

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第6898282号
(P6898282)

(45) 発行日 令和3年7月7日 (2021. 7. 7)

(24) 登録日 令和3年6月14日 (2021. 6. 14)

(51) Int. Cl.	F 1
C 1 2 Q 1/6869 (2018. 01)	C 1 2 Q 1/6869 Z
C 1 2 Q 1/6809 (2018. 01)	C 1 2 Q 1/6809 Z

請求項の数 28 (全 63 頁)

(21) 出願番号	特願2018-168288 (P2018-168288)	(73) 特許権者	000000918
(22) 出願日	平成30年9月7日 (2018. 9. 7)		花王株式会社
(65) 公開番号	特開2019-193612 (P2019-193612A)		東京都中央区日本橋茅場町 1 丁目 1 4 番 1
(43) 公開日	令和1年11月7日 (2019. 11. 7)		〇号
審査請求日	令和1年12月12日 (2019. 12. 12)	(74) 代理人	110000084
(31) 優先権主張番号	特願2018-86972 (P2018-86972)		特許業務法人アルガ特許事務所
(32) 優先日	平成30年4月27日 (2018. 4. 27)	(72) 発明者	松村 奨士
(33) 優先権主張国・地域又は機関	日本国 (JP)		栃木県芳賀郡市貝町赤羽 2 6 0 6 花王株
			式会社研究所内
		(72) 発明者	佐藤 平行
			栃木県芳賀郡市貝町赤羽 2 6 0 6 花王株
			式会社研究所内
		(72) 発明者	大坪 裕紀
			栃木県芳賀郡市貝町赤羽 2 6 0 6 花王株
			式会社研究所内

最終頁に続く

(54) 【発明の名称】 高精度シーケンシング方法

(57) 【特許請求の範囲】

【請求項 1】

DNA のシーケンシング方法であって、

(1) サンプル DNA の断片を調製すること；

(2) 該サンプル DNA の断片を PCR にかけて PCR 産物を得ること、ここで、該サンプル DNA の断片の各々について 2 つ以上の増幅断片が作製される；

(3) 得られた PCR 産物をシーケンシングし、該 PCR 産物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、ここで、該リード配列は、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し、それによって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；

(4) 得られた複数の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、ここで、該 1 つ以上のグループは、平均で 1. 0 5 ～ 3 0 個のリード配列を含む；

(5) 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に、その際、該サンプル DNA 断片の 2 本の相補鎖の各々に由来するリード配列を含むリード配列のグループの中から、該 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、
を含む、方法。

【請求項 2】

10

20

前記（２）において、前記PCRでの初期DNA量が前記サンプルDNAのサイズ1Mbpあたり250amol以下である、請求項1記載の方法。

【請求項3】

前記（３）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、請求項1又は2記載の方法。

【請求項4】

前記（４）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1～3のいずれか1項記載の方法。

【請求項5】

請求項1又は2記載の方法であって、

前記（３）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本の相補鎖のうちの一方の鎖の配列を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、

ここで、該リード配列のペアは、前記サンプル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、

前記（４）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み、ここで、該1つ以上のグループは、平均で1.05～30個のリード配列のペアを含み、

前記（５）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み、その際、該サンプル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含むリード配列のペアのグループの中から、該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、

方法。

【請求項6】

前記（３）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、請求項5記載の方法。

【請求項7】

前記（４）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5又は6記載の方法。

【請求項8】

前記（４）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7記載の方法。

【請求項9】

前記サンプルDNA断片の2本の相補鎖が、それらの5'末端側と3'末端側に異なる標識配列を有し、かつ、前記シーケンシングにより、各リード配列に付随する該標識配列の情報が取得され、

前記リード配列のペアのグループの中のリード配列のペアを、それに含まれる2本のリード配列に付随する該標識配列の情報と、該2本のリード配列の前記参照配列上での互いの位置関係に基づいて、サンプルDNA断片の2本の相補鎖のどちらに由来するか識別す

10

20

30

40

50

ることをさらに含む、請求項8記載の方法。

【請求項10】

前記シーケンシングにおいて、前記PCRでの初期DNA量1amolあたり0.02～ 10×10^6 個のリード配列を取得する、請求項1～4のいずれか1項記載の方法。

【請求項11】

前記シーケンシングにおいて、前記PCRでの初期DNA量1amolあたり0.02～ 10×10^6 個のリード配列のペアを取得する、請求項5～9のいずれか1項記載の方法。

【請求項12】

前記サンプルDNAが、10kbp以上のサイズを有する、請求項1～11のいずれか1項記載の方法。 10

【請求項13】

前記シーケンシングにおいて、前記サンプルDNAの1Mbpあたり0.05～ 1600×10^6 個のリード配列又はリード配列のペアを取得する、請求項1～12のいずれか1項記載の方法。

【請求項14】

前記PCRにかけるサンプルDNAの断片に、個別のサンプルDNAの断片を識別するための標識が連結されていない、請求項1～13のいずれか1項記載の方法。

【請求項15】

前記(4)において、さらに前記リード配列のグループ又は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又はリード配列のペアを集める手順を含む、請求項1～14のいずれか1項記載の方法。 20

【請求項16】

前記PCRにかけるサンプルDNAの断片に、該断片を構成する2本の相補鎖を識別できる標識配列が付加される、請求項1～15のいずれか1項記載の方法。

【請求項17】

ゲノムDNAの変異を検出する方法であって、
細胞中のゲノムDNAをサンプルDNAとして用いて、請求項1～16のいずれか1項記載のDNAのシーケンシング方法を実施し、配列データを作成すること；

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、ここで該参照配列は、該ゲノムDNA中の既知配列である、
を含む、方法。 30

【請求項18】

前記ゲノムDNAが試験物質に曝露した細胞のゲノムDNAを含む、請求項17記載の方法。

【請求項19】

前記ゲノムDNAが試験物質に曝露した細胞のゲノムDNAと、該試験物質に曝露していない対照細胞のゲノムDNAを含み、

該試験物質に曝露した細胞のゲノム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を比較することをさらに含む、
請求項18記載の方法。 40

【請求項20】

前記細胞が大腸菌、サルモネラ菌又は酵母細胞である、請求項17～19のいずれか1項記載の方法。

【請求項21】

前記サルモネラ菌がS. Typhimurium LT-2株、TA100株、TA98株、TA1535株、TA1538株又はTA1537株である、請求項20記載の方法。

【請求項22】

前記細胞が哺乳動物細胞、又は哺乳動物由来培養細胞である、請求項 1 7 ～ 1 9 のいずれか 1 項記載の方法。

【請求項 2 3】

前記哺乳動物がヒト又はマウスである、請求項 2 2 記載の方法。

【請求項 2 4】

前記ゲノム DNA の変異が経時変化によるゲノム DNA の変異を含む、請求項 1 7 記載の方法。

【請求項 2 5】

前記ゲノム DNA が、経時変化した細胞のゲノム DNA と、より経時変化していない細胞のゲノム DNA を含み、

該経時変化した細胞のゲノム DNA について検出した変異と、該より経時変化していない細胞のゲノム DNA について検出した変異とを比較することをさらに含む、請求項 2 4 記載の方法。

【請求項 2 6】

前記細胞が哺乳動物細胞、又は哺乳動物由来培養細胞である、請求項 2 4 又は 2 5 記載の方法。

【請求項 2 7】

前記細胞が培養細胞である、請求項 2 4 又は 2 5 記載の方法。

【請求項 2 8】

前記変異が塩基対置換型変異である、請求項 1 7 ～ 2 7 のいずれか 1 項記載の方法。

【発明の詳細な説明】

【技術分野】

【0 0 0 1】

本発明は、DNA のシーケンシング方法、及びこれを用いた変異解析方法に関する。

【背景技術】

【0 0 0 2】

近年、次世代シーケンシング (NGS) の技術が著しい発展をとげ、ゲノム DNA のシーケンシング等に汎用されている (非特許文献 1)。例えばがん細胞のゲノム変異解析では、NGS は特定の遺伝子領域に絞った解析だけでなく、従来では困難だったゲノム全体の変異の解析を可能にし、様々な新しい知見を生み出している (非特許文献 2、3)。

【0 0 0 3】

これまでに様々なシーケンサーが開発されており、代表的なものとしてイルミナ社のシーケンサーである HiSeq や MiSeq がある。これらシーケンサーでは解析対象となる細胞、組織等の DNA 断片を数百 bp の大きさに細かく断片化し、この断片配列の配列情報 (リード配列) を同時並行で大量に取得してデータ解析に使用する。これらのリード配列一つ一つの読み取り精度はそれほど高くないものの、複数の細胞に由来するリード配列を集め、集めた複数のリード配列の間で高確率に出現する塩基情報を取得することで、ゲノム配列上の同一箇所を複数回シーケンシングし、高精度のシーケンシングが可能になる。得られたシーケンシングデータは、ヒトの一塩基多型やがん細胞の遺伝子変異の同定等の変異解析に利用することができる。

【0 0 0 4】

一方で、上記のようなシーケンシング手法で検出できる変異は、一塩基多型や、がん細胞の変異のような、解析対象となる細胞集団内で比較的多くの細胞に共通して含まれている変異である。NGS での個々のリード配列におけるエラー頻度は約 $1 / 10^3$ bp と言われていることから、これを下回るような低頻度 (例えば 0.1 % 以下) でしか含まれない変異を正確に同定することは、上記のようなシーケンシング手法では通常困難である。このような低頻度な変異の事例として、がん組織中の少数の細胞集団しか有していない稀な変異や、あるいは、解析対象の細胞集団が環境変異原等に暴露されたことによって、集団中の個々の細胞に独立に生じた変異等が挙げられる。これまで、これら低頻度な変異の解析に NGS が用いられた例は少ない。しかし、低頻度な変異を NGS で大規模に同定す

10

20

30

40

50

ることが可能になれば、がんの早期発見や、化学物質等の変異原の毒性評価法として適用可能であることが考えられる。また、ヒトの個々の細胞が有する変異を網羅的に解析することで、対象のヒトがそれまでに暴露された環境変異原を推定し、がんの予防に役立てられるなどの様々な応用が期待できる。

【0005】

低頻度な体細胞変異は、単離した単一細胞を解析することによって同定可能であることが報告されている。例えば、Gundryらは、マウス等の細胞に変異原物質を暴露後、単一の細胞を全ゲノムシーケンシングすることによって、化学物質による変異を同定している（非特許文献4）。しかし、この手法には、単離培養の難しいがん組織には適用が困難な場合があることや、細胞の単離の手間が発生してしまうこと、また単一細胞由来の変異情報しか得ることができないなどの課題があった。

10

【0006】

そこで、単一細胞の単離を介さず、個々のDNA断片のシーケンシング精度を高めることによって、細胞集団中に含まれる稀な変異を正確に同定するシーケンシング方法が報告されている。広く用いられる方法として、イルミナ社のシーケンサーにおいて、ペアエンドシーケンシングで得られたリードペアを重ね合わせる方法がある（非特許文献5）。これは、各断片について得られるリードペアの双方のコンセンサスを取ることによって、シーケンシング精度を高めることが可能な方法である。

【0007】

しかし、NGSでのエラーには、ライブラリ調製や保存の過程で解析対象のDNAに生じる酸化修飾等を原因とするエラーがある（非特許文献6、7）。これらのエラーは、上述のリードペアを重ね合わせる方法では除くことが難しい。そこで、DNAの相補的な2本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、エラーか真の変異かを見分ける方法が開発されている（非特許文献8～10、特許文献1）。DNAの酸化修飾は、基本的にDNA 2本鎖のうち片方の鎖だけに起こるため、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、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。例えば、Duplex Sequencing法（非特許文献8、特許文献1）においては、各DNA断片に12bpからなる断片固有のタグ配列を付加し、PCR増幅及びシーケンシングを行い、シーケンシング後に該タグ配列を参照して同一のDNA断片を構成する2本の相補鎖に由来するリード配列を集めた後、それら相補鎖間のコンセンサスを取ることでシーケンシング精度を高める。また、Pacific Bioscience社のシーケンサーを用いたSMRTbell Template法（非特許文献10）も相補鎖の情報を用いてシーケンシング精度を高めることができる。

20

30

【0008】

しかし、上述した相補鎖情報を利用する既存の方法は、ライブラリ調製にDNA断片固有のタグ配列を付加するなど、特殊なライブラリ調製法を必要とするか、あるいは、SMRTbell Template法のように特定のシーケンサーのみでしか実施できない。そのため、上述の方法は、広く一般的に実施可能な方法として認知されているとは言い難い。一方で、理論上のみではあるが、上記のようなDNA断片固有のタグ配列を付加することなく、DNA断片自体の部分配列を指標として同じDNA領域に由来する配列を集める方法が開示されている（特許文献2）。このアプローチは、ライブラリ調製の手順を簡略化でき、またシーケンサーの種類に依存しないため、汎用性に優れると考えられる。しかしながら、DNA断片固有のタグ配列を付加しない場合、異なるDNA断片に由来する配列情報を同一断片の情報として誤認識するリスクがあることが指摘されている（特許文献1）。

40

【先行技術文献】

【特許文献】

【0009】

【特許文献1】国際公開公報第2013/142389号

【特許文献2】国際公開公報第2012/142213号

50

【非特許文献】

【0010】

【非特許文献1】 Schendure & Ji, Nature Biotechnology, 26(10):1135-1145, 2008

【非特許文献2】 Stratton, Science, 331:1553-1558, 2011

【非特許文献3】 Alexandrov et al., Cell Reports, 3:246-259, 2013

【非特許文献4】 Gundry et al., Nucleic Acid Research, 40(5):2032-2040, 2012

【非特許文献5】 Zhang et al., Bioinformatics, 30(5):614-620, 2014

【非特許文献6】 Costello et al., Nucleic Acid Research, 41(6):e67, 2013

【非特許文献7】 Shibutani et al., Nature, 349:431-434, 1991

【非特許文献8】 Schmitt et al., PNAS, 109(36):14508-14513, 2012

10

【非特許文献9】 Gregory et al., Nucleic Acid Research, 44(3):e22, 2016

【非特許文献10】 Travers et al., Nucleic Acid Research, 38(15):e159, 2010

【発明の概要】

【発明が解決しようとする課題】

【0011】

高精度シーケンシング方法を化学物質の毒性評価法や特定の個人の変異解析などに適用するにあたり、その手順の簡略化は有用である。本発明は、より簡便に高い読み取り精度を実現することができるシーケンシング方法を提供する。さらに本発明は、該シーケンシング方法の最適条件と適用可能条件を提供する。

【課題を解決するための手段】

20

【0012】

本発明は、DNAのシーケンシング方法であって、

(1) サンプルDNAの断片を調製すること；

(2) 該サンプルDNAの断片をPCRにかけてPCR産物を得ること、ここで、該サンプルDNAの断片の各々について2つ以上の増幅断片が作製される；

(3) 得られたPCR産物をシーケンシングし、該PCR産物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；

(4) 得られた複数の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、ここで、該1つ以上のグループは、平均で1.05～30個のリード配列を含む；

30

(5) 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、方法を提供する。

【0013】

また本発明は、DNAのシーケンシング方法であって、

(1) サンプルDNAの断片を調製すること；

(2) 該サンプルDNAの断片をPCRにかけてPCR産物を得ること、ここで、該サンプルDNAの断片の各々について2つ以上の増幅断片が作製され、該PCRでの初期DNA量は、該サンプルDNAのサイズ1Mbpあたり250amol以下である；

40

(3) 得られたPCR産物をシーケンシングし、該PCR産物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；

(4) 得られた複数の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；

(5) 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、方法を提供する。

50

【0014】

さらに本発明は、ゲノムDNAの変異を検出する方法であって、

細胞中のゲノムDNAをサンプルDNAとして用いて、前記DNAのシーケンシング方法を実施し、配列データを作成すること；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、ここで該参照配列は、該ゲノムDNA中の既知配列である、を含む、方法を提供する。

【発明の効果】

【0015】

本発明の方法によれば、当該分野で広く実施されるシーケンシングの手順とほぼ同等のライブラリ調製及び実験操作で、同一のDNA断片に由来する配列情報や、さらには該断片を形成する相補鎖の配列情報を集めて、これらの配列情報を用いたシーケンシングを実施することができる。本発明の方法では、同一のDNA断片や相補鎖の配列情報を集めるために、従来のシーケンシング法のように個別のDNA断片に固有のタグ配列による個別のDNA断片の標識を行う必要がない。本発明の方法は、従来法に比べて技術の簡便さ、及び汎用性に優れる。

【図面の簡単な説明】

【0016】

【図1】PCRでの初期DNA量の異なるライブラリについての10Gbpシーケンシングデータ（コントロール群）における、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の分布。横軸はグループあたりリードペア数、縦軸は各リードペア数を有するグループの数を表す。各曲線は、初期DNA量の異なるライブラリのデータを表す。

【図2】PCRでの初期DNA量の異なるライブラリについての10Gbpシーケンシングデータ（ENU群）における、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の分布。横軸はグループあたりリードペア数、縦軸は各リードペア数を有するグループの数を表す。各曲線は、初期DNA量の異なるライブラリのデータを表す。

【図3】PCRでの初期DNA量の異なるライブラリについての10Gbpシーケンシングデータの解析効率。A：コントロール群、B：ENU群。

【図4】PCRでの初期DNA量の異なるライブラリについての2Gbpシーケンシングデータ（コントロール群）における、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の分布。横軸はグループあたりリードペア数、縦軸は各リードペア数を有するグループの数を表す。各曲線は、初期DNA量の異なるライブラリのデータを表す。

【図5】PCRでの初期DNA量の異なるライブラリについての2Gbpシーケンシングデータの解析効率。

【図6】PCRでの初期DNA量の異なるライブラリについての10Gbpシーケンシングデータの重なり率。A：コントロール群、B：ENU群。

【図7】PCRでの初期DNA量の異なるライブラリ（20000～156amol）についての10Gbpシーケンシングデータから算出した、コントロール群及びENU群でのゲノム変異頻度。

【図8】PCRでの初期DNA量の異なるライブラリ（156～5amol）についての10Gbpシーケンシングデータから算出した、コントロール群及びENU群でのゲノム変異頻度。

【図9】AT塩基対に塩基対置換型変異を導入した1000bp合成DNAにおける各塩基対変異パターンの変異頻度。

【図10】PCR初期DNA量の違いによる重なり率の変化。横軸はPCR初期DNA量。Control：対照サンプル、Mutation：変異サンプル。

【図11】実施例2の10Gbpシーケンシングデータ（コントロール群）からのコンセンサスリード配列の作成効率。横軸はグループあたりリードペア数、縦軸は全リードペア数に対するコンセンサスリード配列の割合を表す。

