

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7146839号
(P7146839)

(45)発行日 令和4年10月4日(2022.10.4)

(24)登録日 令和4年9月26日(2022.9.26)

(51)国際特許分類	F I		
C 4 0 B 50/06 (2006.01)	C 4 0 B	50/06	
C 1 2 Q 1/44 (2006.01)	C 1 2 Q	1/44	
C 1 2 Q 1/6869(2018.01)	C 1 2 Q	1/6869	Z
C 1 2 Q 1/686(2018.01)	C 1 2 Q	1/686	Z
C 1 2 N 9/16 (2006.01)	C 1 2 N	9/16	Z
請求項の数 27 (全61頁) 最終頁に続く			

(21)出願番号	特願2020-77622(P2020-77622)	(73)特許権者	000000918 花王株式会社 東京都中央区日本橋茅場町1丁目14番 10号
(22)出願日	令和2年4月24日(2020.4.24)	(74)代理人	110000084 特許業務法人アルガ特許事務所
(65)公開番号	特開2021-80241(P2021-80241A)	(72)発明者	大坪 裕紀 栃木県芳賀郡市貝町赤羽2606 花王 株式会社研究所内
(43)公開日	令和3年5月27日(2021.5.27)	(72)発明者	松村 奨士 栃木県芳賀郡市貝町赤羽2606 花王 株式会社研究所内
審査請求日	令和3年11月29日(2021.11.29)	審査官	野村 英雄
(31)優先権主張番号	特願2019-207386(P2019-207386)		
(32)優先日	令和1年11月15日(2019.11.15)		
(33)優先権主張国・地域又は機関	日本国(JP)		
早期審査対象出願			
最終頁に続く			

(54)【発明の名称】 シーケンシング用ライブラリの調製方法

(57)【特許請求の範囲】

【請求項1】

ゲノムDNAの変異を検出するためのシーケンシング用ライブラリの調製方法であって、ゲノムDNAを断片化すること；及び、

調製したゲノムDNAの断片を1本鎖特異的ヌクレアーゼで処理し、該断片から1本鎖部分を除去すること、
該1本鎖特異的ヌクレアーゼでの処理の後、該断片の両端へシーケンシングに必要な標識配列を付加すること、
を含み、

該ゲノムDNAが、生細胞から抽出したDNA、凍結細胞から抽出したDNA、又はそれらのDNAの保存サンプルである、
方法。

【請求項2】

前記ゲノムDNAの変異の検出が、試験物質の遺伝毒性、生殖発生毒性、もしくはその他毒性の評価のため、ゲノムDNAに対する経時変化、生活環境、もしくは遺伝的要素の影響の評価のため、又は、培養細胞の品質評価のために行われる、請求項1記載の方法。

【請求項3】

前記ゲノムDNAの変異の検出が試験物質の毒性の評価のために行われ、該ゲノムDNAが該試験物質に暴露した細胞のゲノムDNAであるか、又は、

前記ゲノムDNAの変異の検出がゲノムDNAに対する経時変化の影響の評価のために行

われ、該ゲノムDNAが経時変化した細胞のゲノムDNAであるか、又は、前記ゲノムDNAの変異の検出が培養細胞の品質評価のために行われ、該ゲノムDNAが一定期間培養した培養細胞のゲノムDNAと、同じ種類の培養細胞であって遺伝情報既知の細胞のゲノムDNAである、

請求項1記載の方法。

【請求項4】

前記1本鎖特異的ヌクレアーゼが1本鎖特異的エンドヌクレアーゼ、1本鎖特異的エキソヌクレアーゼ、又はそれらの組み合わせである、請求項1～3のいずれか1項記載の方法。

【請求項5】

前記1本鎖特異的ヌクレアーゼでの処理が、前記ゲノムDNAの断片を前記1本鎖特異的エンドヌクレアーゼで処理した後に、さらに前記1本鎖特異的エキソヌクレアーゼで処理することを含む、請求項4記載の方法。

【請求項6】

前記1本鎖特異的エンドヌクレアーゼがS1 nuclease又はMung Bean Nucleaseである、請求項4又は5記載の方法。

【請求項7】

前記1本鎖特異的エンドヌクレアーゼが、前記ゲノムDNAの断片1ngあたり0.02U/ng以上のS1 nucleaseである、請求項4又は5記載の方法。

【請求項8】

前記1本鎖特異的エンドヌクレアーゼが、前記ゲノムDNAの断片1ngあたり0.02U/ng以上のMung Bean Nucleaseである、請求項4又は5記載の方法。

【請求項9】

前記1本鎖特異的エキソヌクレアーゼがRecJfである、請求項4又は5記載の方法。

【請求項10】

前記1本鎖特異的エキソヌクレアーゼが、前記ゲノムDNAの断片1ngあたり0.10U/ng以上のRecJfである、請求項4又は5記載の方法。

【請求項11】

前記標識配列を付加したゲノムDNAの断片を増幅することをさらに含む、請求項1～10のいずれか1項記載の方法。

【請求項12】

前記増幅がPCRであり、前記1本鎖特異的ヌクレアーゼがS1 nucleaseであり、かつ

前記ゲノムDNAの断片1ngあたりの該S1 nucleaseのユニット数(U/ng)が0.05U/ng以下のとき、該PCRにおける前記ゲノムDNA 1Mbpあたりの初期DNA量が250amol以下であるか、又は

前記ゲノムDNAの断片1ngあたりの該S1 nucleaseのユニット数(U/ng)が0.05U/ngより大きいとき、下記式で算出される指標が60以下である：

指標 = 該PCRにおける初期DNA量 (amol / Mbp ゲノムDNA) × 3^{log S1 nuclease (U/ng)}

(式中、S1 nuclease (U/ng) > 0.05、logは常用対数である)、請求項11記載の方法。

【請求項13】

前記増幅がPCRであり、前記1本鎖特異的ヌクレアーゼがMung Bean Nucleaseであり、かつ

前記ゲノムDNAの断片1ngあたりの該Mung Bean Nucleaseのユニット数(U/ng)が0.05U/ng以下のとき、該PCRにおける前記ゲノムDNA 1Mbpあたりの初期DNA量が250amol以下であるか、又は

前記ゲノムDNAの断片1ngあたりの該Mung Bean Nucleaseのユニ

10

20

30

40

50

ット数 (U/ng) が $0.05 U/ng$ より大きいとき、下記式で算出される指標が 60 以下である：

指標 = 該 PCR における初期 DNA 量 ($amol / Mbp$ ゲノム DNA) $\times 3 \log Mung$ Bean Nuclease (U/ng)

(式中、 $Mung Bean Nuclease (U/ng) > 0.05$ 、 \log は常用対数である)、

請求項 1.1 記載の方法。

【請求項 1.4】

請求項 1 ~ 1.3 のいずれか 1 項記載の方法で調製されたシーケンシング用ライブラリをシーケンシングすることを含む、ゲノム DNA の変異を検出するための DNA のシーケンシング方法。

10

【請求項 1.5】

前記シーケンシング方法が、以下：

(1) 前記ライブラリをシーケンシングし、該ライブラリに含まれる複数の増幅断片の各々について 1 つ以上のリード配列を作成し、該複数の増幅断片についての複数のリード配列を得ること；

(2) 得られた複数のリード配列の中から、該ライブラリの調製に用いた ゲノム DNA 上の同一領域の配列情報を有するリード配列を集めてグループ化することにより、リード配列のグループを 1 つ以上作成すること；及び、

(3) 該リード配列のグループに含まれるリード配列の間で配列情報のコンセンサスを取ること、

20

を含む、請求項 1.4 記載の方法。

【請求項 1.6】

前記 (1) が、前記 ゲノム DNA の断片を構成する 2 本の相補鎖の各々に由来する増幅断片に対して 1 つ以上のリード配列を作成することを含む、請求項 1.5 記載の方法。

【請求項 1.7】

前記 (2) が、参照配列上の同一の位置にマッピングされるリード配列を同じグループに分けることを含む、請求項 1.6 記載の方法。

【請求項 1.8】

前記 (3) が、前記リード配列のグループの中から、前記 ゲノム DNA 断片の 2 本の相補鎖の各々に由来するリード配列を少なくとも 1 つずつ集め、集めたリード配列の間で配列情報のコンセンサスを取ることを含む、請求項 1.7 記載の方法。

30

【請求項 1.9】

前記 (1) において、前記複数のリード配列が、以下からなるリード配列のペアを複数個含む：

リード 1：前記増幅断片を構成する 2 本の相補鎖のうち一方の鎖の配列を 5' 末端側から 3' 側へ読んだ配列に相当する配列情報を含むリード配列、

リード 2：該一方の鎖の配列を 3' 末端側から 5' 側へ読んだ配列に相当する配列情報を含むリード配列、

前記 (2) が、得られたリード配列のペアの中から、該 ゲノム DNA 上の同一領域の配列情報を有するリード配列のペアを集めてグループ化することにより、リード配列のペアのグループを 1 つ以上作成することを含む、

40

前記 (3) が、該リード配列のペアのグループに含まれるリード配列の間で配列情報のコンセンサスを取ることを含む、

請求項 1.5 記載の方法。

【請求項 2.0】

前記 (1) が、前記 ゲノム DNA の断片を構成する 2 本の相補鎖の各々に由来する増幅断片に対して 1 つ以上の前記リード配列のペアを作成することを含む、請求項 1.9 記載の方法。

【請求項 2.1】

50

前記(2)が、前記リード配列のペアのリード1とリード2を参照配列に対してマッピングし、リード1の先頭とリード2の先頭とに挟まれる該参照配列の領域が同一であるリード配列のペアを同じグループに分けることを含む、請求項20記載の方法。

【請求項22】

前記(2)が、前記リード配列のペアに含まれる一方のリード配列の先頭が前記参照配列上の同じ位置に位置するリード配列のペアを集め、次いで集めたリード配列のペアの中から、該リード配列のペアに含まれるもう一方のリード配列の先頭が該参照配列上の同じ位置に位置するリード配列のペアを集めて、集めたリード配列のペアを同じグループに分けることを含む、請求項20記載の方法。

【請求項23】

前記(3)が、前記リード配列のペアのグループの中から、前記ゲノムDNA断片の2本の相補鎖の各々に由来するリード配列のペアを少なくとも1組ずつ集め、集めたリード配列のペアに含まれるリード配列の間で配列情報のコンセンサスを取ることを含む、請求項21又は22記載の方法。

【請求項24】

請求項1～13のいずれか1項記載の方法によりシーケンシング用ライブラリを調製すること；及び、

該シーケンシング用ライブラリをシーケンシングすること、
を含む、ゲノムDNAの変異を検出する方法。

【請求項25】

試験物質の遺伝毒性、生殖発生毒性、もしくはその他毒性の評価のため、ゲノムDNAに対する経時変化、生活環境、もしくは遺伝的要素の影響の評価のため、又は、培養細胞の品質評価のためにゲノムDNAの変異を検出する方法である、請求項24記載の方法。

【請求項26】

試験物質の毒性の評価のためにゲノムDNAの変異を検出する方法であって、該ゲノムDNAが該試験物質に暴露した細胞のゲノムDNAであるか、又は、ゲノムDNAに対する経時変化の影響の評価のためにゲノムDNAの変異を検出する方法であって、該ゲノムDNAが経時変化した細胞のゲノムDNAであるか、又は、培養細胞の品質評価のためにゲノムDNAの変異を検出する方法であって、該ゲノムDNAが一定期間培養した培養細胞のゲノムDNAと、同じ種類の培養細胞であって遺伝情報既知の細胞のゲノムDNAである、
請求項24記載の方法。

【請求項27】

前記シーケンシングが請求項15～23のいずれか1項記載の方法により行われる、請求項24～26のいずれか1項記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、シーケンシング用ライブラリの調製方法に関する。

【背景技術】

【0002】

次世代シーケンシング(NGS)技術は、近年著しい発展を遂げ、がん細胞のゲノム変異解析などにおいて盛んに活用されて新しい知見を生み出している。NGSのためのシーケンサーとしては、イルミナ社のHiSeqやMiSeqなどのシーケンサーが多用されている。これらのシーケンサーでは、解析対象となる細胞や組織から抽出したサンプルDNAを数百bpの長さで断片化した後、該断片の突出末端を平滑化し、次いで両末端にシーケンシングアダプターを結合してライブラリDNAを調製し、これをシーケンシングする。該末端の平滑化では、一般に、T4 DNAポリメラーゼ等の酵素により3'側突出末端が除去され、一方、5'側突出末端は、対となる鎖が合成されて修復される(末端修復)

。

10

20

30

40

50

【 0 0 0 3 】

シーケンシングからのデータには通常、サンプルDNAの性状やライブラリ調製の工程に起因するエラーが含まれ、これらは解析対象の細胞集団内の変異等の正確な同定への障害となる。例えば、サンプルDNA自体が保存中の損傷などにより1本鎖になることで、シーケンシングのエラーの原因となることがある。例えば、ホルマリン固定パラフィン包埋（FFPE）サンプルのDNAは、サンプル調製や保存の間に損傷して1本鎖になることがあり、これら1本鎖DNA同士は、繰り返し配列などの部分で誤って対形成してキメラ断片を生成する（特許文献1、非特許文献1）。また、cell free DNA（cfDNA）は、血中で分解されて1本鎖になることがある。1本鎖DNAのシトシンは、脱アミノ化によりチミジンに変換されやすいため、シーケンシングで高頻度にエラーを引き起こす（特許文献2）。このような1本鎖DNAに由来するエラーを低減するために、ライブラリ調製の過程でサンプルDNAを1本鎖特異的ヌクレアーゼで処理して1本鎖部分を除去する方法が提案されている（特許文献1、2、及び非特許文献1）。例えば、特許文献1及び非特許文献1には、FFPEサンプルのDNAを1本鎖特異的ヌクレアーゼで処理することで、シーケンシングにおけるキメラ断片の検出率及びエラー率が減少したことが報告されている。特許文献2には、cfDNAを1本鎖特異的ヌクレアーゼで処理することで、シトシンの脱アミノ化によるシーケンシングのエラーを低減することが可能であることが記載されている。

10

【 0 0 0 4 】

あるいは、サンプル調製や保存の過程で解析対象のDNAに生じる酸化修飾等は、シーケンシングのエラーの原因となる（非特許文献2）。また、Kennedyらは、シーケンスリードの端部でのエラーの増加が、断片化したDNAの末端修復工程に起因する可能性に触れている（非特許文献3）。

20

【 0 0 0 5 】

近年、相補鎖情報を利用してシーケンシングのエラーを低減する方法が注目されている。例えば、サンプル調製や保存の過程でDNAに生じる酸化修飾等は、2本鎖のうち片方の鎖だけに起こるため、2本の相補鎖間に固定された変異を同定することで酸化修飾等に起因するエラーを除くことができる（特許文献3、4）。しかしながら、塩基の酸化修飾がDNA末端の突出部位に存在する場合には、該修飾された塩基が末端修復工程において誤った塩基とペアを形成し、この誤った塩基がPCR等を経てDNAの2本鎖に固定されることがある。そのため、突出部位に存在する酸化修飾等の起きた塩基は、相補鎖情報を活用したシーケンシングにおいても取り除くことができないエラーとなり得る。非特許文献3には、シーケンスリードの両端から5塩基を削除して解析することで、DNAの末端修復工程に起因するエラーを低減することを提案している。

30

【 先行技術文献 】

【 特許文献 】

【 0 0 0 6 】

【文献】国際公開広報第2015/057985号

国際公開広報第2019/126803号

国際公開広報第2013/142389号

国際公開広報第2019/208827号

40

【 非特許文献 】

【 0 0 0 7 】

【文献】Nucleic Acids Research, 47(2):e12, 2019

Nucleic Acids Research, 41(6):e67, 2013

Nature Protocols, 9(11):2586-2606, 2014

【 発明の概要 】

【 発明が解決しようとする課題 】

【 0 0 0 8 】

本発明は、シーケンシングエラーを低減させるシーケンシング用ライブラリを調製する

50

方法に関する。

【課題を解決するための手段】

【0009】

本発明は、シーケンシング用ライブラリの調製方法であって、
サンプルDNAを断片化すること；及び、
調製したサンプルDNAの断片を1本鎖特異的ヌクレアーゼで処理し、該断片から1本鎖部分を除去すること、
を含み、
該サンプルDNAが、生細胞から抽出したDNA、凍結細胞から抽出したDNA、又はそれらのDNAの保存サンプルである、
方法を提供する。

10

【0010】

また本発明は、前記シーケンシング用ライブラリをシーケンシングすることを含む、DNAのシーケンシング方法を提供する。

【0011】

また本発明は、細胞中のゲノムDNAをサンプルDNAとして用いて、前記シーケンシング用ライブラリの調製方法によりシーケンシング用ライブラリを調製すること；及び
該シーケンシング用ライブラリをシーケンシングすること、
を含む、ゲノムDNAの変異を検出する方法を提供する。

20

【発明の効果】

【0012】

本発明によれば、サンプル調製や保存の過程で生じる解析対象DNAの酸化修飾や損傷に起因するシーケンシングエラーを低減することができる。

【図面の簡単な説明】

【0013】

【図1】DMSO暴露ライブラリにおける6つの変異パターンの変異頻度。データは同一条件で暴露した3サンプルの平均値と標準偏差を示す。

【図2】DMSO暴露ライブラリにおける12変異パターンの変異頻度に対するリードペア両端の塩基の除去の影響。データは同一条件で暴露した3サンプルの平均値と標準偏差を示す。

30

【図3】DMSO暴露ライブラリの6つの変異パターンの変異頻度に対するS1 nuclease処理の影響。

【図4】異なるユニット数のS1 nucleaseで処理したDMSO暴露ライブラリにおける12変異パターンの変異頻度。

【図5】続き。

【図6】DMSO暴露ライブラリの6つの変異パターンの変異頻度に対するMBN処理の影響。

【図7】異なるユニット数のMBNで処理したDMSO暴露ライブラリにおける12変異パターンの変異頻度。

【図8】続き。

40

【図9】DMSO暴露ライブラリの6つの変異パターンの変異頻度に対するRecJf処理の影響。

【図10】異なるユニット数のRecJfで処理したDMSO暴露ライブラリにおける12変異パターンの変異頻度。

【図11】続き。

【図12】変異原処理したサンプルの変異検出に対するS1 nuclease処理の影響。異なるユニット数のS1 nucleaseで処理したDMSO暴露ライブラリ(DMSO control)及び3-MC暴露ライブラリ(3MC)における変異頻度。

【図13】続き。

【図14】変異原処理したサンプルの変異検出に対するMBN処理の影響。異なるユニッ

50

ト数のMBNで処理したDMSO暴露ライブラリ(DMSO control)及び3-MC暴露ライブラリ(3MC)における変異頻度。

【図15】変異原処理したサンプルの変異検出に対するRecJ_f処理の影響。異なるユニット数のRecJ_fで処理したDMSO暴露ライブラリ(DMSO control)及び3-MC暴露ライブラリ(3MC)における変異頻度。

【図16】S1 nuclease処理したDMSO暴露ライブラリのシーケンシングにおけるゲノムに対するカバレッジを示すヒストグラム。横軸はゲノム上の位置、縦軸は約100塩基区間のカバレッジを正規化した値。

【図17】MBN処理したDMSO暴露ライブラリのシーケンシングにおけるゲノムに対するカバレッジを示すヒストグラム。横軸はゲノム上の位置、縦軸は約100塩基区間のカバレッジを正規化した値。

10

【図18】RecJ_f処理したDMSO暴露ライブラリのシーケンシングにおけるゲノムに対するカバレッジを示すヒストグラム。横軸はゲノム上の位置、縦軸は約100塩基区間のカバレッジを正規化した値。

【図19】断片の誤認識率に対するS1 nuclease処理の影響。縦軸は断片の誤認識率(リードペアのグループに異なるindexが含まれる割合(%))を示す。

【図20】断片の誤認識率に対するMBN処理の影響。縦軸は断片の誤認識率(リードペアのグループに異なるindexが含まれる割合(%))を示す。

【図21】断片の誤認識率に対するRecJ_f処理の影響。縦軸は断片の誤認識率(リードペアのグループに異なるindexが含まれる割合(%))を示す。

20

【図22】S1 nuclease処理を行った断片の誤認識率に対する初期DNA量の影響。縦軸は断片の誤認識率(リードペアのグループに異なるindexが含まれる割合(%))を示す。

【図23】断片の誤認識率に対するS1 nuclease+RecJ_f処理の影響。縦軸は断片の誤認識率(リードペアのグループに異なるindexが含まれる割合(%))を示す。横軸は使用したRecJ_fのユニット数を表す。

【図24】S1 nuclease+RecJ_f処理DMSO暴露ライブラリにおける6変異パターンの変異頻度。凡例は、各バーで示すデータに用いたRecJ_fのユニット数を表す。

【発明を実施するための形態】

30

【0014】

(1. 定義)

本明細書において、「変異(又は突然変異)」(mutation)とは、DNAに生じる突然変異をいい、例えば、DNAにおける塩基又は配列の欠失、挿入、置換、付加、逆位、及び転座が挙げられる。本明細書における変異は、1塩基の欠失、挿入、置換、付加、ならびに2以上の塩基からなる配列の欠失、挿入、置換、付加、逆位、及び転座を包含する。また本明細書における変異には、遺伝子のコード領域及び非コード領域における変異が含まれ、また発現するアミノ酸の変化を伴う変異、及び発現するアミノ酸の変化を伴わない変異(サイレント変異)が含まれる。

【0015】

40

本発明において評価される物質の「遺伝毒性」とは、該物質が変異を引き起こす性質(いわゆる変異原性)をいう。

【0016】

本明細書において、「参照配列」とは、解析の対象であるDNA中に含まれる既知の配列である。当該既知の配列としては、公共のデータベース等に登録されている配列を使用することが好ましいが、予めシーケンサー等で配列決定した解析対象DNA中の配列であってもよい。該参照配列の領域や長さ、その数は特に限定されず、解析の目的に応じてDNA中から適宜選択され得る。

【0017】

本明細書において、PCRで得られる「増幅断片」とは、鋳型DNAのPCR増幅によ

50

り得られた 2 本鎖 DNA 断片をいう。

【 0 0 1 8 】

本明細書において、DNA 又はその断片に関する「2本の相補鎖」とは、2本鎖の DNA 又はその断片を構成する互いに相補的な2本の1本鎖をいう。

【 0 0 1 9 】

本明細書において、「生リード配列」とは塩基配列のシーケンシングにより読み出された配列情報をいう。また、本明細書において、「リード配列」とは、生リード配列に対して、PCR やシーケンシング反応のために付加したアダプター配列やクオリティの低い塩基等のトリミングなどを行って、生リード配列からシーケンシング対象である塩基配列の情報を取り出したものをいう。ただし、上記のトリミング等の必要がない場合、生リード配列をそのままリード配列として用いることも可能である。また、生リード配列にシーケンシング対象塩基配列の配列情報が複数含まれる場合、それら個々のシーケンシング対象塩基配列の配列情報を個々のリード配列として取り出すことができ、その場合1つの生リード配列から1つ以上のリード配列が作成され得る。したがって基本的には、本明細書におけるリード配列は、サンプルDNA断片にアダプター配列等が付加される場合でも、該アダプター配列等の配列情報を含まず、サンプルDNA断片に由来する塩基配列の情報のみを含む。リード配列は、シーケンシング対象である塩基配列（例えば、サンプルDNA断片の塩基配列）のいずれかの末端の塩基から始まる塩基配列の情報を有する。リード配列の長さは、通常、シーケンサーの性能や仕様に依存する。したがって、リード配列は、場合によっては、シーケンシング対象である塩基配列の一方の末端の塩基から他方の末端の塩基までの配列（全配列）の情報を有していてもよいが、必ずしもその必要はない。

10

20

【 0 0 2 0 】

本明細書において、リード配列の「先頭」及び「末尾」とは、それぞれ、該リード配列の作成時に最初に読み取られた末端、及び最後に読み取られた末端をいう。本明細書において、リード配列に関する「配列の向き」とは、該リード配列をマッピングしたDNA配列における該リード配列の先頭から末尾への方向をいう。

【 0 0 2 1 】

本明細書において、2個以上のリード配列が「サンプルDNA上の同一領域の配列情報を有する」とは、サンプルDNAの配列（又は参照配列）上においてそれらのリード配列の両末端が配置すると推定される位置が同一であることをいう。該「サンプルDNA上の同一領域の配列情報を有する」とは、該2個以上のリード配列が100%配列同一であることを要求しないが、一方、両末端が配置すると推定される位置が1bpでも異なるリード配列は、「サンプルDNA上の同一領域の配列情報を有する」ものではない。

30

【 0 0 2 2 】

本明細書において、2個以上のリード配列が「参照配列上の同一の位置にマッピングされる」とは、参照配列にマッピングしたときに、それらのリード配列の先頭と末尾の位置がそれぞれ、参照配列上で同一の位置に配置されることをいう。

【 0 0 2 3 】

本明細書において、「リードペア」とは、1つのシーケンシング対象配列から読み取られた2本のリード配列のペアをいう。リードペアに含まれる該2本のリード配列の一方は、該対象配列を5'末端側から3'側へ読んだ配列に相当する配列情報を含むリード配列（本明細書において「リード1」と称する）であり、他方は、同じ一方の鎖の配列を3'末端側から5'側へ読んだ配列に相当する配列情報を含むリード配列（本明細書において「リード2」と称する）である。

40

【 0 0 2 4 】

本明細書において、DNA、配列又は断片の「リード1の先頭とリード2の先頭とに挟まれる領域」とは、リード1とリード2をマッピングした該DNA、配列又は断片における、リード1の先頭が配置される部位からリード2の先頭が配置される部位までの領域（リード1の先頭が配置される部位とリード2の先頭が配置される部位とを含む）をいう。

【 0 0 2 5 】

50

本明細書において、2個以上のリードペアが「サンプルDNA上の同一領域の配列情報を有する」とは、それらのリードペアの間で、サンプルDNA配列（又は参照配列）上の「リード1の先頭とリード2の先頭とに挟まれる領域が同一」であることを意味する。2個以上のリードペアが「サンプルDNA上の同一領域の配列情報を有する」とき、それらのリードペア間でリード配列が100%配列同一であることは必ずしも要求されない。一方、該「リード1の先頭とリード2の先頭とに挟まれる領域」の末端の位置が1bpでも異なるリードペアは、「サンプルDNA上の同一領域の配列情報を有する」ものではない。

【0026】

本明細書において、2個以上のリードペアが「参照配列上の同一の位置にマッピングされる」とは、参照配列にマッピングしたときに、それらのリードペアの両末端がそれぞれ、参照配列上で同一の位置に配置されることをいう。「リードペアの両末端」は、リード1及びリード2の読み込み開始位置に相当する。

10

【0027】

本明細書において、酵素の「ユニット(U)数」とは、該酵素の活性（触媒活性とも言う）の単位を指しており、酵素ごとにその定義は異なり得る。

【0028】

本明細書中で引用された全ての特許文献、非特許文献、及びその他の刊行物は、その全体が本明細書中において参考として援用される。

【0029】

(2. ライブラリ調製方法)

