Vol.2013-BIO-33 No.1
2013/3/21

真正細菌を対象とした SNPとINDEL同時検出ツールの開発

芦田 和重[1,a)]　佐藤 哲大[1]　中村 建介[2]　湊 小太郎[1]

概要：DNA の塩基配列中の変異が原因とされるガンなどの疾病について，変異箇所を特定することで診断や治療が行える可能性が期待されている．解析対象の塩基配列中のどの位置にどのような変異が発生しているかを特定する技術は変異コールと呼ばれており，ゲノム解析の中心となる技術の一つである．
正確な変異箇所の特定のため高精度の変異コールが必要とされているが，現在の変異コールは精度が低く，アライメントの手法次第では INDEL が検出できない場合や，INDEL 検出の精度を上げることにより SNP 検出の精度が下がるなどの問題がある．そのため，解析の手法を変えた複数回の解析が推奨されており，解析に必要な時間やコストが増大する原因となっている．この問題を解決するため，より精度の高い変異コールツールを開発する必要がある．
本研究では，真正細菌の塩基配列中から SNP と INDEL を高精度に同時検出できる変異コールの実現を目的とし，既存のツールによる変異コールの問題点の指摘と，独自のアルゴリズムによる変異コールツールの作成を行った．また，複数の真正細菌を対象とし，作成したツールの精度検証を行った．その結果，SNP 検出の精度を下げない INDEL 検出の実現と，既存のツールによる変異コールでは検出できなかった INDEL の検出に成功した．

キーワード：変異コール，塩基配列解析，真正細菌，INDEL 検出

Development of mutation calling tool which identifies SNP and INDEL targeting eubacteria

Ashida Kazushige[1,a)]　Sato Tetsuo[1]　Nakamura Kensuke[2]　Minato Kotaro[1]

Abstract: Identifying mutations in genome DNA sequences is one of most fundamental methods to diagnose or predict hereditary disease or cancer. The technology of specifying mutation is called ”mutation calling”, and is one of the most important technologies in genomics.
Although the highly precise mutation calling is needed for pinpointing mutation, the present mutation calling process has low accuracy and there are some problems. In a certain alignment algorithm, INDEL is undetectable; and the precision of SNP calling falls by raising accuracy of INDEL calling. Therefore, two or more analyses with different tools are recommended, but it causes increasing time and cost. To solve these problems, it is necessary to develop a mutation calling tool with higher-precision.
The purpose of this research is to build the mutation calling tool which can detect SNP and INDEL simultaneously with high precision. We investigated some problems of the mutation calling by the known procedure and created a new tool with original algorithm. Moreover, accuracy verification of our tool was performed by analysing two eubacteria. As a result, it enabled the INDEL calling which does not lower the accuracy of SNP calling, and identifying of INDEL which was not detectable in the mutation call by the known procedure.

Keywords: mutation calling, base sequence determination, eubacteria, INDEL calling

1. はじめに

次世代シーケンサと呼ばれる技術の進歩により，様々な生物のゲノムの DNA などの塩基配列解読が近年急速に進んでいる．その結果，様々な生物の DNA 配列を解析することが可能になり，塩基配列中にどのような変異が発生しているか簡便に調べられるようになった．変異の解析結果と疾患や薬に対する副作用の感受性などの情報を組み合わせれば，効果が高く副作用の低い「テーラーメード医療」とよばれるような個々人の体質に合わせたきめ細かい医療の実現に大きく貢献できると期待されている [1].