【図12】実施例5のライブラリNo. 1における推定フラグメントについてのグループ

10

20

30

40

50

あたりのリードペア数の分布。各曲線はマウス各染色体についてのデータを表す。横軸はグループあたりリードペア数、縦軸は各リードペア数を有するグループの数を表す。

【図 1 3】マウス DNA における塩基対置換変異の頻度。

【発明を実施するための形態】

【0017】

(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伴わない変異（サイレント変異）が含まれる。

10

【0018】

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。

【0019】

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。

20

【0020】

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 増幅により得られた 2 本鎖 DNA 断片をいう。

【0021】

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 本の一本鎖をいう。

【0022】

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。

30

40

【0023】

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。

50

【0024】

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。

【0025】

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。

10

【0026】

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。

【0027】

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とリード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。

20

【0028】

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。

30

【0029】

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。

【0030】

(2. DNAのシーケンシング方法)

本発明は、高精度なDNAのシーケンシング方法を提供する。基本的には、本発明によるDNAのシーケンシング方法は、サンプルDNAの断片を取得すること；該サンプルDNAの断片をPCRにかけてPCR産物を得ること；得られたPCR産物をシーケンシングし、該PCR産物に含まれる各サンプルDNA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。

40

【0031】

(2-1) サンプルDNA断片の調製及びPCR

本発明によるDNAのシーケンシング方法で用いられる「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

50

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。これらサンプル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。

【0032】

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。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を取得および解析する場合、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。

【0033】

DNAの断片は、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調製することができる。DNAの断片化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やイルミナ社のNexteraテクノロジーなどを用いたトランスポゾン及びトランスポゼース処理が挙げられる。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応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。

【0034】

本発明の方法では、サンプルDNAの断片をPCRにかけてPCR産物を得る。PCRは、市販のPCR用試薬や機器を用いて、常法に従って行うことができる。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。該PCRにおいては、鋳型として使用された該サンプルDNAの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。

【0035】

（2-2）シーケンシング及びリード配列の作成

次いで、得られたサンプルDNA断片のPCR産物をシーケンシングにかける。PCR産物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。

【0036】

該シーケンシングにより、増幅断片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PCR産物に含まれる増幅断片のうちの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PCR産物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PCR産物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。

10

【0037】

（2-3）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、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。

20

【0038】

本発明の方法の一実施形態においては、PCR産物に含まれる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増幅断片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含む増幅断片を調製する；得られた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。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。

30

40

【0039】

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

50

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。

【0040】

(2-4) 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断片についてのコンセンサス配列を表す。

10

【0041】

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。

20

【0042】

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。

【0043】

(2-5) 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変更は、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変更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断片を構成する2本の相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マッピングされる。したがって、ゲノムの同一箇所に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。

30

40

【0044】

したがって、本発明の一実施形態では、相補鎖情報を用いたDNAのシーケンシング方法を提供する。当該方法では、上記(2-2)で述べたPCR産物のシーケンシングの際に、各サンプルDNA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

50

。より詳細には、該PCR産物のシーケンシングにより該PCR産物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。

【0045】

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(2-3)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(2-4)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。

10

【0046】

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の一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

【0047】

リード配列のグループの中から、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。

30

【0048】

上記の手順において、サンプルDNA断片に標識配列を付加する場合、該標識配列は、各サンプルDNA断片の2本の相補鎖に由来するリード配列を互いに識別可能に標識する。一方、該標識配列は、該リード配列がいずれのサンプルDNA断片に由来するかを識別する必要はない。好ましくは、各サンプルDNA断片に付加される該標識配列は同一である。例えば、サンプル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有するようにする。該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。好ましくは、サンプル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、かつサンプルDNA断片の両端の標識配列の領域の両鎖は互いに相補的でない配列を含む。また好ましくは、各サンプルDNA断片の間で、該5'末端側の標識配列には共通の配列が使用され、かつ該3'末端側の標識配列にも共通の配列が使用される。結果、サンプルDNA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し得、かつ該5'末端側の標識配列と該3'末端側の標識配列は各1本鎖間で共通であり得る。このような、リード配列がサンプルDNA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。該

40

50

標識配列が付加されたサンプルDNA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。

【0049】

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断片の2本の相補鎖うちの同じ一本鎖に由来する。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断片の2本の相補鎖うちのもう一方の一本鎖に由来する。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。

【0050】

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。

【0051】

(2-6) リードペアを用いたサンプルDNA配列情報の抽出

本発明の方法の一実施形態においては、上記(2-2)で述べたPCR産物のシーケンシングの際に、PCR産物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。

【0052】

当該方法においては、PCR産物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PCR産物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。

【0053】

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本の相補鎖のうちの一方の鎖の配列を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（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一本鎖）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）。

10

【0054】

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、それらを同じグループに分ける方法が挙げられる。

20

【0055】

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。本発明の方法においては、通常、PCRにかけたサンプル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。

【0056】

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（2-4）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

【0057】

（2-7）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（2-6）で述べたPCR産物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

【0058】

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（2-5）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

50

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センサスを取る具体的な手法は、上記（２－４）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。

【００５９】

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で、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

【００６０】

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

【００６１】

リードペアのグループの中からサンプル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。

30

【００６２】

上記の手順において、サンプルDNA断片に標識配列を付加する場合、例えば、サンプル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有するようにする。好ましくは、サンプル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、かつサンプルDNA断片の両端の標識配列の領域の両鎖は互いに相補的でない配列を含む。また好ましくは、各サンプルDNA断片の間で、該５'末端側の標識配列には共通の配列が使用され、かつ該３'末端側の標識配列にも共通の配列が使用される。結果、サンプルDNA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し得、かつ該５'末端側の標識配列と該３'末端側の標識配列は各１本鎖間で共通であり得る。このような、リード配列がサンプルDNA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。該標識配列が付加されたサンプルDNA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。

40

【００６３】

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

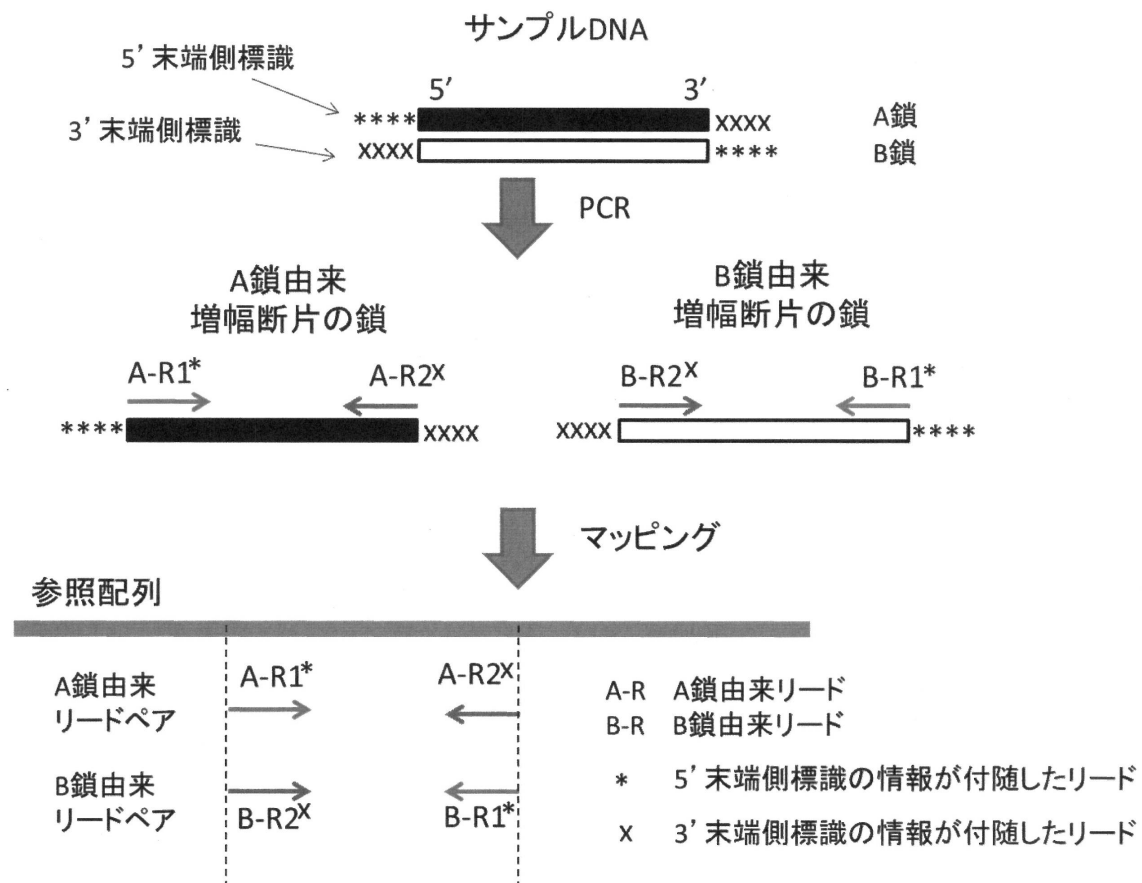
50

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（下図参照）。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。

【0064】

【化1】

標識配列の情報に基づくリードペアの識別



【0065】

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。

10

【0066】

本発明の方法においては、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。また本発明の方法においては、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。一方、本発明の方法においては、リード配列がいずれのサンプルDNA断片に由来するかを標識によって識別する必要はない。したがって、本発明の方法において使用される該標識配列は、サンプルDNAの個別の断片を識別する標識でなくともよい。本発明の方法においては、サンプルDNAの個別の断片を識別するための標識（例えば、非特許文献8及び特許文献1に記載されるサンプルDNA断片固有のタグ配列）を付加する必要はない。本発明の方法では、

20

【0067】

（3．シーケンシング条件の最適化）

上述した本発明によるDNAの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を用いないために、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。例えばがん遺伝子のような特定領域の変異の有無に基づき診断を行うような場合、変異の正確な同定が求められるため、変異の見逃しは重大な問題となり得る。一方、化学物質の変異原性評価や、特定の個人のゲノム全体に生じた変異情報に基づき変異原の推定を行う場合では、サンプルDNA全体での変異の概要を捉えることがより重要であり、必ずしも特定領域の変異の正確な同定を要求しない。この場合、変異の見逃しは、サンプルDNA全体での変異の傾向に影響を与える程度に高頻度に起こらない限り、許容できる。

30

【0068】

これまで、シーケンシング条件の最適化については、Kennedyらによって、Duplex Sequencing法でのライブラリ調製時のPCR条件及びシーケンシングデータ量の最適化の検討がなされている（Nature protocol, 2014 9(11):2586-2606）。例えば、Kennedyらの検討では、各PCR産物のコンセンサスリード配列を作成するにあたり、同一のDNA断片由来のリード配列が少なくとも3つ以上含まれることを要件としている。しかし、この最適条件は、ターゲットとなるゲノム領域の変異の有無を正確に捉える目的で設計されており、サンプルDNA全体での変異の傾向を広く捉えるための本発明の方法での最適条件とは異なる可能性がある。本発明の方法でサンプルDNA全体の変異の概要を捉える場合、必要な精度を得るためには必ずしも同一のDNA断片由来のリード配列を多数必要としない。むしろ、本発明の方法では、同一のDNA断片由来のリード数を減らして代わりにより多数の異なるDNA断片を評価した方が効率的である。したがって、本発明の方法は、Kennedyらの条件と比べてより少ないシーケンス量でサンプルDNA全体の変異の傾向を解析できる可能性がある。

40

50

【0069】

下記実施例2に記載されるとおり、本発明によるDNAのシーケンシング方法の最適条件と、サンプルDNAの変異の傾向の解析に許容できる精度を発揮するための適用可能条件とを調べた。本発明によるDNAのシーケンシング方法における変異の見逃しが、どういった条件でどの程度起こる可能性があるのか、及び、そのような見逃しを回避するためのシーケンシング条件について以下に記載する。

【0070】

本発明のシーケンシング方法の効率及び精度に影響を与える可能性のある因子は以下のとおりである。

〔1〕 サンプルDNAの配列データの得られる効率（解析効率）

10

本発明のシーケンシング方法においては、サンプルDNA断片のPCRを行い、その産物をライブラリ（様々なDNA断片のPCR産物の混合物）としてシーケンシングに供する。従って、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、当該ライブラリ中に同一DNA断片由来のPCR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。すなわち、適切なPCR条件とシーケンシングデータ量の設定が解析効率にとって重要である。

【0071】

さらに、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に影響している。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数が減少すると、サンプルDNAの配列データを作成できる確率が低下し、解析効率は低くなる。一方、グループあたりのリード配列又はリードペア数が多くなりすぎても、一つの配列データ作成に使用されるリード数が多くなりすぎるため、やはり解析効率は低下する。

20

【0072】

〔2〕 異なるDNA断片を同一断片と誤認識すること（断片の誤認識）の割合

断片の誤認識は、ライブラリ中に異なるサンプルDNA断片に由来するが配列が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（PCR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。

30

【0073】

断片の誤認識は、上述したリード配列又はリードペアのグループに、同一のサンプルDNA断片に由来しないリード配列又はリードペアが含まれる割合に依存し、これは(1)ライブラリ中の同一のサンプルDNA断片に由来するPCR産物がシーケンシングされる確率と、(2)配列が偶然一致する（又は一致するとみなされる程度に高い同一性を有する）異なるDNA断片由来のPCR産物がシーケンシングされる確率と、の比率によって決定される。PCRでの初期DNA量を減少させ、PCRのサイクル数を増やすと、全PCR産物中における同一断片由来の産物の割合が増加して、(1)の確率は増加し、逆に(2)の確率は低下するので、断片の誤認識は低下する。一方、(1)と(2)の比率はシーケンシングデータ量の大小によらないので、断片の誤認識率はシーケンシングデータ量に依存しない。

40

【0074】

以上の〔1〕及び〔2〕から、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時のPCR条件、特にPCR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、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。

【0075】

50

適切なPCR初期DNA量は、サンプルDNAのサイズに依存し得る。小さなサイズのサンプルDNAに対してPCR初期DNA量が多すぎる場合、断片の誤認識の確率が高まり、正しく変異を検出できない可能性がある。また、大きなサイズのサンプルDNAに対してPCR初期DNA量が多い場合、後述の解析効率を考慮したときに、必要なシーケンスデータ量が膨大になる可能性がある。一方、大きなサイズのサンプルDNAに対してPCR初期DNA量が少なすぎる場合、サンプルDNAの配列の多様度を十分に反映できなくなるため、ゲノムのカバー率が低下する可能性がある。しかし、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、サンプルDNAのサイズに対して少ないPCR初期DNA量を用いても、解析効率が低下しない範囲であれば、特に問題ではない。このようなケースとしては、サンプルDNA全体のうち、ランダムに一部の領域を選択して、該領域内での変異の頻度を把握したい場合などが考えられる。従って、本発明のシーケンシング方法における、サンプルDNAのサイズに対してのPCR初期DNA量は、断片の誤認識の確率とゲノムのカバー率の観点から、その研究の目的に応じて適切な範囲が選択され得る。本発明のシーケンシング方法におけるPCR初期DNA量は、断片の誤認識の確率を低くすることと適切なシーケンスデータ量との観点から、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PCR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、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、さらに好ましくは1amol以上、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。

【0076】

一例において、本発明のシーケンシング方法におけるPCR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、PCR初期DNA量はPCR反応に使用するDNAサンプル中のDNA量であり、PCRプライマーなどのDNA量は含まない。

【0077】

さらに、適切なPCR初期DNA量は、解析に用いることができるデータ（配列情報）の量に依存し得る。PCR初期DNA量が少なすぎる場合、その中に含まれるサンプルDNA断片数が少なくなり、解析に用いることができるデータ（配列情報）を十分に確保できなくなる。本発明のシーケンシング方法におけるPCR初期DNA量は、好ましくは0.1amol以上、より好ましくは1amol以上、さらに好ましくは5amol以上、なお好ましくは20amol以上、さらになお好ましくは39amol以上、さらになお好ましくは78amol以上である。一方、解析効率の観点からは、PCR初期DNA量は、好ましくは100000amol以下、より好ましくは20000amol以下、さらに好ましくは5000amol以下である。例えば、本発明のシーケンシング方法におけるPCR初期DNA量は、好ましくは0.1~100000amol、より好ましくは1~100000amol、さらに好ましくは5~100000amol、なお好ましくは20~100000amol、さらになお好ましくは20~20000amol、さらになお好ましくは39~20000amol、さらになお好ましくは78~20000amol、さらになお好ましくは20~5000amol、さらになお好ましくは39~5000amol、さらになお好ましくは78~5000amolである。

【0078】

本発明のシーケンシング方法において、適切なシーケンシングデータ量は、PCR初期DNA量に依存し得る。PCR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る。また、本発明の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、PCR初期DNA量1amol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4Mbp、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200bp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（8Mbp）以上、さらに好ましくは 0.08×10^6 個（16Mbp）以上、なお好ましくは 0.16×10^6 個（32Mbp）以上であり、かつ、好ましくは 10×10^6 個（2000Mbp）以下、より好ましくは 5×10^6 個（1000Mbp）以下、さらに好ましくは 2.5×10^6 個（500Mbp）以下、なお好ましくは 2×10^6 個（400Mbp）以下である。例えば、本発明の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、PCR初期DNA量1amol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（4~2000Mbp）、より好ましくは $0.04 \sim 5 \times 10^6$ 個（8~1000Mbp）、さらに好ましくは $0.08 \sim 2.5 \times 10^6$ 個（16~500Mbp）、なお好ましくは $0.16 \sim 2 \times 10^6$ 個（32~400Mbp）である。

【0079】

当該PCR初期DNA量及びシーケンシングデータ量は、それぞれ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に影響する。一方、異なるPCR初期DNA量及びシーケンシングデータ量の条件の間で、最大の解析効率をもたらす条件におけ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の平均値は、ほぼ一定である（後述の実施例、表3を参照）。したがって、本発明のシーケンシング方法においては、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が所定の範囲になるように各種条件を設定することで、最適な解析効率及び精度でのシーケンシングが可能になる。本発明のシーケンシング方法において、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、本発明の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。