20

シーケンシング用ライブラリ調製の過程で、DNA断片の端部の1本鎖突出部位に酸化修飾等の塩基の修飾が生じた場合、末端修復工程における当該修飾塩基の誤った塩基とのペア形成、及び該誤った塩基を有する鎖のPCR増幅により、2本の相補鎖に変異が起きた場合と同じ相補鎖情報を有するライブラリが調製される。このようなライブラリは、相補鎖情報を活用したシーケンシングにおいても取り除くことができないエラーをもたらす得る。本発明者は、相補鎖情報を活用したシーケンシングにおいて、GC TA、GC CGの変異において、C A、C Gに比べて、G T、G Cの変異が高頻度に検出されることを確認した(図1)。これら高頻度の変異は、グアニンが酸化修飾されたことに起因するエラーと考えられた。このエラーの原因として、シーケンシング用ライブラリの調製過程で断片化されたサンプルDNAの末端に1本鎖突出が生じ、該1本鎖突出部位のグアニンが酸化修飾されたためと考えられた(下記概念図1左)。

30

【0030】

こうした末端突出部位のエラーを除くため、本発明者は、従来のアプローチ(非特許文献3)に従って、シーケンシングで得られたリードペアの両端の10~20塩基を除去した。その結果、除去した塩基数に依存して、G T、G Cの変異頻度が減少した(図2)。この結果は、G T、G Cの変異がリードペアの両端部に多く存在していることを示し、これらの変異がDNA断片の末端の1本鎖突出部位のグアニンの酸化修飾に起因するエラーであることを支持した。しかし、この従来のアプローチでは、両端をそれぞれ20塩基除去したとしてもエラーによるグアニンの変異を十分に低減することはできなかった(図2)。リードペアから除去する塩基数の増加によってエラーをより低減できると予想されるが、リードペアの塩基数の減少は、DNA解析の効率や精度を低下させる。

40

【0031】

(2-1) 概要

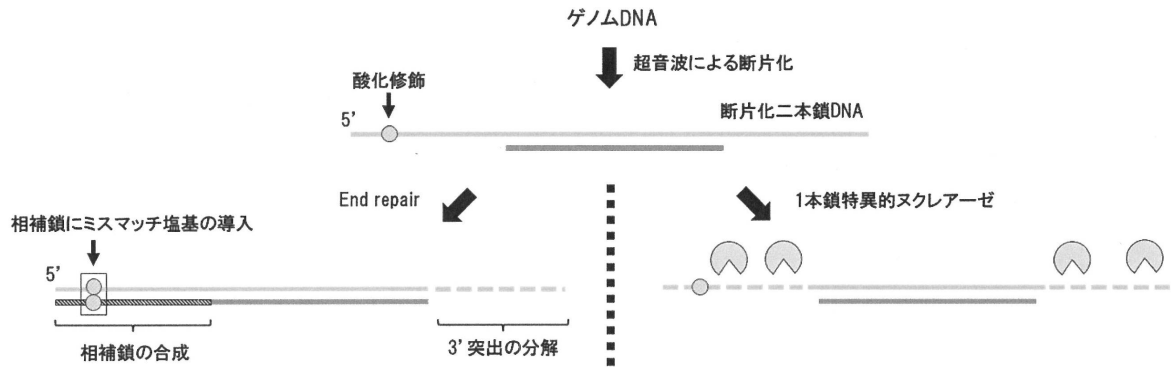
本発明においては、ライブラリ調製の過程で、サンプルDNAを超音波等により断片化してDNA断片を調製したのちに、該サンプルDNA断片を1本鎖特異的ヌクレアーゼで処理して、その1本鎖部分を除去することにより、酸化修飾等によるシーケンシングのエラーを効率的に低減する(概念図1右)。

【0032】

50

【化 1】

概念図 1



10

【 0 0 3 3】

(2 - 2) サンプル DNA

本発明によるライブラリの調製方法で用いられる「サンプル DNA」は、2本鎖 DNA であればよく、その由来は動物、植物、微生物などを含み、特に限定されない。該サンプル DNA の種類としては、ゲノム DNA、ミトコンドリアゲノム DNA、葉緑体ゲノム DNA、プラスミド DNA、ウイルスゲノム DNA、合成 DNA などが挙げられ、限定されないが、ゲノム DNA が好ましい。

20

【 0 0 3 4】

好ましくは、該サンプル DNA は、体内で、又は細胞や組織サンプルの調製及び保存過程で、分解や損傷を受けていないか又は分解や損傷が低頻度である DNA (以下の本明細書において、「新鮮な」DNA ともいう) である。好ましくは、該「新鮮な」DNA は、1本鎖に分解された領域をほとんど含まない、ほぼ完全な2本鎖で存在する DNA をいう。例えば、該「新鮮な」DNA は、生細胞から抽出した DNA (例えば、生体、器官、組織、又はそれらから採取した細胞から固定処置等を経ることなく直接抽出した DNA、細菌等の微生物の細胞から直接抽出した DNA)、凍結細胞から抽出した DNA (例えば、凍結保存した生体から抽出した DNA、生体から採取した器官、組織又は細胞を採取後すみやかに凍結保存したものから抽出した DNA)、及びそれらの抽出した DNA の保存サンプル (例えば、凍結保存、溶剤や不活性ガス中での低温保存をしたサンプル等)、などの DNA の化学修飾や分解を促進する環境に長期間さらされていない DNA であり得る。一方、FFPE サンプル由来 DNA のようなホルマリン固定された細胞由来の DNA、及び cfDNA のような一定期間血液中に存在していた DNA は、該「新鮮な」DNA から除外され得る。あるいは、「新鮮な」DNA は、Agilent 4200 Tape Station、又は Agilent 2200 Tape Station (いずれもアジレント・テクノロジー社製) で分析した DNA Integrity Number (以下、本明細書において「DIN」という) が、好ましくは6以上、より好ましくは7以上、さらに好ましくは7.3以上、さらにより好ましくは7.5以上の DNA として定義することができる。

30

40

【 0 0 3 5】

サンプル DNA は、細胞から当該分野における通常の方法を用いて抽出又は単離することによって取得することができる。該抽出又は単離には、例えば、市販の DNA 抽出キットなどを用いることができる。あるいは、細胞から抽出又は単離後保存されている DNA を取得し、本発明の方法で使用してもよい。合成 DNA は、公知の化学合成法により合成することができる。

【 0 0 3 6】

あるいは、本発明の方法では、2本鎖 DNA の代わりに、2本鎖 RNA を用いてもよい

50

。2本鎖RNAは、それを保有するウイルスや細胞などから、市販のRNA抽出キットなど、当該分野における通常の方法で抽出又は単離することができる。あるいは、抽出又は単離後保存されている2本鎖RNAを取得し、本発明の方法で使用してもよい。本発明の方法においてRNAを取得及び解析する場合、取得されたRNAはPCR前にcDNAに変換され、該cDNA由来のリード配列の塩基Tは塩基Uと読み替えられる。

【0037】

(2-3) DNA断片の調製

サンプルDNAの断片化は、超音波処理、酵素処理など、切断箇所がランダムになる当該分野における通常の方法を用いて実施することができる。DNAの断片化処理の具体的な例としては、コバリス社のDNA Shearingシステムなどを用いた集中超音波処理等が挙げられる。調製する断片の長さは、シーケンサーが精度よく読み取れる長さに応じて適宜選択され得る。一般的には、100~10,000bpが選択され得るが、シーケンサーが精度よく読み取れる限りは10,000bp以上の長さの断片が調製されてもよく、シーケンサーの種類に依存してより適切な範囲が選択され得る。例えば、断片の増幅を行うシーケンシング反作用のシーケンサーにかける場合は、断片の長さは平均長100~1000bpが好ましく、平均長200~500bpがより好ましい。あるいは、より長い断片を調製し、これを後述するPCRにかけ、シーケンシング反応に適切な長さのPCR産物を調製してもよい。

【0038】

(2-4) ヌクレアーゼ処理

本発明の方法では、上述した新鮮なサンプルDNAを断片化した後、得られた断片を1本鎖特異的ヌクレアーゼで処理し、該断片から1本鎖部分を除去する。従来の1本鎖特異的ヌクレアーゼ処理は、FFPEサンプルのDNAやcfDNA等の比較的分解や損傷を受けており、既に断片化した状態でサンプル中に存在するDNAを対象としていた。本発明のように分解の程度が低い新鮮なDNAを、ライブラリ調製のために1本鎖特異的ヌクレアーゼで処理したことはこれまで報告されていない。

【0039】

本発明において、1本鎖特異的ヌクレアーゼによる分解の対象は、主にDNA断片の末端に存在する1本鎖突出部位であり得るが、この限りではない。例えば、DNA断片の両端以外（例えば中心部分）に存在する1本鎖部分も、本発明における1本鎖特異的ヌクレアーゼによる除去の対象であり得、その除去はエラー低減に寄与し得る。例えば、2本鎖DNA断片の片方の鎖にニックが存在する場合、後述する末端修復工程（例えば、End Repairカクテルの酵素での処理）の際にニック以降の鎖が再度合成され、エラー率増加に寄与する可能性がある。当該ニックにおける1本鎖部分を1本鎖特異的ヌクレアーゼで除去することは、エラー低減に寄与し得る。

【0040】

本発明の方法で使用可能な1本鎖特異的ヌクレアーゼは、1本鎖特異的に作用する限り、エンドヌクレアーゼであってもエキソヌクレアーゼであってもよい。1本鎖特異的エンドヌクレアーゼの例としては、S1 nuclease、Mung Bean Nuclease (MBN)などが挙げられ、1本鎖特異的エキソヌクレアーゼの例としてはExonuclease VIIなどが挙げられ、1本鎖特異的5'→3'エキソヌクレアーゼの例としてはRecJfなどが挙げられる。このうち、1本鎖への特異性が高い点及び2本鎖で挟まれた1本鎖も除去できる点から、S1 nuclease及びMBNが好ましく、S1 nucleaseがより好ましい。これらの1本鎖特異的ヌクレアーゼは市販されており、例えば、プロメガ社、タカラバイオ社、New England Biolabs社などから購入することができる。当該1本鎖特異的ヌクレアーゼ処理においては、1種類の酵素のみを用いてもよいが、複数種の酵素を組み合わせる用いてもよい。例えば、サンプルDNA断片をS1 nuclease、MBNなどのエンドヌクレアーゼで処理した後、さらにRecJfなどのエキソヌクレアーゼで処理することが好ましく、その逆も同様である。

10

20

30

40

50

【0041】

サンプルDNA断片の1本鎖特異的ヌクレアーゼによる処理は、通常の手順で、例えば購入元の提供するプロトコルに従って、実施することができる。反応条件は、酵素の至適条件や、基質であるDNA断片の量に従って適宜決定することができる。例えば、反応液におけるサンプルDNA断片1ngあたりの酵素活性単位(ユニット数; U/ng)は、S1 nucleaseでは、シーケンシングエラー低減の観点からは好ましくは0.01 U/ng以上、より好ましくは0.02 U/ng以上、さらに好ましくは0.03 U/ng以上、さらに好ましくは0.05 U/ng以上、さらに好ましくは0.10 U/ngであり、一方、現実的に実施可能な上限値、及び、高濃度時に非特異的に生じる2本鎖DNAの分解の観点からは16.7 U/ng以下が好ましく、反応効率の観点からは、好ましくは5.00 U/ng以下、より好ましくは1.67 U/ng以下、さらに好ましくは1.00 U/ng以下、さらに好ましくは0.30 U/ng以下である。あるいは、酵素反応の効率の観点から好ましいS1 nucleaseの酵素量の範囲は、0.02~5.00 U/ng、より好ましくは0.03~1.67 U/ng、さらに好ましくは0.03~1.00 U/ng、さらに好ましくは0.05~1.00 U/ng、さらに好ましくは0.10~0.30 U/ngである。また例えば、反応液におけるMBNの酵素量は、シーケンシングエラー低減の観点からは好ましくは0.01 U/ng以上、より好ましくは0.02 U/ng以上、さらに好ましくは0.03 U/ng以上、さらに好ましくは0.05 U/ng以上、さらに好ましくは0.10 U/ng以上であり、一方、現実的に実施可能な上限値、及び、高濃度時に非特異的に生じる2本鎖DNAの分解の観点からは16.7 U/ng以下が好ましく、反応効率の観点からは、好ましくは、5.00 U/ng以下、より好ましくは1.67 U/ng以下、さらに好ましくは1.00 U/ng以下、さらに好ましくは0.30 U/ng以下である。あるいは、酵素反応の効率の観点から好ましいMBNの酵素量の範囲は、0.02~5.00 U/ng、より好ましくは0.03~1.67 U/ng、さらに好ましくは0.03~1.00 U/ng、さらに好ましくは0.05~1.00 U/ng、さらに好ましくは0.10~0.30 U/ngである。また例えば、反応液におけるRecJfの酵素量は、シーケンシングエラー低減の観点からは好ましくは0.10 U/ng以上、より好ましくは0.30 U/ng以上であり、一方、現実的に実施可能な上限値の観点からは100 U/ng以下が好ましく、推奨のDNAの量に近い条件(60 ng)で酵素反応を行う観点からは16.7 U/ng以下が好ましく、反応効率の観点からは1.00 U/ngが好ましい。あるいは、酵素反応の効率の観点から好ましいRecJfの酵素量の範囲は、0.10~16.7 U/ng、より好ましくは0.30~1.00 U/ngである。なお本明細書において、酵素活性1単位(1U)は以下の通り定義される：

- ・S1 nuclease：30 mM 酢酸ナトリウム(pH 4.6、25℃)、50 mM NaCl、1 mM ZnCl₂、5%グリセロール、0.5 mg/mL 変性仔牛胸腺DNAの混合溶液中において、37℃で1分間に1 µgの酸可溶性物質を生成する酵素活性。
- ・MBN：熱変性仔牛胸腺DNAを基質として、37℃、pH 5.0において、1分間に1 µgの酸可溶性分解物を生成する酵素活性。
- ・RecJf：全反応液50 µL(1×NE Buffer 2及び1.5 µgの超音波処理[³H]標識1本鎖E. coli DNAを含む)中、37℃、1分間で、0.5 ngのトリクロロ酢酸可溶性デオキシリボヌクレオチドを生成する酵素活性。

【0042】

さらに、反応に用いる酵素のユニット数は、後述する増幅(PCR)工程の初期DNA量と関連し得る。例えばS1 nucleaseでは、ユニット数が0.05 U/ngより大きい場合、下記式で算出される指標：

指標 = 初期DNA量 (amol / Mbp サンプルDNA) × 3^{log S1 nuclease (U/ng)}
(式中、S1 nuclease (U/ng) > 0.05、logは常用対数である)が、好ましくは60以下、より好ましくは30以下、さらに好ましくは15以下、さらにより好ましくは7.5以下である。また、例えばMBNでは、ユニット数が0.05 U/n

10

20

30

40

50

g より大きい場合、下記式で算出される指標：

指標 = 初期 DNA 量 ($a \text{ mol} / \text{Mbp}$ サンプル DNA) $\times 3 \log \text{MBN} (U/ng)$
 (式中、 $\text{MBN} (U/ng) > 0.05$ 、 \log は常用対数である) が、好ましくは 60
 以下、より好ましくは 30 以下、さらに好ましくは 15 以下、さらにより好ましくは 7.5
 以下である。一方、 $0.05 U/ng$ 以下の $S1 \text{ nuclease}$ もしくは MBN 、
 又はユニット数に関わらず RecJf を用いる場合、前記の式は成立せず、後述する増幅 (PCR)
 工程の初期 DNA 量は、サンプル DNA 1 Mbp あたり、好ましくは $250 a \text{ mol}$
 以下、より好ましくは $125 a \text{ mol}$ 以下、さらに好ましくは $62.5 a \text{ mol}$ 以下、
 さらにより好ましくは $31.3 a \text{ mol}$ 以下、なお好ましくは $15.7 a \text{ mol}$ 以下
 である。

10

【 0 0 4 3 】

反応後の酵素は、失活させるか又は洗浄除去することが望ましい。ヌクレアーゼ処理した DNA 断片は、その後の PCR 工程に用いることができるように精製する。DNA の精製には、エタノール沈殿、電気泳動、カラム精製、ビーズ精製、アフィニティー精製などの通常の手段を用いることができる。

【 0 0 4 4 】

(2 - 5) 追加処理

本発明においては、上記サンプル DNA 断片の 1 本鎖特異的ヌクレアーゼ処理以降は、通常の手順に従って、シーケンシング用ライブラリを調製することができる。例えば、1 本鎖特異的ヌクレアーゼ処理した DNA 断片を、必要に応じて、末端修復、末端への塩基付加、増幅などの処理にかけて、ライブラリを調製する。好ましくは、該末端修復、末端への塩基付加、及び増幅が、この順序で全て行われる。該末端修復、末端への塩基付加、及び増幅の工程は、 $\text{TruSeq Nano DNA Library Prep Kit}$ (イルミナ社) などの市販の試薬を用いて実施することができる。

20

【 0 0 4 5 】

(2 - 5 - 1) 末端修復

サンプル DNA 断片は、1 本鎖特異的ヌクレアーゼで処理した後にも、末端に短い 1 本鎖突出部位が残存することがある。末端修復では、該ヌクレアーゼ処理後の DNA 断片において、該残存する 1 本鎖突出部位を有する末端を平滑化する。該平滑化処理では、一般に、 $T4 \text{ DNA}$ ポリメラーゼ等の $3' - 5'$ エキソヌクレアーゼにより $3'$ 側突出末端が除去され、一方、 $5'$ 側突出末端は、 $5' - 3'$ ポリメラーゼにより対となる鎖が合成され、これにより DNA 断片の両端が平滑化される。

30

【 0 0 4 6 】

(2 - 5 - 2) 塩基付加

末端への塩基付加は、末端平滑化した DNA 断片に対して、その両端へのシーケンシングに必要な標識配列の付加や、該標識配列を付加するための $3'$ 末端へのアデニンの付加を行う処理である。標識配列が付加された DNA 断片を増幅し、シーケンシングすることで、該 DNA 断片の配列情報と該標識配列の情報とを取得することができる。また該標識配列の情報に従って、リード配列を識別又は分類することができる。例えば、DNA 断片の両末端に付加した標識配列は、リード配列が該 DNA 断片の全配列の情報を有するかを判断する指標となる。あるいは、DNA 断片の片方の末端に標識配列を付加し、該標識配列を含まない側からシーケンシングすることで、リード配列が該サンプル DNA 断片の全配列の情報を有するかを判断することができる。

40

【 0 0 4 7 】

相補鎖情報を活用したシーケンシングに用いるライブラリの調製においては、サンプル DNA の断片の両末端に、リード配列が該断片の 2 本の相補鎖のいずれに由来するかを識別可能にする標識配列を付加させることが望ましい。例えば、1 つの DNA 断片を構成する 2 本の相補鎖の $5'$ 末端側と $3'$ 末端側にそれぞれ異なる標識配列を付加させる。一実施形態においては、1 つの DNA 断片の両鎖の間で $5'$ 末端側の標識配列は同一であり、両鎖の $3'$ 末端側の標識配列も同一であり、かつ両端の標識配列は互いに相補的でない配列を含

50

む（以下の本明細書において、これを「相補鎖標識配列」と呼ぶ；下記概念図2参照）。好ましくは、該相補鎖標識配列においては、標識した各DNA断片の間で、5'末端側の標識配列は共通であり、かつ3'末端側の標識配列も共通である。よって、各断片を構成する2つの1本鎖は、それぞれ5'末端側及び3'末端側に異なる標識配列を有するが、該5'末端側の標識配列と該3'末端側の標識配列は各1本鎖間で共通である。一方、該相補鎖標識配列は、該リード配列がサンプルDNAのいずれの個別断片に由来するかを識別する必要はない。このような相補鎖標識配列の例としては、イルミナ社のTruSeqに付属のアダプター配列が挙げられる。

【0048】

別の一実施形態においては、サンプルDNAの断片を個別に識別する標識配列（以下の本明細書において、これを「個別断片標識配列」と呼ぶ；例えば、PNAS, 109(36):14508-14513, 2012、又は特許文献1に記載されるような、サンプルDNA断片固有のタグ配列）をDNA断片に付加することができる。このような標識は、リード配列がDNA断片の2本の相補鎖のいずれに由来するかを識別させ、相補鎖情報を活用したシーケンシングを可能にする。ただし、相補鎖情報を活用したシーケンシングの効率の観点からは、特にサンプルDNAのサイズが大きい場合、相補鎖標識配列を用いることが好ましい。

【0049】

（2-5-3）増幅

DNA断片の増幅には、PCR等の既存の方法を用いることができる。得られた増幅断片は、必要に応じて通常の手順で精製し、シーケンシング用ライブラリとして用いることができる。PCRは、市販のPCR用試薬や機器を用いて、常法に従って実施することができる。あるいは、PCR増幅装置を備えたシーケンサーを用いてもよい。サンプルDNAの断片のPCR増幅をその工程に含む高スループットシーケンサーとしては、HiSeq（イルミナ社製）、MiSeq（イルミナ社製）などが上市されている。

【0050】

好ましくは、当該PCRにおいては、鋳型として使用されたDNA断片の各々について、2つ以上の増幅断片がそれぞれ作製される。このとき、鋳型として用いたサンプルDNAの断片の少なくとも一部の各々について2つ以上の増幅断片が調製されればよい。一方、該PCRで全部の鋳型サンプルDNAの断片について2つ以上の増幅断片を得てもよいが、その必要はない。サンプルDNAの断片のPCR増幅をその工程に含む高スループットシーケンサーでは、シーケンシング反応に用いるPCR産物量を一定量用いることがシーケンシング効率の点で推奨されている。そのため、PCRにかけるサンプルDNA量（PCRでの初期DNA量）に応じてPCRのサイクル数を変更し、PCR産物量を推奨量にあわせることが好ましい。

【0051】

（3．シーケンシング方法）

上記の手順で得られたライブラリを用いてシーケンシングを実施することができる。本発明で得られたライブラリは、各種シーケンシング方法に適用可能である。好ましくは、本発明で得られたライブラリは、相補鎖情報を活用したシーケンシング（例えば、特許文献4に記載のシーケンシング方法）に用いられる。以下に、特許文献4を参考に、本発明で得られたライブラリを用いた、相補鎖情報を活用したシーケンシング方法（以下、本シーケンシング方法という）の概要を説明する。

【0052】

（3-1）概要

本シーケンシング方法は、基本的には、本発明で得られたライブラリをシーケンシングし、該ライブラリに含まれる各サンプルDNAの断片由来の複数の増幅断片の各々について1つ以上の読み取り結果（リード配列）を作成し、複数の増幅断片についての複数のリード配列を得ること；該シーケンシングで得られたリード配列の中から、該サンプルDNA上の同一領域の配列情報を有するリード配列を集めること；集めたリード配列の情報をを用いて、該サンプルDNAの配列情報を構築すること、を含む。

10

20

30

40

50

【 0 0 5 3 】

(3 - 2) シーケンシング及びリード配列の作成

ライブラリのシーケンシングは、解析等に必要な部分、例えば後述する変異解析の場合、参照配列との配列比較に使用すべき部分について行えば足りる。例えば、その配列の少なくとも一部、好ましくは全体が、参照配列の DNA 領域に対応する断片をシーケンシングすればよい。哺乳動物細胞等の場合には、エクソン領域等を選択的にシーケンシングしてもよい。領域の選択には、SureSelect (アジレント・テクノロジー社製) 等のキットが上市されている。

【 0 0 5 4 】

該シーケンシングにより、ライブラリについての生リード配列が取得される。該生リード配列から、PCR やシーケンシング反応のために付加したアダプター配列やクオリティの低い塩基等のトリミングなどを行ってサンプル DNA の断片に由来する配列を取り出すことで、リード配列が作成される。あるいは、上記トリミング等の必要がない場合、生リード配列をそのままリード配列として用いてもよい。該生リード配列又はリード配列が作成される増幅断片は、該ライブラリに含まれる増幅断片のうち少なくとも一部である複数の増幅断片であればよい。一方、該ライブラリに含まれる全増幅断片についてリード配列を取得してもよいが、その必要はない。該リード配列は、該複数の増幅断片の各々に対して1つ以上作成される。それらのリード配列は、該増幅断片 (すなわちそれが由来するサンプル DNA の断片) の2本の相補鎖のいずれかについての配列情報を有する。したがって、該ライブラリのシーケンシングにより、複数のリード配列が得られる。なお、この段階で得られた該複数のリード配列を含むデータを、本明細書において「シーケンシングデータ」と呼ぶことがある。

【 0 0 5 5 】

(3 - 3) リード配列のグループ化

次いで、得られた複数のリード配列の中から、各リード配列の配列情報に基づいて、サンプル DNA 上の同一領域の配列情報を有するリード配列を集める。集めたリード配列は、グループ化される。したがって、本発明の方法で作成される「リード配列のグループ」とは、サンプル DNA 上の同一領域の配列情報を有するリード配列の集合であり、言い換えると、同一のサンプル DNA 断片に由来すると推定されるリード配列の集合である。本発明の方法においては、通常、ライブラリ調製の際に PCR にかけたサンプル DNA 断片の数とシーケンシングデータの量に依存して、1つ以上のリード配列のグループが作成され得る。

【 0 0 5 6 】

本発明の方法の一実施形態においては、ライブラリに含まれる1増幅断片に対して、1本以上のリード配列が作成され、得られたリード配列は上述のようにグループ化される。好ましい実施形態においては、上述したリード配列のグループの作成に利用されるリード配列は、元のサンプル DNA の断片 (すなわち該リード配列が由来する増幅断片の元となるサンプル DNA の断片) の全配列の情報を有するリード配列である。シーケンシングで得られたリード配列の中から元のサンプル DNA の断片の全配列の情報を有するリード配列を選抜する手順としては、リード配列の末尾の塩基の読み取り精度 (クオリティ値) が高いリード配列を選別する方法、末端に標識配列を付加したライブラリを調製し、これをシーケンシングし、該標識配列の情報の有無に基づいてリード配列を選別する方法、などが挙げられる。このうち、標識配列を用いた方法についてより具体的な手順の例を説明する: まず、サンプル DNA の断片の両末端にそれぞれ異なる標識配列を付加し、これを PCR 増幅することにより、両末端に該標識配列を有する増幅断片を含むライブラリを調製する; 得られたライブラリをシーケンシングし、該増幅断片由来のリード配列と、それに付随する該標識配列の情報を取得する。該両末端の標識配列の両方の情報が付随するリード配列は、元のサンプル DNA の断片の全配列の情報を有するリード配列とみなされる。別の例では、サンプル DNA の断片の片方の末端に標識配列を付加し、これを PCR 増幅して該標識配列を含む増幅断片を調製する; 得られた増幅断片を、該標識配列のない末端

10

20

30

40

50

の側からシーケンシングする。該標識配列の情報が付随するリード配列は、元のサンプルDNA断片の全配列の情報を有するリード配列とみなされる。ここで該標識配列の情報は、生リード配列から取得してもよく、又はシーケンシングプライマーの配列情報から取得してもよい。

