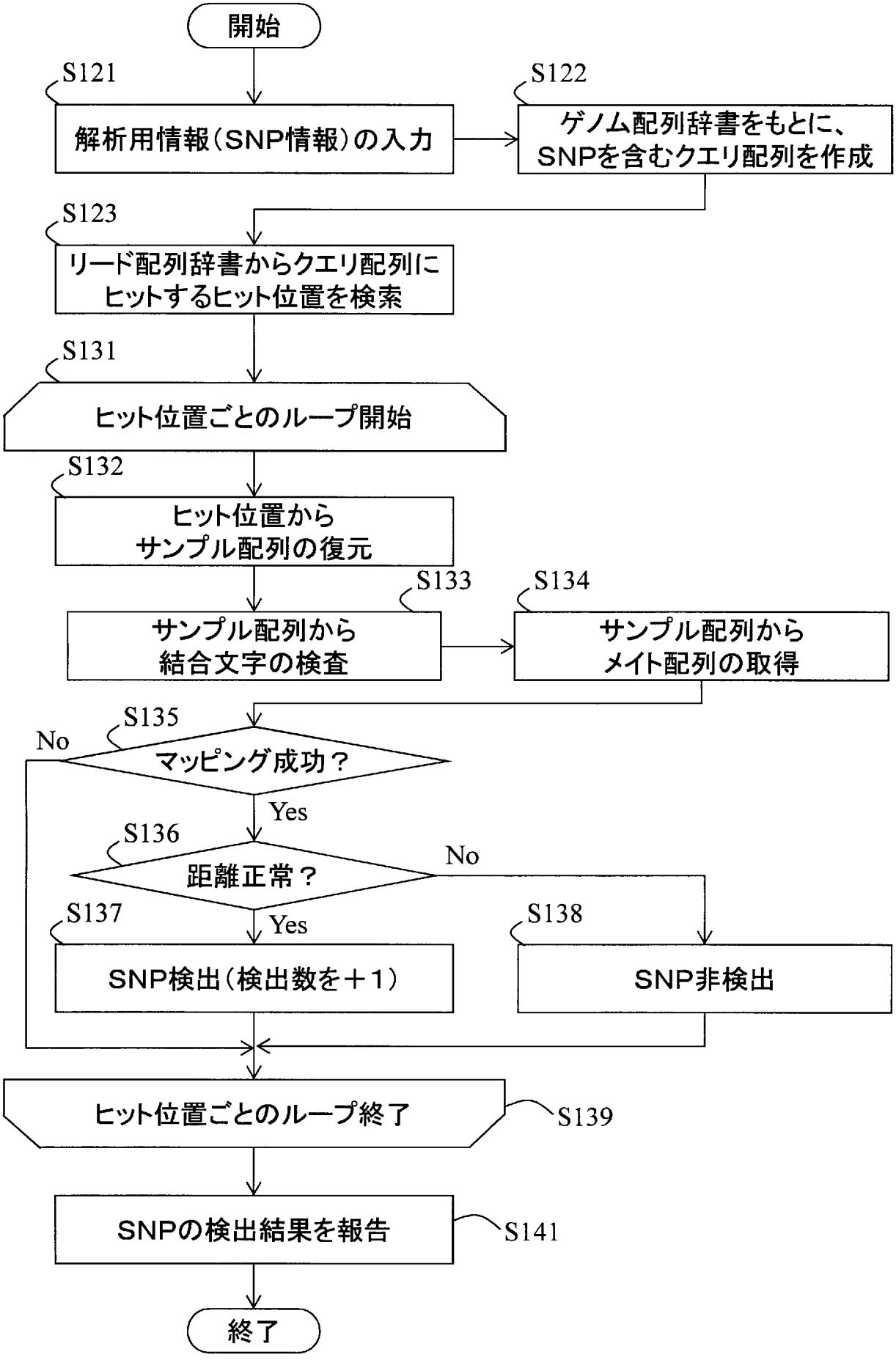
(a)


$&ACGTN
321
320
324
W
S
325
AT
$&CGN
M
K
$&C
GN
C
P
331
C
$&

(b)