【0080】

上記のことから、適切なシーケンシングデータ量はまた、サンプルDNAのサイズに依存し得る。大きなサイズのサンプルDNAに対しては、そのゲノムのカバー率を十分に高くするためにはより多くのPCR初期DNA量が必要となる。従って、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少なすぎる場合、配列データを作成するための十分なリードペア数が確保できなくなり解析効率が低下する可能性がある。しかし、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、サンプルDNAのサイズに対して少ないシーケンシングデータ量を用いても、解析効率が低下しない範囲であれば、特に問題ではない。一方、小さなサイズのサンプルDNAに対しては、より少ないPCR初期DNA量で、そのゲノムのカバー率を高くすることができる。従って、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。本発明の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1Mbpあたりのリード配列又はリードペア数で、好ましくは 0.05×10^6 個(10Mbp)以上、より好ましくは 0.1×10^6 個(20Mbp)以上、さらに好ましくは 0.2×10^6 個(40Mbp)以上、なお好ましくは 0.5×10^6 個(100Mbp)以上、さらになお好ましくは 1×10^6 個(200Mbp)以上、さらになお好ましくは 2×10^6 個(0.4Gbp)以上であり、かつ、好ましくは 1600×10^6 個(320Gbp)以下、より好ましくは 800×10^6 個(160Gbp)以下、さらに好ましくは 400×10^6 個(80Gbp)以下、なお好ましくは 200×10^6 個(40Gbp)以下、さらになお好ましくは 100×10^6 個(20Gbp)以下、さらになお好ましくは 50×10^6 個(10Gbp)以下である。例えば、本発明の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1Mbpあたりリード配列又はリードペア数で、好ましくは $0.05 \sim 1600 \times 10^6$ 個($0.01 \sim 320$ Gbp)、より好ましくは $0.1 \sim 800 \times 10^6$ 個($0.02 \sim 160$ Gbp)、さらに好ましくは $0.2 \sim 400 \times 10^6$ 個($0.04 \sim 80$ Gbp)、なお好ましくは $0.5 \sim 200 \times 10^6$ 個($0.1 \sim 40$ Gbp)、さらになお好ましくは $1 \sim 100 \times 10^6$ 個($0.2 \sim 20$ Gbp)、さらになお好ましくは $2 \sim 50 \times 10^6$ 個($0.4 \sim 10$ 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、本発明の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1Mbpあたりのリード配列又はリードペア数で、 0.05×10^6 個(10Mbp)未満であってもよい。例えば、ゲノムサイズ約3Gbp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1Mbpあたりのリード配列またはリードペア数で、好ましくは $0.00003 \sim 16 \times 10^6$ 個($0.006 \sim 3200$ Mbp)、より好ましくは $0.00007 \sim 8 \times 10^6$ 個($0.014 \sim 1600$ Mbp)、さらに好ましくは $0.0001 \sim 4 \times 10^6$ 個($0.02 \sim 800$ Mbp)、なお好ましくは $0.0003 \sim 2 \times 10^6$ 個($0.06 \sim 400$ Mbp)、さらになお好ましくは $0.0005 \sim 1 \times 10^6$ 個($0.1 \sim 200$ Mbp)、さらになお好ましくは $0.001 \sim 0.5 \times 10^6$ 個($0.2 \sim 100$ Mbp)である。

【0081】

本発明のシーケンシング方法の有効性は、サンプルDNAのサイズに依存し得る。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。このことは、サンプルDNAのサイズ低下に伴って、PCRの初期DNA量を少なくする必要がある、よって得られる最終的な配列データが少なくなってしまうことを意味する。したがって、本発明のシーケンシング方法をゲノム変異解析に応用する場合、サンプルDNAのサイズに依存して、十分な変異データを解析できなくなる可能性がある。本発明の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10kbp以上、より好ましくは100kbp以上、さらに好ましくは1Mbp以上、なお好ましくは4Mbp以上である。なお、サンプルDNAのサイズ

は、該サンプルDNAの由来する生物のゲノムDNAのサイズであることが好ましく、該ゲノムDNAのうちのシーケンシング可能領域のDNAのサイズであることがより好ましい。一方、サンプルDNAのサイズが、該ゲノムDNAのうちのシーケンシング対象領域のDNAのサイズ、又は該ゲノムDNAのうちの解析対象領域のDNAのサイズであることも好ましい。該シーケンシング可能領域としては、PCR反応やシーケンシング反応が可能な領域が挙げられ、該シーケンシング対象領域としては、哺乳動物細胞等におけるエクソン領域などが挙げられ、該解析対象領域としては、解析に用いる参照配列の領域や、リード配列又はリードペアのグループを作成することができた参照配列の領域などが挙げられる。

【0082】

10

本発明の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 $10 \sim 1250 \text{ amol}$ 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 $0.2 \sim 12500 \times 10^6$ 個($0.04 \sim 2500 \text{ Gbp}$)、好ましくは $0.4 \sim 6250 \times 10^6$ 個($0.08 \sim 1250 \text{ Gbp}$)、より好ましくは $0.8 \sim 3125 \times 10^6$ 個($0.16 \sim 625 \text{ Gbp}$)、さらに好ましくは $1.6 \sim 2500 \times 10^6$ 個($0.32 \sim 500 \text{ Gbp}$)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 $20 \sim 625 \text{ amol}$ 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 $0.4 \sim 6250 \times 10^6$ 個($0.08 \sim 1250 \text{ Gbp}$)、好ましくは $0.8 \sim 3125 \times 10^6$ 個($0.16 \sim 625 \text{ Gbp}$)、より好ましくは $1.6 \sim 1563 \times 10^6$ 個($0.32 \sim 313 \text{ Gbp}$)、さらに好ましくは $3.2 \sim 1250 \times 10^6$ 個($0.64 \sim 250 \text{ Gbp}$)である。

20

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 $39 \sim 313 \text{ amol}$ 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 $0.78 \sim 3130 \times 10^6$ 個($0.156 \sim 626 \text{ Gbp}$)、好ましくは $1.56 \sim 1565 \times 10^6$ 個($0.312 \sim 313 \text{ Gbp}$)、より好ましくは $3.12 \sim 783 \times 10^6$ 個($0.624 \sim 157 \text{ Gbp}$)、さらに好ましくは $6.24 \sim 626 \times 10^6$ 個($1.248 \sim 125 \text{ Gbp}$)である。

本発明の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、 $1.05 \sim 30$ 、好ましくは $1.1 \sim 20$ 、さらに好ましくは $1.2 \sim 10$ 、なお好ましくは $1.4 \sim 5$ である。

30

【0083】

本発明の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 $10 \sim 5000 \text{ amol}$ 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 $0.2 \sim 50000 \times 10^6$ 個($0.04 \sim 10000 \text{ Gbp}$)、好ましくは $0.4 \sim 25000 \times 10^6$ 個($0.08 \sim 5000 \text{ Gbp}$)、より好ましくは $0.8 \sim 12500 \times 10^6$ 個($0.16 \sim 2500 \text{ Gbp}$)、さらに好ましくは $1.6 \sim 10000 \times 10^6$ 個($0.32 \sim 2000 \text{ Gbp}$)である。

40

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 $20 \sim 2500 \text{ amol}$ 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 $0.4 \sim 25000 \times 10^6$ 個($0.08 \sim 5000 \text{ Gbp}$)、好ましくは $0.8 \sim 12500 \times 10^6$ 個($0.16 \sim 2500 \text{ Gbp}$)、より好ましくは $1.6 \sim 6250 \times 10^6$ 個($0.32 \sim 1250 \text{ Gbp}$)、さらに好ましくは $3.2 \sim 5000 \times 10^6$ 個($0.64 \sim 1000 \text{ Gbp}$)である。

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 $39 \sim 1250 \text{ amol}$ 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 $0.78 \sim 12500 \times 10^6$ 個($0.156 \sim 2500 \text{ Gbp}$)、好ましくは 1

50

． 56 ～ 6250 × 10⁶ 個 (0.312 ～ 1250 Gbp)、より好ましくは 3.12 ～ 3125 × 10⁶ 個 (0.624 ～ 625 Gbp)、さらに好ましくは 6.24 ～ 2500 × 10⁶ 個 (1.248 ～ 500 Gbp) である。

本発明の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 3 Gbp 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。

【0084】

サイズ約 5 Mbp のサンプル DNA の例としては、サルモネラ菌のゲノム (約 4.86 Mbp) が挙げられる。サルモネラ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。

【0085】

(4. シーケンシング方法の応用)

本発明による DNA のシーケンシング方法で得られた配列データは、シーケンシングにおける読み取りエラーや、サンプル DNA の酸化修飾等に起因する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、本発明による DNA のシーケンシング方法は、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 をサンプル DNA として本発明のシーケンシング方法を行い、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。

【0086】

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 DNA のシーケンシング方法を実施し、配列データを作成すること；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、を含む。

【0087】

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、試験物質に曝露した細胞のゲノム DNA である。好ましくは、該ゲノム DNA は、試験物質に曝露した細胞 (被験細胞) のゲノム DNA と、該試験物質に曝露していない細胞 (対照細胞) のゲノム DNA であり、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、試験物質の曝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、サルモネラ菌細胞、及び大腸菌細胞が挙げられるが、これらに限定されない。サルモネラ菌の好ましい例は、上述したとおりである。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、WP2 uvrA 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D T 4 0などが挙げられる。

【0088】

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。細胞を試験物質に曝露する手段は、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

【0089】

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であり、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

【0090】

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。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。

30

【0091】

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。

40

【0092】

好ましい実施形態において、本発明では、解析対象ゲノムDNA全体における変異の傾向を調べるために、ゲノムの特定部位の変異の有無を厳密に同定するよりも、ゲノムのより広い領域の変異を検出し、それらの変異パターンを解析して、広いゲノム領域内での変

50

異の傾向を解析することが優先される。

【0093】

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、本発明によるDNAのシーケンシング方法で取得された配列データが参照配列と比較されて、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。

【0094】

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行うことができる。

【0095】

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。

【0096】

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进行分类し（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。

【0097】

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、T

CC、TCG、及びTC T」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。

10

【0098】

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。

【0099】

20

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。

30

【0100】

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。

【0101】

<1> DNAのシーケンシング方法であって、

(1) サンプルDNAの断片を調製すること；

(2) 該サンプルDNAの断片をPCRにかけてPCR産物を得ること、ここで、該サンプルDNAの断片の各々について2つ以上の増幅断片が作製される；

40

(3) 得られたPCR産物をシーケンシングし、該PCR産物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；

(4) 得られた複数の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、ここで、該1つ以上のグループは、平均で1.05～30個のリード配列を含む；

(5) 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、方法。

50

< 2 > DNAのシーケンシング方法であって、

(1) サンプルDNAの断片を調製すること；

(2) 該サンプルDNAの断片をPCRにかけてPCR産物を得ること、ここで、該サンプルDNAの断片の各々について2つ以上の増幅断片が作製され、該PCRでの初期DNA量は、該サンプルDNAのサイズ1Mbpあたり250amol以下である；

(3) 得られたPCR産物をシーケンシングし、該PCR産物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；

(4) 得られた複数の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；

10

(5) 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、方法。

< 3 > 好ましくは、前記(3)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 >又は< 2 >記載の方法。

< 4 > 好ましくは、前記(4)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 >～< 3 >のいずれか1項記載の方法。

< 5 > 好ましくは、前記(5)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、< 3 >又は< 4 >記載の方法。

20

【0102】

< 6 > 前記< 1 >記載の方法であって、好ましくは、

前記(3)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本の相補鎖のうちの一方の鎖の配列を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、

30

前記(4)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み、ここで、該1つ以上のグループは、平均で1.05～30個のリード配列のペアを含み、

前記(5)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、方法。

< 7 > 前記< 2 >記載の方法であって、好ましくは、

前記(3)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：

40

リード1：前記増幅断片を構成する2本の相補鎖のうちの一方の鎖の配列を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(4)が、得られたリード配列のペア中から、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み、

前記(5)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、

50

方法。

< 8 > 好ましくは、前記 (3) 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、< 6 > 又は < 7 > 記載の方法。

< 9 > 好ましくは、前記 (4) 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、< 6 > ~ < 8 > のいずれか 1 項記載の方法。

< 10 > 好ましくは、前記 (4) 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、< 9 > 記載の方法。

< 11 > 好ましくは、前記サンプル DNA 断片の 2 本の相補鎖が、それらの 5' 末端側と 3' 末端側に異なる標識配列を有し、かつ、前記シーケンシングにより、各リード配列に付随する該標識配列の情報が取得され、

前記リード配列のペアのグループの中のリード配列のペアを、それに含まれる 2 本のリード配列に付随する該標識配列の情報と、該 2 本のリード配列の前記参照配列上での互いの位置関係に基づいて、サンプル DNA 断片の 2 本の相補鎖のどちらに由来するか識別することをさらに含む、< 10 > 記載の方法。

< 12 > 好ましくは、前記 (5) 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、< 8 > ~ < 11 > のいずれか 1 項記載の方法。

【0103】

< 13 > 前記 (4) で得られた前記 1 つ以上のリード配列のグループに含まれるリード配列の数が、該グループ間の平均で、

好ましくは 1.05 ~ 30 個

より好ましくは 1.1 ~ 20 個、

さらに好ましくは 1.2 ~ 10 個、

なお好ましくは 1.4 ~ 5 個、

である、< 1 > ~ < 5 > のいずれか 1 項記載の方法。

< 14 > 前記シーケンシングにおいて、前記 PCR での初期 DNA 量 1 amol あたり、

好ましくは 0.02 ~ 10×10^6 個、

より好ましくは 0.04 ~ 5×10^6 個、

さらに好ましくは 0.08 ~ 2.5×10^6 個、

なお好ましくは 0.16 ~ 2×10^6 個、

のリード配列を取得するか、

あるいは、

好ましくは 4 ~ 2000 Mb p、

より好ましくは 8 ~ 1000 Mb p、

さらに好ましくは 16 ~ 500 Mb p、

なお好ましくは 32 ~ 400 Mb p、

のリード配列を取得する、

< 1 > ~ < 5 > 及び < 13 > のいずれか 1 項記載の方法。

【0104】

< 15 > 前記 (4) で得られた前記 1 つ以上のリード配列のペアのグループに含まれるリード配列のペアの数が、該グループ間の平均で、

好ましくは 1.05 ~ 30 個

好ましくは 1.1 ~ 20 個、

10

20

30

40

50

さらに好ましくは1.2～10個、
なお好ましくは1.4～5個、
である、＜6＞～＜12＞のいずれか1項記載の方法。
＜16＞前記シーケンシングにおいて、前記PCRでの初期DNA量1amolあたり、
好ましくは0.02～10×10⁶個、
より好ましくは0.04～5×10⁶個、
さらに好ましくは0.08～2.5×10⁶個、
なお好ましくは0.16～2×10⁶個、
のリード配列のペアを取得するか、
あるいは、

10

好ましくは4～2000Mbp、
より好ましくは8～1000Mbp、
さらに好ましくは16～500Mbp、
なお好ましくは32～400Mbp、
のリード配列のペアを取得する、
＜6＞～＜12＞及び＜15＞のいずれか1項記載の方法。

【0105】

＜17＞前記（2）において、前記PCRでの前記サンプルDNAのサイズ1Mbpあたりの初期DNA量が、

20

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、

30

である、＜1＞～＜16＞のいずれか1項記載の方法。

＜18＞前記（2）において、前記PCRでの前記サンプルDNAのサイズ1Mbpあたりの初期DNA量が、

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、

40

である、＜1＞～＜17＞のいずれか1項記載の方法。

＜19＞前記（2）において、前記PCRでの初期DNA量が、

50

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、 $\langle 1 \rangle \sim \langle 18 \rangle$ のいずれか1項記載の方法。
 $\langle 20 \rangle$ 前記(2)において、前記PCRでの初期DNA量が、
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、 $\langle 1 \rangle \sim \langle 19 \rangle$ のいずれか1項記載の方法。
 $\langle 21 \rangle$ 前記サンプルDNAが、
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、
のサイズを有する、 $\langle 1 \rangle \sim \langle 20 \rangle$ のいずれか1項記載の方法。
 $\langle 22 \rangle$ 前記シーケンシングにおいて、前記サンプルDNAの 1 Mbp あたり、
好ましくは $0.05 \sim 1600 \times 10^6$ 個、
より好ましくは $0.1 \sim 800 \times 10^6$ 個、
さらに好ましくは $0.2 \sim 400 \times 10^6$ 個、
なお好ましくは $0.5 \sim 200 \times 10^6$ 個、
さらになお好ましくは $1 \sim 100 \times 10^6$ 個、
さらになお好ましくは $2 \sim 50 \times 10^6$ 個、
のリード配列又はリード配列のペアを取得するか、
あるいは、
好ましくは $0.01 \sim 320 \text{ Gbp}$ 、
より好ましくは $0.02 \sim 160 \text{ Gbp}$ 、
さらに好ましくは $0.04 \sim 80 \text{ Gbp}$ 、
なお好ましくは $0.1 \sim 40 \text{ Gbp}$ 、
さらになお好ましくは $0.2 \sim 20 \text{ Gbp}$ 、
さらになお好ましくは $0.4 \sim 10 \text{ Gbp}$ 、
のリード配列又はリード配列のペアを取得する、
 $\langle 1 \rangle \sim \langle 21 \rangle$ のいずれか1項記載の方法。
 $\langle 23 \rangle$ 好ましくは、前記サンプルDNAが哺乳動物由来ゲノムDNAであり、かつ前記
シーケンシングにおいて、該サンプルDNAの 1 Mbp あたり、
好ましくは $0.00003 \sim 16 \times 10^6$ 個、
より好ましくは $0.00007 \sim 8 \times 10^6$ 個、
さらに好ましくは $0.0001 \sim 4 \times 10^6$ 個、
なお好ましくは $0.0003 \sim 2 \times 10^6$ 個、
さらになお好ましくは $0.0005 \sim 1 \times 10^6$ 個、
さらになお好ましくは $0.001 \sim 0.5 \times 10^6$ 個、
のリード配列又はリード配列のペアを取得するか、
あるいは、
好ましくは $0.006 \sim 3200 \text{ Mbp}$ 、
より好ましくは $0.014 \sim 1600 \text{ Mbp}$ 、
さらに好ましくは $0.02 \sim 800 \text{ Mbp}$ 、
なお好ましくは $0.06 \sim 400 \text{ Mbp}$ 、
さらになお好ましくは $0.1 \sim 200 \text{ Mbp}$ 、