【0057】

集めたリード配列からリード配列のグループを作成する手段としては、例えば、参照配列上の同一の位置にマッピングされるリード配列を集める方法、少なくとも両末端領域の配列が同等であるリード配列を集める方法、などが挙げられる。なお、「少なくとも両末端領域の配列が同等」とは、アラインさせたリード配列が、少なくとも両末端領域において配列同一性が80%以上、好ましくは90%以上、より好ましくは95%以上、さらに好ましくは97%以上であり、かつ両末端が同じ位置にアラインすることをいう。該「末端領域」の長さは適宜選択することができ、例えば末端を含め、10塩基以上、好ましくは10~30塩基程度であればよい。あるいは、配列全体の同一性が80%以上、好ましくは90%以上、より好ましくは95%以上、さらに好ましくは97%以上であり、かつ両末端が同じ位置にアラインするリード配列を集めることで、リード配列のグループを作成してもよい。

10

【0058】

(3-4) リード配列のグループからのサンプルDNA配列情報の抽出

次に、得られたリード配列のグループから、サンプルDNAの配列情報を抽出する。詳細には、該リード配列のグループに含まれるリード配列の情報を用いて1つの配列データを導き出す。得られた配列データは、該グループのリード配列が由来する特定のサンプルDNAの断片についてのコンセンサス配列を表す。

20

【0059】

例えば、リード配列のグループに含まれるリード配列の間で配列情報のコンセンサスを取ることで、1つの配列データを作成することができる。リード配列間でのコンセンサスを取る具体的な手法としては、以下が挙げられる：リード配列をアライメントし、アライメントした全てのリード配列の対応する塩基が一致した場合にその塩基を"コンセンサス塩基"とする方法；リード配列をアライメントした後、配列上の各位置で最大の頻度で出現する塩基を決定し、"コンセンサス塩基"として抽出する方法；リード配列をアライメントした後、対応する位置にある塩基の中でシーケンサーでの読み取り精度(クオリティ値)の最も高い塩基を"コンセンサス塩基"として採用する方法；リード配列をアライメントした後、クオリティ値や塩基の出現頻度等を基に、確率論的に"コンセンサス塩基"を決定する方法；あるいは、これらを組み合わせた方法、など。

30

【0060】

リード配列間でのコンセンサスを取る際には、リード配列のグループに含まれる全てのリード配列が用いられてもよいが、該グループ内の一部のリード配列のみが用いられてもよい。リード配列間でのコンセンサスを取ることで、シーケンシングにおける読み取りエラーなどのエラーを除外することができるので、高精度な読み取り結果を得ることができる。得られた配列データは、サンプルDNAの一領域の配列を示す最終的な配列データとして取得することができる。

40

【0061】

(3-5) 相補鎖情報に基づくシーケンシング

シーケンシングエラーを引き起こす、DNAの酸化修飾等による塩基の置換は、基本的にはDNA2本鎖のうち片方の鎖だけに起こる。したがって、DNAの2本の相補鎖それぞれについてのシーケンシング情報を用いることで、片方の鎖にのみ発生した塩基の置換を変異として検出することなく、2本鎖に固定された真の変異のみを同定することが可能となる。DNAの2本の相補鎖の配列は、相補的であるものの、互いに等価の情報を有する。従って理論上は、シーケンシングで得られたリード配列の中から等価の情報を有する配列を探すことにより、相補鎖の情報を得ることが可能である。例えば、ある生物種のゲノム配列からサンプルDNAを調製した場合、サンプルDNAの断片を構成する2本の相

50

補鎖それぞれに由来する読み取り領域が同一である2つのリード配列は、解析対象となる生物種の参照配列にマッピングした場合には、ゲノムの同一箇所にもマッピングされる。したがって、ゲノムの同一箇所にもマップされ得るリード配列を集めて、それらリード配列をその由来する相補鎖によって選抜することで、2本の相補鎖のそれぞれに由来するリード配列を取得することができる。さらにそれら2本の相補鎖に由来するリード配列間でのコンセンサスをとることにより、相補鎖の情報を反映させた高精度なリード情報を得ることが可能である。

【0062】

本シーケンシング方法では、上記(3-2)で述べたライブラリのシーケンシングの際に、各サンプルDNAの断片を構成する2本の相補鎖の各々に対してリード配列を作成する。より詳細には、ライブラリのシーケンシングにより該ライブラリに含まれる該複数の増幅断片の各々についてのリード配列を作成する際に、各サンプルDNAの断片を構成する2本の相補鎖の各々に由来する増幅断片に対して、それぞれ1つ以上のリード配列が作成される。すなわち、1個のサンプルDNAの断片に対して2つ以上のリード配列が取得され、それらのリード配列はそれらが由来するサンプルDNAの断片の2本の相補鎖の一方及び他方についての配列情報を有する。

10

【0063】

次いで、得られた複数のリード配列から、1つ以上のリード配列のグループを作成する。リード配列のグループを作成する手段は、上記(3-3)で述べたとおりである。ここで得られるリード配列のグループには、特定のサンプルDNAの断片についての2本の相補鎖の一方及び他方の配列情報を有するリード配列が含まれている。したがって、該リード配列のグループに含まれるリード配列の間で配列情報のコンセンサスを取ることで、相補鎖の情報を反映させた配列データを作成することができる。リード配列間でのコンセンサスを取る具体的な手法は、上記(3-4)で述べたとおりである。リード配列間でのコンセンサスを取る際には、リード配列のグループに含まれる全てのリード配列が用いられてもよいが、該グループ内の一部のリード配列のみが用いられてもよい。

20

【0064】

好ましくは、上記リード配列間でのコンセンサスを取る工程は、リード配列のグループの中から、サンプルDNAの断片の2本の相補鎖の各々に由来するリード配列を少なくとも1つずつ集め、集めたリード配列の間で配列情報のコンセンサスを取ることを含む。これにより、相補鎖情報を用いたコンセンサスデータ(本明細書において「相補鎖間コンセンサスリード配列」ともいう)を得ることができる。得られた相補鎖間コンセンサスリード配列は、シーケンシングにおける読み取りエラーやDNA酸化修飾等に起因するエラーなどの片方の鎖にのみ生じるエラーが除外された高精度な読み取り結果であり、サンプルDNAの断片についての配列を示す最終的な配列データとして取得することができる。

30

【0065】

リード配列のグループの中から、サンプルDNAの断片の2本の相補鎖の各々に由来するリード配列を集める手順としては、例えば、以下の手順が挙げられる：予めサンプルDNAの断片に2本の相補鎖を識別できる標識配列を付加することにより、該標識配列を含む増幅断片を調製する；次いで、該増幅断片をシーケンシングし、該増幅断片由来のリード配列と、それに付随する該標識配列の情報を取得する；得られたリード配列から、リード配列のグループを作成する；次いで、リード配列に付随する標識配列の情報を利用して、リード配列のグループの中から互いに相補的な鎖に由来するリード配列を集める。

40

【0066】

上記の手順においては、好ましくはサンプルDNAの断片に、上記(2-5-2)で述べたリード配列が該断片の2本の相補鎖のいずれに由来するかを識別可能にする標識配列(例えば、相補鎖標識配列又は個別断片標識配列)を付加する。好ましくは、相補鎖標識配列が用いられる。該標識配列が付加されたサンプルDNAの断片から得られた増幅断片をシーケンシングすることで、該増幅断片由来のリード配列とそれに付随する該標識配列の情報を取得することができる。

50

【 0 0 6 7 】

次に、当該標識配列の情報を利用して、リード配列のグループの中から互いに相補的な鎖に由来するリード配列を集める際の好ましい手順を説明する。リード配列のグループに含まれるリード配列を参照配列にマッピングするとき、5'末端側の標識配列の情報が付随し、かつその先頭が末尾に対して参照配列上でより5'側に位置するリード配列と、3'末端側の標識配列の情報が付随し、かつその先頭が末尾に対して参照配列上でより3'側に位置するリード配列は、サンプルDNAの断片の2本の相補鎖のうち同じ1本鎖に由来する。一方、3'末端側の標識配列の情報が付随し、かつその先頭が末尾に対して参照配列上でより5'側に位置するリード配列と、5'末端側の標識配列の情報が付随し、かつその先頭が末尾に対して参照配列上でより3'側に位置するリード配列は、サンプルDNAの断片の2本の相補鎖のうちのもう一方の1本鎖に由来する。したがって、参照配列にマッピングされたリード配列の参照配列に対する配置と、それに付随する標識配列の情報に基づいて、リード配列のグループ内の各リード配列がサンプルDNAの断片を構成する2本の相補鎖のどちらに由来するかを識別することができる。あるいは、増幅断片の末端に特定の標識配列が付加しているときにのみ開始するシーケンシング反応を行うことにより、標識配列の情報に基づいて、サンプルDNAの断片の特定の1本鎖に由来するリード配列を識別することができる。このようにサンプルDNAの断片の同じ1本鎖に由来するリード配列を予め識別しておくことで、リード配列のグループの中から互いに相補的な鎖に由来するリード配列を集めることができる。

10

【 0 0 6 8 】

上述したリード配列のグループから相補鎖間コンセンサスリード配列を得る手順の具体的な例としては、リード配列のグループの中から、サンプルDNAの断片の2本の相補鎖の各々に由来する2本のリード配列を選択し、それら2本のリード配列の間で配列情報のコンセンサスを取ることが挙げられる。さらに、該手順を繰り返して複数の相補鎖間コンセンサスリード配列を作成した後、さらにそれらの間でのコンセンサスを取り、1つの相補鎖間コンセンサスリード配列を作成してもよい。あるいは、該相補鎖間コンセンサスリード配列を得る手順の別の具体的な例としては、リード配列のグループに含まれるリード配列を、サンプルDNAの断片の2本の相補鎖の一方に由来する群と他方に由来する群とに分け、各群のリード配列の間でコンセンサスを取り、得られた2つのコンセンサスデータの間でさらにコンセンサスを取り、1つの相補鎖間コンセンサスリード配列を作成することが挙げられる。あるいはサンプルDNAの断片の2本の相補鎖に由来するリード配列を特に区別せず、リード配列のグループに含まれるリード配列の間でコンセンサスを取り、コンセンサスリード配列を作成することが挙げられる。

20

30

【 0 0 6 9 】

(3 - 6) リードペアを用いたサンプルDNA配列情報の抽出

本シーケンシング方法の一実施形態においては、上記(3 - 2)で述べたライブラリのシーケンシングの際に、該ライブラリに含まれる該複数の増幅断片の各々に対して1本のリード配列を作成する代わりに、2本のリード配列からなるリード配列のペア(すなわち「リードペア」)が1つ作成される。作成されたリードペアから、上記と同様の原理で、サンプルDNAの配列情報が抽出される。

40

【 0 0 7 0 】

当該方法においては、ライブラリのシーケンシングにより、各増幅断片に対して1つ以上のリードペアが作成される。また該1つ以上のリードペアの作成は、各サンプルDNAの断片に由来する2つ以上の該増幅断片について行われる。それらのリードペアは、該サンプルDNAの断片の2本の相補鎖のいずれかについての配列情報を有する。したがって、本実施形態においては、上述したライブラリのシーケンシングで得られる複数のリード配列は、複数個のリードペアを含む。

【 0 0 7 1 】

該リードペアを構成する2本のリード配列の一方は、該増幅断片を構成する2本の相補鎖のうち一方の鎖の配列を5'末端側から3'側へ読んだ配列に相当する配列情報を含む

50

リード配列（すなわち「リード1」）であり、他方は、同じ一方の鎖の配列を3'末端側から5'側へ読んだ配列に相当する配列情報を含むリード配列（すなわち「リード2」）である。リード1とリード2は、オリジナルの鎖（増幅断片を構成する1本鎖）に対して互いに逆向きに配置する。すなわち、該オリジナルの鎖に対してマッピングした場合、リード1の先頭は、その末尾に比べてより該オリジナルの鎖の5'側に配置し、一方、リード2の先頭は、その末尾に比べてより該オリジナルの鎖の3'側に配置する（後出の模式図1参照）。

【0072】

次いで、得られたシーケンシングデータ中の複数のリードペアの中から、サンプルDNA上の同一領域の配列情報を有するリードペアを選抜する。集めたリードペアはグループ化される。該リードペアのグループを作成する手段としては、例えば、リードペアのリード1とリード2を参照配列に対してマッピングし、リード1の先頭とリード2の先頭とに挟まれる該参照配列の領域が同一であるリード配列のペアを集めて、同じグループに分ける方法が挙げられる。より詳細な手順の例としては、まず、リードペアに含まれる一方のリード配列（リード1又は2）の先頭が参照配列上の同じ位置に位置するリードペアを集め、次いで集めたリード配列のペアの中から、該リードペアに含まれるもう一方のリード配列（リード2又は1）の先頭が参照配列上の同じ位置に位置するリード配列のペアを集めて、それらと同じグループに分ける方法が挙げられる。

10

【0073】

したがって、本発明の方法で作成される「リード配列のペア（リードペア）のグループ」とは、サンプルDNA上の同一領域の配列情報を有する（すなわち、同一のサンプルDNA断片に由来する）と推定されるリードペアの集合である。本方法においては、通常、ライブラリ調製に用いたサンプルDNAの断片の数とシーケンシングデータの量に依存して、1つ以上のリードペアのグループが作成され得る。

20

【0074】

次いで、得られたリードペアのグループに含まれるリード配列の情報を用いて、サンプルDNAの配列情報を抽出する。例えば、リードペアのグループに含まれるリード配列の間で配列情報のコンセンサスを取ることで、1つの配列データを作成することができる。リード配列間でのコンセンサスを取る具体的な手法は、上記（3-4）で述べたとおりである。リード配列間でのコンセンサスを取る際には、リードペアのグループに含まれる全てのリードペアのリード配列が用いられてもよいが、該グループ内の一部のリードペアのリード配列のみが用いられてもよい。得られた配列データは、サンプルDNAの断片についての配列を示す最終的な配列データとして取得することができる。

30

【0075】

（3-7）リードペアを用いた相補鎖情報に基づくシーケンシング

上述したリードペアを用いて、相補鎖情報を用いたDNAのシーケンシング方法を行うことができる。当該方法では、上記（3-6）で述べたライブラリのシーケンシングの際に、各サンプルDNAの断片を構成する2本の相補鎖の各々に由来する増幅断片に対して、1つ以上のリードペアが作成される。すなわち、1個のサンプルDNAの断片に対して2つ以上のリードペアが取得され、それらのリードペアは、該サンプルDNAの断片の2本の相補鎖の一方及び他方についての配列情報を有する。したがって、本実施形態においては、上述したシーケンシングで得られる複数のリード配列は、複数個のリードペアを含む。

40

【0076】

次いで、得られた複数のリードペアから、1つ以上のリードペアのグループを作成する。リードペアのグループを作成する手段は、上記（3-5）で述べたとおりである。ここで得られるリードペアのグループには、特定のサンプルDNAの断片についての2本の相補鎖の一方及び他方の配列情報を有するリードペアが含まれている。したがって、該リードペアのグループに含まれるリード配列の間で配列情報のコンセンサスを取ることで、相補鎖の情報を反映させた配列データを作成することができる。リード配列間でのコン

50

センサスを取る具体的な手法は、上記(3-4)で述べたとおりである。リード配列間でのコンセンサスを取る際には、リードペアのグループに含まれる全てのリードペアのリード配列が用いられてもよいが、該グループ内の一部のリードペアのリード配列のみが用いられてもよい。

【0077】

次いで、得られたリードペアのグループに含まれるリード配列の情報を用いて、1つの配列データを導き出す。例えば、リードペアのグループに含まれるリード配列の間で配列情報のコンセンサスを取ることで、1つの配列データを作成することができる。得られた配列データは、該グループのリード配列が由来する特定のサンプルDNAの断片についての配列である。リードペアのグループに、サンプルDNAの断片についての2本の相補鎖の配列情報を有するリード配列が含まれている場合には、それらの間でのコンセンサスを取ることにより、シーケンシングにおける読み取りエラーやDNA酸化修飾等に起因するエラーなどの片方の鎖にのみ生じるエラーを除外することができる。

10

【0078】

好ましくは、リードペアのグループに含まれるリード配列間でのコンセンサスを取る工程は、リードペアのグループの中から、サンプルDNAの断片の2本の相補鎖の各々に由来するリードペアを少なくとも1組ずつ集め、集めたリードペアに含まれるリード配列の間で配列情報のコンセンサスを取ることを含む。これにより、相補鎖間コンセンサスリード配列を得ることができる。得られた相補鎖間コンセンサスリード配列は、サンプルDNAの断片についての配列を示す最終的な配列データとして取得することができる。

20

【0079】

リードペアのグループの中からサンプルDNAの断片の2本の相補鎖の各々に由来するリードペアを集める手順としては、例えば、以下の手順が挙げられる：予めサンプルDNAの断片に2本の相補鎖を識別できる標識配列を付加しておくことにより、該標識配列を含む増幅断片を調製する；次いで、該増幅断片をシーケンシングし、該増幅断片由来のリードペアと、それに付随する該標識配列の情報を取得する；得られたリードペアから、リードペアのグループを作成する；次いで、リードペアに付随する標識配列の情報を利用して、リードペアのグループの中から、互いに相補的な鎖に由来するリードペアを集める。

【0080】

上記の手順においては、好ましくはサンプルDNAの断片に、上記(2-5-2)で述べたリード配列が該断片の2本の相補鎖のいずれに由来するかを識別可能にする標識配列(例えば、相補鎖標識配列又は個別断片標識配列)を付加する。好ましくは、相補鎖標識配列が用いられる。該標識配列が付加されたサンプルDNAの断片から得られた増幅断片をシーケンシングすることで、該増幅断片由来のリードペアと、それに含まれる各リード配列に付随する該標識配列の情報を取得することができる。この場合、各リードペアのリード1とリード2には、いずれか一方に5'末端側の標識配列の情報が、他方に3'末端側の標識配列の情報が、それぞれ付随する。

30

【0081】

次に、当該標識配列の情報を利用して、リードペアのグループの中から互いに相補的な鎖に由来するリードペアを集める際の好ましい手順を説明する。リードペアのグループに含まれるリードペアを参照配列にマッピングするとき、5'末端側の標識配列の情報が付随するリード配列の先頭が、参照配列上で、リードペアのもう一方のリード配列の先頭よりも5'側に位置する(すなわち、3'末端側の標識配列の情報が付随するリード配列の先頭が、参照配列上で、もう一方のリード配列の先頭よりも3'側に位置する)リードペアと、5'末端側の標識配列の情報が付随するリード配列の先頭が、参照配列上で、リードペアのもう一方のリード配列の先頭よりも3'側に位置する(すなわち、3'末端側の標識配列の情報が付随するリード配列の先頭が、参照配列上で、もう一方のリード配列の先頭よりも5'側に位置する)リードペアに分かれる。前者のリードペアと後者のリードペアは、サンプルDNA断片の2本の相補鎖のそれぞれに由来する(下記概念図2参照)。したがって、リードペアに含まれる2本のリード配列に付随する標識配列の情報と、該2本のリード

40

50

配列の参照配列上での互いの位置関係に基づいて、リードペアのグループ内の各リードペアがサンプルDNAの断片を構成する2本の相補鎖のどちらに由来するかを識別することができる。あるいは、増幅断片の末端に特定の標識配列が付加しているときにのみ開始するシーケンシング反応を行うことにより、標識配列の情報に基づいて、サンプルDNAの断片の特定の1本鎖に由来するリードペアを識別することができる。このようにサンプルDNA断片の同じ1本鎖に由来するリードペアを予め識別しておくことで、リードペアのグループの中から互いに相補的な鎖に由来するリードペアを集めることができる。

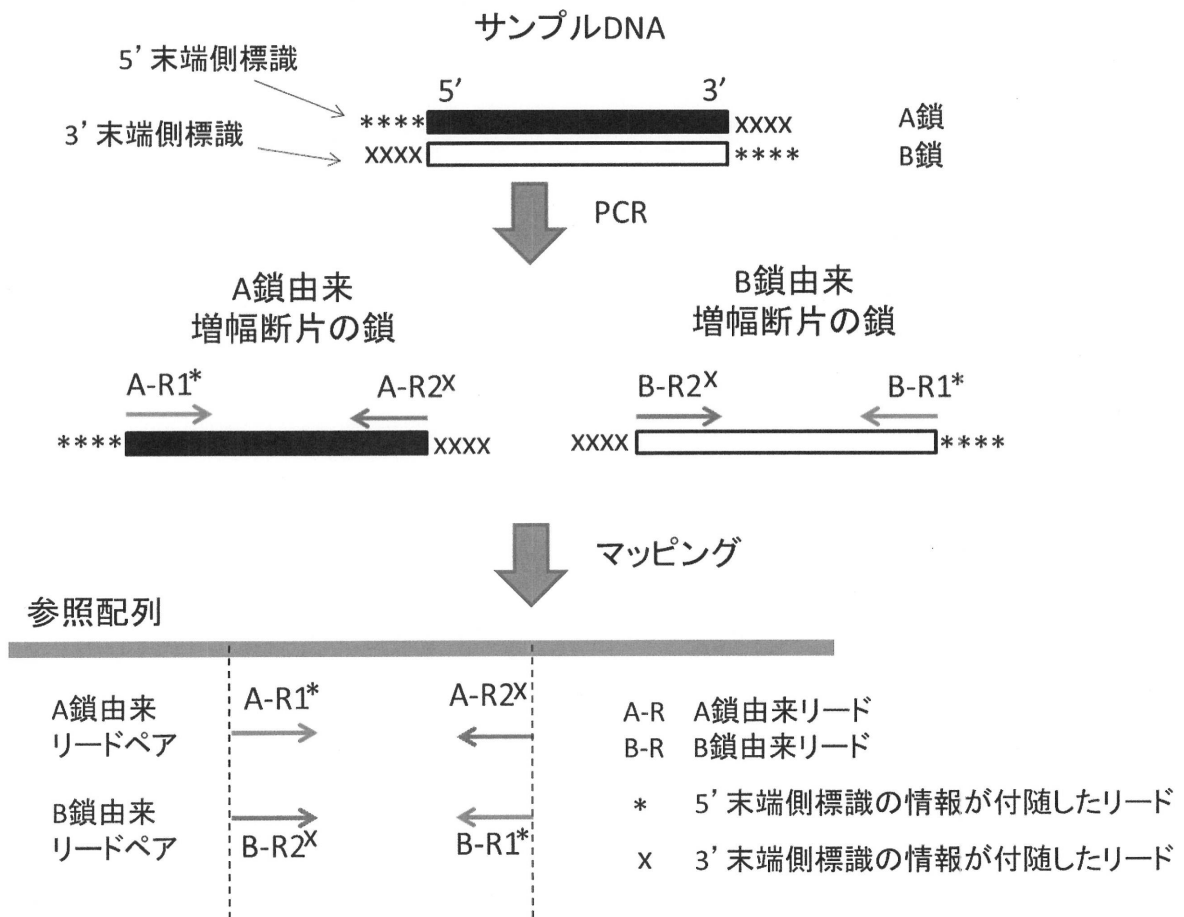
【0082】

【化2】

概念図2

10

標識配列の情報に基づくリードペアの識別



20

30

【0083】

上述したリードペアのグループから相補鎖間コンセンサスリード配列を得る手順の具体的な例としては、リードペアのグループの中から、サンプルDNAの断片の2本の相補鎖の各々に由来する2組のリードペアを選択し、それらのリードペアに含まれるリード配列の間で配列情報のコンセンサスを取ることが挙げられる。さらに、該手順を繰り返して複数の相補鎖間コンセンサスリード配列を作成した後、さらにそれらの間でのコンセンサスを取り、1つの相補鎖間コンセンサスリード配列を作成してもよい。あるいは、該相補鎖間コンセンサスリード配列を得る手順の別の具体的な例としては、リードペアのグループに含まれるリードペアを、サンプルDNAの断片の2本の相補鎖の一方に由来する群と他方に由来する群とに分け、各群のリードペアに含まれるリード配列の間でコンセンサスを取り、得られた2つのコンセンサスデータの間でさらにコンセンサスを取り、1つの相補

40

50

鎖間コンセンサスリード配列を作成することが挙げられる。あるいはサンプルDNAの断片の2本の相補鎖に由来するリード配列を特に区別せず、リードペアのグループに含まれるリード配列の間でコンセンサスを取り、コンセンサスリード配列を作成することが挙げられる。リードペアのグループから相補鎖間コンセンサスリード配列を作成する手順のより具体的な例は、以下の実施例1（模式図3）に説明されている。

【0084】

上述したリード配列又はリードペアのグループ分けは、リード配列に含まれるサンプルDNA自体の配列情報に基づいて行うことができる。互いに相補的な鎖の配列情報を有するリード配列は、サンプルDNAの断片に付加した標識配列の配列情報等に基づいて識別され得る。

10

【0085】

（3-8）個別断片標識配列を用いた相補鎖情報の抽出

あるいは、上述した個別断片標識配列を用いることで、サンプルDNAの断片の2本の相補鎖にそれぞれ由来するリード配列を識別することができる。この場合、必ずしもリード配列又はリードペアのグループを作成する必要はなく、個別の標識配列の情報に基づいて、1つのDNA断片の2本の相補鎖に由来するリード配列を抽出することができる。抽出したリード配列間でのコンセンサスをとることにより、相補鎖の情報を反映させた高精度なリード情報を得ることが可能である。

【0086】

（4．シーケンシングのための最適条件）

本シーケンシング方法では、サンプルDNAの個別の断片を識別するための標識（個別断片標識配列）を用いない場合、本来異なるDNA断片に由来する配列を誤って同一断片として誤認識する可能性があり、そのため本来変異として検出されるべきものがエラーと見なされて見逃される可能性がある。

20

【0087】

シーケンシングデータからサンプルDNAの配列データ（例えば、リード配列のグループ内でのコンセンサスデータや、相補鎖間コンセンサスリード配列）が得られる効率（解析効率）は、ライブラリ中に同一DNA断片由来の増幅産物（順鎖及び相補鎖を含む）が含まれている割合と、該ライブラリを用いたシーケンシングデータの量（リード数又はbp）に依存する。例えば、ライブラリ調製での増幅工程（例えばPCR）での初期DNA量と、シーケンシングデータ量は、上述したリード配列又はリードペアのグループに含まれるリード配列又はリードペアの数、ひいては解析効率に影響する。

30

【0088】

断片の誤認識は、ライブラリ中に、異なるサンプルDNA断片に由来する配列の重複する断片が含まれており、かつそれら断片の双方がシーケンシングされている場合に発生し得る。したがって、断片の誤認識には、ライブラリ中のDNA配列の多様度（増幅工程での初期DNA量）が関係する。また、ライブラリ中のDNA配列の多様度には、サンプルDNAの配列の多様度が影響し、サンプルDNAの配列の多様度はサンプルDNAのサイズ（総bp）に概ね依存する。よって、サンプルDNAのサイズも断片の誤認識に影響する。また、1本鎖特異的ヌクレアーゼ処理も断片の誤認識に影響し得る。1本鎖特異的ヌクレアーゼの配列特異性に起因して、該ヌクレアーゼ処理後のDNA断片には、該ヌクレアーゼが除去しにくい配列が残ることがある。すなわち、該ヌクレアーゼ処理後のDNA断片では、断片の両端の配列が同一となる確率が高まる可能性があり、これにより断片の誤認識の割合が増加し得る。