311
ATACTG$$&&
WWWSWSSSSS
341
001101101100
340
W
S
342
344
345
343
ATAT
0110
10100000
CG$$&&
MKMMMM
M
K
346
347
C$$&&
YPPPP
011101
0
G
P
350
351
$$&&
0011

[図6]


開始
S121
解析用情報(SNP情報)の入力
S122
ゲノム配列辞書をもとに、
SNPを含むクエリ配列を作成
S123
リード配列辞書からクエリ配列に
ヒットするヒット位置を検索
S131
ヒット位置ごとのループ開始
S132
ヒット位置から
サンプル配列の復元
S133
サンプル配列から
結合文字の検査
S134
サンプル配列から
メイト配列の取得
S135
No
マッピング成功？
Yes
S136
距離正常？
No
S137
Yes
SNP検出(検出数を＋1)
S138
SNP非検出
ヒット位置ごとのループ終了
S139
SNPの検出結果を報告
S141
終了

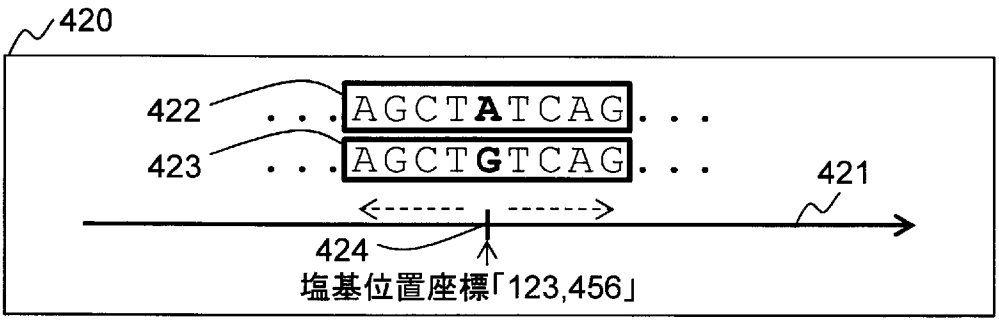
[図7]

(a)

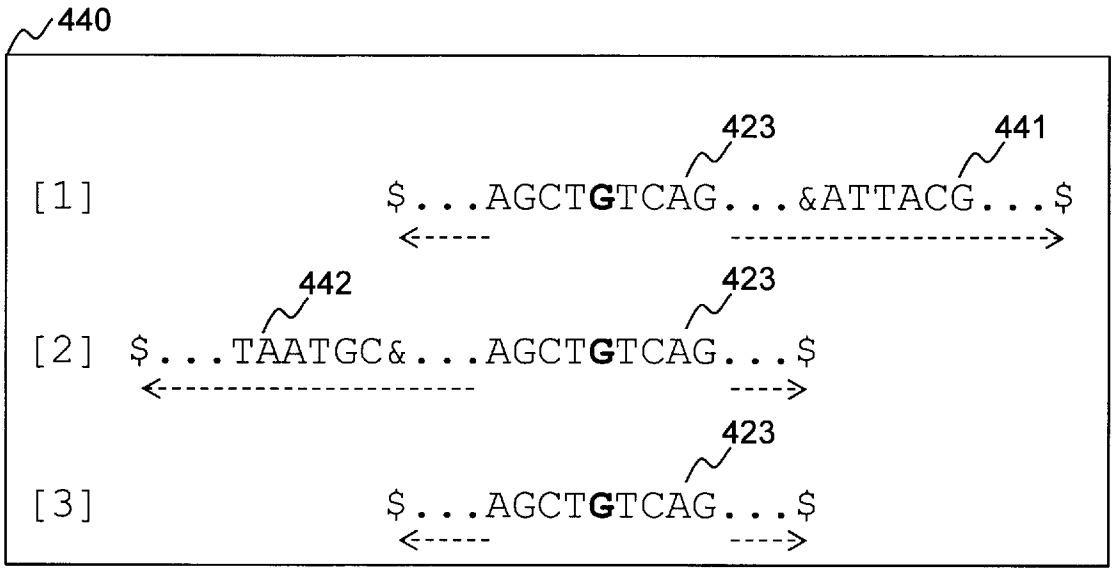
400

染色体名	塩基位置座標	標準塩基種類	変異塩基種類
chr7	123,456	A	G
chrX	345,678	C	T
...			