さらになお好ましくは0.2～100Mbp、
のリード配列又はリード配列のペアを取得する、
<1>～<21>のいずれか1項記載の方法。
<24>好ましくは、前記PCRにかけるサンプルDNAの断片に、個別のサンプルDNAの断片を識別するための標識が連結されていない、<1>～<23>のいずれか1項記載の方法。

【0106】

<25>ゲノムDNAの変異を検出する方法であって、

細胞中のゲノムDNAをサンプルDNAとして用いて、<1>～<24>のいずれか1項記載のDNAのシーケンシング方法を実施し、配列データを作成すること；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、ここで該参照配列は、該ゲノムDNA中の既知配列である、を含む、方法。

10

<26>好ましくは、前記ゲノムDNAが試験物質に曝露した細胞のゲノムDNAを含む、<25>記載の方法。

<27>好ましくは、

前記ゲノムDNAが試験物質に曝露した細胞のゲノムDNAと、該試験物質に曝露していない対照細胞のゲノムDNAを含み、

該試験物質に曝露した細胞のゲノム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を比較することをさらに含む、

20

<26>記載の方法。

<28>前記細胞が、

好ましくは微生物細胞であり、

より好ましくは大腸菌細胞、サルモネラ菌細胞、又は酵母細胞であり、

さらに好ましくはサルモネラ菌細胞である、

<25>～<27>のいずれか1項記載の方法。

<29>好ましくは、前記サルモネラ菌がS. typhimurium LT-2株、TA100株、TA98株、TA1535株、TA1538株又はTA1537株である、

<28>記載の方法。

<30>好ましくは、前記細胞が哺乳動物細胞、又は哺乳動物由来培養細胞である、<25>～<27>のいずれか1項記載の方法。

30

<31>好ましくは、前記哺乳動物がヒト又はマウスである、<30>記載の方法。

<32>好ましくは、前記ゲノムDNAの変異が経時変化によるゲノムDNAの変異を含む、<25>記載の方法。

<33>好ましくは、

前記ゲノムDNAが、経時変化した細胞のゲノムDNAと、より経時変化していない細胞のゲノムDNAを含み、

該経時変化した細胞のゲノムDNAについて検出した変異と、該より経時変化していない細胞のゲノムDNAについて検出した変異とを比較することをさらに含む、

<32>記載の方法。

40

<34>好ましくは、前記細胞が哺乳動物細胞、又は哺乳動物由来培養細胞である、<32>又は<33>記載の方法。

<35>好ましくは、前記細胞が培養細胞である、<32>又は<33>記載の方法。

<36>好ましくは、前記変異が塩基対置換型変異である、<25>～<35>のいずれか1項記載の方法。

【実施例】

【0107】

以下、実施例を示し、本発明をより具体的に説明する。

【0108】

以下の実施例1では、サンプルDNAの一断片に対して複数のリード配列の情報をを用い

50

ることで、個別のDNA断片を識別する標識を用いることなく、サンプルDNAの配列情報のみに基づいて高精度なシーケンシングを可能にするアルゴリズムを構築した。

【0109】

さらに、実施例2～5では、当該アルゴリズムを用いたシーケンシング方法の最適条件を検討した。上述したとおり、シーケンシングの効率及び精度に影響を与える因子として、主にライブラリ調製時のPCR条件、シーケンシング量、そして解析対象ゲノムのサイズが重要と考えられた。そこで、これらの因子を調整して、サンプルDNAの相補鎖情報を含む複数のリード配列を用いるシーケンシング方法について、化学物質等によるゲノム全体の変異の傾向を捉える上で必要な精度を達成し、効率の最大化が得られる最適条件の探索を行った。本実施例では、シーケンサーに、現在最も汎用されているイルミナ社のHiSeqを用いた。

10

【0110】

実施例1 複数のリード配列を用いたシーケンシングアルゴリズム

1) DNA相補鎖情報を用いたシーケンシングアルゴリズム

本アルゴリズムでは、個別のDNA断片を識別する標識（断片分子固有のタグ配列等）を用いることなく、同一DNA断片のPCR産物に由来すると推定されるリード配列（リードペア）を集め、次いで、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、配列データとして取得する。得られた相補鎖間コンセンサスリード配列は、解析対象DNAの変異解析などに利用することができる。

20

【0111】

当該相補鎖間コンセンサスリード配列の取得の基本的なフローを以下に説明する。

i) 本実施例で用いるライブラリプールには、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、このライブラリプールがシーケンシングされた場合、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が作成され得る。

【0112】

ii) 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、非対称の標識DNA配列を末端に含むアダプター配列を連結する。全てのサンプルDNA断片の両末端に、それぞれ同一のアダプター配列が付加される。これにより、全てのサンプルDNA断片のA鎖の5'末端とB鎖の5'末端、及びA鎖の3'末端とB鎖の3'末端には、それぞれ同じ標識配列が付加され、一方、1つの鎖の中で5'末端と3'末端には異なる標識配列が付加される。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プールとして用いる。増幅産物中に含まれるアダプター配列が、後述のシーケンシングの際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、フローセル上に増幅産物が結合される。通常片側の鎖に由来する増幅産物のみをフローセル上に存在させた後に、シーケンシング反応を行う。

30

【0113】

iii) 得られたライブラリプールについてシーケンシングを行う。シーケンシングでは、ライブラリプール中の各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）のペアが取得される。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）である。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。なお参考のため、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該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ることができる。

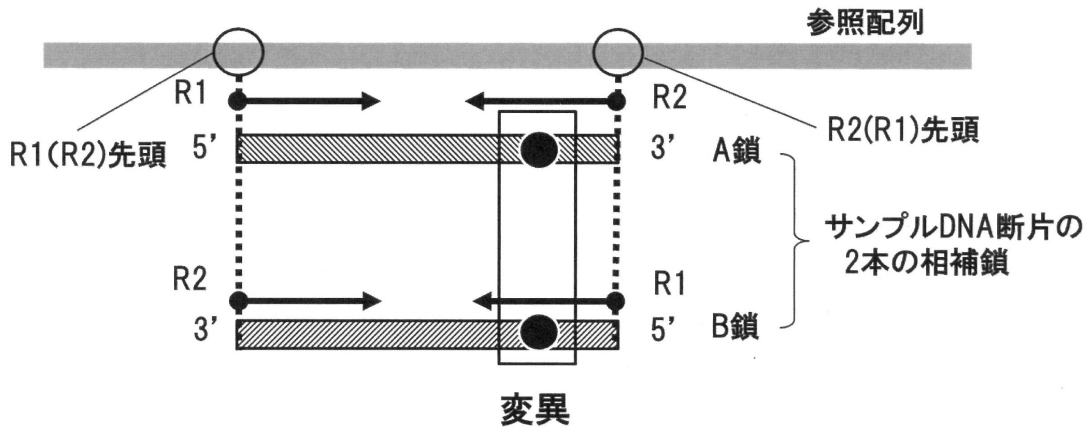
40

50

【0114】

【化2】

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

【0115】

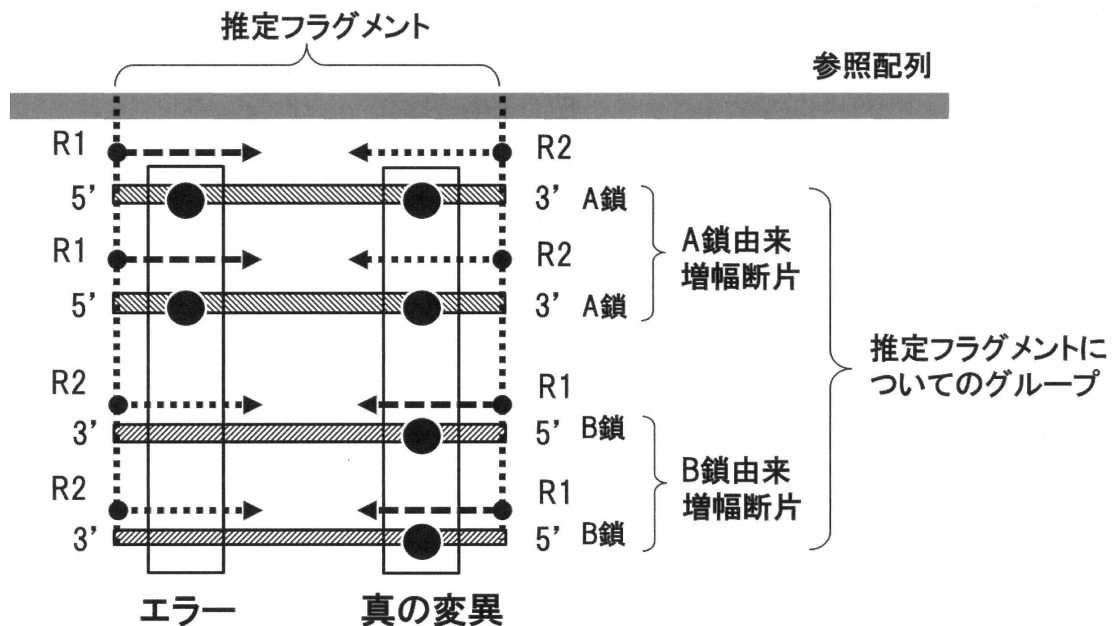
なお、以下の実施例において、用語を以下のように定義する：参照配列上における、マッピングしたリード1（リード2）の先頭からリード2（リード1）の先頭までの領域、言い換えると、リードペア（リード1、リード2）を参照配列上にマッピングしたときに、リード1の先頭とリード2の先頭とに挟まれる該参照配列の領域を、「推定フラグメント」と称する。推定フラグメントが共通するリードペアの群を、推定フラグメントについての「グループ」と称する（模式図2）。

20

【0116】

【化3】

模式図2: 推定フラグメント



30

40

【0117】

iv) 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する。

【0118】

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断片の塩基修飾などに起

50

因する塩基の変更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、変異による塩基置換を両鎖に 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。したがって、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、真の変異を抽出することができる。

【0119】

本アルゴリズムでは、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する。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る。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する。A 鎖由来、または 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。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する。

【0120】

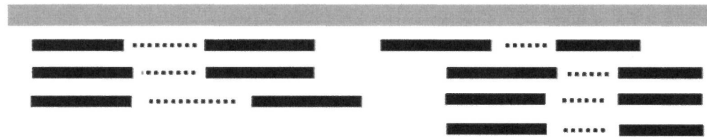
リードペアの集合化から相補鎖間コンセンサスリード配列作成までのより具体的な手順の例を、以下の模式図 3 に示す。模式図 3 の手順では、まず、各相補鎖由来のリードペアを参照配列にマッピングする（1）。このとき、参照配列上で左端（参照配列上の最も 5' 側に配置する端）が同じ位置にあるリードペアの群を第一集合として取得する（2）。次いで、該第一集合から、参照配列上で右端（参照配列上の最も 3' 側に配置する端）が同じ位置にあるリードペアの群を分け、第二集合として取得する（3）。この第二集合は、推定フラグメントの共通するリードペアの集合である。次いで、第二集合を、A 鎖に由来する群（F 群）と、B 鎖に由来する群（R 群）とに分ける（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る。本実施例においては、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る。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することができる（5）。

【0121】

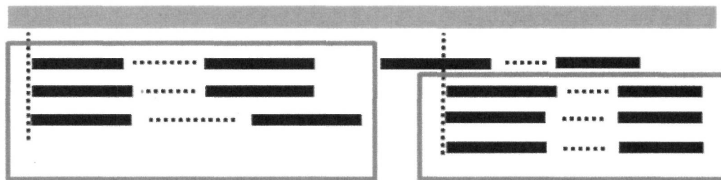
【化 4】

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

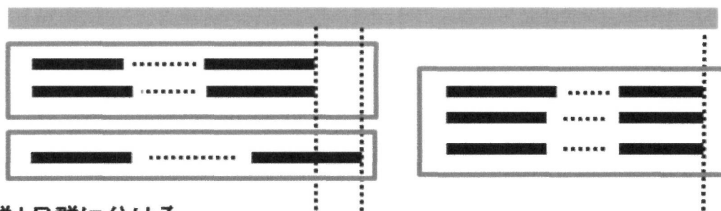
(1) 参照配列へのリードペアのマッピング



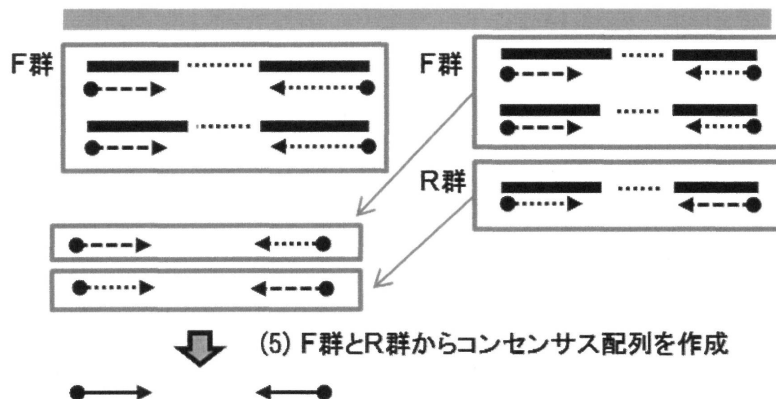
(2) 左端の位置が同じペアを第一集合として取得



(3) 右端の位置が同じペアを第二集合として取得



(4) F群とR群に分ける



(5) F群とR群からコンセンサス配列を作成



【0122】

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変異はエラーとして除外し、両方の鎖に共通して存在する変異を真の変異として取得することができる。この手順により、相補鎖情報を用いて真の変異から解析エラーを除外し、シーケンシングの精度を向上させることが理論的に可能となる。

【0123】

2) 相補鎖間コンセンサスリード配列に基づく変異解析

1) で得られた相補鎖間コンセンサスリード配列は、各種解析に利用することができる。代表的な例として、相補鎖間コンセンサスリード配列を用いたゲノムDNAの変異解析について以下に説明する。

【0124】

基本的には、まず、上記1)の手順で解析対象のゲノムDNAをシーケンシングし、相補鎖間コンセンサスリード配列を作成する。次いで、得られた相補鎖間コンセンサスリード配列を用いて、通常の手順により、解析対象ゲノムの変異を検出する。例えば、相補鎖間コンセンサスリード配列を再度参照配列上にマッピングすることで、解析対象ゲノムの変異を検出することができる。以下の実施例における実際のデータの変異解析は、模式図

10

20

30

40

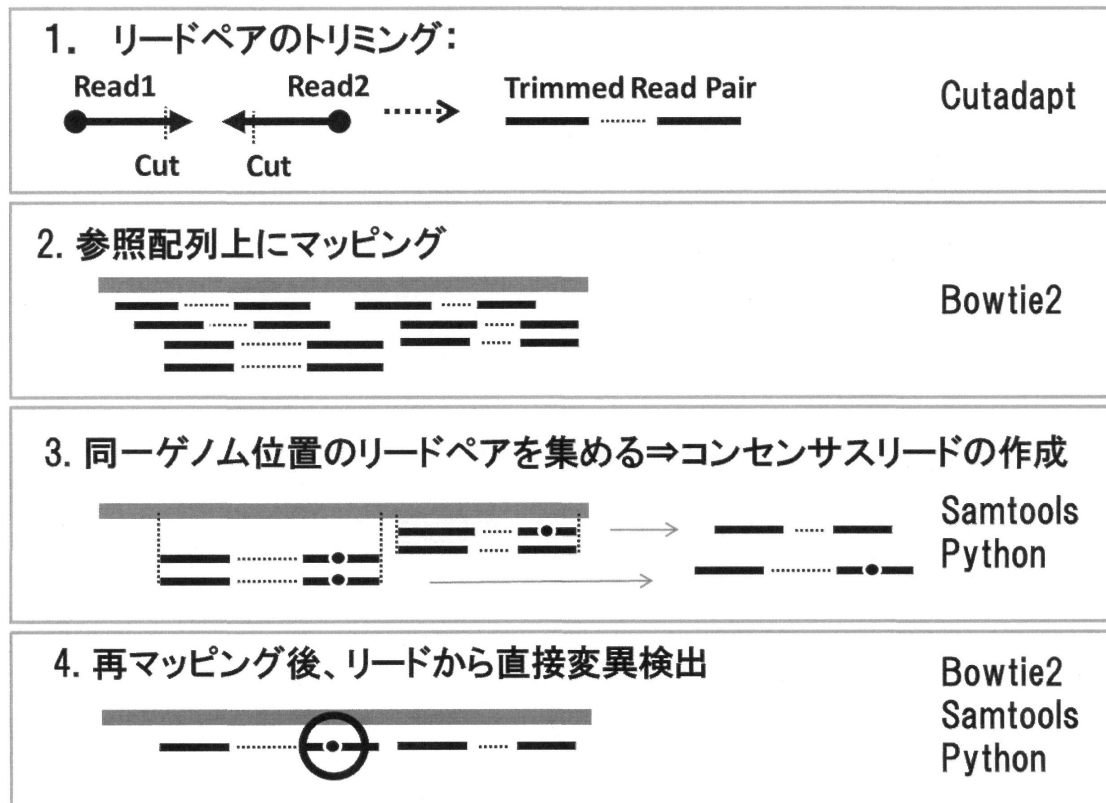
50

4に示したフローに従って行った。解析には、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。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（その全体を本明細書に援用する）に記載された手順に従って実施することができる。具体的な検出の手順は後述の実施例2及び3に記載する。

【0125】

【化5】

模式図4：解析フロー



【0126】

実施例2 変異原暴露DNAサンプルを用いた最適シーケンシング条件の探索

本実施例では、ゲノム変異解析用の配列データを得るための、実施例1に記載したアルゴリズムによるシーケンシングにおける最適条件を探索した。サンプルDNAには、代表的な変異原であるEthyl nitrosourea (ENU、CASRN. 759-73-9)を暴露したSalmonella typhimurium (S. typhimurium) LT-2 TA100株（以下、単に「TA100株」とも称する）のゲノムDNAを用いた。

【0127】

1) TA100株のゲノムDNAサンプルの調製

ENUへの細胞の曝露は、Ames試験のプレインキュベーション法 (K. Mortelmans et al., Mutat. Res. - Fundam. Mol. Mech. Mutagen., 455:29-60, 2000) に準拠して実施した。TA100株を2mLのニュートリエントブイヨン No. 2 (Oxoid社製) に植菌し、37℃、180rpmで4時間振とう培養し、O.D. 660値が1.0以上の前培養液を得た。ENU (シグマアルドリッチ社製) は、ジメチルスルホキシド (DMSO; 和光純薬工業製) で希釈した。試験管内に、ENU溶液100μL、0.1Mリン酸バッファー500μL、及び前培養液100μLを添加し (ENU量: 62.5、125、250、500、及び1000μg/tube)、37℃のウォーターバス中で2