40

【0089】

したがって、シーケンシングの効率及び精度に影響を与え得る因子としては、主にライブラリ調製の増幅工程（例えばPCR）での初期DNA量、シーケンシングデータ量、そしてサンプルDNAのサイズが重要である。さらに、1本鎖特異的ヌクレアーゼ処理の反応液中におけるDNAの単位重量当たりの1本鎖特異的ヌクレアーゼのユニット数（U/ng）も、シーケンシングの効率及び精度に影響を与え得る因子として考慮することが望

50

まれる。さらに、これらの因子に依存するリード配列又はリードペアのグループに含まれるリード配列又はリードペアの数は、シーケンシングの効率及び精度を判断する指標となり得る。

【0090】

ライブラリ調製の増幅（例えばPCR）工程における初期DNA量（以下、単に初期DNA量という）の適切な範囲は、サンプルDNAのサイズに依存し得るが、例えばサンプルDNAの1Mbpあたり、好ましくは250amol以下、より好ましくは125amol以下、さらに好ましくは62.5amol以下、なお好ましくは31.3amol以下、さらになお好ましくは15.6amol以下、さらになお好ましくは7.8amol以下、さらになお好ましくは3.9amol以下、さらになお好ましくは1.7amol以下、さらになお好ましくは0.83amol以下、さらになお好ましくは0.42amol以下、さらになお好ましくは0.21amol以下である。一方で、該初期DNA量は、ゲノムの網羅性を担保する観点から、サンプルDNAの1Mbpあたり、好ましくは0.0003amol以上、より好ましくは0.0007amol以上、さらに好ましくは0.002amol以上、なお好ましくは0.005amol以上、さらになお好ましくは0.01amol以上、さらになお好ましくは0.03amol以上、さらになお好ましくは0.05amol以上、さらになお好ましくは0.1amol以上、さらになお好ましくは0.3amol以上、さらになお好ましくは1amol以上、さらになお好ましくは2amol以上、さらになお好ましくは3.9amol以上、さらになお好ましくは7.8amol以上である。

【0091】

一例において、本シーケンシング方法における初期DNA量は、以下のとおりである：ゲノムサイズ約5Mbpの細菌の場合、サンプルDNAの1Mbpあたり、好ましくは0.1~250amol、より好ましくは0.3~250amol、さらに好ましくは1~250amol、なお好ましくは2~125amol、さらになお好ましくは3.9~62.5amol、さらになお好ましくは7.8~31.3amolである；ゲノムサイズ約10Mbpの酵母の場合、サンプルDNAの1Mbpあたり、好ましくは0.05~250amol、より好ましくは0.1~250amol、さらに好ましくは0.3~125amol、なお好ましくは1~62.5amol、さらになお好ましくは2~31.3amolである；ゲノムサイズ約100Mbpの線虫の場合、サンプルDNAの1Mbpあたり、好ましくは0.005~31.3amol、より好ましくは0.01~31.3amol、さらに好ましくは0.03~15.6amol、なお好ましくは0.1~7.8amol、さらになお好ましくは0.3~3.9amolである；ゲノムサイズ約3Gbpのマウスの場合、サンプルDNAの1Mbpあたり、好ましくは0.0003~1.7amol、より好ましくは0.0007~1.7amol、さらに好ましくは0.002~1.7amol、なお好ましくは0.005~0.83amol、さらになお好ましくは0.01~0.42amol、さらになお好ましくは0.03~0.21amol、である。なお、本願明細書において、初期DNA量は増幅工程に使用するDNAサンプル中のDNA量であり、プライマーなどのDNA量は含まない。

【0092】

ただし、上記のサンプルDNA 1Mbpあたりの初期DNA量の範囲は、上述したライブラリ調製での1本鎖特異的ヌクレアーゼ処理における該ヌクレアーゼのユニット数に依存し得る。例えば、該1本鎖特異的ヌクレアーゼ処理で0.05U/ng以下のS1 nucleaseを用いる場合、断片の誤認識への影響が十分に小さいので、初期DNA量の適切な範囲は上記のとおりである。

一方、該1本鎖特異的ヌクレアーゼ処理で0.05U/ngより大きいユニット数でS1 nucleaseを用いる場合、ユニット数の増加に伴い断片の誤認識率が増加し得る。そのため、該ヌクレアーゼ処理での反応液中におけるS1 nucleaseのユニット数(U/ng)に応じて初期DNA量を設定することが望ましい。S1 nucleaseのユニット数(>0.05U/ng)と初期DNA量の適切な条件は、下記の式よ

り算出される指標で表され、

$$\text{指標} = \text{初期DNA量 (amol / Mbp)} \times 3^{\log S1 \text{ nuclease (U/ng)}}$$

(式中、 $S1 \text{ nuclease (U/ng)} > 0.05$ 、 \log は常用対数である。)

当該指標は、好ましくは60以下、より好ましくは30以下、さらに好ましくは15以下、さらにより好ましくは7.5以下である。

【0093】

例えば、該1本鎖特異的ヌクレアーゼ処理で0.05 U/ng以下のMBNを用いる場合、断片の誤認識への影響が十分に小さいので、初期DNA量の適切な範囲は上記のとおりである。

一方、該1本鎖特異的ヌクレアーゼ処理で0.05 U/ngより大きいユニット数でMBNを用いる場合、ユニット数の増加に伴い断片の誤認識率が増加し得る。そのため、該ヌクレアーゼ処理での反応液中におけるMBNのユニット数(U/ng)に応じて初期DNA量を設定することが望ましい。MBNのユニット数($> 0.05 \text{ U/ng}$)と初期DNA量の適切な条件は、下記の式より算出される指標で表され、

$$\text{指標} = \text{初期DNA量 (amol / Mbp)} \times 3^{\log \text{MBN (U/ng)}}$$

(式中、 $\text{MBN (U/ng)} > 0.05$ 、 \log は常用対数である。)

当該指標は、好ましくは60以下、より好ましくは30以下、さらに好ましくは15以下、さらにより好ましくは7.5以下である。

【0094】

一方、該1本鎖特異的ヌクレアーゼ処理でRecJfを用いる場合、ユニット数に関わらず断片の誤認識への影響は十分に小さいので、初期DNA量の適切な範囲は上記のとおりである。

【0095】

一方、解析に十分な量のデータ(配列情報)を確保するためには、本シーケンシング方法における初期DNA量は、好ましくは0.1 amol以上、より好ましくは1 amol以上、さらに好ましくは5 amol以上、なお好ましくは20 amol以上、さらになお好ましくは39 amol以上、さらになお好ましくは78 amol以上である。解析効率の観点からは、初期DNA量は、好ましくは100000 amol以下、より好ましくは20000 amol以下、さらに好ましくは5000 amol以下である。例えば、本シーケンシング方法における初期DNA量は、好ましくは0.1 ~ 100000 amol、より好ましくは1 ~ 100000 amol、さらに好ましくは5 ~ 100000 amol、なお好ましくは20 ~ 100000 amol、さらになお好ましくは20 ~ 20000 amol、さらになお好ましくは39 ~ 20000 amol、さらになお好ましくは78 ~ 20000 amol、さらになお好ましくは20 ~ 5000 amol、さらになお好ましくは39 ~ 5000 amol、さらになお好ましくは78 ~ 5000 amolである。

【0096】

初期DNA量に対して大きすぎる又は少なすぎるシーケンシングデータ量は、解析効率を低下させ得る。本シーケンシング方法におけるシーケンシングデータ量は、初期DNA量1 amolあたりのリードペア数又はリード配列数で、好ましくは 0.02×10^6 個(リード配列又はリードペアの塩基対量で4 Mbp、これはリード配列の長さ、又はリードペアに含まれるリード配列の合計長の平均が200 bpの場合の値であり、該リード配列長さ又はリードペアに含まれるリード配列の合計長の平均値に合わせて変化し得る値である、以下同)以上、より好ましくは 0.04×10^6 個(8 Mbp)以上、さらに好ましくは 0.08×10^6 個(16 Mbp)以上、なお好ましくは 0.16×10^6 個(32 Mbp)以上であり、かつ、好ましくは 10×10^6 個(2000 Mbp)以下、より好ましくは 5×10^6 個(1000 Mbp)以下、さらに好ましくは 2.5×10^6 個(500 Mbp)以下、なお好ましくは 2×10^6 個(400 Mbp)以下である。例えば、本シーケンシング方法におけるシーケンシングデータ量は、初期DNA量1 amolあたりのリードペア数又はリード配列数で、好ましくは $0.02 \sim 10 \times 10^6$ 個(4 ~ 2000 Mbp)、より好ましくは $0.04 \sim 5 \times 10^6$ 個(8 ~ 1000 Mbp)、さらに好ましくは0.

10

20

30

40

50

0.8 ~ 2.5 × 10⁶個 (16 ~ 500 Mbp)、なお好ましくは0.16 ~ 2 × 10⁶個 (32 ~ 400 Mbp) である。

【0097】

本シーケンシング方法において、最大の解析効率をもたらすリード配列又はリードペアのグループに含まれるリード配列又はリードペアの数の平均値は、初期DNA量及びシーケンシングデータ量にかかわらずほぼ一定である(特許文献4参照)。本シーケンシング方法における、リード配列のグループに含まれるリード配列数、又はリードペアのグループに含まれるリードペア数は、該グループ間の平均で、好ましくは1.05以上、より好ましくは1.1以上、さらに好ましくは1.2以上であり、なお好ましくは1.4以上であり、かつ、好ましくは30以下、より好ましくは20以下、さらに好ましくは10以下、なお好ましくは5以下である。例えば、本シーケンシング方法において、リード配列又はリードペアのグループに含まれるリード配列又はリードペアの数は、該グループ間の平均で、好ましくは1.05 ~ 30、より好ましくは1.1 ~ 20、さらに好ましくは1.2 ~ 10、なお好ましくは1.4 ~ 5である。

10

【0098】

適切なシーケンシングデータ量は、サンプルDNAのサイズに依存し得る。より大きなサイズのサンプルDNAに対しては、より多くの初期DNA量が必要となる。一方、サンプルDNAのサイズに対してシーケンシングデータ量が多すぎる場合、解析効率が低下する。本シーケンシング方法におけるシーケンシングデータ量は、サンプルDNAの1 Mbpあたりのリード配列又はリードペア数で、好ましくは0.05 × 10⁶個 (10 Mbp) 以上、より好ましくは0.1 × 10⁶個 (20 Mbp) 以上、さらに好ましくは0.2 × 10⁶個 (40 Mbp) 以上、なお好ましくは0.5 × 10⁶個 (100 Mbp) 以上、さらになお好ましくは1 × 10⁶個 (200 Mbp) 以上、さらになお好ましくは2 × 10⁶個 (0.4 Gbp) 以上であり、かつ、好ましくは1600 × 10⁶個 (320 Gbp) 以下、より好ましくは800 × 10⁶個 (160 Gbp) 以下、さらに好ましくは400 × 10⁶個 (80 Gbp) 以下、なお好ましくは200 × 10⁶個 (40 Gbp) 以下、さらになお好ましくは100 × 10⁶個 (20 Gbp) 以下、さらになお好ましくは50 × 10⁶個 (10 Gbp) 以下である。例えば、本シーケンシング方法におけるシーケンシングデータ量は、サンプルDNAの1 Mbpあたりリード配列又はリードペア数で、好ましくは0.05 ~ 1600 × 10⁶個 (0.01 ~ 320 Gbp)、より好ましくは0.1 ~ 800 × 10⁶個 (0.02 ~ 160 Gbp)、さらに好ましくは0.2 ~ 400 × 10⁶個 (0.04 ~ 80 Gbp)、なお好ましくは0.5 ~ 200 × 10⁶個 (0.1 ~ 40 Gbp)、さらになお好ましくは1 ~ 100 × 10⁶個 (0.2 ~ 20 Gbp)、さらになお好ましくは2 ~ 50 × 10⁶個 (0.4 ~ 10 Gbp) である。なお、哺乳動物由来ゲノムDNAなどのようにサンプルDNAのサイズが大きい場合で、かつサンプルDNAの配列全体に対しての配列データの網羅性が特に問題とならない場合、本シーケンシング方法におけるシーケンシングデータ量は、サンプルDNAの1 Mbpあたりのリード配列又はリードペア数で、0.05 × 10⁶個 (10 Mbp) 未満であってもよい。例えば、ゲノムサイズ約3 Gbpのマウスの場合のシーケンシングデータ量は、サンプルDNAの1 Mbpあたりのリード配列又はリードペア数で、好ましくは0.00003 ~ 16 × 10⁶個 (0.006 ~ 3200 Mbp)、より好ましくは0.00007 ~ 8 × 10⁶個 (0.014 ~ 16000 Mbp)、さらに好ましくは0.0001 ~ 4 × 10⁶個 (0.02 ~ 8000 Mbp)、なお好ましくは0.0003 ~ 2 × 10⁶個 (0.06 ~ 4000 Mbp)、さらになお好ましくは0.0005 ~ 1 × 10⁶個 (0.1 ~ 2000 Mbp)、さらになお好ましくは0.001 ~ 0.5 × 10⁶個 (0.2 ~ 1000 Mbp) である。

20

30

40

【0099】

サンプルDNAのサイズが小さすぎると、シーケンシング用のライブラリ中の配列の多様性が低下して断片の誤認識の確率が高くなることがある。本シーケンシング方法におけるサンプルDNAのサイズは、好ましくは10 kbp以上、より好ましくは100 kbp以上、さらに好ましくは1 Mbp以上、なお好ましくは4 Mbp以上であるが、サンプル

50

DNAの由来する生物のゲノムDNAのサイズ等に依存し得る。

【0100】

本シーケンシング方法の好ましい一実施形態においては、サンプルDNAのサイズは約5Mbpであり、PCR初期DNA量は、好ましくは10~1250amolであり、シーケンシングデータ量は、リード配列又はリードペア数で0.2~12500×10⁶個(0.04~2500Gbp)、好ましくは0.4~6250×10⁶個(0.08~1250Gbp)、より好ましくは0.8~3125×10⁶個(0.16~625Gbp)、さらに好ましくは1.6~2500×10⁶個(0.32~500Gbp)である。

より好ましくは、サンプルDNAのサイズは約5Mbpであり、PCR初期DNA量は20~625amolであり、シーケンシングデータ量は、リード配列又はリードペア数
10
で0.4~6250×10⁶個(0.08~1250Gbp)、好ましくは0.8~3125×10⁶個(0.16~625Gbp)、より好ましくは1.6~1563×10⁶個(0.32~313Gbp)、さらに好ましくは3.2~1250×10⁶個(0.64~250Gbp)である。

さらに好ましくは、サンプルDNAのサイズは約5Mbpであり、PCR初期DNA量は39~313amolであり、シーケンシングデータ量は、リード配列又はリードペア数で0.78~3130×10⁶個(0.156~626Gbp)、好ましくは1.56~1565×10⁶個(0.312~313Gbp)、より好ましくは3.12~783×10⁶個(0.624~157Gbp)、さらに好ましくは6.24~626×10⁶個(1.248~125Gbp)である。
20

本シーケンシング方法の別の好ましい一実施形態においては、サンプルDNAのサイズは約5Mbpであり、リード配列又はリードペアのグループあたりのリード配列又はリードペアの数は、該グループ間の平均で、1.05~30、好ましくは1.1~20、さらに好ましくは1.2~10、なお好ましくは1.4~5である。

上述したとおり、上記PCR初期DNA量は、ライブラリ調製での1本鎖特異的ヌクレアーゼ処理における該ヌクレアーゼのユニット数に依存し得る。

【0101】

本シーケンシング方法のさらに別の好ましい一実施形態においては、サンプルDNAのサイズは約3Gbpであり、PCR初期DNA量は、好ましくは10~5000amolであり、シーケンシングデータ量は、リード配列又はリードペア数で0.2~50000×10⁶個(0.04~10000Gbp)、好ましくは0.4~25000×10⁶個(0.08~5000Gbp)、より好ましくは0.8~12500×10⁶個(0.16~2500Gbp)、さらに好ましくは1.6~10000×10⁶個(0.32~2000Gbp)である。
30

より好ましくは、サンプルDNAのサイズは約3Gbpであり、PCR初期DNA量は20~2500amolであり、シーケンシングデータ量は、リード配列又はリードペア数で0.4~25000×10⁶個(0.08~5000Gbp)、好ましくは0.8~12500×10⁶個(0.16~2500Gbp)、より好ましくは1.6~6250×10⁶個(0.32~1250Gbp)、さらに好ましくは3.2~5000×10⁶個(0.64~1000Gbp)である。
40

さらに好ましくは、サンプルDNAのサイズは約3Gbpであり、PCR初期DNA量は39~1250amolであり、シーケンシングデータ量は、リード配列又はリードペア数で0.78~12500×10⁶個(0.156~2500Gbp)、好ましくは1.56~6250×10⁶個(0.312~1250Gbp)、より好ましくは3.12~3125×10⁶個(0.624~625Gbp)、さらに好ましくは6.24~2500×10⁶個(1.248~500Gbp)である。

本シーケンシング方法のなお別の好ましい一実施形態においては、サンプルDNAのサイズは約3Gbpであり、リード配列又はリードペアのグループあたりのリード配列又はリードペアの数は、該グループ間の平均で、1.05~30、好ましくは1.1~20、さらに好ましくは1.2~10、なお好ましくは1.4~5である。
50

上述したとおり、上記PCR初期DNA量は、ライブラリ調製での1本鎖特異的ヌクレアーゼ処理における該ヌクレアーゼのユニット数に依存し得る。

【0102】

サイズ約5MbpのサンプルDNAの例としては、サルモネラ属細菌のゲノム(約4.86Mbp)が挙げられる。サルモネラ属細菌の好ましい例としては、Ames試験に使用されるS. typhimurium LT-2株、TA100株、TA98株、TA1535株、TA1538株、TA1537株等が挙げられる。

【0103】

(5.シーケンシング方法の応用)

本発明のライブラリを用いたシーケンシングで得られた配列データは、DNA断片の1本鎖部分の酸化修飾等に起因するシーケンシングエラーが除外された高精度な配列データである。したがって、本発明のライブラリを用いたシーケンシングは、これに限定されないが、変異解析に応用することができる。より詳細には、例えば、ゲノムDNAの変異解析による、試験物質の遺伝毒性の評価や、生殖発生毒性等のその他毒性の評価、ゲノムDNAに対する経時変化、生活環境、遺伝的要素などの影響の評価、培養細胞の品質評価などに応用することができる。これらの応用においては、変異解析の対象であるゲノムDNAから本発明のライブラリを調製し、これをシーケンシングして配列データを取得する。次いで、得られた配列データを用いて変異解析を行い、解析対象ゲノムDNAの変異を検出する。

【0104】

したがって、本発明はまた、ゲノムDNAの変異を検出する方法を提供する。当該方法は、細胞中のゲノムDNAをサンプルDNAとして用いて、本発明によるシーケンシング用ライブラリの調製方法によりシーケンシング用ライブラリを調製すること、該シーケンシング用ライブラリをシーケンシングすること、を含む。該シーケンシングにより、該ゲノムDNAについての配列データが作成される。該配列データを参照配列と比較して、該配列データと該参照配列とで塩基がマッチしない部位を変異部位として検出することで、該ゲノムDNAの変異を検出することができる。

【0105】

一実施形態において、本発明によるゲノムDNAの変異を検出する方法は、試験物質の遺伝毒性の評価に利用される。本実施形態では、該ゲノムDNAは、試験物質に暴露した細胞のゲノムDNAである。好ましくは、該ゲノムDNAは、試験物質に暴露した細胞(被験細胞)のゲノムDNAと、該試験物質に暴露していない細胞(対照細胞)のゲノムDNAである。好ましくは、これらのゲノムDNAは新鮮なDNAである。該新鮮なDNAは、好ましくはDINが6以上のDNAであり、より好ましくはDINが7以上のDNAであり、より好ましくはDINが7.3以上のDNA、さらにより好ましくはDINが7.5以上のDNAである。本実施形態では、該被験細胞のゲノムDNAについて検出した変異と、該対照細胞のゲノムDNAについて検出した変異とが比較される。例えば、該被験細胞でのみ検出された変異を、試験物質の暴露により生じた変異として同定することができる。本実施形態において使用される細胞は、特に限定されず、微生物細胞、動物細胞、植物細胞を含み得る。動物の例としては、好ましくは哺乳動物、鳥類、カイコ、線虫などが挙げられ、微生物の例としては、大腸菌、サルモネラ菌、酵母などが挙げられるが、これらに限定されない。本実施形態において使用される細胞の好ましい例としては、サルモネラ属細胞、及び大腸菌細胞が挙げられるが、これらに限定されない。サルモネラ属細胞の好ましい例としては、Ames試験に使用されるSalmonella typhimurium LT-2株、TA100株、TA98株、TA1535株、TA1538株、TA1537株等が挙げられる。大腸菌の好ましい例としては、分子生物学研究で汎用されるK-12株や、Ames試験に使用されるWP2株、WP2uvrA株等が挙げられる。本実施形態において使用される細胞の別の好ましい例としては、生体から採取した哺乳動物細胞、及び哺乳動物由来培養細胞が挙げられる。哺乳動物の好ましい例としては、マウス、ラット、ハムスター、チャイニーズハムスター、ウサギ、ヒトなどが挙げ

10

20

30

40

50

られ、このうちマウス及びヒトが好ましい。本実施形態において使用される細胞の別の好ましい例としては、生体から採取した鳥類細胞、及び鳥類由来培養細胞が挙げられる。鳥類の好ましい例としては、ニワトリが挙げられ、鳥類由来培養細胞の例としてはDT40などが挙げられる。

【0106】

該試験物質の例としては、その遺伝毒性を評価したい物質であれば特に制限されない。例えば、遺伝毒性を有すると疑われる物質、又は遺伝毒性の有無を確認したい物質、どのような変異を誘発するかを調べたい物質などが挙げられる。試験物質は、天然に存在する物質であっても、化学的もしくは生物学的方法等で人工的に合成した物質であってもよく、又は化合物であっても、組成物もしくは混合物であってもよい。あるいは、該試験物質は、紫外線や放射線などであってもよい。細胞を試験物質に暴露する手段は、試験物質の種類に応じて適宜選択すればよく、特に限定されない。例えば、細胞を含む培地に試験物質を添加する方法、細胞を試験物質の存在する雰囲気下に置く方法などが挙げられる。

10

【0107】

別の一実施形態において、本発明によるゲノムDNAの変異を検出する方法は、ゲノムDNAに対する経時変化、生活環境、遺伝的要素などの影響の評価に利用される。経時変化としては、細胞や個体の成長、加齢、老化、継代培養などが挙げられ、生活環境としては、食生活、運動などの生活習慣、居住地などが挙げられ、遺伝的要素としては、性別、種、特定の遺伝子の欠損や塩基対置換などが挙げられるが、これらに限定されない。本実施形態の好適な例は、ゲノムDNAに対する経時変化の影響の評価であり、該ゲノムDNAには、経時変化した細胞のゲノムDNAが用いられる。より好ましくは、該ゲノムDNAは、経時変化した細胞（被験細胞）のゲノムDNAと、より経時変化していない細胞（対照細胞）のゲノムDNAである。好ましくは、これらのゲノムDNAは新鮮なDNAである。該新鮮なDNAは、好ましくはDINが6以上のDNAであり、より好ましくはDINが7以上のDNAであり、より好ましくはDINが7.3以上のDNA、さらにより好ましくはDINが7.5以上のDNAである。本実施形態では、該被験細胞のゲノムDNAについて検出した変異と、該対照細胞のゲノムDNAについて検出した変異とが比較される。該対照細胞として用いられる、より経時変化していない細胞としては、成長、加齢、老化又は継代培養の程度が被験細胞より少ない細胞（例えば、より若い細胞、老化処理していない細胞、継代していないか継代数の少ない細胞など）が挙げられる。例えば、該被験細胞でのみ検出された変異を、経時変化により生じた変異として同定することができる。本実施形態において使用される細胞の好ましい例としては、生体から採取した哺乳動物細胞、及び哺乳動物由来培養細胞が挙げられる。哺乳動物の好ましい例としては、上述したとおりである。

20

30

【0108】

別の一実施形態において、本発明によるゲノムDNAの変異を検出する方法は、培養細胞の品質評価に利用される。本実施形態で用いられる該ゲノムDNAは、変異の有無を調べたい培養細胞のゲノムDNAであればよい。該変異の有無を調べたい培養細胞の例としては、ある一定期間培養した細胞であって、その変異の傾向を確認したいものが挙げられる。好ましくは、該ゲノムDNAは、該変異の有無を調べたい培養細胞（被験細胞）のゲノムDNAと、対照細胞のゲノムDNAである。対照細胞としては、例えば、同じ種類の培養細胞であって、遺伝情報既知の（例えば変異の有無及びその変異タイプが確認されている）細胞が用いられる。好ましくは、これらのゲノムDNAは新鮮なDNAである。該新鮮なDNAは、好ましくはDINが6以上のDNAであり、より好ましくはDINが7以上のDNAであり、より好ましくはDINが7.3以上のDNA、さらにより好ましくはDINが7.5以上のDNAである。本実施形態では、該被験細胞のゲノムDNAについて検出した変異と、該対照細胞のゲノムDNAについて検出した変異とが比較される。例えば、該被験細胞でのみ検出された変異を、培養中に生じた変異として同定することができる。

40

【0109】

50

本発明によるゲノムDNAの変異を検出する方法で検出される変異としては、塩基対置換型変異、及び短い挿入/欠失変異が挙げられる。塩基対置換型変異とは、DNAの塩基対情報を別の塩基対に変化させる変異であり、例えば、1塩基対置換型変異、及び2塩基対又は3塩基対以上が置換した多塩基対置換型変異を含む。本発明では、好ましくは1塩基対置換型変異が検出される。一方、短い挿入/欠失変異とは、DNAの配列中に短い塩基配列の挿入又は欠失を引き起こす変異であり、好ましくは挿入又は欠失した塩基の長さが10bp以下、より好ましくは1~5bpの挿入又は欠失変異をいう。

【0110】

塩基対置換型変異、及び短い挿入/欠失変異の検出は、WO/2018/150513 (その全体を本明細書に援用する)に記載の順序に従って実施することができる。その例として、以下に、解析対象ゲノムDNAにおける1塩基対置換型変異のパターンを検出する場合の好ましい手順を記載する。塩基対置換型変異の検出においては、シーケンシングで取得された配列データが参照配列と比較され、該配列データと該参照配列とで塩基がマッチしない部位が変異部位として検出される。検出された部位は、塩基対置換型変異を有する変異部位として取得される。本発明においては、変異解析の目的に応じて、該参照配列との比較に、取得された配列データの一部を用いてもよく、又は全部を用いてもよい。