(b)



(c)

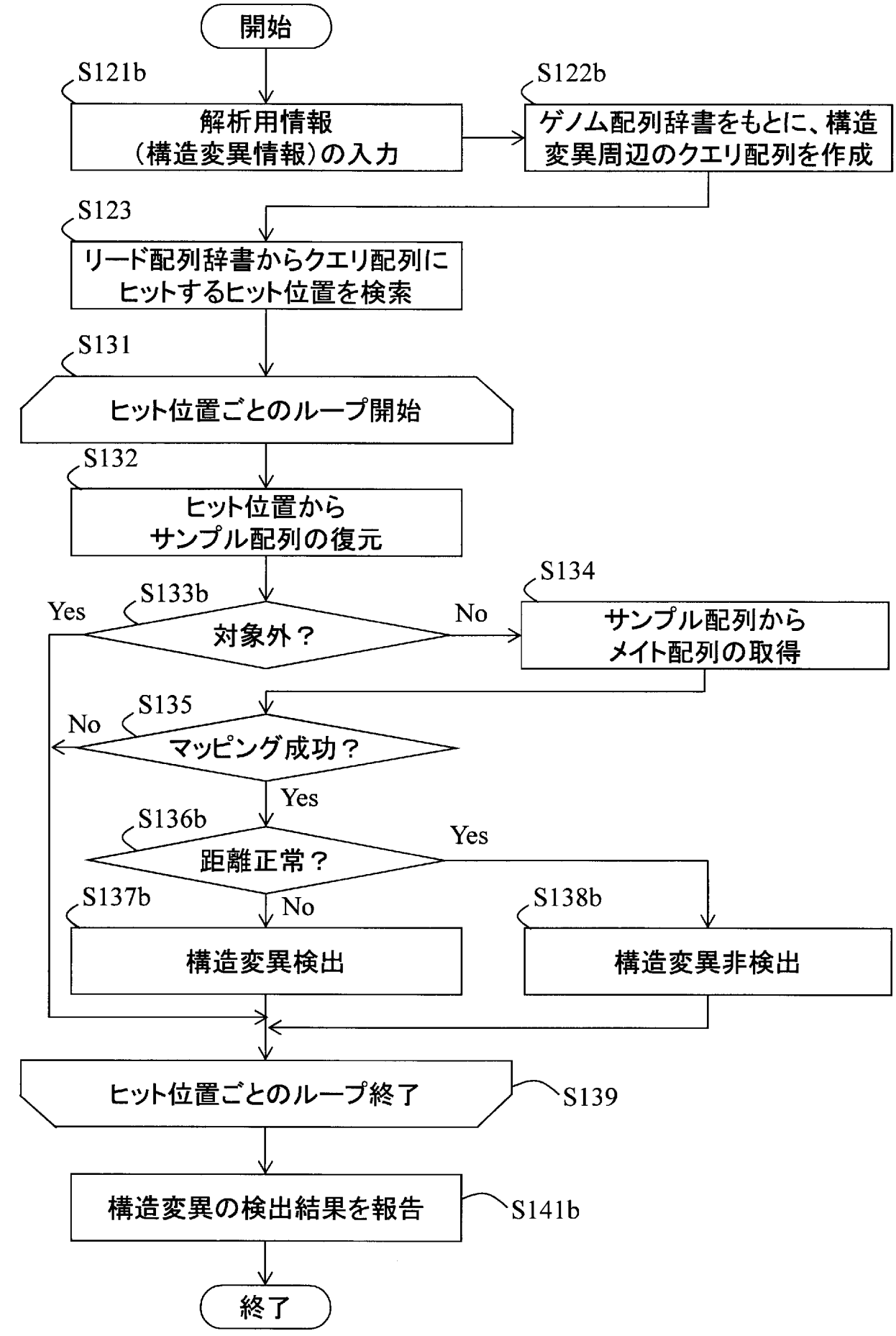


(d)

460

染色体名	塩基位置座標	標準塩基種類	変異塩基種類	標準塩基検出数	変異塩基検出数
chr7	123,456	A	G	23	19
chrX	345,678	C	T	43	0
...					