0 分間、1 0 0 r p mで振とう培養した（E N U群）。コントロール群には、E N U溶液の代わりに溶媒（D M S O）1 0 0 μ Lを添加した。2 0 分間振とう培養後、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2 m LのN u t r i e n t B r o t h溶液に培養液5 0 μ Lを添加し、インキュベーター内で3 7 ℃、1 8 0 r p mで1 4 時間追培養した後、菌懸濁液を回収し、7 5 0 0 r p mで5 分間遠心し、上清を除去し、菌体を回収した。

【0 1 2 8】

また、A m e s 試験用に、上記と同様の条件で、プレインキュベーション法でE N Uを曝露した菌懸濁液を作製し、4 5 ℃に加温した2 m Lのt o p a g a r（1 % N a C l、1 % a g a r、0 . 0 5 m M H i s t i d i n e及び0 . 0 5 m M B i o t i n eを含む）を添加し、ボルテックスで懸濁した後、最小グルコース寒天培地（テスメディア（登録商標）A N；オリエンタル酵母工業製）上に重層した。得られたプレートで3 7 ℃で4 8 時間培養後、観察されたコロニーを計数した。

【0 1 2 9】

2) T o t a l D N Aの回収と濃度測定

1) で得られた菌体（コントロール群、およびE N U群）から、D N e a s y B l o o d & T i s s u e K i t（キアゲン社製）を用い、推奨プロトコールに従って、T o t a l D N Aを回収した。得られたD N Aサンプルの2 本鎖D N Aの濃度を、Q u b i t 3 . 0 F l u o r o m e t e r（T h e r m o F i s h e r S c i e n t i f i c社製）を用いて、付属のQ u b i tTM d s D N A B R A s s a y K i tで測定した。

【0 1 3 0】

3) シーケンシング用ライブラリの調製

2) で濃度を測定した各D N Aサンプル（コントロール群及びE N U群）の1 0 0 n g相当量をそれぞれ複数サンプル用意し、各々をD N A S h e a r i n gシステム M E 2 2 0（コバリス社製）で推奨プロトコールに従って平均約3 5 0 b pの長さに断片化した。得られた断片化D N AをT r u S e q N a n o D N A L i b r a r y P r e p K i t（イルミナ社製、以下T r u S e qと略記する）を用いて、推奨プロトコールに従ってライブラリ調製に供した。

【0 1 3 1】

ライブラリ調製では、得られたD N A断片を、実施例1で説明した標識配列を末端に含むアダプター配列で標識した。それらを用いて表1記載の異なる条件でP C Rを行い、シーケンシング用ライブラリを調製して、条件の異なるライブラリをそれぞれ標識した。アダプター配列は、イルミナ社のT r u S e qに付属のアダプター配列を使用した。なお、表1のとおり、使用するアダプター配列はライブラリごとに変更し、さらに各ライブラリには、それぞれ、異なるインデックス配列により区別される2種類の異なるアダプター（1st i n d e x及び2nd i n d e x）を使用した。したがって各ライブラリの各断片には、2種類のアダプターのどちらかが結合した。すなわち、各サンプル（コントロール群及びE N U群）のライブラリ調製液に、表1に従って、1st i n d e x又は2nd i n d e xのアダプターを添加して、それぞれ別々にライゲーション反応に供した。得られたライゲーション反応液を推奨プロトコールに従って精製した後、A g i l e n t 4 2 0 0 T a p e S t a t i o n（アジレント・テクノロジー社製）のH i g h S e n s i t i v i t y D 5 0 0 0 キットを用いてライゲーション産物の濃度測定を行った。測定濃度を基に、各サンプルを表1に従って、T r u S e qのR e s u s p e n s i o n B u f f e rで段階的に希釈し、1st i n d e xと2nd i n d e xのアダプターのリゲーション産物それぞれの希釈液各1 2 . 5 μ Lを混合し、2 5 μ Lとした。得られた混合液を、推奨プロトコールに従い、P C Rによる精製に供した。P C Rでの初期D N A量の差を調整するため、各サンプルを表1記載のP C Rサイクル数で増幅し、ライブラリのD N A量を調整した。

【0 1 3 2】

【表 1】

ライブラリ No.	1	2	3	4	5	6	7	8
DNA濃度 (pmol/L)	800	400	200	100	50	25	12.5	6.25
PCR input (amol)	20000	10000	5000	2500	1250	625	313	156
PCR input (amol)/ Mbp (genome)	4000	2000	1000	500	250	125	62.5	31.3
1 st Index*	2	4	5	6	7	12	13	10
2 nd Index*	15	16	18	19	1	3	8	9
PCR サイクル数	7	8	9	10	11	12	13	14

ライブラリ No.	9	10	11	12	13	14
DNA濃度 (pmol/L)	6.25	3.13	1.56	0.78	0.39	0.20
PCR input (amol)	156	78	39	20	10	5
PCR input (amol)/ Mbp (genome)	31.3	15.6	7.8	3.9	2	1
1 st Index*	2	4	5	6	7	12
2 nd Index*	13	14	15	16	18	19
PCR サイクル数	14	15	16	17	18	19

*: Indexの識別番号

【0133】

4) シーケンシング

3) で調製したライブラリを、HiSeq2500 (イルミナ社製) を用いて、表1記載のライブラリNo. 1～8については2×125bpのリード長で、ライブラリNo. 9～14については2×100bpのリード長でシーケンシングした。ライブラリあたり、平均で約10Gbpのシーケンシングデータを得た。

【0134】

5) リード配列の編集、相補鎖情報の抽出、及び変異解析

シーケンシングによって得られた生リード配列の編集及び変異解析は、上述の模式図4に示す解析フローに従って実施した。まず、Cutadaptソフトウェアを用いて、生リード配列からアダプター配列、及びクオリティの低い塩基等のトリミングを行いリード配列のペアを作成した。次に、得られたトリミング後のリードペアのFastqファイルについて、アダプター配列のインデックス情報を元に、同一ライブラリに由来するFastqファイルを1つのファイルにまとめた。なお、アダプター配列のインデックス情報は、リード配列情報にリンクされている情報であるが、リード配列に含まれる情報ではない。一方、より少ないシーケンシングデータ量での解析のため、約2Gbp (10×10⁶個 (10M) のリードペア) のシーケンシングデータの解析を実施した。すなわち、上記の約10GbpのシーケンシングデータのFastqファイル (10Gbp Fastqファイル) の先頭と末尾から、それぞれ5×10⁶個 (計10×10⁶個、10Mリードペア) のリードペアを抽出して1つのファイルにまとめ、約2Gbpでの解析用のFastqファイル (2Gbp Fastqファイル) を作成した。その後、これらのFastqファイル (10Gbp及び2Gbp) について、Bowtie2ソフトウェアを用いて参照配列へのマッピングを行い、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実施した後、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て、実施例1に示し

10

20

30

40

50

たアルゴリズムに従って、推定フラグメントが共通する2本の相補鎖に由来するリードペアのセットを抽出し、該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（模式図3参照）。得られた相補鎖間コンセンサスリード配列は、新たなFastq形式のファイルとして出力した。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を用いて作成したプログラムを用いて、変異解析を行った。なお、Bowtie2ソフトウェアでマッピングする参照配列には、PCT/JP2017/005700に記載されるS. typhimurium TA100株のゲノム配列を用いた。

【0135】

10

6) リードペア数分布の算出

5)で作成した各ライブラリからの10Gbp Fastqファイル及び2Gbp Fastqファイルそれぞれについて、マッピング結果を基に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、グループあたりリードペア数の分布を求めた。また、この分布から、各ライブラリにおける平均リードペア数/グループを算出した。

【0136】

7) 重なり率の算出

5)で作成した各ライブラリの10Gbp Fastqファイル及び2Gbp Fastqファイルそれぞれから、マッピング結果を基に、推定フラグメントについてのグループのうち2つ以上のリードペアが含まれるグループを抽出した。抽出したグループを対象として、各リードペアについてのアダプター配列のインデックス情報を元に、異なるDNA断片に由来するリードペア（同一断片の2本の相補鎖に由来しないリードペア）を含むグループの割合（重なり率、Overlap rate）を算出した。すなわち、各リードペアのインデックス配列（1st index又は2nd index）を確認し、1st index付リードペアと2nd index付リードペアの両方を含むグループの数、及び抽出した2つ以上のリードペアが含まれるグループ全体に対するそれらの割合（重なり率）を算出した（模式図5）。

20

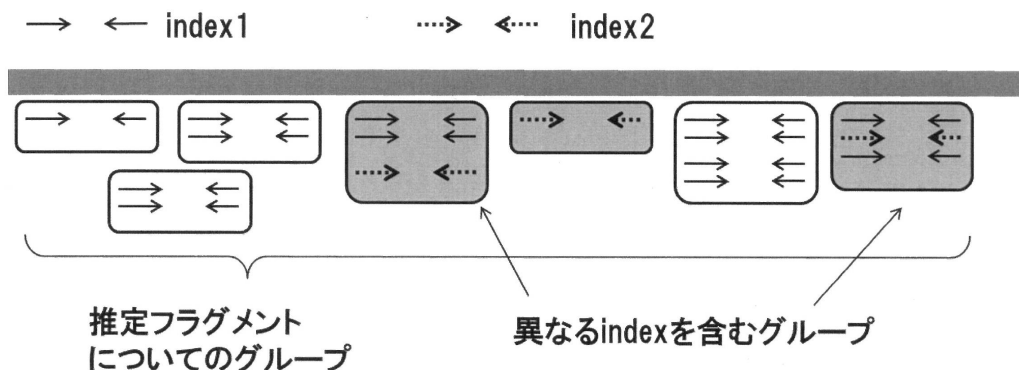
重なり率（％）＝（1st index付リードペアと2nd index付リードペアの両方を含むグループの数）／（2つ以上のリードペアが含まれるグループの数）×100

30

【0137】

【化6】

模式図5



40

【0138】

8) 変異頻度、及び解析効率の算出

5)で相補鎖間コンセンサスリード配列を参照配列にマッピングした後、PCT/JP2017/005700に記載された手順に従って、変異した塩基を検出した。すなわち、Pythonを用いたプログラムを用いて、各ライブラリについて、参照配列に対して

50

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によって4群に分けた。次いで各群の塩基の総数、及び参照配列に対して変異した塩基を検出した。検出した変異した塩基を参照配列の塩基と比較することで、コントロール群及びE N U群それぞれについて、解析対象塩基中におけるA T塩基対、G C塩基対の各 10^6 bpにおける各変異パターン（A T→T A、A T→C G、A T→G C、及びG C→T A、G C→C G、G C→A T）、及び各変異パターンの変異頻度を算出した。また、変異解析の際に用いた各ライブラリについてのコンセンサスリード配列中の解析対象塩基の総量（bp）と、各ライブラリのシーケンシングで読み取ったシーケンシングデータ量（bp）から、各ライブラリの解析効率（Data efficiency）を算出した。

$$\text{解析効率 (\%)} = \frac{\text{(相補鎖間コンセンサスリード配列中の解析対象塩基の総量)}}{\text{(シーケンシングデータ量)} \times 100}$$

【0139】

9) 結果と考察

I) Ames試験の復帰突然変異体数

表2にE N U曝露後の復帰突然変異体コロニー数を示す。データは3枚のプレートの測定値と、その平均値を示す。E N U曝露により、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、E N U曝露によりT A 100株のゲノム中に変異が導入されたことが確認できた。

【0140】

【表2】

Ames試験における復帰変異体コロニー数

ENU濃度 (μ g/plate)	復帰突然変異体数			
	1	2	3	Ave.
0 (DMSO)	116	129	114	119.7
62.5	220	192	210	207.3
125	430	391	323	381.3
250	1482	1359	1326	1389.0
500	4095	4006	3751	3950.7
1000	1659	3206	2737	2534.0

【0141】

II) シーケンシングにおける解析効率とリードペア数分布に対する初期DNA量の影響

i) 10Gbpシーケンシングデータ

6)で算出した、10Gbp Fastqファイルにおける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の分布を、図1～2に示す。図1にコントロール（DMSO曝露）群、図2にE N U（E N U曝露）群のデータを、PCRでの初期DNA量の異なるライブラリ（ライブラリNo. 1、4～14）ごとに示す。コントロール群及びE N U群のいずれにおいても、PCRでの初期DNA量を減らすにつれて、グループあたりのリードペア数が増加する傾向がみられた。また、PCRでの初期DNA量ごとに、8)で算出した解析効率を図3に示す。コントロール群において、初期DNA量 $20000 \sim 156 \text{amol}$ （約 $4000 \sim 31.3 \text{amol/Mbp (genome)}$ ）ライブラリのシーケンシングでは、 156amol で最も効率が良かったが、初期DNA量をさらに減少させることでさらに効率が上がることが推測できた。そこで、さらに初期DNA量 $156 \sim 5 \text{amol}$ （約 $31.3 \sim 1 \text{amol/Mbp (genome)}$ ）のライブラリをシーケンシングした結果、 78amol が最も効率が良かったことから、 78amol （約 $15.6 \text{amol/Mbp (genome)}$ ）が最も解析効率が良い初期DNA量と考えられた。

【0142】

ii) 2Gbpシーケンシングデータ

2 G b p F a s t q ファイルについて、i) と同様の解析を行った。2 G b p のシーケンシングデータの解析は、コントロール群についてのみ実施した。P C Rでの初期DNA量を減らすにつれて、1 0 G b p データと同様に、グループあたりのリードペア数が増加する傾向がみられた（図4）。解析効率は、1 0 G b p データとは異なり、2 0 a m o l の初期DNA量で最大となった（図5）。

【0143】

iii) 解析効率とリードペア数の関係性

グループに含まれる平均的なリードペア数が少なければ、相補鎖間コンセンサスリード配列を作成できるグループが少なくなり、解析効率は低くなると考えられる。一方、グループあたりの平均的なリードペア数が多すぎれば、一つの相補鎖間コンセンサスリード配列作成に使用されるリード数が多くなり、逆に効率が低下するものと考えられる。1 0 G b p 及び2 G b p シーケンシングデータについての初期DNA量ごとの解析効率と推定フラグメントのグループあたり平均リードペア数を調べた。結果を表3に示す。1 0 G b p 及び2 G b p シーケンシングデータのいずれにおいても、平均リードペア数は、基本的には初期DNA量の減少に従って増加した。1 0 G b p シーケンシングデータでは、最も効率のよい7 8 a m o l における平均リードペア数は、コントロール群で2.35、ENU群で2.32であった。2 G b p シーケンシングデータでは、最も効率のよい2 0 a m o l における平均リードペア数は、2.21で、1 0 G b p の場合と近い値であった。この結果から、推定フラグメントについてのグループに含まれるリードペア数が平均約2になる辺りで、最適な解析効率を得られることが示された。

【0144】

【表3】

10 Gbp sequencing data

Library No.		1	4	5	6	7	8	9	10	11	12	13	14
PCR input DNA amount (amol)		20000	2500	1250	625	313	156	156	78	39	20	10	5
Control	Data efficiency (%)	0.40	0.89	1.42	2.32	3.96	5.50	4.01	4.95	4.17	2.57	0.99	0.56
	Ave. Number of read pairs/group	1.03	1.06	1.09	1.16	1.32	1.60	1.61	2.35	4.12	7.38	21.04	35.28
ENU	Data efficiency (%)	0.44	1.25	1.48	2.81	4.02	5.13	3.63	5.53	5.29	2.76	1.90	0.66
	Ave. Number of read pairs/group	1.03	1.08	1.09	1.20	1.32	1.52	1.41	2.32	3.15	7.58	11.84	32.72

2 Gbp sequencing data

Library No.		1	4	5	6	7	8	9	10	11	12	13	14
PCR input DNA amount (amol)		20000	2500	1250	625	313	156	156	78	39	20	10	5
Control	Data efficiency (%)	0.16	0.35	0.60	1.02	1.97	3.08	2.37	3.88	6.26	7.68	6.74	3.46
	Ave. Number of read pairs/group	1.01	1.02	1.03	1.04	1.09	1.14	1.14	1.25	1.55	2.21	4.12	8.59

【0145】

III) シーケンシングデータの重なり率に対する初期DNA量の影響

i) 1 0 G b p シーケンシングデータ

7) で1 0 G b p F a s t q ファイルについて算出した、1つの推定フラグメントについてのグループが異なるDNA断片に由来するリードペアを含む割合（重なり率）を、図6及び表4～5に示す。図6A及び表4にDMSO曝露（コントロール）群、図6B及

び表5にENU曝露(ENU)群のデータを示す。重なり率は、異なる細胞からの別々のDNA断片由来の異なるリード同士が、偶然等価の配列情報を有することで、同一断片由来のリードとして誤認識された割合を表す。本実施例では2種類のインデックスを用いたことから、異なるDNA断片が同一断片として誤認識される事象のうち、およそ半数が検出されたと推定される。したがって、得られた重なり率の約2倍の値が、異なるDNA断片の配列が偶然等価の配列情報を有する実際の割合(真の重なり率)であると推定される。10Gbpシーケンシングデータでは、PCRでの初期DNA量が減少するに従って重なり率は低下した。最も解析効率の高い初期DNA量78amol(約15.6amol/Mbp(genome))での重なり率は、コントロール群で0.59%、ENU群で0.75%であったことから、78amol(約15.6amol/Mbp(genome))付近は、解析効率が高いと同時に、真の重なり率を1%程度に抑えられる初期DNA量であると考えられた。

【0146】

【表 4】

Control

Library No.	1	4	5	6	7	8	9	10	11	12	13	14
PCR input DNA amount (amol)	20000	2500	1250	625	313	156	156	78	39	20	10	5
PCR input / Mbp (genome)	4000	500	250	125	62.6	31.3	31.3	15.6	7.8	4	2	1
Total number of groups with >2 read pairs	789990	1565934	2264112	3589346	5762666	8080584	7391538	8968099	6788263	3952934	1817281	826902
Number of groups with 1 st Index	310691	763538	1086879	1599464	2643429	3993483	4090814	4509930	3426909	1814305	892466	399114
Number of groups with 2 nd Index	263240	590018	991787	1826132	2990486	3989365	3211417	4404941	3343135	2132718	922665	426879
Number of groups with both indexes	216059	212378	185446	163750	128751	97736	89307	53228	18219	5911	2150	909
Overlap rate (%)	27.35	13.56	8.19	4.56	2.23	1.21	1.21	0.59	0.27	0.15	0.12	0.11