10

【0111】

次いで、検出した変異部位の塩基と変異前の塩基の種類に基づいて、各変異を塩基の変異パターンに従って分類する。さらに、該塩基の変異パターンの各々について、出現頻度を決定することができる。これらの手順は、Python等のプログラミング言語を用いて作成したプログラム等を用いて実施することができる。

20

【0112】

より詳細な例においては、配列データに含まれる各塩基を、下記(i)~(iv)に分ける。

- (i) 参照配列上の塩基がAである位置に存在する塩基
- (ii) 参照配列上の塩基がTである位置に存在する塩基
- (iii) 参照配列上の塩基がGである位置に存在する塩基
- (iv) 参照配列上の塩基がCである位置に存在する塩基

上記(i)及び(ii)は、参照配列の塩基対がATであった部位に存在する塩基であり、上記(iii)及び(iv)は、参照配列の塩基対がGCであった部位に存在する塩基である。これらの塩基の中から、参照配列と塩基がマッチしない(すなわち塩基対置換変異している)ものを検出する。次いで、検出された変異部位の各々について、参照配列と配列データの配列情報に基づいて変異前及び後の塩基対を求める。これらのデータから、各変異を、変異前の塩基対がATであった場合について[AT TA、AT CG、及びAT GC]の3パターン、変異前の塩基対がGCであった場合について[GC TA、GC CG、及びGC AT]の3パターンの、全部で6つの塩基対の変異パターンに分類することができる。さらに、各変異パターンに属する変異の総数、及び解析した塩基の総数に基づいて、各変異パターンの出現頻度を決定することができる。例えば、AT、GC塩基対それぞれについての解析した塩基の総数に基づいて、各々の塩基対ごとに3種類の変異パターンの出現頻度を算出することができる。

30

【0113】

さらに、上記の各変異パターンを、変異検出の際にリード配列がマッピングされた参照配列上の塩基によってさらに2パターンに分類することができる。例えば、変異パターンがGC TAの変異であれば、参照配列上のG上でTが検出される場合とC上でAが検出される場合に分けられる。これらをそれぞれGからTへの変異(G T)、CからAへの変異(C A)と定義する。したがって、G T及びC Aに分けて変異頻度を算出することができる。AT TA、AT CG、AT GC、GC CG、及びGC ATについても同様である。2本鎖DNAに固定された真の変異ならば、これら2パターンの変異頻度は同等になる。一方、これら2パターン間で変異頻度に偏りが認められる場合、リード配列の由来するサンプルDNAの2本の鎖間で変異頻度が異なることを意味し、この変異は、酸化修飾等による塩基の変異に起因するエラーである可能性が高い。したがっ

40

50

て、上記のような2パターンへの分類は、シーケンシングエラーの検出に利用することができる。

【0114】

本発明においては、多塩基対置換型変異を解析することもできる。多塩基対置換型変異としては、例えば、2塩基対置換型変異及び3塩基対置換型変異が挙げられる。多塩基対置換型変異の解析の場合には、例えば、変異前の塩基配列に応じて変異パターンを分類し（例えば2塩基対置換型においては $4 \times 4 = 16$ 通り）、次いで、各変異パターンに属する変異の総数、及び解析した変異の総数に基づいて、各変異パターンの出現頻度を決定することができる。

【0115】

本発明においては、1塩基対置換型変異のシーケンスコンテキスト解析を行うこともできる。この解析では、上記手順で1塩基対置換型変異を検出した後、検出した各変異について、参照配列に基づいて、変異前の塩基と、該変異前の塩基の上流及び下流に隣接する塩基とを含む配列（いわゆるコンテキスト）を決定する。続いて、各変異を、塩基対の変異パターン及び該コンテキストに従ってタイプ分けする。すなわち、検出した変異を、上述した手順で6つの塩基対の変異パターン[AT TA, AT CG, AT GC, GC TA, GC CG、及びGC AT]に分ける。一方で、検出した各変異を、コンテキストに従って分類する。例えば、変異部位の両隣の1塩基ずつを含めた3塩基長のコンテキストは、 4×4 の16群[例えば、Cからの変異の場合、ACA、ACC、ACG、ACT、CCA、CCC、CCG、CCT、GCA、GCC、GCG、GCT、TCA、TCC、TCG、及びTCT]に分類される。結果、各変異は、塩基対の変異パターンとコンテキストに従って、全部で96（ $4 \times 6 \times 4$ ）のタイプに分類される。さらに長いコンテキストを解析することも可能である。例えば、変異部位の両隣の2塩基ずつを含めた5塩基長のコンテキストに従うと、各変異は256群（ $4 \times 4 \times 4 \times 4$ ）に分類され、この分類と6つの塩基対パターンにより、各変異は最終的に全部で1536（ $4 \times 4 \times 6 \times 4 \times 4$ ）のタイプに分類される。さらに変異部位の両隣のn塩基ずつを含めた $2n + 1$ 塩基長のコンテキストに従うと、各変異は 4^{2n} 群に分類され、この分類と6つの塩基対パターンにより、各変異は最終的に全部で $4^{2n} \times 6$ 個のタイプに分類される。次いで、各変異タイプに属する変異の総数、及び解析した塩基の総数に基づいて、上記変異タイプの各々の変異頻度を決定することができる。

【0116】

次に、解析対象ゲノムDNAにおける短い挿入/欠失変異を検出する場合の好ましい手順を記載する。短い挿入/欠失変異の検出においては、配列データをそれぞれ参照配列と比較することによって、各配列データにおける該参照配列に対して塩基が挿入又は欠失されている部位を検出する。該参照配列との比較には、取得された配列データの一部を用いてもよく、又は全部を用いてもよい。検出される挿入又は欠失部位としては、好ましくは挿入又は欠失した塩基の長さが10bp以下、より好ましくは1~5bpである部位がよいが、これに限定されない。検出された部位は、挿入又は欠失変異を有する変異部位として取得される。

【0117】

さらに、取得された各変異について、変異のタイプ（挿入変異か又は欠失変異か）、該挿入又は欠失部位の塩基長、あるいは挿入又は欠失した塩基の種類を決定することができる。特定の塩基長の挿入又は欠失部位を検出する手順は、上述したPython等のプログラミング言語を用いて作成したプログラムを用いて行うことができる。さらに、各配列データと参照配列との比較によって、挿入又は欠失した塩基の種類を同定することができる。これらにより、各配列データにおける挿入又は欠失部位の塩基長、あるいは挿入又は欠失部位の塩基の種類を決定することができる。さらに、挿入又は欠失の頻度を、塩基長及び/又は塩基の種類ごとに決定してもよい。例えば、各リード配列について取得した挿入又は欠失変異を塩基長ごとに分類し、それぞれの頻度を決定することができる。また例えば、挿入又は欠失した塩基をその種類（A、T、G、及びC）ごとに分類し、それぞれ

10

20

30

40

50

の頻度を決定することができる。さらに、該塩基長及び塩基の種類による分類を組み合わせたより細かい変異の分類を行い、それぞれの頻度を決定することができる。

【0118】

本発明の例示的实施形態として、さらに以下の物質、製造方法、用途、方法等を本明細書に開示する。ただし、本発明はこれらの実施形態に限定されない。

【0119】

〔1〕シーケンシング用ライブラリの調製方法であって、

サンプルDNAを断片化すること；及び、

調製したサンプルDNAの断片を1本鎖特異的ヌクレアーゼで処理し、該断片から1本鎖部分を除去すること、

を含む、

方法。

〔2〕前記サンプルDNAが、

好ましくはホルマリン固定細胞のDNA又はcfDNAではなく、より好ましくは、生細胞から抽出したDNA、凍結細胞から抽出したDNA、又はそれらのDNAの保存サンプルであり、かつ

好ましくは、DINが6以上、さらに好ましくは7以上、さらに好ましくは7.3以上、さらにより好ましくは7.5以上である、

〔1〕記載の方法。

〔3〕前記1本鎖特異的ヌクレアーゼが、

好ましくは、1本鎖特異的エンドヌクレアーゼ、1本鎖特異的エキソヌクレアーゼ、又はそれらの組み合わせであり、

より好ましくは、S1 nuclease、Mung Bean Nuclease (MBN)、RecJf、及びExonuclease VIIからなる群より選択される少なくとも1種である、

〔1〕又は〔2〕記載の方法。

〔4〕好ましくは、前記1本鎖特異的ヌクレアーゼでの処理が、前記サンプルDNAの断片を1本鎖特異的エンドヌクレアーゼで処理した後に、さらに1本鎖特異的エキソヌクレアーゼで処理することを含むか、又は1本鎖特異的エキソヌクレアーゼで処理した後に、さらに1本鎖特異的エンドヌクレアーゼで処理することを含む、〔3〕記載の方法。

〔5〕好ましくは、前記1本鎖特異的エンドヌクレアーゼがS1 nucleaseであり、

前記サンプルDNAの断片1ng当たりのS1 nucleaseのユニット数(U/ng)が、

好ましくは0.01U/ng以上、より好ましくは0.02U/ng以上、さらに好ましくは0.05U/ng以上であり、かつ好ましくは16.7U/ng以下、より好ましくは5.00U/ng以下、さらに好ましくは1.67U/ng以下であるか、又は、

好ましくは0.02~5.00U/ng、より好ましくは0.05~1.67U/ngである、〔3〕又は〔4〕記載の方法。

〔6〕好ましくは、前記1本鎖特異的エンドヌクレアーゼがMBNであり、

前記サンプルDNAの断片1ng当たりのMBNのユニット数(U/ng)が、

好ましくは0.01U/ng以上、より好ましくは0.02U/ng以上、さらに好ましくは0.03U/ng以上、さらに好ましくは0.05U/ng以上、さらに好ましくは0.10U/ng以上であり、かつ好ましくは16.7U/ng以下、より好ましくは5.00U/ng以下、さらに好ましくは1.67U/ng以下、さらに好ましくは1.00U/ng以下、さらに好ましくは0.30U/ng以下であるか、又は、

好ましくは0.02~5.00U/ng、より好ましくは0.03~1.67U/ng、さらに好ましくは0.03~1.00U/ng、さらに好ましくは0.05~1.00U/ng、さらに好ましくは0.10~0.30U/ngである、

〔3〕又は〔4〕記載の方法。

10

20

30

40

50

〔 7 〕 好ましくは、前記 1 本鎖特異的エキソヌクラーゼが $RecJ_f$ であり、
前記サンプル DNA の断片 1 ng 当たりの $RecJ_f$ のユニット数 (U / ng) が、
好ましくは $0.10 U / \text{ng}$ 以上、より好ましくは $0.30 U / \text{ng}$ 以上であり、かつ好ましくは $100 U / \text{ng}$ 以下、より好ましくは $16.7 U / \text{ng}$ 以下、さらに好ましくは $1.00 U / \text{ng}$ 以下であるか、又は、

好ましくは $0.10 \sim 16.7 U / \text{ng}$ 、より好ましくは $0.30 \sim 1.00 U / \text{ng}$ である、

〔 3 〕 又は〔 4 〕 記載の方法。

〔 8 〕 好ましくは、前記 1 本鎖特異的ヌクラーゼで処理した前記サンプル DNA の断片を、末端修復、末端への塩基付加、及び増幅からなる群より選択されるいずれか 1 つ以上の処理に供することをさらに含み、

10

より好ましくは、前記 1 本鎖特異的ヌクラーゼで処理した前記サンプル DNA の断片を、末端修復、末端への塩基付加、及び増幅に供することをさらに含む、

〔 1 〕 ~ 〔 7 〕 のいずれか 1 項記載の方法。

〔 9 〕 好ましくは、前記末端への塩基付加が、前記サンプル DNA の断片の両末端への標識配列の付加である、〔 8 〕 記載の方法。

〔 10 〕 好ましくは、前記増幅が PCR である、〔 8 〕 又は〔 9 〕 記載の方法。

〔 11 〕 前記 1 本鎖特異的ヌクラーゼが $S1 \text{ nuclease}$ であり、前記サンプル DNA の断片 1 ng 当たりの該ヌクラーゼのユニット数 (U / ng) が $0.05 U / \text{ng}$ 以下のとき、前記 PCR における該サンプル DNA 1 Mbp 当たりの初期 DNA 量が、好ましくは 250 amol 以下、より好ましくは 125 amol 以下、さらに好ましくは 62.5 amol 以下、さらにより好ましくは 31.3 amol 以下、なお好ましくは 15.7 amol であるか；

20

前記 1 本鎖特異的ヌクラーゼが $S1 \text{ nuclease}$ であり、前記サンプル DNA の断片 1 ng 当たりの該ヌクラーゼのユニット数 (U / ng) が $0.05 U / \text{ng}$ より大きいとき、下記式で算出される指標：

指標 = PCR における初期 DNA 量 (amol / Mbp サンプル DNA) $\times 3^{\log S1 \text{ nuclease } (U/\text{ng})}$

(式中、 $S1 \text{ nuclease } (U / \text{ng}) > 0.05$ 、 \log は常用対数である)

が、好ましくは 60 以下、より好ましくは 30 以下、さらに好ましくは 15 以下、さらにより好ましくは 7.5 以下であるか；

30

前記 1 本鎖特異的ヌクラーゼが MBN であり、前記サンプル DNA の断片 1 ng 当たりの該ヌクラーゼのユニット数 (U / ng) が $0.05 U / \text{ng}$ 以下のとき、前記 PCR における該サンプル DNA 1 Mbp 当たりの初期 DNA 量が、好ましくは 250 amol 以下、より好ましくは 125 amol 以下、さらに好ましくは 62.5 amol 以下、さらにより好ましくは 31.3 amol 以下、なお好ましくは 15.7 amol であるか；

前記 1 本鎖特異的ヌクラーゼが MBN であり、前記サンプル DNA の断片 1 ng 当たりの該ヌクラーゼのユニット数 (U / ng) が $0.05 U / \text{ng}$ より大きいとき、下記式で算出される指標：

40

指標 = PCR における初期 DNA 量 (amol / Mbp サンプル DNA) $\times 3^{\log MBN (U/\text{ng})}$

(式中、 $MBN (U / \text{ng}) > 0.05$ 、 \log は常用対数である)

が、好ましくは 60 以下、より好ましくは 30 以下、さらに好ましくは 15 以下、さらにより好ましくは 7.5 以下である、

〔 10 〕 記載の方法。

【 0120 】

〔 12 〕 前記〔 1 〕 ~ 〔 11 〕 のいずれか 1 項記載の方法で調製されたシーケンシング用ライブラリをシーケンシングすることを含む、DNA のシーケンシング方法。

〔 13 〕 好ましくは、前記シーケンシング方法が、以下：

50

(1) 前記ライブラリをシーケンシングし、該ライブラリに含まれる複数の増幅断片の各々について1つ以上のリード配列を作成し、該複数の増幅断片についての複数のリード配列を得ること；

(2) 得られた複数のリード配列の中から、該ライブラリの調製に用いたサンプルDNA上の同一領域の配列情報を有するリード配列を集めてグループ化することにより、リード配列のグループを1つ以上作成すること；及び、

(3) 該リード配列のグループに含まれるリード配列の間で配列情報のコンセンサスを取ること、

を含む、〔 1 2 〕記載の方法。

〔 1 4 〕好ましくは、前記(1)が、前記サンプルDNAの断片を構成する2本の相補鎖の各々に由来する増幅断片に対して1つ以上のリード配列を作成することを含む、〔 1 3 〕記載の方法。

〔 1 5 〕好ましくは、前記(2)が、参照配列上の同一の位置にマッピングされるリード配列を同じグループに分けることを含む、〔 1 4 〕記載の方法。

〔 1 6 〕好ましくは、前記(3)が、前記リード配列のグループの中から、前記サンプルDNA断片の2本の相補鎖の各々に由来するリード配列を少なくとも1つずつ集め、集めたリード配列の間で配列情報のコンセンサスを取ることを含む、〔 1 5 〕記載の方法。

〔 1 7 〕好ましくは、

前記(1)において、前記複数のリード配列が、以下からなるリード配列のペアを複数個含む：

リード1：前記増幅断片を構成する2本の相補鎖のうち一方の鎖の配列を5'末端側から3'側へ読んだ配列に相当する配列情報を含むリード配列、

リード2：該一方の鎖の配列を3'末端側から5'側へ読んだ配列に相当する配列情報を含むリード配列、

前記(2)が、得られたリード配列のペアの中から、該サンプルDNA上の同一領域の配列情報を有するリード配列のペアを集めてグループ化することにより、リード配列のペアのグループを1つ以上作成することを含む、

前記(3)が、該リード配列のペアのグループに含まれるリード配列の間で配列情報のコンセンサスを取ることを含む、

〔 1 3 〕記載の方法。

〔 1 8 〕好ましくは、前記(1)が、前記サンプルDNAの断片を構成する2本の相補鎖の各々に由来する増幅断片に対して1つ以上の前記リード配列のペアを作成することを含む、〔 1 7 〕記載の方法。

〔 1 9 〕好ましくは、前記(2)が、前記リード配列のペアのリード1とリード2を参照配列に対してマッピングし、リード1の先頭とリード2の先頭とに挟まれる該参照配列の領域が同一であるリード配列のペアを同じグループに分けることを含む、〔 1 8 〕記載の方法。

〔 2 0 〕好ましくは、前記(2)が、前記リード配列のペアに含まれる一方のリード配列の先頭が前記参照配列上の同じ位置に位置するリード配列のペアを集め、次いで集めたリード配列のペアの中から、該リード配列のペアに含まれるもう一方のリード配列の先頭が該参照配列上の同じ位置に位置するリード配列のペアを集めて、集めたリード配列のペアを同じグループに分けることを含む、〔 1 8 〕記載の方法。

〔 2 1 〕好ましくは、前記(3)が、前記リード配列のペアのグループの中から、前記サンプルDNA断片の2本の相補鎖の各々に由来するリード配列のペアを少なくとも1組ずつ集め、集めたリード配列のペアに含まれるリード配列の間で配列情報のコンセンサスを取ることを含む、〔 1 9 〕又は〔 2 0 〕記載の方法。

【 0 1 2 1 〕

〔 2 2 〕ゲノムDNAをサンプルDNAとして用いて、前記〔 1 〕～〔 1 1 〕のいずれか1項記載の方法によりシーケンシング用ライブラリを調製すること；及び、

該シーケンシング用ライブラリをシーケンシングすること、

10

20

30

40

50

を含む、ゲノムDNAの変異を検出する方法。

〔23〕好ましくは、前記シーケンシングが前記〔13〕～〔21〕のいずれか1項記載の方法により行われる、〔22〕記載の方法。

〔24〕好ましくは、前記変異が塩基対置換型変異である、〔22〕又は〔23〕記載の方法。

【実施例】

【0122】

以下、実施例を示し、本発明をより具体的に説明する。

【0123】

参考例1 シーケンシング及び変異解析

後述の比較例及び実施例で用いたシーケンシング方法及び変異解析のフローを以下に説明する。基本的には、特許文献4に記載される相補鎖情報を活用した高精度シーケンシング法を用いた。具体的には、ライブラリをシーケンシングし、同一のDNA断片に由来すると推定されるリードペアを集めた。次いで、該DNA断片の2本の相補鎖（以下、A鎖及びB鎖と称する）のそれぞれに由来すると推定されるリード配列間でのコンセンサスリード配列（相補鎖間コンセンサスリード配列）を作成した。得られた相補鎖間コンセンサスリード配列は変異解析に使用した。

【0124】

1) 相補鎖情報を活用したシーケンシング

シーケンサーにはイルミナ社のHiSeqを用いた。HiSeqシーケンサー用のライブラリには、サンプルDNA断片の2本の相補鎖の双方に由来するPCR産物が含まれる。したがって、このライブラリをシーケンシングすることで、該2本の相補鎖のそれぞれについてリード1とリード2を作成した。

【0125】

互いに相補的な鎖のリード配列を識別するため、PCRの前に、サンプルDNA断片の両末端に、相補鎖標識配列（イルミナ社のTruSeqに付属のアダプター配列）を連結した。次いで、該アダプター配列に特異的に結合するプライマーを用いたPCRにより、該アダプター配列を含むPCR産物を生成し、シーケンシング用のライブラリとして用いた。HiSeqシーケンサーにおいては、該アダプター配列がシーケンシングに使用されるフローセル上のオリゴDNA断片とアニーリングすることで、該フローセル上に増幅産物が結合され、シーケンシングされる。

【0126】

シーケンシングでは、ライブラリ中の各PCR産物に含まれる個々の増幅断片（サンプルDNA断片に由来する）に対して、それぞれ2本のリード配列（リード1、リード2）のペアが取得された。このとき、該増幅断片の一方の鎖の配列を5'側から3'側へ読んだ配列情報を含むリード配列がリード1（R1）であり、同じ鎖の配列を3'側から5'側へ読んだ配列に相当する配列情報を含むリード配列がリード2（R2）であった。

【0127】

2) リード配列の編集、及び相補鎖情報の抽出

1)で得られたリード配列を、アダプター配列及びクオリティの低い塩基等のトリミングを行った後、参照配列へマッピングした。サンプルDNA断片の2本の相補鎖由来のリードペアを参照配列上にマッピングしたときの、参照配列に対する各リードペアの配置の概念図を模式図1に示す。参考のため、模式図1には、各リードペアが由来するサンプルDNA断片の2本の相補鎖を図示する。互いに相補的な鎖に由来するリードペアの間では、リード1の先頭とリード2の先頭とに挟まれる参照配列の領域は同一である。したがって参照配列上でのリードペアのマッピング位置に基づいて、同じサンプルDNA断片に由来すると考えられるリードペアを集めた。

【0128】

10

20

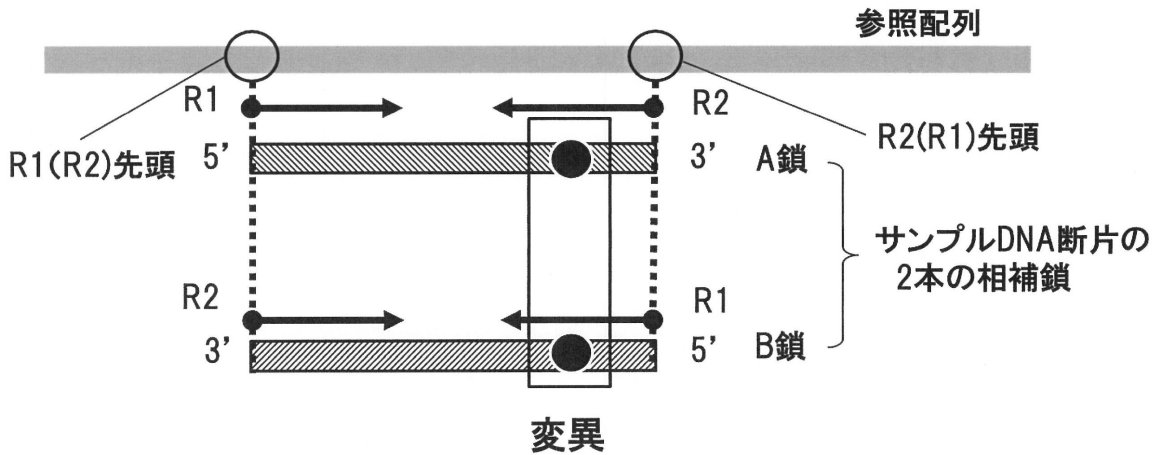
30

40

50

【化3】

模式図1: 参照配列に対する相補鎖由来リードペアのマッピングの概念図。



10

【0129】

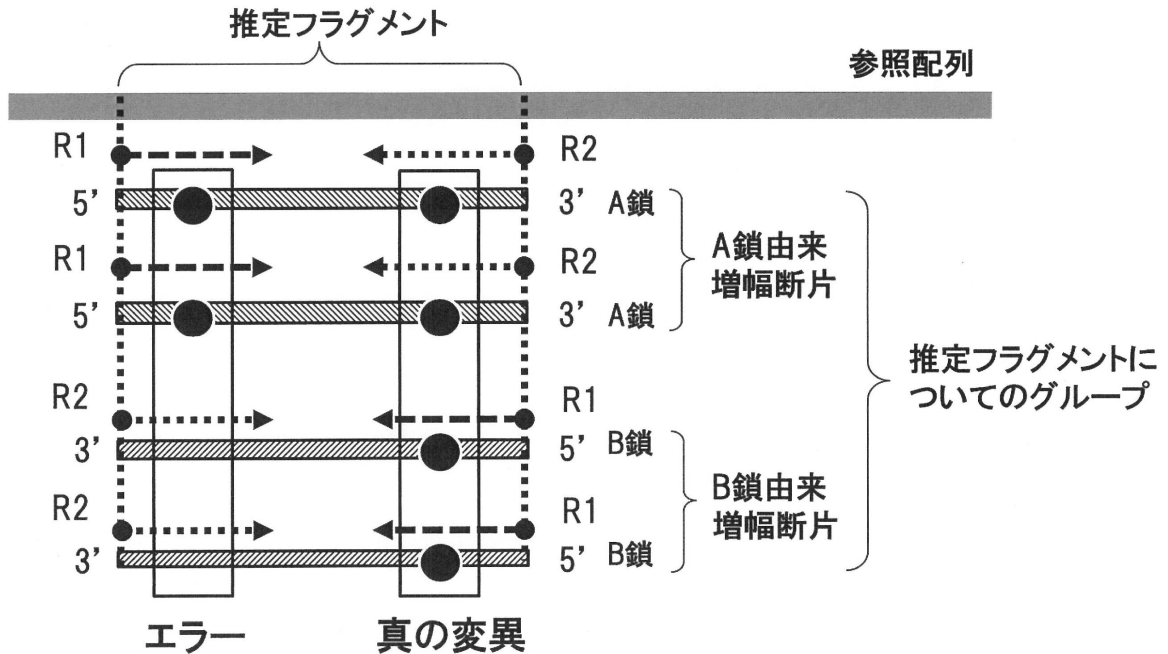
なお本方法に関し、参照配列上における、マッピングしたリード1（リード2）の先頭からリード2（リード1）の先頭までの領域、言い換えると、リードペア（リード1、リード2）を参照配列上にマッピングしたときに、リード1の先頭とリード2の先頭とに挟まれる該参照配列の領域を、「推定フラグメント」と称する。推定フラグメントが共通するリードペアの群を、推定フラグメントについての「グループ」と称する（模式図2）。

20

【0130】

【化4】

模式図2: 推定フラグメント



30

40

【0131】

次いで、推定フラグメントについてのグループから、互いに相補的な2本の鎖のそれぞれに由来するリードペアの組み合わせを、リードペアのセットとして取得した。

【0132】

50

サンプルDNA断片から得られた増幅断片は、サンプルDNA断片に元々含まれる変異を両鎖に保有するのに加えて、片方の鎖のみに、サンプルDNA断片の酸化修飾などに起因する塩基の置換を有することがある。このようなケースを模式図1、2に例示する。模式図1に示すサンプルDNA断片は、変異による塩基の置換（真の変異）を両鎖に1つずつ保有する。一方、模式図2に示した該サンプルDNA断片由来の増幅断片は、変異による塩基の置換（真の変異）を両鎖に保有するのに加え、片方の鎖のみにサンプル調製過程で生じた塩基の置換（エラー）を有する。これらの真の変異及びエラーは、各リードペアのリード1とリード2に読み取られている。本方法では、相補鎖に由来するリードペアのセットの有する配列情報から、両鎖に固定された真の変異と片方の鎖のみに生じたエラーとを区別し、真の変異を抽出した。