[図8]


開始
S121b
解析用情報
(構造変異情報)の入力
S122b
ゲノム配列辞書をもとに、構造
変異周辺のクエリ配列を作成
S123
リード配列辞書からクエリ配列に
ヒットするヒット位置を検索
S131
ヒット位置ごとのループ開始
S132
ヒット位置から
サンプル配列の復元
S133b
対象外？
Yes
No
S134
サンプル配列から
メイト配列の取得
S135
マッピング成功？
No
Yes
S136b
距離正常？
Yes
No
S137b
構造変異検出
S138b
構造変異非検出
ヒット位置ごとのループ終了
S139
構造変異の検出結果を報告
S141b
終了

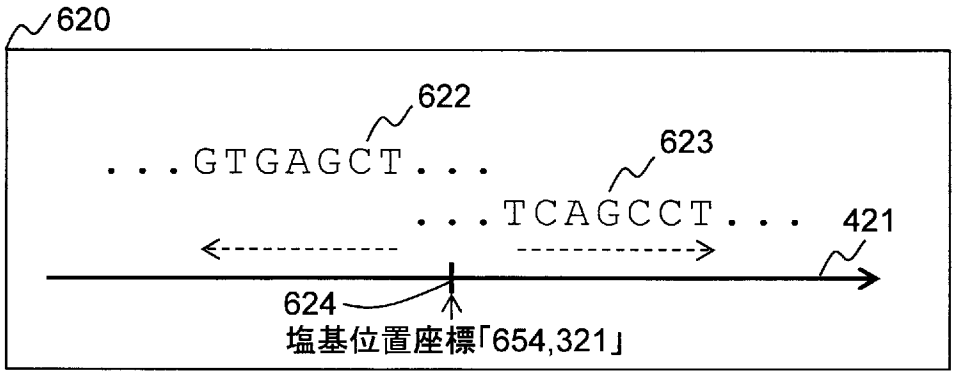
[図9]

(a)

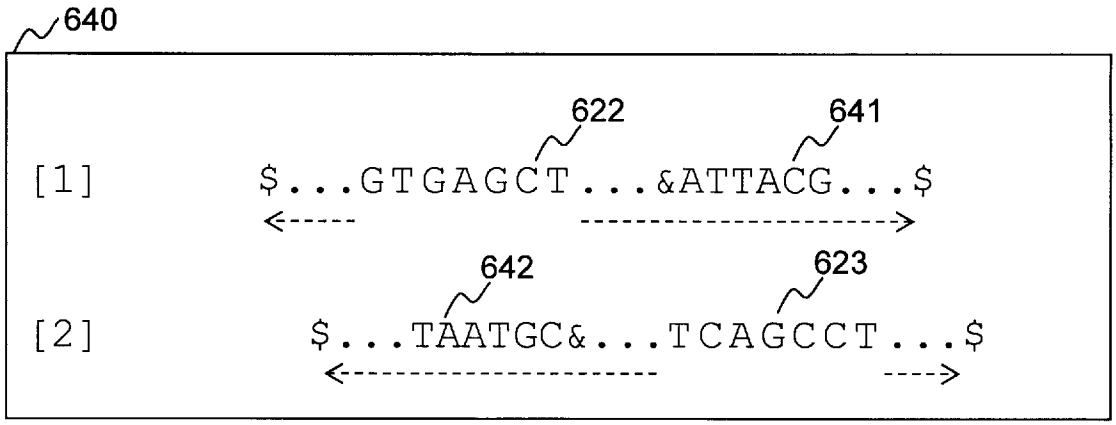
600

染色体名	塩基位置座標	変異タイプ	変異長
chr3	654,321	欠失	500塩基
chr12	111,222	挿入	100塩基
...			