【 0 1 4 7 】

10

20

30

40

【表 5】

ENU

Library No.	1	4	5	6	7	8	9	10	11	12	13	14
PCR input DNA amount (amol)	20000	2500	1250	625	313	156	156	78	39	20	10	5
PCR input / Mbp (genome)	4000	500	250	125	62.6	31.3	31.3	15.6	7.8	4	2	1
Total number of groups with >2 read pairs	871423	3013638	2607096	5283784	6981568	7530681	6585067	10920911	8109174	4734379	2622344	1235504
Number of groups with 1 st Index	315282	1490191	1192767	2821092	2936126	3354008	3452811	5486021	4155822	2294235	1290880	631743
Number of groups with 2 nd Index	305534	1070116	1163017	2218867	3851879	4068711	3003870	5352910	3922071	2431960	1328239	602016
Number of groups with both indexes	250607	453331	251312	243825	193563	107962	128386	81980	31281	8184	3225	1745
Overlap rate (%)	28.76	15.04	9.64	4.61	2.77	1.43	1.95	0.75	0.39	0.17	0.12	0.14

【0148】

ii) 2 Gbp シーケンシングデータ

2 Gbp Fastq ファイルについてのコントロール群での重なり率を表6に示す。
 10 Gbp シーケンシングデータと比較して、重なり率がやや低下する傾向が認められたが、大きな変化は確認されなかった。

10

20

30

40

50

【0149】

【表6】

Control

Library No.	1	4	5	6	7	8	9	10	11	12	13	14
PCR input DNA amount (amol)	20000	2500	1250	625	313	156	156	78	39	20	10	5
PCR input / Mbp (genome)	4000	500	250	125	62.6	31.3	31.3	15.6	7.8	4	2	1
Total number of groups with >2 read pairs	71875	128267	200278	324282	588306	882830	775318	1180509	1748845	1987957	1522713	804308
Number of groups with 1 st Index	29721	60245	91129	150217	263525	417418	339891	592836	865164	996323	770114	389044
Number of groups with 2 nd Index	27270	52691	93816	159290	311138	453231	424834	579061	877875	988514	751552	414910
Number of groups with both indexes	14884	15331	15333	14775	13643	12181	10593	8612	5806	3120	1047	354
Overlap rate (%)	20.71	11.95	7.66	4.56	2.32	1.38	1.37	0.73	0.33	0.16	0.07	0.04

【0150】

IV) 変異頻度に対する初期DNA量の影響

8) で求めた、各ライブラリからの10Gbpシーケンシングデータをもとに検出した

10

20

30

40

50

変異頻度を図7～8に示す。ENU群では、いずれのライブラリにおいてもGC→AT変異の頻度が最も高かった。これは、既知のENUによる変異の変異スペクトル解析の結果と一致していた(Matsuda et al., Genes and Environment, 2015, 37:15-24)。PCRでの初期DNA量が20000～156 amol (約4000～31.3 amol/Mbp (genome))のライブラリでは、初期DNA量が低下するに従ってENU群での変異頻度が上昇する傾向があった(図7)。一方、156～5 amol (約31.3～1 amol/Mbp (genome))のライブラリでは、初期DNA量による変異頻度の上昇はみられなかった(図8)。特に重なり率が1%を下回るPCR初期DNA量が78 amol～5 amolの条件においては、重なり率の低下に伴う変異頻度の上昇が認められないことから、重なりによる変異頻度への影響が実験誤差の範囲に収まっており、これらの条件が示す付近に真の変異頻度があるものと考えられた。そこで、これら条件のうち、GC→AT変異頻度が最も高かった20 amol (約4 amol/Mbp (genome))ライブラリを基準にした各ライブラリのGC→AT変異の相対変異頻度を算出し、重なり率とともに表7に示した。S. typhimuriumのゲノム(ゲノムサイズ約4.86 Mbp)においては、およそ1250 amol (約250 amol/Mbp (genome))以下の初期DNA量、又はDNAの多様度がこれと同等以下のライブラリを用いれば、少なくとも70%以上の精度で変異を検出できることが示唆された。一方、PCRでの初期DNA量が多くなるにつれ変異頻度は低下した。表7に示した各ライブラリの重なり率から、PCRでの初期DNA量が多い条件では、重なり率が大きいため本来変異として検出されるべきものがエラーとして誤認識され、正確な変異頻度が算出されていなかったと推測された。

【0151】

【表7】

Library No.	1	4	5	6	7	8	9	10	11	12	13	14
PCR input DNA amount (amol)	20000	2500	1250	625	313	156	156	78	39	20	10	5
PCR input / Mbp (genome)	4000	500	250	125	62.6	31.3	31.3	15.6	7.8	4	2	1
GC to AT mutation frequency/ 10^6 bp	10.48	37.70	46.50	55.30	56.57	60.42	61.01	63.62	63.38	63.89	63.35	60.05
Relative mutation frequency to 20 amol	16.41	59.01	72.79	86.56	88.55	94.57	95.50	99.59	99.20	100.00	99.16	94.00
Overlap rate (%)	28.76	15.04	9.64	4.61	2.77	1.43	1.95	0.75	0.39	0.17	0.12	0.14

【0152】

V) 最適なシーケンシング条件の検討

本実施例でのシーケンシングにおける各種条件を表8に示す。表8の各列の数値は、以下の理論値、設定値、及び測定値を示す。

A. PCRでの初期DNA量 (amol)

B. サンプルDNA 1 MbpあたりのPCRでの初期DNA量 (amol/Mbp (sample DNA))

C. PCR開始時のDNA断片の総数 (個: 6.02×10^{23} 個 $\times$ A)

D. PCRのサイクル数

E. PCRにおける各断片の増幅率 (2のC乗)

F. PCR増幅後の断片の総数 (B $\times$ D)

G. Eのうち、同一DNA断片由来のPCR産物が占める割合 (D/E)

H. シーケンシングされた塩基の総数 (シーケンシングデータ量)

I. シーケンシングデータ中のリードペア数

J. 重なり率 (異なるDNA断片が同一断片として誤認識される割合) の測定値

K. 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の測定値

L. 解析効率の測定値

【0153】

【表8】

A	B	C	D	E	F	G	H	I	J	K	L
PCR input DNA amount (amol)	PCR input /Mbp (sample DNA)	Initial no. of distinct fragments	PCR cycles	Amplification rate	No. of amplified fragments	Ratio of distinct fragments	No. of base pairs sequenced (Gbp)	No. of read pairs sequenced (M reads)	Overlap rate (%)	Ave. Number of read pairs/group	Data efficiency (%)
20000	4000	1.20E+10	7	2 ⁷	1.54E+12	8.31E-11	11.12 2	44 10	27.35 20.71	1.03 1.01	0.40 0.16
2500	500	1.51E+09	10	2 ¹⁰	1.54E+12	6.64E-10	11.16 2	45 10	13.56 11.95	1.06 1.02	0.89 0.35
1250	250	7.53E+08	11	2 ¹¹	1.54E+12	1.33E-09	10.61 2	42 10	8.19 7.66	1.09 1.03	1.42 0.60
625	125	3.76E+08	12	2 ¹²	1.54E+12	2.66E-09	10.64 2	43 10	4.56 4.56	1.16 1.04	2.32 1.02
313	62.5	1.88E+08	13	2 ¹³	1.54E+12	5.31E-09	10.66 2	43 10	2.23 2.32	1.32 1.09	3.96 1.97
156	31.3	9.39E+07	14	2 ¹⁴	1.54E+12	1.06E-08	11.44 2	46 10	1.21 1.37	1.61 1.14	4.01 2.37
78	15.6	4.70E+07	15	2 ¹⁵	1.54E+12	2.13E-08	14.37 2	72 10	0.59 0.73	2.35 1.25	4.95 * 3.88
39	7.8	2.35E+07	16	2 ¹⁶	1.54E+12	4.26E-08	14.27 2	71 10	0.27 0.33	4.12 1.55	4.17 6.26
20	4	1.20E+07	17	2 ¹⁷	1.58E+12	8.31E-08	13.31 2	67 10	0.15 0.16	7.38 2.21	2.57 7.68 *
10	2	6.02E+06	18	2 ¹⁸	1.58E+12	1.66E-07	16.80 2	84 10	0.12 0.07	21.04 4.12	0.99 6.74
5	1	3.01E+06	19	2 ¹⁹	1.58E+12	3.32E-07	13.02 2	65 10	0.11 0.04	35.28 8.59	0.56 3.46

*: Most efficient condition

【0154】

i) 解析効率の最大化

表3に示したデータから、シーケンシングにおける解析効率は、解析に使用したシーケ

10

20

30

40

50

ンシングデータ量とPCRでの初期DNA量に依存していることが分かった。本実施例では、相補鎖間コンセンサスリード配列を抽出する際、1つの推定フラグメントについてのグループに含まれるリードペア数に特に制限を設けなかった。グループ内に2本の相補鎖にそれぞれ由来する2つのリードペアのセットが少なくとも1個含まれていれば、相補鎖間コンセンサスリード配列を作成した。その結果、推定フラグメントについてのグループに含まれるリードペア数が、平均約2になる辺りの初期DNA量で、最適な解析効率が得られることが分かった(表3)。

【0155】

ii) 重なり率の最少化

表4、表5及び表6に示したデータでは、重なり率(異なるDNA断片が同一断片として誤認識される割合)は、PCRの初期DNA量に依存し、解析に使用したシーケンシングデータ量には依存しなかった。重なり率は、(1)ライブラリプール中の同一DNA断片に由来するPCR産物がシーケンシングされる確率と、(2)等価な配列情報を有する異なるDNA断片由来のPCR産物がシーケンシングされる確率と、の比率によって決定される。表8の値を基にすると、PCRでの初期DNA量を減少させ、PCRのサイクル数を増やすと、全PCR産物中における同一断片由来の産物の割合が増加して、(1)の確率は増加し、逆に(2)の確率は低下し、これによって誤認識の割合が低減されるので、より正確に解析を実施できる。一方、シーケンシングされる断片は、表8のEに示すような大過剰のDNA分子を含むPCR産物からランダムにピックアップされるため、シーケンシングデータ中での(1)と(2)の比率は、データ量の大小によらずPCR産物と同等になる。これが、2Gbpと10Gbpのシーケンシングで重なり率が同等になった理由であると推測された。

【0156】

iii) 変異解析のためのシーケンシング条件

本実施例で最大の解析効率をもたらした条件(初期DNA量78amol、シーケンシングデータ量10Gbp)での変異解析の精度(表7)は、サルモネラ菌(ゲノムサイズ約4.86Mbp)の変異原性物質の曝露により生じる変異頻度の測定にとって充分高いと判断された。したがって、上記V)i)において算出した最大の解析効率をもたらすシーケンシング条件は、化学物質等によって導入されたゲノム変異の全体像を捉える上で、十分な条件であると考えられた。

【0157】

本実施例での解析効率とPCR初期DNA量、シーケンシングデータ量の関係を表9に示す。約2Gbp(10Mリードペア)シーケンシングの解析において、解析効率の高かった(>4%)39、20、及び10amolのライブラリにおいては、PCR初期DNA量1amolあたりのシーケンシングデータ量は、それぞれ0.26、0.50、及び1.00Mリードペア(リードペアの塩基対量で51、100、及び200Mbp)/amolであった。一方、約10Gbpのシーケンシングでの解析において効率の高かった(>4%)、156(1回目)、156(2回目)、78、及び30amolにおいては、PCR初期DNA量1amolあたりのシーケンシングデータ量は、それぞれ0.29、0.41、0.92、及び1.83Mリードペア(リードペアの塩基対量で73、82、184、及び366Mbp)/amolであった。従って、比較的高い解析効率(>4%)を得るためのシーケンシングデータ量は、PCR初期DNA量1amolあたり0.2~2M個(40~400Mbp)リードペアと算出された。また、解析効率が最大となった2Gbpシーケンシングでの20amolライブラリと、10Gbpシーケンシングでの78amolライブラリでのシーケンシングデータ量がそれぞれ0.50M及び0.92Mリードペア/molであったことから、最大の解析効率を得るためのシーケンシングデータ量は、PCR初期DNA量1amolあたり0.5~1M個(100~200Mbp)リードペアと算出された。

【0158】

【表 9】

シーケンシングデータ量とPCR初期DNA量の関係
2Gbpデータの解析

	PCR input DNA amount (amol)														
PCR input (amol)	20000	2500	1250	625	313	156	156	78	39	20	10	5			
PCR input / Mbp (sample DNA)	4000	500	250	125	62.5	31.3	31.3	15.6	7.8	4	2	1			
Data efficiency (%)	0.16	0.35	0.60	1.02	1.97	3.08	2.37	3.88	6.26	7.68	6.74	3.46			
Ave. Number of read pair / group	1.01	1.02	1.03	1.04	1.09	1.14	1.14	1.25	1.55	2.21	4.12	8.59			
reads pair sequenced (M read pair)	10	10	10	10	10	10	10	10	10	10	10	10			
base pair sequenced (Gbp)*	2	2	2	2	2	2	2	2	2	2	2	2			
read pair (M reads) / PCR input (amol)	0.001	0.004	0.01	0.02	0.03	0.06	0.06	0.13	0.26	0.50	1.00	2.00			
base pair (Mbp) / PCR input (amol)	0.10	0.80	1.60	3.20	6.39	12.82	12.82	25.64	51.28	100.00	200.00	400.00			

*: 2 × 100bpと仮定した場合

10Gbpデータの解析

	PCR input DNA amount (amol)														
PCR input (amol)	20000	2500	1250	625	313	156	156	78	39	20	10	5			
PCR input / Mbp (sample DNA)	4000	500	250	125	62.5	31.3	31.3	15.6	7.8	4	2	1			
Data efficiency (%)	0.40	0.89	1.42	2.32	3.96	5.50	4.01	4.95	4.17	2.57	0.99	0.56			
Ave. Number of read pair / group	1.03	1.06	1.09	1.16	1.32	1.60	1.61	2.35	4.12	7.38	21.04	35.28			
reads pair sequenced (M read pair)	44	45	42	43	43	46	64	72	71	67	84	65			
base pair sequenced (Gbp)	11.12	11.15	10.61	10.64	10.66	11.44	12.87	14.37	14.27	13.31	16.80	13.02			
read pair (M reads) / PCR input (amol)	0.002	0.02	0.03	0.07	0.14	0.29	0.41	0.92	1.83	3.33	8.40	13.02			
base pair (Mbp) / PCR input (amol)	0.56	4.46	8.49	17.03	34.05	73.30	82.49	184.24	366.01	665.44	1679.67	2603.45			

【0159】

上記IV)の検討では、1250amol(約250amol/Mbp(genome))以下のPCR初期DNA量、又はDNAの多様度がこれと同等以下のライブラリを用いれば、誤認識率(真の重なり率)を約20%以下まで低くした状態でサルモネラ菌の変異解析が可能であることが示された(表7)。そこで、PCR初期DNA量に対して最大効率でシーケンシングした場合に、サルモネラ菌のゲノム情報の取得できる配列データ量を決定した。初期DNA量1250amol(約250amol/Mbp(genome))

10

20

30

40

50

）において最適なシーケンシングデータ量は約 625～1250M 個（125～250 Gbp）リードペアであり、比較的効率の高い（＞4％）シーケンシングデータ量は約 250～2500M 個（50～500 Gbp）リードペアと算出された。このシーケンシングデータ量から実施例 1 のアルゴリズムで得られる配列データ（相補鎖間コンセンサスリード配列）量は、本実施例と同様に解析効率約 5％（表 3）、誤認識率（真の重なり率）約 20％以下と仮定すれば、最大効率の条件の場合、6.25 Gbp～12.5 Gbp と算出される。また、多少の効率の低下を許容し、例えば 500 Gbp でシーケンシングした場合、上述のように効率が＞4％であると仮定すれば、約 20 Gbp の配列データを取得することができる。これが、本発明に基づくサルモネラ菌のゲノム変異解析に利用できる最大データ量である。

10

【0160】

iv) 変異解析可能なゲノムサイズの推計

上記 IV) のとおり、サルモネラ菌のゲノム解析に利用可能な初期 DNA 量は最大 1250 amol（約 250 amol/Mbp（genome））であったことから（表 7）、高い解析効率、及び低い誤認識率（真の重なり率）を達成するための初期 DNA 量の上限は、ゲノム 1 Mbp あたり 250 amol と計算された。一方で、本実施例で検出した ENU 群での変異頻度が $1/10^6 \sim 1/10^5$ bp のオーダーであったことから、変異解析には少なくとも 10^6 bp のデータ量が必要であると仮定された。これらの値をもとに、本発明の方法を用いた変異解析を適用可能な最も小さいゲノムサイズを推計した。表 10 に、異なるサイズのゲノムについて、適用可能な PCR 初期 DNA 量の最大値、最適シーケンシングデータ量、及び取得可能な配列データ量の最大値（解析効率 5％と仮定）を示す。表 10 より、約 10 kbp 以上のゲノムから 10^6 bp 以上のデータ量が得られることから、ゲノムサイズ約 10 kbp 以上のサンプルであれば、本発明に基づく変異解析法を適用可能であると考えられる。

20

【0161】

【表 10】

解析対象 ゲノムサイズ	適用可能な PCR 初期 DNA 量の最大値	最適シーケンシング データ量	取得可能な配列データ 量の最大値
5 Mbp	1250 amol	125 - 250 Gbp	6.3 - 12.5 Gbp
1 Mbp	250 amol	25 - 50 Gbp	1250 - 2500 Mbp
100 kbp	25 amol	2.5 - 5 Gbp	125 - 250 Mbp
10 kbp	2.5 amol	250 - 500 Mbp	12.5 - 25 Mbp
1 kbp	0.25 amol	25 - 50 Mbp	1.25 - 2.5 Mbp

30

【0162】

実施例 3 合成 DNA 配列を用いた変異解析

本実施例では、塩基対置換型変異を含む 1000 bp の合成 DNA 配列を解析対象とした変異解析により、実施例 1 に記載したシーケンシングアルゴリズムの有用性と、異なる細胞由来のフラグメントが偶然重なる事象が変異解析に及ぼす影響を検討した。