10

【0133】

本方法では、集めたリードペアのセットから相補鎖間コンセンサスリード配列を作成した。相補鎖間コンセンサスリード配列の作成においては、まず、推定フラグメントの共通するリードペアを集め、それらをA鎖由来のリードペアとB鎖由来のリードペアとに分けた。次いで、1つ以上のA鎖由来のリードペアと1つ以上のB鎖由来のリードペアとの組み合わせをリードペアのセットとして取得し、それらを用いて相補鎖間コンセンサスリード配列を作成した。リードペアのセットに含まれるA鎖由来又はB鎖由来のリードペアの数は特に限定されず、A鎖由来とB鎖由来双方のリードペアが少なくとも1つ以上含まれていれば良いとした。例えば、A鎖由来のリードペアが2つで、B鎖由来のリードペアが2つの場合や、A鎖由来のリードペアが3つで、B鎖由来のリードペアが1つの場合でも、それらの間でコンセンサスを取ることで相補鎖間コンセンサスリード配列を作成した。

20

【0134】

リードペアの集合化から相補鎖間コンセンサスリード配列作成までのより具体的な手順の例を、以下の模式図3に示す。模式図3のとおり、本方法では、まず、各相補鎖由来のリードペアを参照配列にマッピングした(1)。このとき、参照配列上で左端(参照配列上の最も5'側に配置する端)が同じ位置にあるリードペアの群を第一集合として取得した(2)。次いで、該第一集合から、参照配列上で右端(参照配列上の最も3'側に配置する端)が同じ位置にあるリードペアの群を分け、第二集合として取得した(3)。この第二集合は、推定フラグメントの共通するリードペアの集合であった。次いで、第二集合を、A鎖に由来する群(F群)と、B鎖に由来する群(R群)とに分けた(4)。このとき、A鎖に由来する群であるかB鎖に由来する群であるかは、シーケンシングの際に取得される標識配列の情報に基づいて識別することができた。本方法においては、サンプルDNA断片に付加されたアダプター配列中の標識配列を認識し結合するフローセルを用いてシーケンシング反応を行った。フローセル内での断片の増幅後、5'側に付加されたアダプター配列中の標識配列を特異的に切断することにより、各増幅断片のリード1、リード2のシーケンシングの方向性を統一することで、標識配列の情報に基づいてリードペアをF群とR群とに分けた。該F群とR群は、それぞれ、DNA断片を構成する2本の相補鎖のいずれか一方に由来するリードペアの集合であった。したがって、該F群とR群との間でコンセンサスを取ることで、相補鎖間コンセンサスリード配列を作成した(5)。

30

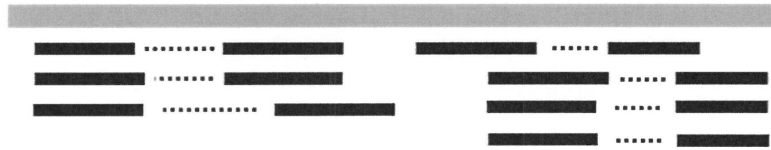
【0135】

40

【化5】

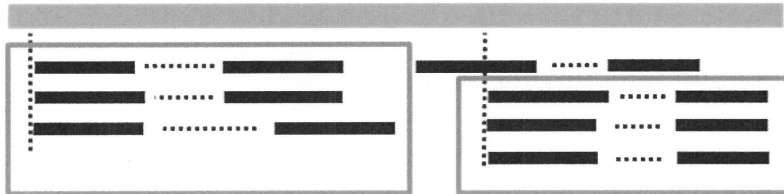
模式図3: リードペアからの相補鎖間コンセンサスリード配列作成手順

(1) 参照配列へのリードペアのマッピング



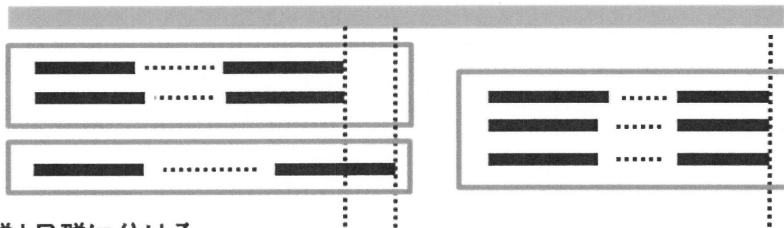
(2) 左端の位置が同じペアを第一集合として取得

10



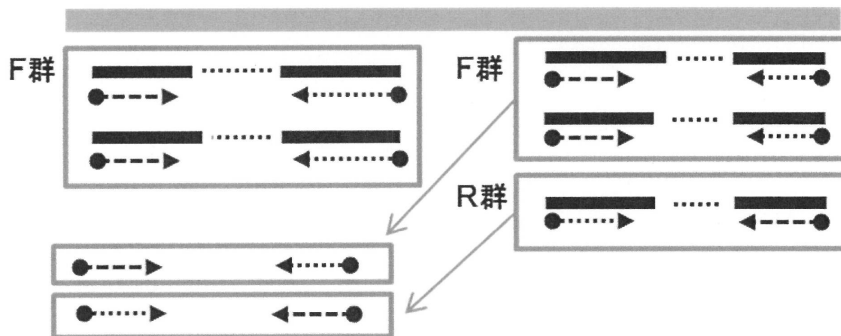
(3) 右端の位置が同じペアを第二集合として取得

20



(4) F群とR群に分ける

30



(5) F群とR群からコンセンサス配列を作成



【0136】

相補鎖間コンセンサスリード配列を作成することにより、片方の鎖にのみ生じた置換はエラーとして除外し、両方の鎖に共通して存在する置換を真の変異として取得した。

40

【0137】

3) 変異解析

2) で得られた相補鎖間コンセンサスリード配列を参照配列上に再度マッピングすることで、解析対象ゲノムの変異を検出した。参照配列に再マッピングした相補鎖間コンセンサスリード配列から変異した塩基を検出するための具体的な手順は、PCT/JP2017/005700に記載された手順に従った。

【0138】

4) ソフトウェア、プログラム

リード配列の編集、相補鎖情報の抽出、及び変異解析のフローを模式図4に示す。解析

50

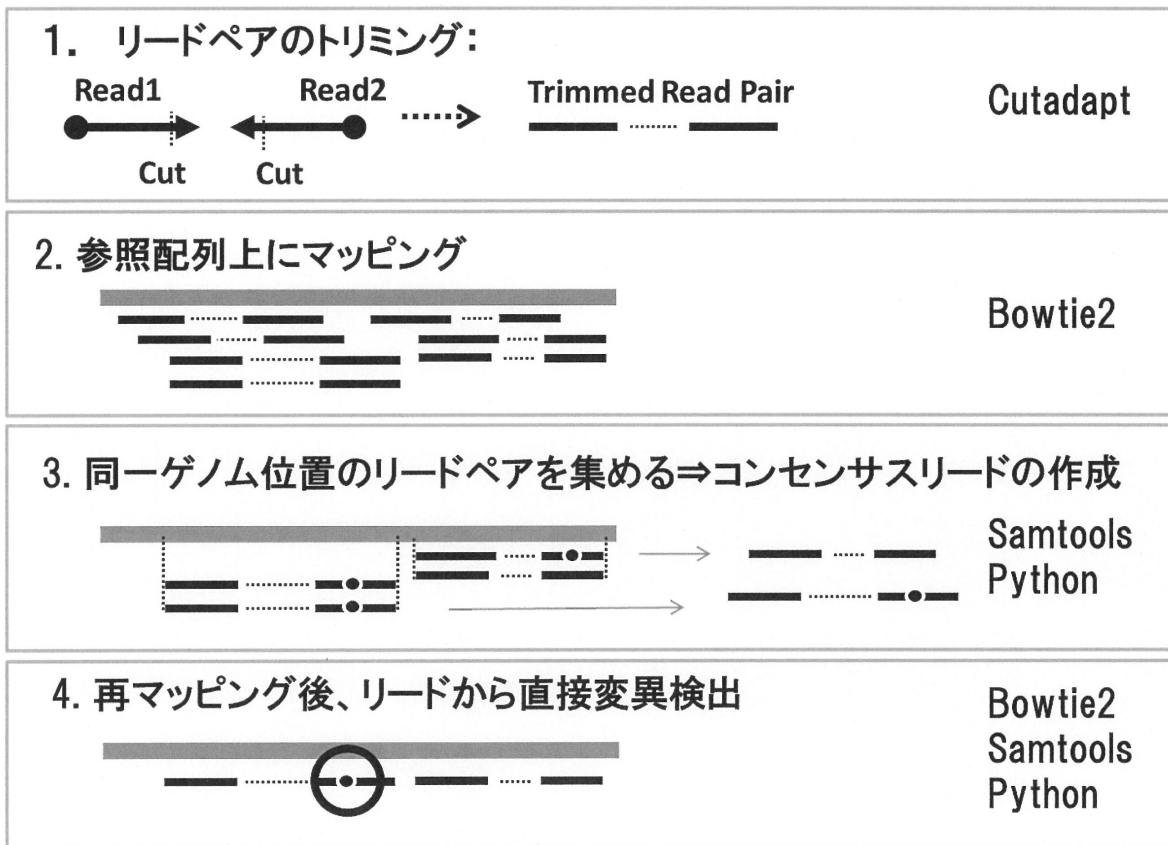
には、Cutadaptソフトウェア、Bowtie2ソフトウェア、Samtoolsソフトウェア、及びプログラミング言語Pythonを用いて作成したプログラムを用いた。まず、各ライブラリ由来のFastqファイル(リード1、及びリード2)に対して、Cutadaptソフトウェアを用いて、アダプター配列及びクオリティの低い塩基等のトリミングを行った。その後、各ライブラリ由来のFastqファイルを、Bowtie2ソフトウェアを用いて参照配列へマッピングし、Samフォーマットのファイルを得た。Samtoolsソフトウェアを用いてSamフォーマットのファイルのリードの並び替えを行い、次いで、プログラミング言語Pythonで作成したプログラムを用いて、推定フラグメントについてのグループを作成し、その中からリードペアのセットを集め、相補鎖間コンセンサスリード配列を作成した。得られた相補鎖間コンセンサスリード配列を、再度Bowtie2ソフトウェアで参照配列にマッピングし、Samtoolsソフトウェア、及び、プログラミング言語Pythonで作成したプログラムを用いて、変異解析を行った。

10

【0139】

【化6】

模式図4: 解析フロー



20

30

40

【0140】

比較例1 リードペアの両端の塩基の除去によるエラー低減

参考例1のシーケンシング法を用いて、新鮮なゲノムDNAの断片における末端1本鎖突出部位に由来するエラーの存在を検証した。また、末端部エラーに対する既存の改善法であるリードペアの両端から塩基を除く方法によるエラーの低減効果を検討した。

【0141】

1) サンプルDNA

サンプルDNAとして、ジメチルスルホキシド(DMSO; 和光純薬工業製)を暴露したSalmonella typhimurium LT-2 TA100株(以下、単に

50

「TA100株」とも称する)のゲノムDNAを用いた。

【0142】

TA100細胞株のDMSOへの暴露は、Ames試験のブレインキュベーション法に準拠して実施した(Mol. Mech. Mutagen., 455:29-60, 2000, Sci. Rep. 8(1):9583)。2mLのニュートリエントブイオンNo. 2(Oxoid社製)にTA100株を植菌し、37、180rpmで4時間振とう培養し、OD660値が1.0以上の前培養液を得た。試験管内に、DMSO100 μ L、S9 mix(家田貿易社製)500 μ L、及び前培養液100 μ Lを添加し、37のウォーターバス中で20分間、100rpmで振とう培養した(DMSO暴露細胞)。20分間の振とう培養後、培養液を含む試験管をウォーターバスから取り出し、予め分注しておいた2mLのNutrient Broth溶液(S9 mixを18.5%含む)に培養液50 μ Lを添加し、インキュベーター内で37、180rpmで14時間追培養した。培養後、菌懸濁液を回収し、7500rpmで5分間遠心し、上清を除去して細胞を回収した。DMSO暴露細胞から、DNeasy Blood & Tissue Kit(キアゲン社製)を用い、推奨プロトコルに従って、Total DNAを回収した。得られたDNAサンプルの2本鎖DNAの濃度を、Qubit 3.0 Fluorometer(Thermo Fisher Scientific社製)を用いて、付属のQubitTM dsDNA BR Assay Kitで測定した。

10

【0143】

2)シーケンシング用ライブラリの調製

20

サンプルDNAからのライブラリ調製には、TruSeq Nano DNA Library Prep Kit(イルミナ社製、以下TruSeqと略記する)を用いた。TruSeqの推奨プロトコルは、DNAの断片化、End Repair(2本鎖DNA断片の1本鎖突出末端の平滑化)、A-tailing(2本鎖DNA断片の3'末端へのアデニンの付加)、Adapter ligation(2本鎖DNA断片両末端へのアダプターの付加)、及びPCR enrichment(PCR増幅によるライブラリDNAの濃縮)から構成される。1)で得たDMSO暴露細胞由来DNAの120ng相当量を複数サンプル用意し、それらをDNA Shearingシステム ME220(コバリス社製)で推奨プロトコルに従って平均約350bpの長さで断片化した。得られた断片化DNAに、End Repair、A-tailing、Adaptor Ligationを実施した。得られたAdaptor Ligationの反応液を推奨プロトコルに従って精製し、2本鎖DNA断片の両末端にアダプターが付加されたDNA(アダプター付加DNA)を得た。Agilent 4200 TapeStation(アジレント・テクノロジー社製)のHigh Sensitivity D5000キットを用いてアダプター付加DNAの濃度を測定した。

30

【0144】

特許文献4に基づいて、PCRに用いるアダプター付加DNAの初期量(初期DNA量)の最適条件を78amol(15.6amol/Mbp)と推定した。これに従い、TruSeqに付属のResuspension bufferで段階的に希釈し、78amolのアダプター付加DNAを含む希釈液25 μ Lを得た。得られた希釈液を、推奨プロトコルに従いPCR enrichmentに供した。78amolの初期DNA量とシーケンシングに必要なDNA量を考慮して、15サイクルのPCRを実施した。反応液から推奨プロトコルに従ってDNAを精製し、ライブラリとした。Agilent 4200 TapeStationのHigh Sensitivity D1000キットを用いてライブラリDNAの濃度を測定した。

40

【0145】

3)シーケンシング及び変異解析

2)で調製したライブラリを、2x100bpのリード長でシーケンシングし、ライブラリあたり、平均で約10Gbp(約50Mリードペア)のシーケンシングデータを得た。得られたシーケンシングデータから相補鎖間コンセンサスリード配列を作成し、参照配

50

列にマッピングした後、変異した塩基を検出した。シーケンシング、相補鎖間コンセンサスリード配列の作成、及び変異解析は参考例 1 の手順に従って実施した。なお、参照配列には、GenBank (www.ncbi.nlm.nih.gov/genbank/) から取得した *Styphimurium* LT-2 株 (以下、単に LT-2 株とも略記する。) のゲノム配列を用いた (GenBank assembly accession: GCA_000006945.2)。

【0146】

4) 変異頻度の算出

Python で作成したプログラムを用いて、各ライブラリについて、参照配列にマッピングされた全相補鎖間コンセンサスリード配列中の全解析対象塩基を、対応する参照配列の塩基 (A、T、G、及び C) によって 4 群に分けた。そして、各群の塩基の総数と参照配列に対して変異した塩基を検出した。検出された変異を、6 つの変異パターン (ATTA、ATCG、ATGC、及び GCTA、GCCG、GCAT) に分類し、各変異パターンにおける変異頻度を算出した。さらに、各変異パターンを、リード配列がマッピングされた参照配列上の塩基によって、さらに 2 パターンの変異に分類して、各々の変異頻度を算出した。すなわち、ATTA は A T 及び T A に、ATCG は A C 及び T G に、ATGC は A G 及び T C に、GCTA は G T 及び C A に、GCCG は G C 及び C G に、GCAT は G A 及び C T に分類して、これら 12 種の変異パターンそれぞれについて変異頻度を算出した。

【0147】

5) リードペアの両端からの塩基の除去によるエラー低減

参考例 1 の手順に従って、3) で得た相補鎖間コンセンサスリード配列を再度参照配列にマッピングして Sam フォーマットのファイルを作成した。該 Sam フォーマットファイル中で、リードペアの両端の 0 塩基 (control)、10 塩基、又は 20 塩基を、Python で作成したプログラムを用いてクオリティ値を下げることで、変異解析の対象から除外した。その後、参考例 1 の手順に従って変異解析を行った。変異頻度は、4) に示した 12 種の変異パターンについて算出した。

【0148】

6) 結果と考察

4) で算出したサンプル DNA における 6 つの変異パターンについての変異頻度を図 1 に示す。AT 塩基対の変異頻度に比べて GC 塩基対の変異頻度が大きいことから、グアニンの酸化修飾によるエラーの存在が推測された。また、5) で算出した両末端を除去したリードペアから求めた 12 種の変異パターンについての変異頻度を図 2 に示す。GC 塩基対の変異 (GCTA、GCCG) において、CA、CG に比べて、GT、GC の変異が高頻度に検出された。真の変異は、G、C の両塩基で同等の頻度で検出されるはずである。高頻度のグアニンの変異が検出されたことは、これが真の変異ではなく、酸化修飾等による塩基の変異に起因するエラーであることを示唆する。また、GT、GC の変異頻度は、リードペアの両端から除去した塩基数に依存して減少した。この結果は、該グアニンの変異によるエラーがリードペアの両端部に多く存在していることを示した。したがって、DNA 断片の末端 1 本鎖部位における酸化修飾等によるグアニンの変異が、該エラーの主な原因となっていると考えられた。

【0149】

次いで GCTA 及び GCCG の変異について、G の変異と C の変異の間での変異頻度の差を算出し、下記式に基づいて、リードペアの両端除去によるエラーの減少率を求めた。

$$\text{エラー減少率 (\%)} = (A - B) / A \times 100$$

A : 両端の塩基を除去しないとき (control) の GC 間の変異頻度の差

B : 両端から塩基を除いたときの GC 間の変異頻度の差

エラー減少率を表 1 に示す。エラー減少率は、10 塩基の除去で < 30% であり、20 塩基の除去でも 40% 程度であった。なお Kennedy ら (非特許文献 3) が報告した

両端から5塩基除く方法は、10塩基除くよりもさらにエラー低減効果が小さいと推測された。これらの結果は、両端20塩基の除去ではDNA断片の末端1本鎖突出部分を十分に削除できなかったことを表す。除去する塩基数を増加することによりエラーをより低減できると予想されるが、リードペアからの多数の塩基の削除は、変異解析に充てられる塩基数が減少するため解析効率を低下させる。結果、DNA断片の末端1本鎖部位における酸化修飾等に起因するエラーの改善にとって、リードペアの両端の変異解析対象からの除去は有効なアプローチとは言えない。

【0150】

【表1】

変異パターン	エラー減少率(%)	
	両端部10塩基除去	両端部20塩基除去
GC to TA	27.5	40.2
GC to CG	26.2	40.0

10

【0151】

実施例1 1本鎖特異的ヌクレアーゼを用いたライブラリ調製法によるエラー低減

DNA断片の1本鎖特異的ヌクレアーゼ処理によるエラー低減効果を評価した。

【0152】

1) サンプルDNA

比較例1の1)と同様の手順で、DMSO暴露細胞を調製した、また同様の手順で、TA100株を3-Methylcholanthrene(3-MC)に暴露した。3-MC(シグマアルドリッチ社製、CASRN.56-49-5)は、DMSOに溶解した。試験管内に、3-MC溶液100 μ L、S9mix(家田貿易社製)500 μ L、及びTA100株の前培養液100 μ Lを添加し(3-MC量:1000 μ g/tube)、37 $^{\circ}$ Cのウォーターバス中で20分間、100rpmで振とう培養した(3-MC暴露細胞)。比較例1の1)と同様の手順で菌懸濁液から細胞を回収し、DNAを抽出した。

20

【0153】

2) Ames試験

Ames試験用に、上記と同様の条件で3-MCを暴露した菌懸濁液を調製した。これに、45 $^{\circ}$ Cに加温した2mLのtop agar(1%NaCl、1%agar、0.05mM Histidine及び0.05mM Biotinを含む)を添加し、ボルテックスで攪拌した後、最小グルコース寒天培地(テスメディア(登録商標)AN;オリエンタル酵母工業製)の上に重層した。得られたプレートを37 $^{\circ}$ Cで48時間培養後、観察されたコロニーを計数した。

30

【0154】

3) シーケンシング用ライブラリの調製

I) サンプルDNAの断片化

DMSO暴露細胞又は3-MC暴露細胞由来DNAの60ng又は100ng相当量を複数サンプル用意し、それらをDNA Shearingシステム ME220で平均約350bpの長さに断片化した。各サンプルの断片を2群に分けた。ヌクレアーゼで処理しない群(非処理群)については、次の工程のEnd Repairを行うために、推奨プロトコルに従って、TruSeqに付属のResuspension bufferでDNA断片を懸濁し、60 μ Lの溶出液を得た。ヌクレアーゼで処理する群(処理群)については、DNA断片をTruSeqに付属のSample Purification Beads(以下、単にビーズとも略記する)に吸着させ、80%エタノール水で2回洗浄し、乾燥させるステップを推奨プロトコルに従って行い、精製した。その後の溶出操作では、Distilled water(DW、ニッポンジーン社製)でビーズを懸濁し、30 μ LのDNA断片を含むDNA溶出液を得た。

40

【0155】

50

II)ヌクレアーゼ処理

1本鎖特異的ヌクレアーゼには、S1 nuclease (プロメガ社、カタログ番号：M5761)、Mung Bean Nuclease (MBN) (タカラバイオ社、カタログ番号：2420A)、又はRecJf (New England Biolabs社、カタログ番号：M0264L)を用いた。各酵素の活性値(ユニット数)は以下の通り定義した。

・S1 nuclease：30mM酢酸ナトリウム(pH4.6、25℃)、50mM NaCl、1mM ZnCl₂、5%グリセロール、0.5mg/mL変性仔牛胸腺DNAの混合溶液中において、37℃で1分間に1μgの酸可溶性物質を生成する酵素活性を1Uとした。

・MBN：熱変性仔牛胸腺DNAを基質として、37℃、pH5.0において、1分間に1μgの酸可溶性分解物を生成する酵素活性を1Uとした。

・RecJf：全反応液50μL(1×NE Buffer 2及び1.5μgの超音波処理[³H]標識1本鎖E. coli DNAを含む)中、37℃、1分間で、0.5ngのトリクロロ酢酸可溶性デオキシリボヌクレオチドを生成するために必要な酵素量を1Uとした。

【0156】

II-1)S1 nuclease処理

I)で得られたDNA溶出液に、S1 nucleaseに付属の10×Reaction Bufferを4μL添加した。1×Reaction BufferでS1 nucleaseを適宜希釈し、1、3、10、30、100、300UをDNA溶出液に添加し、DWを添加し、全量を40μLとした。S1 nucleaseを1000U添加するサンプルについては、ビーズ精製で得られた30μLのDNA溶出液に4.6μLの10×Reaction Bufferを添加し、S1 nucleaseの原液を12μL加えて全量を46μLとした。S1 nucleaseを加えた反応液を攪拌し、30℃で30分間インキュベートした。反応液中にライブラリ調製開始時のDNA全量(60ng)が存在すると考えると、1ng当たりのDNAに対するS1 nucleaseのユニット数はそれぞれ、0.02、0.05、0.17、0.50、1.67、5.00、16.7U/ngであった。反応液中のS1 nucleaseの失活のために、0.5M EDTA(pH8.0)(ニッポンジーン社製)を3μL添加し、70℃で10分間インキュベートした。失活させた反応液からDNAを精製するため、反応液と等量のTruSeqに付属のビーズを添加し、推奨プロトコルに従って精製操作を進め、TruSeqに付属のResuspension bufferで懸濁し、60μLの溶出液を得た(S1 nuclease処理群)。

【0157】

II-2)MBN処理

I)で得られたDNA溶出液に、MBNに付属の10×Mung Bean Nuclease Bufferを5μL添加した。1×Mung Bean Nuclease BufferでMBNを適宜希釈し、3、10、30、100UをDNA溶出液に添加し、全量を50μLとした。MBNを加えた反応液を攪拌し、37℃で10分間インキュベートした。酵素反応液中のMBNの失活のために、0.5M EDTA(pH8.0)を3μL添加し、65℃で10分間インキュベートした。反応液中に100ngのDNA断片が存在すると考えると、1ng当たりのDNAに対するユニット数はそれぞれ、0.03、0.1、0.3、1.0U/ngであった。失活させた反応液からDNAを精製するため、反応液と等量のTruSeqに付属のビーズを添加し、推奨プロトコルに従って精製操作を進め、TruSeqに付属のResuspension bufferで懸濁し、60μLの溶出液を得た(MBN処理群)。

【0158】

II-3)RecJf処理

I)で得られたDNA溶出液に、RecJfに付属の10×NE Buffer 2を5

10

20

30

40

50

μL 添加した。1 × NE Buffer 2 で Rec J_f を適宜希釈し、3、10、30、100 ユニットを DNA 溶出液に添加し、全量を 50 μL とした。Rec J_f を加えた反応液を攪拌し、37 °C で 60 分間インキュベートした。反応液中に 100 ng の DNA 断片が存在すると考えると、1 ng 当たりの DNA に対するユニット数はそれぞれ、0.03、0.1、0.3、1.0 U/ng であった。酵素反応液中の Rec J_f の失活のために、65 °C で 20 分間インキュベートした。失活させた反応液から DNA を精製するため、反応液と等量の TruSeq に付属のビーズを添加し、推奨プロトコルに従って精製操作を進め、TruSeq に付属の Resuspension buffer で懸濁し、60 μL の溶出液を得た (Rec J_f 処理群)。

【0159】

III) End Repair、A-tailing、Adaptor Ligation 及び PCR enrichment

II) で得られた非処理群、S1 nuclease 処理群、MBN 処理群、及び Rec J_f 処理群に、比較例 1 の 2) と同様の手順で、TruSeq の推奨プロトコルに従って End Repair、A-tailing、Adaptor Ligation を実施した。得られた Adaptor Ligation の反応液を推奨プロトコルに従って精製し、2 本鎖 DNA 断片の両末端にアダプターが付加された DNA (アダプター付加 DNA) を得た。Agilent 4200 TapeStation (アジレント・テクノロジー社製) の High Sensitivity D5000 キットを用いてアダプター付加 DNA の濃度を測定した。次いで、比較例 1 の 2) と同様の手順で PCR enrichment を実施し、ライブラリを得た。

【0160】

4) シーケンシング及び変異解析

3) で調製したライブラリを、2 × 150 bp のリード長でシーケンシングし、ライブラリあたり、平均で約 15 Gbp (約 50 Mリードペア) のシーケンシングデータを得た。得られたシーケンシングデータから、相補鎖間コンセンサスリード配列の作成、及び変異検出を実施した。シーケンシング、相補鎖間コンセンサスリード配列の作成、及び変異解析は参考例 1 の手順に従って実施した。