(b)



(c)



(d)

660

染色体名	塩基位置座標	変異タイプ	変異長	変異有り検出数	変異無し検出数
chr3	654,321	欠失	200塩基	31	29
chr12	111,222	挿入	50塩基	57	0
...					

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2015/057348

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/22(2011.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F19/22

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Jitsuyo Shinan Toroku Koho 1996-2015
Kokai Jitsuyo Shinan Koho 1971-2015 Toroku Jitsuyo Shinan Koho 1994-2015

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubMed, JSTPlus/JMEDPlus/JST7580(JDreamIII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Heng LI et al, The Sequence Alignment/Map format and SAMtools, Bioinformatics, 2009.06.08, Vol.25, No.16, pp.2078-2079	1-8
A	JP 2008-547080 A (Agency for Science, Technology and Research), 25 December 2008 (25.12.2008), paragraphs [0007] to [0172] & US 2006/0281097 A1 & WO 2006/135342 A1 & EP 1899860 A1 & CA 2611951 A1 & KR 10-2008-0026153 A & CN 101233509 A & AU 2006258264 A1	1-8
A	Heng LI et al, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 2009.07.15, Vol.25, No.14, pp.1754-1760	1-8

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
22 May 2015 (22.05.15)	02 June 2015 (02.06.15)
Name and mailing address of the ISA/ Japan Patent Office 3-4-3,Kasumigaseki,Chiyoda-ku, Tokyo 100-8915,Japan	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2015/057348

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Koichi KIMURA et al, LOCALIZED SUFFIX ARRAY AND ITS APPLICATIONTO GENOME MAPPING PROBLEMS FOR PAIRED-ENDSHORT READS, Genome Informatics, 2009.10, Vol.23, No.1, pp.60-71	1-8
A	JP 2009-116559 A (Hitachi, Ltd.), 28 May 2009 (28.05.2009), paragraphs [0018] to [0061] (Family: none)	1-8
A	JP 2014-502513 A (Keygene N.V.), 03 February 2014 (03.02.2014), paragraphs [0033] to [0067] & US 2013/0331277 A1 & WO 2012/096579 A2 & EP 2663655 A2 & AU 2012205884 A1 & CA 2823815 A1 & CN 103476946 A & KR 10-2014-0040697 A	1-8

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１５／０５７３４８

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/22(2011.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/22

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１５年
日本国実用新案登録公報	１９９６－２０１５年
日本国登録実用新案公報	１９９４－２０１５年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

PubMed
JSTPlus/JMEDPlus/JST7580(JDreamⅢ)

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	Heng LI et al, The Sequence Alignment/Map format and SAMtools, Bioinformatics, 2009.06.08, Vol.25, No.16, pp.2078-2079	１－８
A	JP 2008-547080 A（エイジェンシー・フォー・サイエンス，テクノロジー・アンド・リサーチ）2008.12.25，［０００７］－［０１７２］ & US 2006/0281097 A1 & WO 2006/135342 A1 & EP 1899860 A1 & CA 2611951 A1 & KR 10-2008-0026153 A & CN 101233509 A & AU 2006258264 A1	１－８

☑　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
２２．０５．２０１５	０２．０６．２０１５

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｌ　３６５８
松野　広一
電話番号　０３－３５８１－１１０１　内線　３５６２

Ｃ（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
Ａ	Heng LI et al, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 2009.07.15, Vol.25, No.14, pp.1754-1760	１－８
Ａ	Koichi KIMURA et al, LOCALIZED SUFFIX ARRAY AND ITS APPLICATIONTO GENOME MAPPING PROBLEMS FOR PAIRED-ENDSHORT READS, Genome Informatics, 2009.10, Vol.23, No.1, pp.60-71	１－８
Ａ	JP 2009-116559 A（株式会社日立製作所）2009.05.28,［００１８］－［００６１］（ファミリーなし）	１－８
Ａ	JP 2014-502513 A（キージーン・エン・フェー）2014.02.03,［００３３］－［００６７］ & US 2013/0331277 A1 & WO 2012/096579 A2 & EP 2663655 A2 & AU 2012205884 A1 & CA 2823815 A1 & CN 103476946 A & KR 10-2014-0040697 A	１－８