【0163】

40

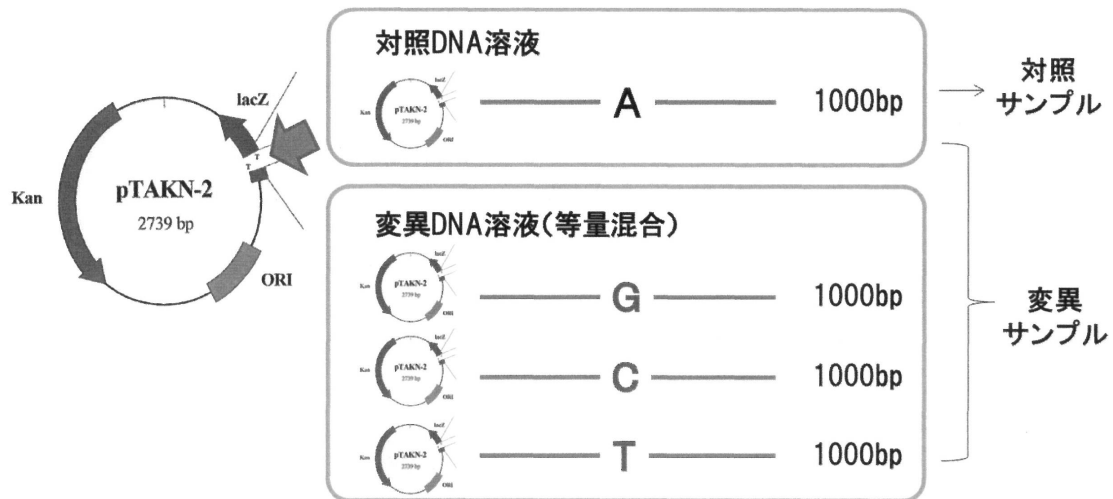
1) サンプル DNA の調製

1000 bp のランダム配列 DNA（以下、対照 DNA とする）と、該対照 DNA に塩基対置換型変異（A：T 塩基対）を導入した DNA（以下、変異 DNA とする）を作製した。これらを混合して、A：T 塩基対における 3 種類の塩基対置換パターン（A：T＞G：C、A：T＞C：G、及び A：T＞T：A）を既知の頻度で等しく含むサンプル DNA を調製した。模式図 6 に、サンプル DNA 調製手順の概念図を示す。

【0164】

【化 7】

模式図6: サンプルDNA調製手順



10

変異

20

Transversion (プリン⇄ピリミジン)	A:T > C:G 又は A:T > T:A
Transition (プリン⇄プリン 又はピリミジン⇄ピリミジン)	A:T > G:C

【0165】

以下に手順の詳細を説明する。G : C、A : T塩基対をそれぞれ約50%含むランダムな配列を有する1000bpのDNA（対照DNA、配列番号1）を合成し、pTAKN-2ベクターに組み込んで増幅させた。得られたベクターのTEバッファー（pH8.0、和光純薬工業社製）溶液（10ng/μL）を調製した（対照DNA溶液）。対照DNAをベースに、変異DNAを作製した。変異DNAとしては、対照DNAの配列の中心（502番目）に位置するA : T塩基対を他の塩基対（A : T > G : C、C : G又はT : A）で置換した3種類の配列を作製した。各変異DNAをpTAKN-2ベクターに組み込んで増幅させ、得られたベクターのTEバッファー溶液（10ng/μL）を調製した。各溶液を等量混合し、変異DNA溶液を調製した。

30

【0166】

対照DNA溶液に、変異DNA溶液を表11に示した比率で混合し、各塩基対置換が均等の頻度で認められ、かつ総変異頻度が1/10⁴bpのDNAサンプル（変異サンプル）を調製した。対照サンプルとして、変異DNA溶液を含まない（対照DNA溶液のみの）DNAサンプルを調製した。得られた変異サンプル及び対照サンプルを、ライブラリ調製及びシーケンシングに供した。

40

【0167】

【表 1 1】

DNAサンプルにおける変異DNA溶液比と総変異頻度

	変異サンプル	対照サンプル
総変異頻度 (/bp)	1/10 ⁴	0
対照DNA溶液	81 μ L	90 μ L
変異DNA溶液 (塩基対置換型変異A→G, C, T)	9 μ L	-
変異DNA溶液比	10 %	0 %

10

【0 1 6 8】

2) ライブラリ調製及びシーケンシング

1) で調製した変異サンプル及び対照サンプルを用いて実施例 2 と同様の方法でライブラリを調製し、得られたライブラリを用いてシーケンシングを行った。ライブラリ調製の際の P C Rでの初期 D N A 量と、アダプターのインデックス配列、及び P C R サイクル数を表 1 2 に示す。各 P C R 初期 D N A 量のサンプルあたり、約 1 0 G b p のシーケンシングデータを得た。

【0 1 6 9】

【表 1 2】

対照サンプル

ライブラリ No.	1	2	3
DNA濃度 (pmol/L)	3.13	0.78	0.20
PCR input (amol)	78	20	5
1 st Index*	4	6	12
2 nd Index*	14	16	19
PCR サイクル数	15	17	19

20

30

変異サンプル

ライブラリ No.	4	5	6
DNA濃度 (pmol/L)	3.13	0.78	0.20
PCR input (amol)	78	20	5
1 st Index*	3	9	11
2 nd Index*	21	23	27
PCR サイクル数	15	17	19

40

*: Indexの識別番号

【0 1 7 0】

3) 変異解析用配列データの作成、及び変異解析

シーケンシングデータからの変異解析は、実施例 2 と同様に実施した。すなわち、実施例 1 に記載のアルゴリズムを用いて相補鎖間コンセンサスリード配列を含む配列データを作成し、これを用いて、P C T / J P 2 0 1 7 / 0 0 5 7 0 0 に記載された手順に従って塩基置換を検出した。

【0 1 7 1】

50

変異解析用配列データの編集は、上述の模式図4に示す解析フローに従って実施した。まず、Cutadaptソフトウェアを用いて、生リード配列から、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いリード配列のペアを作成した。得られたトリミング後のリードペアのFastqファイルについて、アダプター配列のインデックス情報を元に、同一ライブラリに由来するFastqファイルを1つのファイルにまとめた。得られたFastqファイルの先頭と末尾からそれぞれ 5×10^4 個（合計 10×10^4 個）のリードペアを抽出して1つのファイルにまとめ、約20Mbpのシーケンシングデータ解析用のFastqファイルを作成した。

【0172】

実施例1記載のアルゴリズムによる相補鎖情報を用いた解析では、得られた約20Mbpデータ解析用Fastqファイルについて、Bowtie2ソフトウェアを用いてリードペアの参照配列へのマッピングを行い、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実施した後、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て、実施例1に示したアルゴリズムに従って、参照配列（対照DNAを挿入したpTAKN-2ベクターの配列）にマッピングされたリードペアの情報から、相補鎖間コンセンサスリード配列の抽出を行った。相補鎖間コンセンサスリード配列は、新たなFastqファイルとして出力した。

【0173】

変異解析では、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を用いて作成したプログラムを用いて、PCT/JP2017/005700に記載された手順に従って、変異した塩基を検出した。すなわち、Pythonを用いたプログラムを用いて、マッピングされた全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。次いで、各群の塩基の総数、及び参照配列に対して変異した塩基を検出した。これらの解析は、参照配列中の1000bp対照DNAの領域にマッピングされたリード配列の塩基に対してのみ実施した。

【0174】

4) 変異頻度、及び変異頻度増加量の算出

次いで、変異サンプル及び対照サンプルのそれぞれについて、対応する参照配列の塩基に基づいて、検出した変異した塩基の変異パターン（AT→TA、AT→CG、AT→GC、及びGC→TA、GC→CG、GC→AT）を決定し、さらに各変異パターンの解析対象塩基中における 10^6 bpあたりの変異頻度を算出した。本変異解析では、解析対象の1000bp領域のうち、変異導入していない266番目のA:T塩基対において塩基対置換（A:T>C:G変異）が多く検出された。これは、対照DNA中の自然発生的な変異によるものと推測された（データは示さず）。従って、本変異解析においては、266番塩基上にマッピングされたリード塩基を除いて、変異頻度を算出した。

【0175】

本発明に基づく実施例1のアルゴリズムによる相補鎖情報を用いた解析で得られた、各サンプルの変異頻度を図9に示す。PCR初期DNA量が78amol及び20amolの条件については、いずれの変異パターンについても対照サンプル及び変異サンプルともに変異頻度は0であった。一方、PCR初期DNA量が5amolの条件では、変異サンプルにのみ、約 $1/10^5$ bp（A:T>T:A）～ 5×10^5 bp（A:T>G:C）の頻度で塩基対置換が検出された。これは、78及び20amolの条件においては、異なるDNA断片の偶然の重なる割合が高く、変異がエラーと誤認識されてしまっている一方で、PCR初期DNA量を低減することによって、変異が検出できるようになるということを示していると考えられた。実際に、PCR初期DNA量を78～5amolへと減らしていくことによって、対照サンプル及び変異サンプルのいずれにおいても、異なるインデックスの重なる割合が低下していく様子が観察された（図10）。本実施例では

各塩基対置換変異パターンが、 $1/500$ 〔A T塩基対の変異率〕 $\times 1/3$ 〔変異後の塩基タイプ数〕 $\times 1/10$ 〔変異DNA溶液希釈率〕 $= 6.7 \times 10^{-5}$ bpの頻度で認められる設計である。変異サンプルの5 amolの条件においては、異なるインデックスの重なり率が約30%であることから、A:T>T:Aの変異頻度が設計よりも約1/7程度に低くなっているのは、異なる断片の重なりが影響している可能性が考えられる。

【0176】

以上のことから、本実施例のように解析対象のゲノムサイズが約数kbpである小さな領域の場合には、より大きなゲノムサイズの領域を対象とする場合に比べ、異なる断片の偶然の重なりの影響が大きくなる可能性があることが分かった。しかし、PCR初期DNA量を低減させていくことによって、ゲノムサイズが小さい場合にも本発明による相補鎖情報を用いた解析により、エラー頻度を飛躍的に低減した解析が実施できることが分かった。

【0177】

実施例4 相補鎖間コンセンサスリード配列の作成効率：Duplex Sequencingとの比較

本実施例では、実施例2のコントロールサンプルの10Gbpシーケンスデータにおける、相補鎖間コンセンサスリード配列の作成効率を算出した。実施例2の表3に示したコントロールサンプルの10Gbpシーケンスデータの各ライブラリについて、模式図3に示したフローに従って相補鎖間コンセンサスリード配列を取得した。全リードペア数に対するコンセンサスリードペアの数の割合を相補鎖間コンセンサスリード配列の作成効率として算出し、推定フラグメントのグループあたりの平均リードペア数に対してプロットした(図11)。図11では、表3に示したコントロールサンプルのライブラリNo. 1、4-8をExp. 1、ライブラリNo. 9-14をExp. 2として示した。結果、相補鎖間コンセンサスリード配列の作成効率は、最大で7.5% (Exp. 1) 及び5.9% (Exp. 2) であった。

【0178】

一方、Duplex Sequencingにおいては、最大で約1.4%程度の効率が得られると報告されている(Nature protocol, 2014 9(11):2586-2606)。本方法ではその4~5倍の効率が得られた。これは、Duplex Sequencingのコンセンサス配列の作成においては、サンプルDNA断片の片方の鎖に由来するリード配列が3つ以上含まれるグループのリード配列を用いているのに対し、本方法においては、2本鎖双方に各々由来する2つのリード配列を最低限含む推定フラグメントのグループを全て相補鎖間コンセンサスリードの作成に利用したためと考えられる。Duplex Sequencingでの相補鎖間コンセンサス配列の作成条件は、本方法に比べて1つのコンセンサスリードの作成により多くのリードを用いるため、より正確な相補鎖間コンセンサスリード配列を取得するのに有効であるが、比較的短い領域の解析に適しており、ゲノムの広範囲の解析には効率が悪い。一方、本方法では、相対的に少数のリードを用いて1つコンセンサスリードを作成するため、特定のゲノム領域についてのコンセンサスリードが得られる確率は低下するが、上述したとおり、ゲノムの広範囲の解析の効率は向上する。したがって、本方法は、ゲノム全体の変異解析を実施する場合により効率的な方法であると

【0179】

実施例5 マウスDNAサンプルを用いた変異解析への応用

本実施例では、実施例2で探索した最適条件による相補鎖情報を用いたシーケンシングをマウスのゲノムDNAにおける変異解析に適用した。サンプルDNAには、変異原としてENU(Ethyl nitrosourea, CASRN. 759-73-9)を暴露したマウス(C57BL/6J JmsSlc-Tg(gpt delta)、以下TGマウスとも表す)から調製したゲノムDNAを用いた。現在、動物を用いるin vivo遺伝子突然変異試験には、指標遺伝子を導入した遺伝子改変マウスやラットが使用されている。本発明を用いた変異解析法がマウス等に適用可能であれば、遺伝子改変動物を用いる必要

性がないため、一般毒性試験等への変異原性評価の組み込みが可能となり、実験動物数の削減につながると期待される。また得られる詳細な変異情報は、発がん性予測に有用と考えられ、動物を用いない発がん性予測法の開発に有用と考えられる。なお、本実施例における動物実験は、花王株式会社研究開発部門動物実験委員会による審議を受け、「花王株式会社研究開発部門動物実験に関する規定」に従って実施された。

【0180】

1) マウスからのゲノムDNAサンプルの調製

投与開始時7～9週齢の雄のTGマウスを用いた。ENU (Toronto Research Chemicals社製) は15.0 mg/mLの濃度で生理食塩水 (Saline) に溶解した。ENU溶液を、プラスチック製注射筒および25 G注射針を用いて、1日1回、10 mL (ENU 150 mg/kg/day) を5日間、マウスに腹腔内投与した (ENU群)。コントロール群には生理食塩水 (Saline) を投与した。各群のENUの用量と個体数を表13に示した。最終投与後7日目に動物を安楽死させて大腿骨を摘出し、RecoverEase DNA Isolation Kit (アジレント・テクノロジーズ社製) を用いて、推奨プロトコールに従って、骨髄からゲノムDNAを抽出した。得られたDNA中の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。

【0181】

【表13】

群	動物数	性	ENU 用量 (mg/kg/day)	溶液中 ENU 濃度 (mg/mL)	投与容量 (mL/kg)
Control	5	雄	0 (Saline)	0	10
ENU	5	雄	150	15.0	10

【0182】

2) gpt 遺伝子を用いた遺伝子突然変異頻度の算出

1) で得たマウスの骨髄からフェノール/クロロホルムを用いてDNAを抽出した。抽出したDNAを用いて、OECD Guidelines for the Testing of Chemicals 488 (OECD TG488) に従って、TGマウスに導入されたgpt 遺伝子を用いた遺伝子突然変異頻度の算出を行った (gpt-assay)。より詳細には、得られたゲノムDNAを、Transpack (Agilent Technologies) 製品添付のInstruction Manualに従って入ファージへパッケージングした。得られたパッケージング溶液を、予め培養しておいた大腸菌懸濁液 (YG6020株) と混合し、培養した後、トッパアガーと混合してクロラムフェニコール入りの寒天培地 (タイター用培地) および、クロラムフェニコールと6-チオグアニン入りの寒天培地 (選択培地) 上に重層して、形成されるコロニー数を計数した。選択培地上に形成されたコロニー (変異体コロニー) 数と、タイター用培地上に形成されたコロニー (総コロニー) 数を基に、突然変異頻度 ($\times 10^{-6}$) を算出した。各DNAサンプルの変異体コロニーについては、一部のコロニーをピックアップし、タイター用培地および選択培地を用いてConfirmationを実施した。突然変異頻度は得られたConfirmationの結果に基づき補正した。

【0183】

3) シーケンシング用ライブラリの調製

コントロール群及びENU投与群で、各2匹の動物からのDNAサンプルを実験に供した。各DNAサンプルをDNA Shearingシステム ME220 (コバリス社製) で推奨プロトコールに従って平均約350 bpの長さに断片化した。得られた断片化DNAをTruSeq Nano DNA Library Prep Kit (イルミナ社製、以下TruSeqと略記する) を用いて、推奨プロトコールに従ってライブラリ調製に供した。ライブラリ調製は実施例2と同様の方法で行った。ライブラリ調製の際の溶

液のDNA濃度、PCRでの初期DNA量と、アダプターのインデックス配列、及びPCRサイクル数を表14に示す。

【0184】

【表14】

群	Control		ENU	
ライブラリ No.	1	2	3	4
DNA 濃度 (pmol/L)	3.13	3.13	3.13	3.13
PCR input (amol)	78	78	78	78
1 st Index*	2	4	10	11
2 nd Index*	13	14	25	27
PCR サイクル数	15	15	15	15

*: Index の識別番号

【0185】

4) シーケンシング

3) で調製したライブラリを、HiSeq2500 (イルミナ社製) を用いて、2×100bpのリード長でシーケンシングした。ライブラリあたり、約20Gbpシーケンシングデータを得た。

【0186】

5) リード配列の編集、相補鎖情報の抽出、及び変異解析

シーケンシングによって得られた生リード配列の編集、相補鎖情報の抽出、及び変異解析は、上述の模式図3及び4に示す解析フローに従って実施例2の5)と同様の手順で実施した。なお、Bowtie2ソフトウェアでマッピングする参照配列には、C57BL/6Jのゲノム配列(GCA_000001635.6)を用いた。

【0187】

6) リードペア数分布の算出

5) で作成した各ライブラリについてのマッピング結果を基に、実施例2の6)と同様の手順で推定フラグメントについてのグループあたりリードペア数の分布を求めた。マウス各染色体についてのグループあたりリードペア数の分布、及び平均リードペア数/グループを算出した。

【0188】

7) 重なり率の算出

実施例2の7)と同様の手順で、マウスの各染色体に対応する推定フラグメントについてのグループのうち、2つ以上のリードペアが含まれるグループを抽出し、抽出したグループを対象として染色体ごとの重なり率(Overlap rate)を算出した。