【0161】

5) 変異頻度の算出

比較例 1 の 4) と同様の手順で、6 つの変異パターン及び 12 種の変異パターンについて変異頻度を算出した。次いで GC TA 及び GC CG の変異について、G の置換と C の置換の間の変異頻度の差を算出し、下記式に基づいて、ヌクレアーゼ処理群でのエラーの減少率を求めた。

$$\text{エラー減少率 (\%)} = (A - B) / A \times 100$$

A : 非処理群 (0 U/ng) での GC 間の変異頻度の差

B : 各ユニット数でのヌクレアーゼ処理群での GC 間の変異頻度の差

【0162】

6) 解析効率の算出

変異解析の際に用いた各ライブラリの相補鎖間コンセンサスリード配列中のリードペア数 (本) と各ライブラリのシーケンシングで読み取ったリードペアの総数 (シーケンシングデータ量) (本) から、各ライブラリの解析効率を算出した。

$$\text{解析効率 (\%)} = (\text{相補鎖間コンセンサスリード配列中のリードペア数}) / (\text{シーケンシングデータ量}) \times 100$$

【0163】

7) グループあたりの平均リードペア数

4) で作成した相補鎖間コンセンサスリード配列について、推定フラグメントについてのグループあたりのリードペア数を計数し、リードペア数が等しいグループの数を集計して、平均リードペア数を算出した。

$$\text{平均リードペア数} = \{ \sum_i (i \times (\text{i本のリードペアを含むグループ数})) \} / (\text{グル}$$

10

20

30

40

50

ープの総数)

(i はグループに含まれるリードペアの本数を指す。)

【 0 1 6 4 】

8) 結果と考察

I) Ames 試験の復帰突然変異体数

表 2 に 3 - M C 暴露後の復帰突然変異体コロニー数を示す。データは 3 枚のプレートでの測定値と、その平均値を示す。3 - M C 暴露により復帰突然変異体コロニー数の増加が認められたことから、3 - M C 暴露により T A 1 0 0 株のゲノム中に変異が導入されたことが確認された。

【 0 1 6 5 】

【表 2】

Ames試験における復帰変異体コロニー数

3-MC濃度 (μ g/plate)	復帰突然変異体数			
	1	2	3	平均
0 (DMSO)	112	93	97	101
10	887	881	1279	1016
20	1539	1484	1509	1511
100	1519	1539	1797	1618

10

【 0 1 6 6 】

II) 1 本鎖特異的ヌクレアーゼによるエラー低減効果

II - 1) S 1 n u c l e a s e

DMSO 暴露ライブラリにおける 6 変異パターンの変異頻度を図 3 に示す。非処理群 (S 1 n u c l e a s e 0 U / n g) では、比較例 1 と同じように G C 塩基対の変異頻度が高かった。一方で、S 1 n u c l e a s e 処理群 (S 1 n u c l e a s e 0 . 2 ~ 1 6 . 7 U / n g) では、ユニット数依存的に変異頻度が減少し、0 . 1 7 U / n g でエラー低減効果が飽和した。続いて、同じライブラリでの 1 2 種類の変異パターンの頻度を図 4 ~ 5 に示す。非処理群では比較例 1 と同じように、C A、C G に比べて、G T、G C の変異を高頻度に検出した。そして、S 1 n u c l e a s e 処理群では、ユニット数の増加に伴って G T、G C の変異頻度が減少した。G C T A、G C C G についてのエラー減少率を表 3 に示す。0 . 1 7 U / n g 以上で変異頻度の減少が飽和し、G C 間の変異頻度の偏りが大きく改善された。これは、S 1 n u c l e a s e がサンプル DNA の断片中の 1 本鎖部位を特異的に分解し、該 1 本鎖部位に存在していた酸化修飾されたグアニンを除去したためと考えられた。0 . 1 7 U / n g 以上の S 1 n u c l e a s e 処理により、DNA 断片の末端 1 本鎖部位の塩基の酸化修飾に起因するエラーを取り除くことができることが確認された。

20

【 0 1 6 7 】

【表 3】

エラー減少率 (%)

S1 nuclease (U / ng 反応液中 DNA)	0.02	0.05	0.17	0.50	1.67	5.00	16.7	
変異パターン	GC to TA	52.8	35.3	83.1	69.9	81.7	84.5	79.6
	GC to CG	30.4	50.6	84.9	87.5	86.0	98.3	98.1

40

【 0 1 6 8 】

II - 2) M u n g B e a n N u c l e a s e

図 6 に M B N 処理時の DMSO 暴露ライブラリにおける 6 変異パターンの変異頻度を示した。M B N 処理群 (0 . 0 3 ~ 1 . 0 0 U / n g) において、ユニット数依存的に変異頻度が減少した。続いて、同じライブラリでの 1 2 種類の変異パターンの頻度を図 7 ~ 8 に示す。M B N 処理群では、G C の変異頻度が大きく減少し、G C、C G 間の変異

50

頻度の差が大きく減少した。G Tの変異頻度の減少は認められたが、S1 nucleaseと比較すると小さく、G T、C A間の変異頻度の差はユニット数が大きくなっても残っていた。GC TA、GC CGについてのエラー減少率を表4に示す。GC CGに関しては、0.03 U/ng以上でエラー低減効果があり、0.10 U/ng以上でGC間の変異頻度の偏りが大きく改善された。一方、GC TAに関しては、GC間の変異頻度の差は低減したものの、効果は小さかった。これは、DMSO暴露ライブラリにおけるGC TAの変異頻度がII-1で示した結果よりも低かったことが一因と考えられた。同一条件でDMSOを暴露して調製したDNA (n=3)におけるG T及びC Aの変異頻度の平均値はそれぞれ 0.177×10^{-6} 及び 0.042×10^{-6} であった。該平均値に対するエラー減少率は11.4% (0.03 U/ng)、40.2% (0.10 U/ng)、15.6% (0.30 U/ng)、57.8% (1.00 U/ng)となった。したがって、S1 nucleaseと比較すると小さいが、MBNのエラー低減効果は認められた。

【0169】

【表4】

エラー減少率 (%)

MBN (U/ng反応液中DNA)		0.03	0.10	0.30	1.00
変異パターン	GC to TA	-33.5	9.8	-27.2	36.4
	GC to CG	63.3	90.6	81.2	97.7

【0170】

II-3) RecJf

図9にRecJf処理時のDMSO暴露ライブラリにおける6変異パターンの変異頻度を示した。なお、非処理群の結果はMBN処理群と共通である。RecJf処理群(0.03~1.00 U/ng)において、ユニット数依存的に変異頻度が減少した。続いて、同じライブラリでの12種類の変異パターンの頻度を図10~11に示す。RecJf処理群では、G T、G Cの変異頻度の減少が認められ、G T、C A間及びG C、C G間の変異頻度の差も減少したが、S1 nucleaseと比較するとその効果は小さかった。GC TA、GC CGについてのエラー減少率を表5に示す。また、II-2の時と同様、DMSO暴露ライブラリのGC TAの変異頻度が低いことを考慮し、同一条件でDMSOを暴露して調製したDNA (n=3)におけるG T及びC Aの変異頻度の平均値と比較した。これらの平均値を用いて算出したエラー減少率は-10.8% (0.03 U/ng)、35.2% (0.10 U/ng)、54.1% (0.30 U/ng)、62.3% (1.00 U/ng)となった。したがって、RecJfは、GC TAに関してはMBNと同等のエラー低減効果を示し、GC CGに関しては、S1 nuclease、MBNと比較すると小さいが、エラー低減効果は認められた。また、GC TA、GC CGともに0.10 U/ng以上でエラー低減効果があると考えられた。

【0171】

【表5】

エラー減少率 (%)

RecJf (U/ng反応液中DNA)		0.03	0.10	0.30	1.00
変異パターン	GC to TA	-66.9	2.3	30.8	43.3
	GC to CG	-30.6	8.4	51.5	60.0

【0172】

III) 3-MCの変異頻度の上昇率の改善

III-1) S1 nuclease

10

20

30

40

50

5)の方法でDMSO暴露ライブラリ(DMSO control)、及び、3-MC暴露ライブラリ(3MC)における6変異パターンの変異頻度をS1 nucleaseのユニット数ごとに算出した結果を図12~13に示す。非処理群(control、0 U/ng)では、DMSO controlと比較した3-MCにおける変異頻度の明確な上昇はいずれの変異パターンにおいても検出されなかったが、S1 nuclease処理群では、3-MCでGC→TAの変異頻度の明確な増加が見られた。この変異パターンは、3-MCに暴露された遺伝子組換えマウスの肝臓で検出された変異パターンと一致していた(Environ. Mol. Mutagen., 2000, 36:266-273)。これらの結果は、S1 nuclease処理により1本鎖上のグアニン由来のシーケンシングエラーが減少した一方、真の変異は検出されたためと考えられた。表6に、DMSO controlに対する3-MCでのGC→TA変異頻度の上昇率(SN ratio)を示す。0.17 U/ng以上のS1 nuclease処理により、シーケンシングエラーが低減することで、変異原処理により誘発される低頻度な変異が検出可能になることが示唆された。

【0173】

【表6】

GC→TA変異頻度の上昇率(3MC/DMSO control)

S1 nuclease (U/ng反応液中DNA)	0	0.02	0.05	0.17	0.50	1.67	5.00	16.7
SN ratio	1.27	2.1	1.97	2.72	3.75	3.35	1.63	3.59

【0174】

III-2) Mung Bean Nuclease

III-1同様、MBN処理群における結果を図14に示す。MBN処理群では、S1 nuclease処理群と同様に、3-MCにおいてGC→TAの変異頻度の増加が見られた。表7に、DMSO controlに対する3-MCでのGC→TA変異頻度の上昇率(SN ratio)を示す。本実験での非処理群(0 U/ng)におけるSN ratioは、III-1に比べて高かった。これは、III-1に比べてDMSO controlのGC→TAの変異頻度が低く、3-MCでのGC→TAの変異頻度が大きいためであった。そこで、同一条件でMBN処理なしのDMSO control及び3-MC(それぞれn=3)を調製し、各々についてGC→TAの変異頻度の平均値を算出し、それらの平均値からSN ratioを求めた。その結果、DMSO control、及び3-MCのGC→TAの平均値はそれぞれ、 0.109×10^{-6} 、 0.176×10^{-6} となり、SN ratioは1.61となった。したがって、0.10 U/ng以上のMBNでSN ratioが改善することが推測された。

【0175】

【表7】

GC→TA変異頻度の上昇率(3MC/DMSO control)

MBN (U/ng反応液中DNA)	0	0.03	0.10	0.30	1.00
SN ratio	2.04	1.60	1.81	1.34	1.92

【0176】

III-3) RecJf

III-1同様、RecJf処理群における結果を図15に示す。S1 nuclease及びMBN処理群と同様に、RecJf処理群でも、3-MCにおいてGC→TAの変異頻度の増加が見られた。表8に、DMSO controlに対する3-MCでのGC→TA変異頻度の上昇率(SN ratio)を示す。III-2)で算出した変異頻度の平均値のSN ratio(1.61)を考慮すると、0.10 U/ng以上のRecJfにエラー低減効果があると考えられた。

【 0 1 7 7 】

【 表 8 】

GC→TA変異頻度の上昇率 (3MC/DMSO control)

RecJ _f (U / ng反応液中DNA)	0	0.03	0.10	0.30	1.00
SN ratio	2.04	1.48	2.14	1.68	1.94

【 0 1 7 8 】

IV) 解析効率と平均リードペア数

本実施例でのシーケンシングは最適条件と推定される初期DNA量78amolの条件(特許文献4参照)で実施されたが、ヌクレアーゼ処理がシーケンシング最適条件に影響を及ぼしている可能性がある。そこで、シーケンシングの解析効率と平均リードペア数(特許文献4)に基づいて、本実施例でのシーケンシングが最適条件下でなされたか否かを評価した。表9~11に各ユニット数のS1 nuclease、MBN、及びRecJ_fで処理したライブラリにおける解析効率と平均リードペア数の算出結果を示す。特許文献4で算出されたシーケンシングの最適条件は、解析効率が5~10%程度、平均リードペア数が約2本であり、本実施例でも近い結果が得られた。したがって、ヌクレアーゼ処理によるシーケンシング条件への影響は小さく、本実施例でもほぼ最適条件でシーケンシングが行われたと考えられた。

【 0 1 7 9 】

【 表 9 】

DMSO暴露ライブラリ

S1 nuclease (U / ng反応液中DNA)	0	0.02	0.05	0.17	0.50	1.67	5.00	16.7
解析効率 (%)	8.0	8.3	8.3	8.0	8.5	8.8	8.9	8.9
平均リードペア数 (本)	1.9	2.0	2.0	2.2	1.9	2.0	2.1	1.9

3-MC暴露ライブラリ

S1 nuclease (U / ng反応液中DNA)	0	0.02	0.05	0.17	0.50	1.67	5.00	16.7
解析効率 (%)	8.3	7.9	7.7	8.2	8.5	8.5	8.8	8.0
平均リードペア数 (本)	2.4	2.2	2.0	2.0	2.1	2.2	2.3	2.2

【 0 1 8 0 】

【 表 1 0 】

DMSO 暴露ライブラリ

MBN (U / ng反応液中DNA)	0	0.03	0.10	0.30	1.00
解析効率 (%)	8.2	8.8	8.6	8.9	9.0
平均リードペア数 (本)	2.4	2.4	2.1	2.2	2.3

3-MC 暴露ライブラリ

MBN (U / ng反応液中DNA)	0	0.03	0.10	0.30	1.00
解析効率 (%)	8.3	8.6	8.5	8.8	8.7
平均リードペア数 (本)	2.7	2.5	2.1	2.2	2.2

【 0 1 8 1 】

【表 1 1】

DMSO 暴露ライブラリ

RecJ _f (U / ng反応液中DNA)	0	0.03	0.10	0.30	1.00
解析効率 (%)	8.2	8.6	7.4	8.6	8.8
平均リードペア数 (本)	2.4	2.4	2.8	2.6	2.3

3-MC 暴露ライブラリ

RecJ _f (U / ng反応液中DNA)	0	0.03	0.10	0.30	1.00
解析効率 (%)	8.2	8.4	7.4	8.4	8.5
平均リードペア数 (本)	2.7	2.4	2.5	2.4	2.3

【0182】

実施例 2 変異解析に対する影響の評価

本実施例では、DNA断片の1本鎖特異的ヌクレアーゼ処理が変異解析に与える影響を評価するため、1)相補鎖間コンセンサスリード配列のLT-2株のゲノムに対する網羅性、及び、2)異なるDNA断片の同一断片としての誤認識(断片の誤認識)を調べた。実施例1で得られた各サンプルのリードペア、相補鎖間コンセンサスリード配列を用いた。1本鎖特異的ヌクレアーゼにはS1 nuclease、MBN、及びRecJ_fを用いた。

【0183】

1)相補鎖間コンセンサスリード配列のLT-2株のゲノムに対する網羅性

シーケンシングでのゲノム全体のカバレッジを調べ、ゲノムの特定の部位が特異的にシーケンシングされていないか評価した。DMSO暴露ライブラリの非処理群及びS1 nuclease処理群、MBN処理群、及びRecJ_f処理群の相補鎖間コンセンサスリード配列から各ゲノム位置におけるカバレッジの情報を抽出し、プログラミング言語Pythonで作成したプログラムにより、ゲノム領域をおよそ100塩基ごとに区切り、各領域におけるカバレッジを求め、正規化し(カバレッジの総和が1となる)、ヒストグラムを作成した。さらに、LT-2株のゲノムにマッピングした際のcovered rate(カバレッジが1以上になったゲノム位置の割合)、平均カバレッジ(mean coverage)、カバレッジの標準偏差(SD of coverage)、及び変動係数(CV)を算出した。

$$\text{変動係数 (CV) (\%)} = (\text{カバレッジの標準偏差}) / (\text{平均カバレッジ}) \times 100$$

【0184】

非処理群、及びシーケンシングエラー低減効果が明確に現れた0.17U/ng以上のS1 nuclease処理群でのカバレッジのヒストグラムを図16に示した。全データに共通してみられるゲノム位置800000から900000番目あたりのカバレッジがない部分は、TA100株におけるuvrB遺伝子の欠損部位である(J. Appl. Toxicol., 2017, 37: 1125-1128)。いずれのユニット数のS1 nuclease処理群においても、非処理群と比べてヒストグラムに大きな変化は見られなかった。表12上には、非処理群及びS1 nuclease処理群の相補鎖間コンセンサスリード配列をLT-2株のゲノムにマッピングした際の、各群でのcovered rate、mean coverage、SD of coverage、及びCVを示す。S1 nucleaseのユニット数が増加しても、covered rateやCVは非処理群と大きく変わらないことが確認された。また、非処理群、及び1.00U/ngでのMBN処理群及びRecJ_f処理群でのカバレッジのヒストグラムを図17~18に示した。MBN処理群、及びRecJ_f処理群のどちらにおいても、非処理群と比べてヒストグラムに大きな変化は見られなかった。表12下には、非処理群、MBN処理群及びRecJ_f処理群の相補鎖間コンセンサスリード配列をLT-2株のゲノムにマッピン

グした際の、各群での covered rate、mean coverage、SD of coverage、及び CV を示す。MBN 処理群、及び RecJ_f 処理群のどちらも、covered rate や CV は非処理群と大きく変わらないことが確認された。以上の結果より、サンプル DNA 断片の 1 本鎖特異的ヌクレアーゼ処理によりシーケンシングされるゲノム領域が偏ることはおおむねないものと考えられた。

【0185】

【表12】

S1 nuclease (U / ng 反応液中DNA)	Covered rate (%)	Mean coverage	SD of coverage	CV (%)
0	97.51	140.5	37.07	26.38
0.17	97.52	171.8	43.97	25.59
0.50	97.51	176.9	45.97	25.99
1.67	97.50	184.8	48.63	26.31
5.00	97.52	218.4	58.62	26.84
16.7	97.50	187.8	57.56	30.64

10

Nuclease (U / ng 反応液中DNA)	Covered rate (%)	Mean coverage	SD of coverage	CV (%)
0U	97.54	263.1	66.81	25.40
1.00 (MBN)	97.51	261.6	67.55	25.82
1.00 (RecJ _f)	97.51	231.7	59.45	25.66

20

【0186】

2) 断片の誤認識

相補鎖間コンセンサスリード配列の作成の際、異なる細胞由来のリードペアが偶然に参照配列上の同一の位置にマッピングされると、同じ 2 本鎖 DNA 断片由来のリードペアとして誤認識される。このとき、ある細胞の DNA から変異の入ったリードペアが得られ、別の細胞の DNA から変異のないリードペアが得られていた場合、真の変異がエラーとして除かれてしまう。こうした異なる DNA 断片の同一断片としての誤認識（断片の誤認識）は、ライブラリ調製での DNA 断片増幅過程で初期 DNA 量を解析対象のゲノムサイズに応じて調整し、ライブラリ中のアダプター付加 DNA の多様性を調整することで最小限に抑えられる。実施例 1 のライブラリは、全て初期 DNA 量が 78 amol であることから、断片の誤認識は通常無視できるレベルである。本実施例では、断片の誤認識が 1 本鎖特異的ヌクレアーゼでの処理により増加しないか調べた。

30

【0187】

本解析では、サンプル DNA の識別のため、アダプター配列内の index 情報を利用した。異なる index 情報を持つアダプター配列を用いて DMSO 暴露ライブラリと 3-MC 暴露ライブラリを調製し、シーケンシングデータを得た。それぞれのライブラリの Fastq ファイル（リード 1、リード 2）の先頭から 25 M リードずつを抽出し、リード 1 同士、及び、リード 2 同士で 1 つにまとめ、2 種類の index 情報を含む 50 M リードの Fastq ファイルをリード 1、リード 2 それぞれ作成した。このようにして、1 本鎖特異的ヌクレアーゼのユニット数ごとに、ゲノム DNA の由来の異なるリードペアが混合されたシーケンシングデータを作成した。このデータを参照配列にマッピングし、参考例 1 の方法に従ってリードペアのグループを作成した。これらのグループのうち、2 つ以上のリードペアが含まれるグループを抽出し、各グループ中のリードペアの index 情報をもとに、ゲノム DNA の由来の異なるリードペアが含まれる割合（異なる index が含まれる割合 = 断片の誤認識率）を算出した。

40

異なる index が含まれる割合 (%) = (異なる index 情報が含まれるグループ数) / (2 つ以上のリードペアが含まれるグループ数) × 100

【0188】

50

I) S1 nuclease

各ユニット数のS1 nuclease処理群での異なるindexが含まれる割合、即ち断片の誤認識率を図19及び表13に示す。S1 nucleaseのユニット数の増加に伴い、異なるindexが含まれる割合は増加していた。本実施例では、2種類のindex情報を用いたことから、実際に起こった断片の誤認識のうちのおよそ半分が検出されたと推定され、したがって、算出された異なるindexが含まれる割合の約2倍の値が、実際の誤認識率と推定された。シーケンシングエラーが大きく低減される0.17 U/ng以上でのS1 nuclease処理では、断片の誤認識率はおよそ7%以上で、変異頻度への影響が懸念されるレベルであった。

【0189】

10

【表13】

S1 nuclease (U / ng 反応液中DNA)	0	0.02	0.05	0.17	0.50	1.67
異なるindexが含まれる割合 (%)	1.53	1.60	1.77	3.56	7.44	10.32

【0190】

II) MBN

各ユニット数のMBN処理群での断片の誤認識率を図20及び表14に示す。MBNにおいても、ユニット数の増加に伴い、異なるindexが含まれる割合は増加した。0.10 U/ng以上では断片の推定誤認識率(異なるindexが含まれる割合の約2倍の値)はおよそ6%以上で、変異検出への影響が懸念されるレベルであった。

20

【0191】

【表14】

MBN (U / ng 反応液中DNA)	0	0.03	0.10	0.30	1.00
異なるindexが含まれる割合 (%)	1.58	1.83	3.16	6.30	12.65

【0192】

30

III) RecJf

各ユニット数のRecJf処理群での断片の誤認識率を図21及び表15に示す。ユニット数が増加に伴い、異なるindexが含まれる割合が僅かに増加したが、変異検出に影響するほどではなかった。

【0193】

【表15】

RecJf (U / ng 反応液中DNA)	0	0.03	0.10	0.30	1.00
異なるindexが含まれる割合 (%)	1.58	1.42	1.29	1.52	1.81

40

【0194】

3) 結果と考察

実施例1の結果から、末端修復の前にDNAをS1 nuclease、MBN又はRecJfで処理することで、シーケンシングにおけるエラーを低減できることが確認できた。したがって、エラー低減効果は1本鎖特異的ヌクレアーゼに共通することが示された。エラー低減効果は、S1 nuclease > MBN > RecJfの順で大きかった。この理由の1つとして、両側が2本鎖である1本鎖部分は、1本鎖特異的エキソヌクレアーゼ(RecJf)では分解できないが、1本鎖特異的エンドヌクレアーゼ(S1 nuclease、及びMBN)では分解できること考えられた。一方、実施例2の結果から、S1

50

nuclease及びMBNにおいては高精度シーケンシング法を併用すると、断片の誤認識率が増え、変異頻度に影響があることが明らかとなった。誤認識率が増えた原因は、S1 nuclease及びMBN活性の配列特異性によるものと考えられた。すなわち、DNA断片の末端にS1 nuclease及びMBNで分解されにくい1本鎖配列が残ったことで、リードペアの両末端部位が偶然に一致し、参照配列上の同一の位置にマッピングされる可能性が上昇することにより誤認識率が増加したと推測された。この問題を解決するためには、(i)初期DNA量をさらに減少させる、又は(ii)S1 nucleaseもしくはMBN処理後、断片を特異性の異なる1本鎖特異的ヌクレアーゼでさらに処理する、という2つの手段が考えられた。これらの手段の有効性について、S1 nucleaseを用いて、この後の実施例で検討した。一方、RecJfに関しては、誤認識率が大きく増えなかったことから、変異検出への影響を受けずに使用することができると考えられた。

10

【0195】

実施例3 断片の誤認識率に対する初期DNA量の影響

断片の誤認識率、すなわちリードペアが偶然に参照配列上の同一の位置にマッピングされる可能性は、ライブラリ中のサンプルDNAの多様性を減少させることで抑えられる。そこで本実施例では、ライブラリ調製における初期DNA量を78amolよりもさらに減少させることでリードの偶然の重なりを低下させることができるか検討した。

【0196】

1) シーケンシング用ライブラリの調製

20

比較例1及び実施例1で調製したDMSO暴露細胞及び3-MC暴露細胞由来のゲノムDNAをサンプルDNAとした。120ng相当量のDNAをそれぞれ複数サンプル用意し、実施例1に記載の方法でS1 nuclease処理したライブラリを調製した。S1 nucleaseのユニット数は、シーケンシングエラーの低減と断片の誤認識率を考慮して0.08U/ng(DNA)及び0.25U/ng(DNA)とした。アダプター付加DNAのPCR enrichmentの過程では、初期DNA量を39及び20amolとし、PCR産物のDNA量を考慮して、39amolのDNAは16サイクル、20amolのDNAは17サイクルでPCR増幅してライブラリを調製した。

【0197】

2) シーケンシング及び断片の誤認識率の算出

30

実施例1と同様にライブラリをシーケンシングした。次いで実施例2と同様の手順で断片の誤認識率(異なるindexが含まれる割合)を算出した。

【0198】

3) 結果と考察

断片の誤認識率を図2.2及び表1.6に示す。実施例2の2)と同様、表1.6の値の約2倍の値が、実際の誤認識率と推定された。特許文献4の実施例に記載のとおり、初期DNA量を減少させることで断片の誤認識率を減少させることができた。そして、0.08U/ngのS1 nucleaseで処理する場合は、初期DNA量を39amol以下にすれば、実際の誤認識率はおよそ5%以下となり、変異を見逃す懸念をできる限り小さくすることができた。同様に、0.25U/ngのS1 nucleaseの場合は、初期DNA量を20amol以下にすると誤認識率が5%以下となった。

40

【0199】

【表1.6】

異なるindexが含まれる割合(%)

S1 nuclease (U / ng 反応液中DNA)		0	0.08	0.25
初期DNA量	39amol	0.68	2.69	4.32
	20amol	0.31	1.29	2.05

【0200】

50

実施例 2 の断片の誤認識と本実施例の結果から、酵素量 (U/ng) に応じて断片の誤認識が増加するが、適切な初期 DNA 量を選択することで断片の誤認識を減少させることが可能であることが示された。S1 nuclease 処理下での断片の誤認識の増加率は、下記式で定義され、かつ実施例 2 の結果に基づいて酵素量ごとに表 17 のように算出された。

断片の誤認識の増加率 = [S1 nuclease (U/ng) 処理時の断片の誤認識率 (%)] / [S1 nuclease 非処理時の断片の誤認識率 (%)]

【0201】

【表 17】

S1 nuclease (U/ng 反応液中 DNA)	0.017	0.050	0.167	0.500	1.67
断片の誤認識の増加率	1.04	1.16	2.32	4.85	6.72

10

【0202】

表 17 に示すとおり、酵素量 0.05 U/ng 以下では断片の誤認識率への影響は無視できるレベルであったが、酵素量が 0.05 U/ng より大きい条件では、断片の誤認識が増加した。例えば酵素量が 0.05 U/ng より大きく 0.167 U/ng 以下の範囲では、断片の誤認識の増加率は、S1 nuclease 非処理時の 2 倍程度と推測された。上記のとおり断片の誤認識率は初期 DNA 量に依存するため、酵素量が 0.05 U/ng より大きい場合、適切な初期 DNA 量の範囲は S1 nuclease 非処理時の 2 分の 1 程度、すなわち、250 amol / Mbp の 2 分の 1 である 125 amol / Mbp 以下と考えられた。同様に、酵素量が 0.167 U/ng より大きく 0.5 U/ng 以下の場合、断片の誤認識の増加率は S1 nuclease 非処理時の 4 倍程度と推測でき、適切な初期 DNA 量の範囲は、62.5 amol / Mbp 以下と考えられた。酵素量が 0.5 U/ng より大きい場合、断片の誤認識の増加率は S1 nuclease 非処理時の 8 倍以下と推測でき、適切な初期 DNA 量の範囲は、31.3 amol / Mbp 以下と見積もることができた。

20

【0203】

ライブラリ調製とシーケンシングにおける適切な条件は、S1 nuclease の処理濃度の増加率と断片の誤認識の増加率の関係、及び、初期 DNA 量と断片の誤認識の増加率の関係を組み合わせることで導出することができた。例えば、表 17 に示すとおり、S1 nuclease 処理における酵素量が、0.17 U/ng から 1.67 U/ng に 10 倍増えたとき、断片の誤認識率はおよそ 3 倍増えた。したがって、S1 nuclease の酵素量の増加による断片の誤認識率は $[3^{\log S1 \text{ nuclease (U/ng)}}]$

30

(式中、S1 nuclease (U/ng) > 0.05、log は常用対数である) で表すことができる。一方、本実施例の結果から、初期 DNA 量が 2 倍に増えると、断片の誤認識率も 2 倍に増える傾向があった。以上の 2 つの結果を考慮して、S1 nuclease の酵素量が 0.05 U/ng より大きい場合のライブラリ調製とシーケンシングにおける条件は下記の式で表される指標に反映される：

指標 = PCR における初期 DNA 量 (amol / Mbp サンプル DNA) $\times 3^{\log S1 \text{ nuclease (U/ng)}}$

40

(式中、S1 nuclease (U/ng) > 0.05、log は常用対数である)。各条件における上記の指標の数値を表 18 に示す。上記実施例で調べた適切な条件範囲を考慮すると、好ましい条件でのシーケンシングを可能にする指標の値は 60 以下、より好ましくは 30 以下、さらに好ましくは 15 以下、さらにより好ましくは 7.5 以下であると考えられた。

【0204】

50

【表 18】

S1 nuclease (U/ng 反応液中 DNA)		0.08	0.17	0.25	0.50	1.67
初期DNA量 (amol) / サンプル DNA (Mbp)	500	152.8	212.7	258.1	359.2	638.0
	250	76.4	106.3	129.0	179.6	319.0
	125	38.2	53.2	64.5	89.8	159.5
	62.5	19.1	26.6	32.3	44.9	79.7
	31.3	9.55	13.3	16.1	22.5	39.9
	15.6	4.77	6.65	8.06	11.2	19.9
	7.81	2.39	3.32	4.03	5.61	9.97
	3.91	1.19	1.66	2.02	2.81	4.98
	1.95	0.60	0.83	1.01	1.40	2.49

【0205】

また、実施例 2 の結果から、MBNにおいても、S1 nuclease と同等のユニット数で同等の誤認識率を示したので、上記で導出した関係式と適切な条件範囲を、そのまま適用できると考えられた。一方、実施例 2 で示したとおり、Rec J_f のユニット数が断片の誤認識率に及ぼす影響は無視できるレベルであった。

【0206】

実施例 4 断片の誤認識率に対する異なるヌクレアーゼ処理の影響

S1 nuclease 処理による断片の誤認識率の増加は、DNA 断片の末端における S1 nuclease で分解されにくい 1 本鎖の残存が原因と推測された。このため、S1 nuclease で処理後、DNA 断片を特異性の異なる 1 本鎖特異的ヌクレアーゼでさらに処理することで、誤認識率が改善されると考えられた。エンドヌクレアーゼである S1 nuclease と異なり、Rec J_f は 1 本鎖の 5' 末端から分解する 5' 3' エキソヌクレアーゼ活性を有する。本実施例では、S1 nuclease 処理後に DNA 断片を Rec J_f でさらに処理することによる断片の誤認識率への影響を調べた。

【0207】

1) シーケンシング用ライブラリの調製

比較例 1 及び実施例 1 で調製した DMSO 暴露細胞及び 3-MC 暴露細胞由来のゲノム DNA をサンプル DNA とした。100 ng 相当量の DNA をそれぞれ複数サンプル用意し、実施例 1 の 3) I) に記載の方法で、30 μL のサンプル DNA の断片を含む DNA 溶出液を得た。次いで、実施例 1 の 3) II-1) に記載の方法で断片を 30 U (0.3 U/ng) の S1 nuclease で処理した。EDTA の添加と熱失活の後、ビーズを添加し、反応液から DNA を精製し、2 群に分けた。Rec J_f 非処理群は、TruSeq に付属の Resuspension buffer で懸濁し、60 μL の溶出液に調製した。Rec J_f 処理群は、Distilled water でビーズを懸濁して 30 μL の溶出液を得た後、実施例 1 の 3) II-3) に記載の方法で Rec J_f (3 (0.03)、10 (0.1)、30 (0.3)、100 (1.0) U (U/ng)) 処理した。熱失活の後、DNA の精製のため、反応液にビーズを添加し、TruSeq に付属の Resuspension buffer で懸濁し、60 μL の溶出液を得た。得られた溶出液から TruSeq の推奨プロトコルに基づいてライブラリを調製した。アダプター付加 DNA の PCR enrichment の過程では、初期 DNA 量を 78 amol とし、15 サイクルで増幅した。

【0208】

2) シーケンシング及び断片の誤認識率の算出

実施例 1 と同様にライブラリをシーケンシングした。次いで実施例 2 と同様の手順で断片の誤認識率 (異なる index が含まれる割合) を算出した。

【0209】

3) 結果と考察

10

20

30

40

50

0.30 U/ng の S1 nuclease で処理後、各ユニット数の RecJ_f で処理した断片での誤認識率を図 23 及び表 19 に示す。RecJ_f のユニット数の増加に伴い、断片の誤認識率が僅かだが減少した。これは、S1 nuclease が分解しきれなかった 1 本鎖部分を配列特異性の異なる RecJ_f が分解したことによるものと考えられた。したがって、配列特異性の異なる 1 本鎖特異的ヌクレアーゼの組合せ処理により、断片の誤認識率を低減できると考えられた。

【0210】

【表 19】

RecJ _f (U/ng 反応液中 DNA)	0	0.03	0.10	0.30	1.00
異なる index が含まれる割合 (%)	6.03	6.14	5.85	5.80	5.43

10

【0211】

図 24 に、DMSO 暴露ライブラリにおける 6 変異パターンの変異頻度を示す。本実験では 0.30 U/ng の S1 nuclease 処理のみ (RecJ_f 0 U/mg) でもエラーが十分に低減しており、このため、RecJ_f の追加処理によって誤認識率がそれほど低減しなかったのだと考えられた。

20

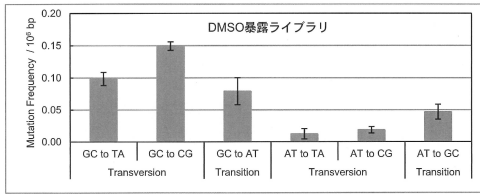
30

40

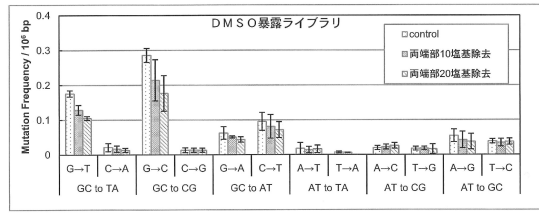
50

【図面】

【図 1】

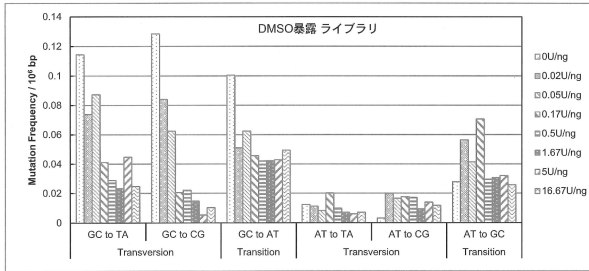


【図 2】

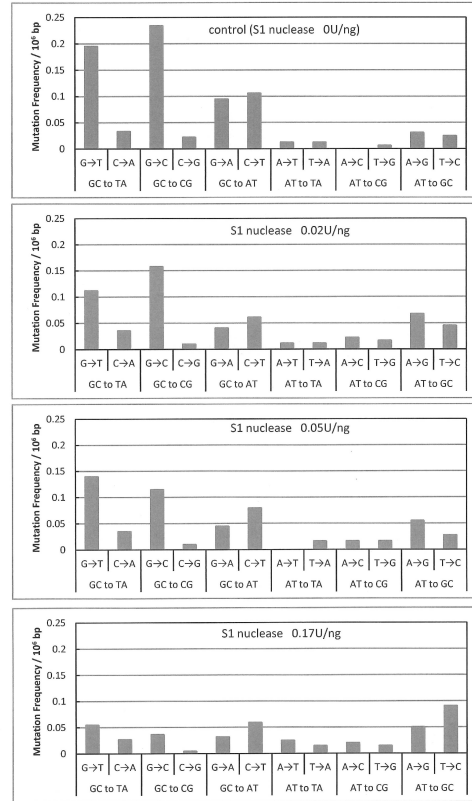


10

【図 3】



【図 4】



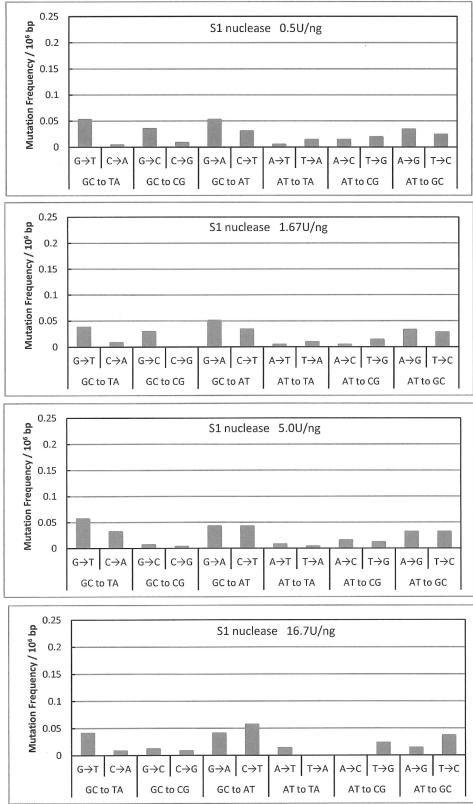
20

30

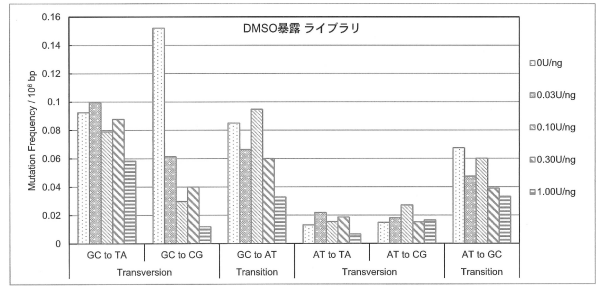
40

50

【 図 5 】

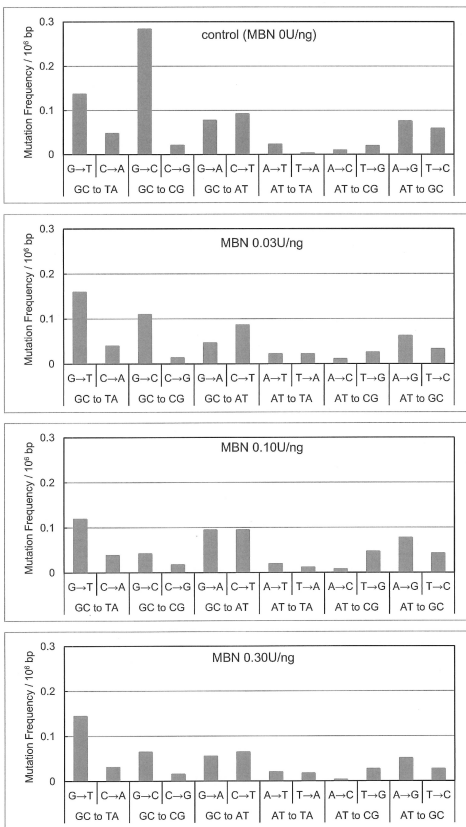


【 図 6 】

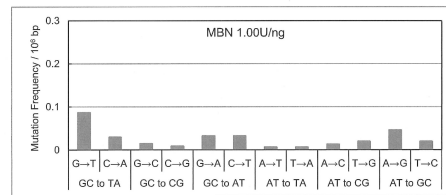


10

【 図 7 】



【 図 8 】

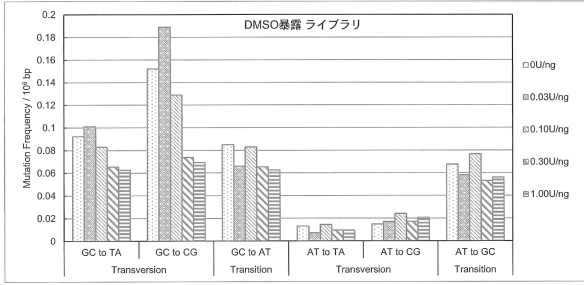


30

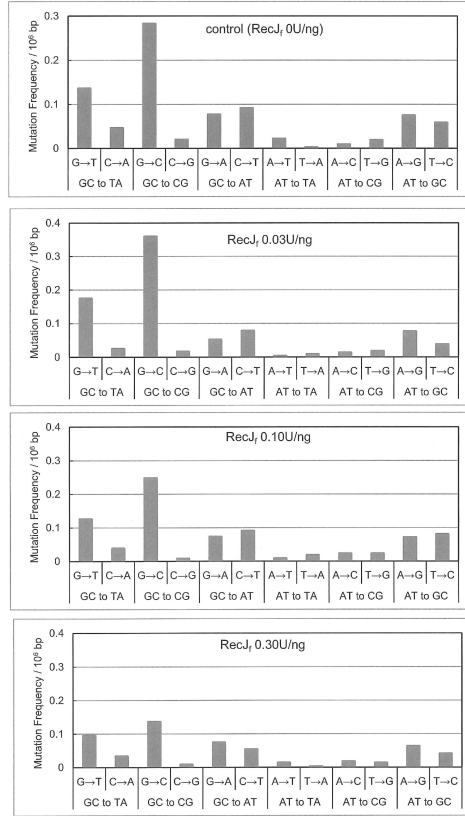
40

50

【 9 】



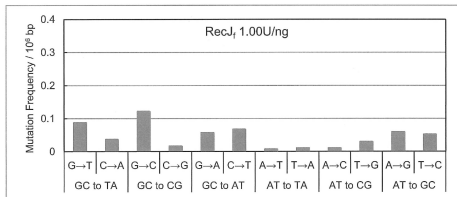
【 10 】



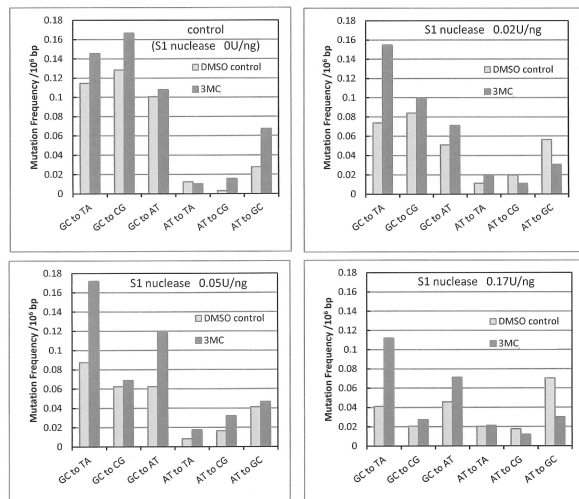
10

20

【 11 】



【 12 】

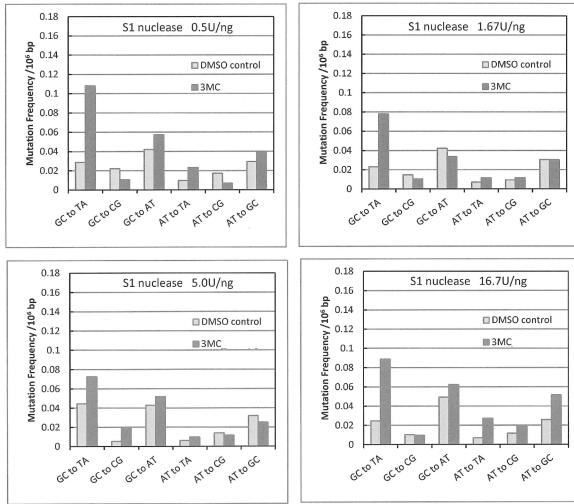


30

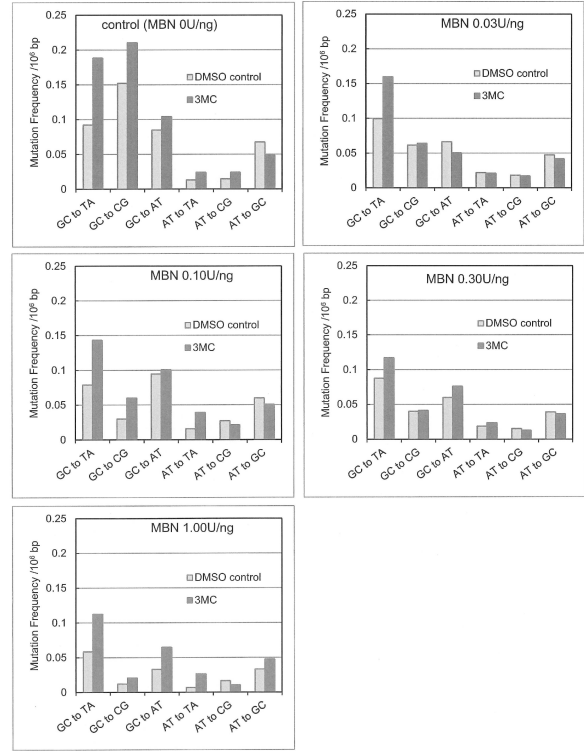
40

50

【 1 3 】



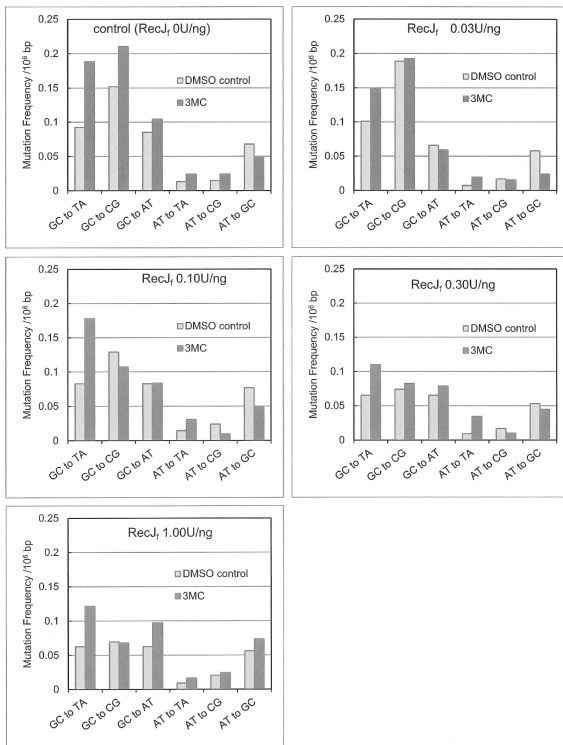
【 1 4 】



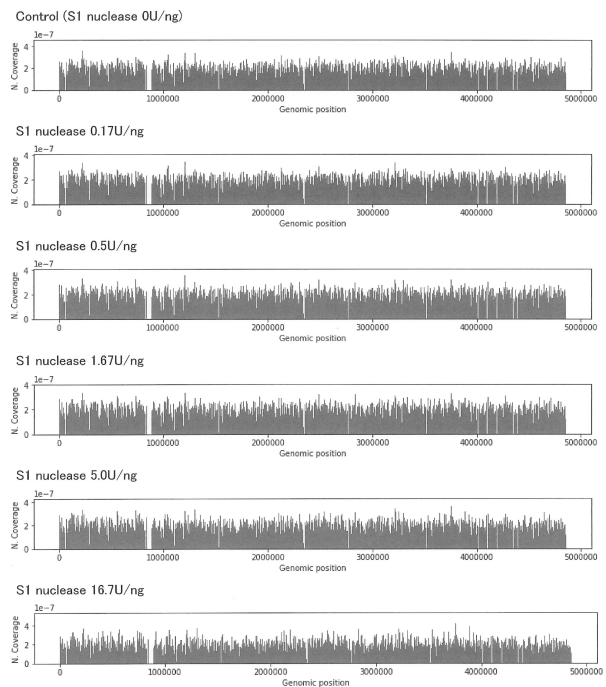
10

20

【 1 5 】



【 1 6 】

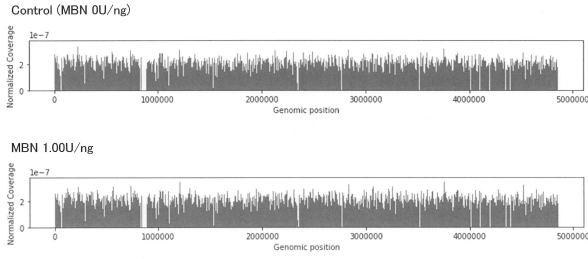


30

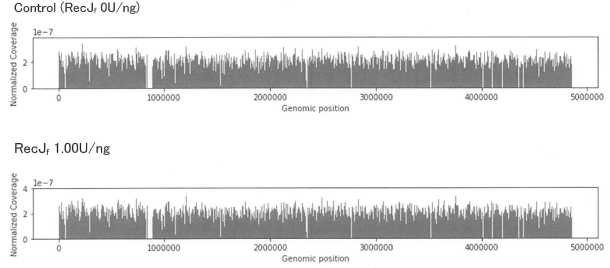
40

50

【 図 17 】

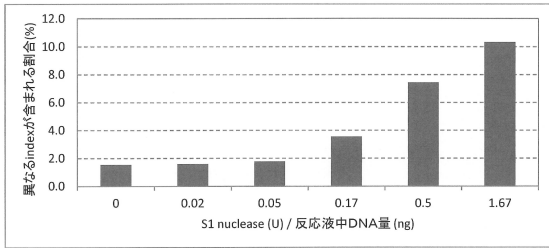


【 図 18 】

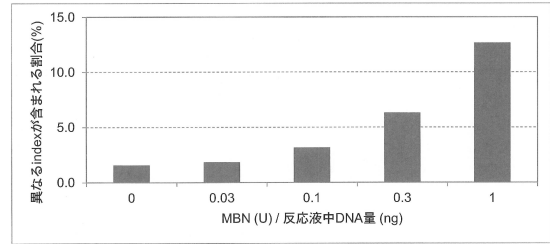


10

【 図 19 】

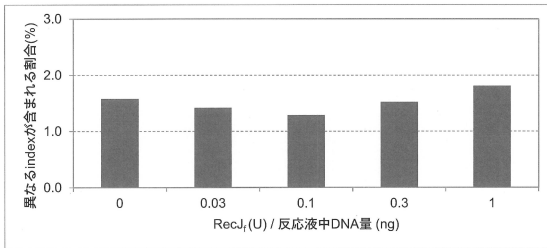


【 図 20 】

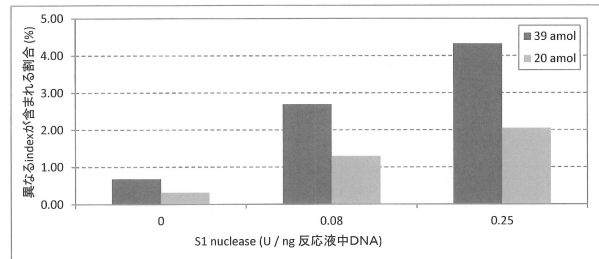


20

【 図 21 】



【 図 22 】

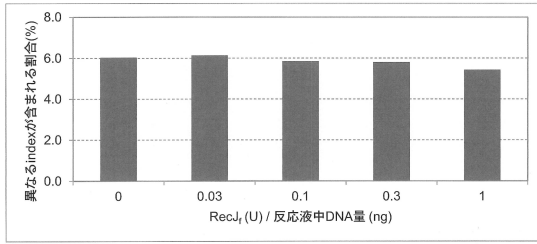


30

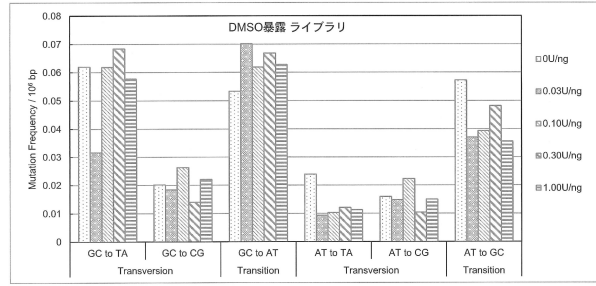
40

50

【 図 2 3 】



【 図 2 4 】



10

20

30

40

50

フロントページの続き

(51)国際特許分類

C 4 0 B 40/06 (2006.01)

F I

C 4 0 B 40/06

(56)参考文献

国際公開第 2 0 0 5 / 0 4 2 7 8 1 (W O , A 2)

特開 2 0 1 9 - 1 9 3 6 1 2 (J P , A)

特開 2 0 0 8 - 1 8 7 9 0 9 (J P , A)

国際公開第 2 0 1 2 / 0 0 2 5 4 1 (W O , A 1)

特表 2 0 1 3 - 5 1 2 6 6 5 (J P , A)

特表 2 0 1 7 - 5 0 9 3 2 4 (J P , A)

(58)調査した分野 (Int.Cl., DB名)

C 1 2 Q 1 / 0 0 - 3 / 0 0

C 1 2 N 1 5 / 0 0 - 1 5 / 9 0

C 1 2 N 9 / 0 0 - 9 / 9 9

J S T P l u s / J M E D P l u s / J S T 7 5 8 0 (J D r e a m I I I)

C A p l u s / M E D L I N E / E M B A S E / B I O S I S (S T N)

P u b M e d