【0189】

8) 変異頻度の算出

5) で相補鎖間コンセンサスリード配列を参照配列にマッピングした後、実施例2の8)と同様の手順で変異した塩基を検出した。すなわち、PCT/JP2017/005700に記載された手順に従い、Pythonを用いたプログラムを用いて、各ライブラリについて、参照配列に対して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によって4群に分けた。次いで各群の塩基の総数、及び参照配列に対して変異した塩基を検出した。検出した変異した塩基を参照配列の塩基と比較することで、コントロール群及びENU群それぞれについて、解析対象塩基におけるAT塩基対、GC塩基対の各10⁶bpにおける各変異パターン(AT→TA、AT→CG、AT→GC、及びGC→TA、GC→CG、GC→AT)、及び各変異パターンの変異頻度を算出した。なお各マウス個体が有するゲノム配列の多型(SNP等)が変異解析結果に影響すると予想されたことから、本解析では多型を有している可能性が高いと考えられるゲノム位置にマッピングされたリード配列の塩基は変異解析の対象から外

10

20

30

40

50

した。具体的には、参照配列の同一の位置（塩基）にマップされたリード塩基のなかに同一の塩基置換が2つ以上認められた場合、その参照配列の位置にマップされたリード塩基は解析対象外とした。また、データベース上に登録されるマウスの変異情報に基づいて、多型が報告されている参照配列上の塩基は解析対象から外した。すなわち、当該参照配列上の多型が報告されている塩基に対してマップされたリード塩基は解析対象外とした。マウスの多型のデータは `ensembl[asia.ensembl.org/Mus_musculus/Info/Index]` から取得した。

【0190】

9) 結果

I) リードペア数分布の算出

ライブラリ No. 1 から得られたデータに基づくマウス各染色体での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の分布を図12に示す。また、各染色体についてのグループあたりの平均リードペア数を表15に示す。各サンプルの平均リードペア数／グループの値は、いずれも2～4の範囲内であった。したがって、マウスゲノムDNAを用いた場合でも、実施例2で算出された本方法の最適条件のPCR初期DNA量及びシーケンシングデータ量でのシーケンシングで高い解析効率が得られると考えられた。

【0191】

【表15】

Average number of read pairs/group

Library No.	Chromosome										
	1	2	3	4	5	6	7	8	9	10	11
1	3.41	3.37	3.43	3.36	3.34	3.39	3.32	3.36	3.35	3.39	3.30
2	3.05	3.02	3.07	3.00	2.99	3.03	2.97	3.01	3.00	3.03	2.96
3	2.85	2.82	2.87	2.81	2.80	2.84	2.78	2.81	2.80	2.84	2.77
4	2.52	2.50	2.54	2.49	2.48	2.51	2.47	2.49	2.48	2.51	2.45
Library No.	Chromosome										
	12	13	14	15	16	17	18	19	X	Y	
1	3.39	3.39	3.41	3.38	3.42	3.33	3.39	3.34	3.47	2.95	
2	3.03	3.03	3.05	3.02	3.05	2.99	3.04	2.99	3.10	2.72	
3	2.84	2.84	2.85	2.83	2.85	2.79	2.84	2.80	2.90	2.44	
4	2.51	2.51	2.52	2.50	2.53	2.47	2.52	2.48	2.56	2.25	

【0192】

II) 重なり率の算出

7) で調べた染色体ごとの重なり率を表16に示す。本解析においては、Y染色体はマッピングされたリード数が少なかったため、対象から除外した。いずれの染色体においても重なり率は同様の値であり、染色体間の平均は、コントロール群（ライブラリ No. 1、No. 2）で0.025%及び0.025%、ENU群（ライブラリ No. 3、No. 4）で0.023%及び0.020%であった。実施例2で算出した *S. typhimurium* のゲノム配列を用いた検討において、PCR初期DNA量が78amolの条件における重なり率は、コントロール群で0.59%、ENU群で0.75%、平均0.67%であった。したがって、同一のPCR初期DNA量を用いた場合、サンプルDNAのサイズがより大きいほど重なり率が低下し、より高精度な解析が実施できると考えられた。一方、マウスのゲノムサイズが、約3Gbp、*S. typhimurium* のゲノムサイズが約5Mbpとすると、実施例2の結果から理論上求められるマウスのゲノム解析における重なり率は、 $[0.67\% \times 5\text{Mbp} / 3000\text{Mbp} = 0.0011\%]$ と推定された。すなわち、実際に得られた重なり率は理論値より20倍程度高かった。マウス等の哺乳類のゲノムのシーケンシングにおいては、ヘテロクロマチンの存在やPCRにおけるバイアスによる次世代シーケンサーでのライブラリ調製への影響のため、実際にシーケンシングされる領域のサイズはゲノムサイズより小さくなる。これが実測値と理論値のずれ

の原因であると考えられた。

【0193】

【表16】

Overlap rate (%)

Library No.	Chromosome										
	1	2	3	4	5	6	7	8	9	10	11
1	0.023	0.028	0.025	0.026	0.022	0.023	0.023	0.023	0.026	0.022	0.023
2	0.027	0.028	0.023	0.023	0.026	0.028	0.019	0.025	0.023	0.024	0.026
3	0.022	0.025	0.025	0.022	0.024	0.022	0.020	0.022	0.021	0.017	0.026
4	0.018	0.022	0.017	0.016	0.019	0.019	0.017	0.020	0.019	0.020	0.022
Library No.	Chromosome										
	12	13	14	15	16	17	18	19	X	Ave.	
1	0.024	0.028	0.022	0.025	0.021	0.030	0.024	0.026	0.026	0.025	
2	0.024	0.024	0.028	0.026	0.026	0.027	0.028	0.026	0.021	0.025	
3	0.024	0.024	0.027	0.020	0.027	0.024	0.020	0.022	0.020	0.023	
4	0.018	0.023	0.021	0.018	0.020	0.020	0.021	0.020	0.019	0.020	

10

【0194】

III) 変異解析結果

コントロール群及びENU群の各塩基対置換の変異頻度の解析結果を図13に示す($n = 2$)。ENU群では、コントロール群に比べて、 $A : T > T : A$ 、 $A : T > G : C$ 、および $G : C > A : T$ の頻度が顕著に高く、これは既存のENUの変異スペクトルの知見 (Proc Natl Acad Sci USA, 1994, 91(14):6564-6568) と一致していた。本方法により、マウスのゲノムにおいても高感度に変異解析ができることが示された。

20

【0195】

IV) 突然変異頻度の算出 (gpt-assay)

2) で算出したコントロール群及びENU群における $g p t - a s s a y$ の結果を表17に示す(コントロール群: $n = 5$ 、ENU投与群: $n = 4$)。ENU群では、コントロール群に比べて、遺伝子突然変異頻度の増加が顕著に認められた。ENUの曝露によってTGマウスのゲノムに突然変異が導入されたことが確認された。

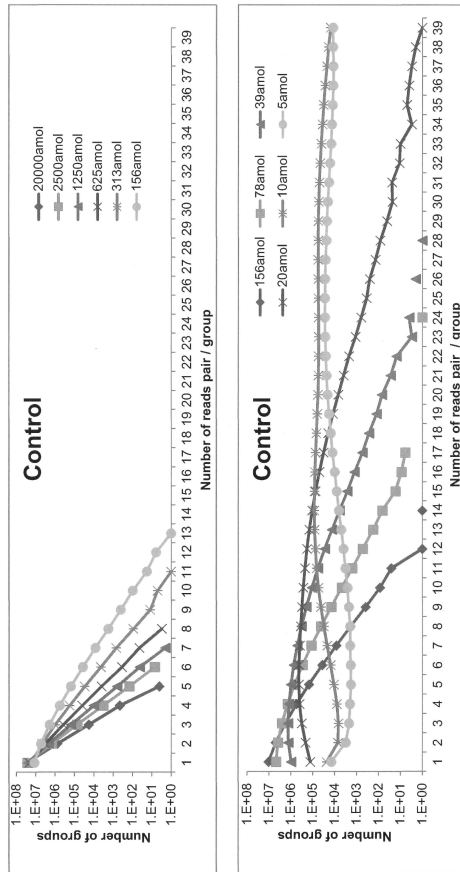
【0196】

【表17】

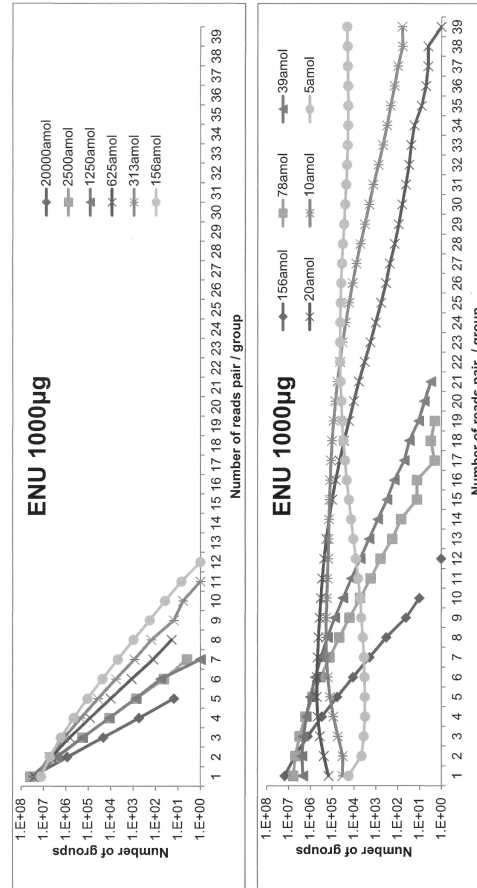
群	ENU 投与量 (mg/kg/day)	Mean Mutant Frequency	S.D.	N
Control	0 (Saline)	2.36	1.26	5
ENU	150	126.33	38.64	4

30

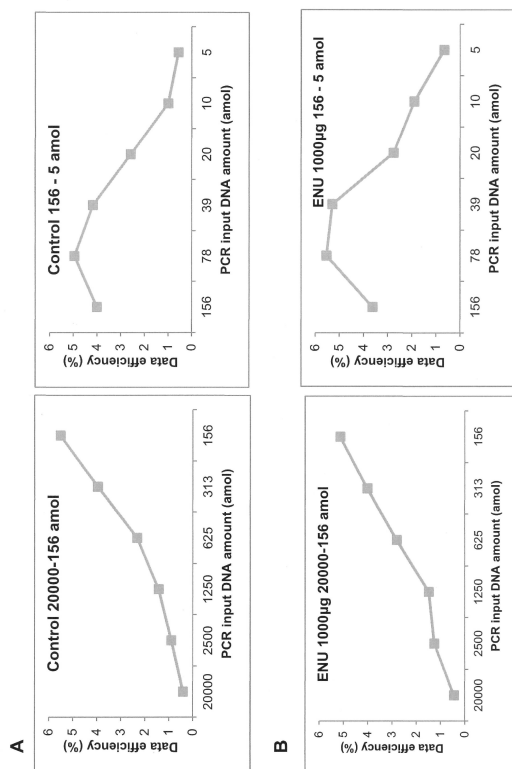
【図 1】



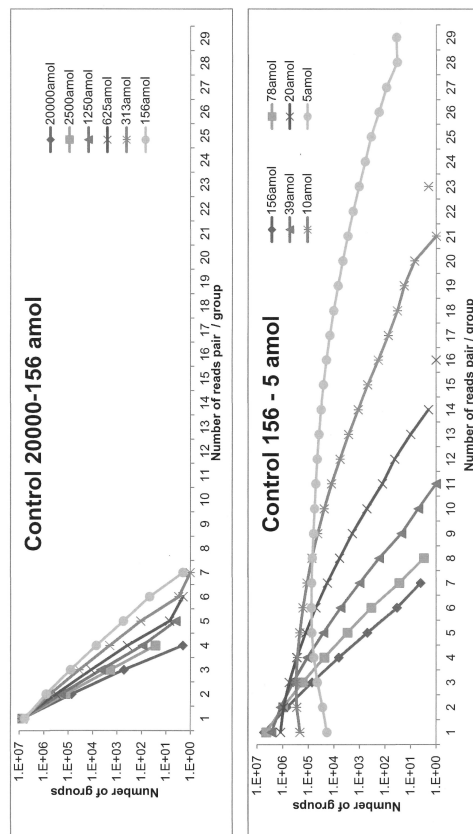
【図 2】



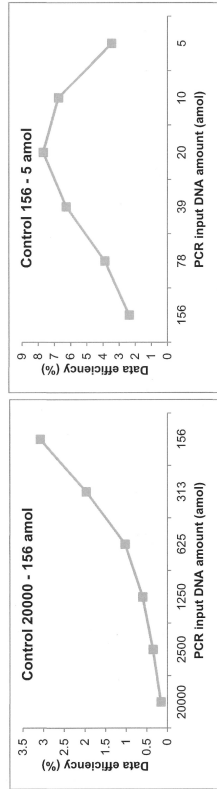
【図 3】



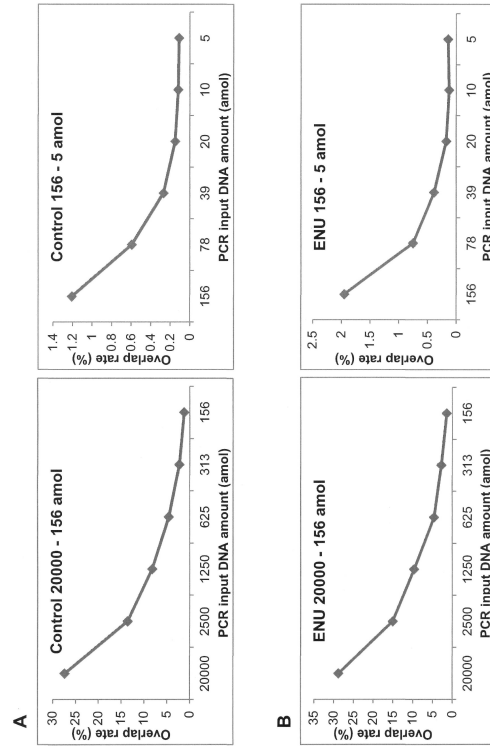
【図 4】



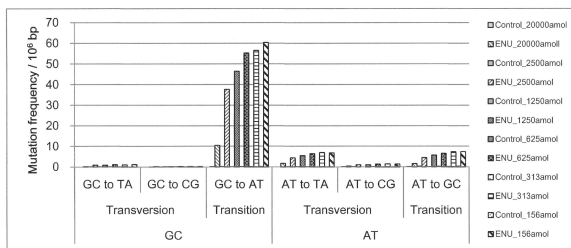
【図 5】



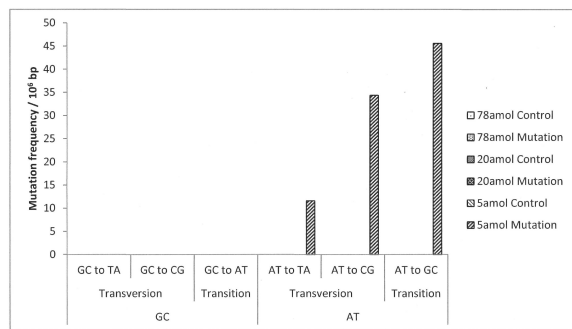
【図 6】



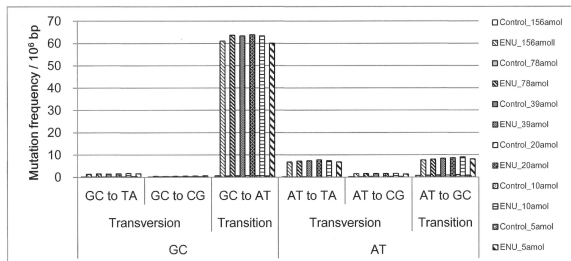
【図 7】



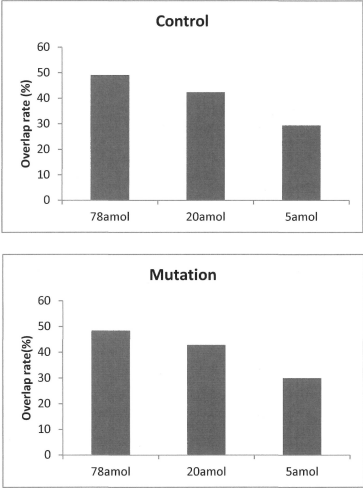
【図 9】



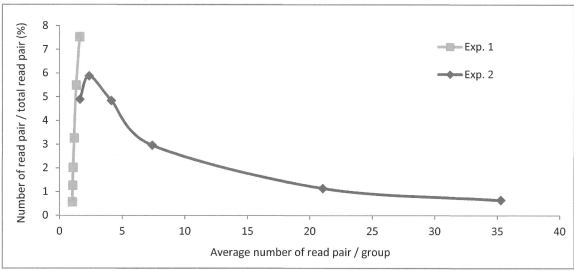
【図 8】



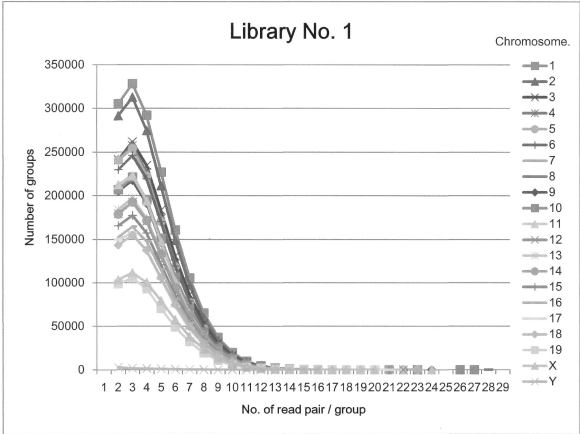
【図 1 0】



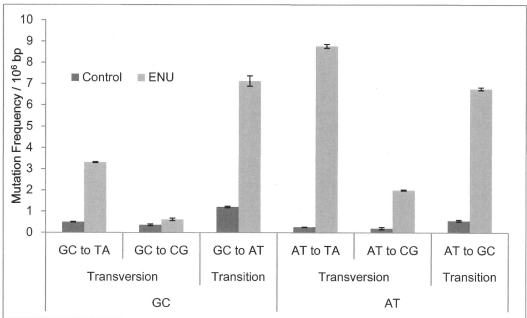
【図 1 1】



【図 1 2】



【図 1 3】



フロントページの続き

審査官 松浦 安紀子

(56)参考文献 PNAS, 2012, vol. 109, no.36, pp.14508-14513, & Supporting Information
J. Appl. Toxicol. 2017, vol. 37, pp.1125-1128

(58)調査した分野(Int.Cl., DB名)

C12Q 1/6869

C12Q 1/6809

JSTPlus/JMEDPlus/JST7580 (JDreamIII)

CAPLUS/MEDLINE/EMBASE/BIOSIS (STN)