塩基配列の変異には大別して置換 (SNP:Single Nucleotide Polymorphism)，挿入 (Insertion)，欠失 (Deletion) の 3 種類があり，挿入と欠失はまとめて INDEL と呼ばれる [2]．それらがどの位置に発生しているかを特定する技術は変異コールと呼ばれており，ゲノム解析の中心となる技術の一つに挙げられる．

変異コールには前段階の解析であるマッピングによって得られたデータを用いるため，マッピングツールのアライメント手法によって変異コールの結果が異なるという特徴がある．正確な変異箇所の特定のためには高精度の変異コールが必要だが，アライメントの手法次第では特定の変異が検出できない場合や，リードの量によって変異コールの結果が変化する問題がある．そのため，解析の手法を変えた複数回の解析が推奨されており，必要な時間やコストが増大する原因となっている [3]．これらの問題を解決するため，より精度の高い変異コールツールを開発する必要がある．

本研究では，真正細菌の塩基配列中から SNP と INDEL を高精度に同時検出できる変異コールの実現を目的とし，既存のツールによる変異コールの問題点の指摘と，独自のアルゴリズムによる変異コールツールの作成を行った．また，複数の真正細菌を対象とし，作成したツールの検証実験を行った．

2. 塩基配列解析の手順

本章では，塩基配列解析の手順について解説する．なお，本研究では Illumina 社の提供する Genome Analyzer IIx という次世代シーケンサを対象とした．

2.1 シーケンス

シーケンサに配列解析を行いたい対象のゲノムの DNA を入力すると，DNA の断片化と化学反応による一塩基単位での配列情報の取得を行う．塩基配列解析は，シーケンスによって得られたデータを対象として段階的に解析を行う．

2.2 ベースコール

ベースコールとはシーケンスによって得られた大量のデータを解析し，各塩基ごとに 4 種類の塩基のうちいずれかに決定する作業である [4]．ベースコールにより得られた TGCTACGAT…という配列データはリードと呼ばれ，一度のシーケンスで数百万から数千万個のリードが得られる．

シーケンス時の化学反応の失敗により，塩基配列決定が正確に行えない可能性がある．塩基配列決定の精度はリードの後半になるにつれて低くなるため，長いリードの塩基配列決定を行う場合，リードの後半ではクオリティが低くなる傾向がある [5].

2.3 マッピング

マッピングとは，前節で述べたベースコールの結果として出力されるリードデータを，レファレンスのどの部分にあたるものか 1 リードずつ検証し，アライメントを行う解析である．マッピング用のツールは複数あり，現在の主流となっているものに”Bowtie”と”BWA”がある．Bowtie は現在，世界で最も使用されているマッピングツールであり [7]，高速なアライメントが可能である．しかし，ギャップアライメントを行わないため，変異のうち挿入と欠失を検出できないという欠点があり，それらの変異箇所を同定するには不向きなツールである．

BWA はギャップアライメントを行っており，Bowtie よりも挿入・欠失箇所の同定に適している [8]．基本的なアライメント手法は Bowtie と同一だが，ギャップアライメントを行うため，アライメント処理に要する時間は Bowtie よりも長い．

本研究で作成した変異コールには mpsmap というマッピングツールを用いた [6]．mpsmap はギャップアライメントを行わないマッピングツールである．同じくギャップアライメントを行わないツールである Bowtie と比較して，ミスマッチの許容数を任意に設定できるという特徴がある．本研究では，ミスマッチの許容数を任意に設定できるという点に注目した．

2.4 変異

変異には大別して置換・挿入・欠失の 3 種類がある．

置換とは，塩基配列中の 1 塩基が別の塩基に置き換わる突然変異のことである．この変異は SNP(Single Nucleotide Polymorphism：一塩基多型) と呼ばれる．

欠失 (Insertion) とは，塩基配列の一部が失われることであり，”IN”と略される．また，挿入 (Deletion) とは，塩基配列の途中に塩基が付加されることであり，”DEL”と略される．欠失と挿入はコドンの構成が変異箇所以降で変化し

[1] 奈良先端科学技術大学院大学
Nara Institute Science and Technology
[2] 前橋工科大学
Maebashi Institute of Technology
[a)] kazushige-a@is.naist.jp

てしまうので，フレームシフト突然変異などの深刻な帰結をもたらす場合がある [9][10].

これらの変異が発生している箇所を検出するためには変異コールを行う必要がある．変異コールとは，どの位置にどのような変異が発生しているかを解析する技術である．変異コールを行うツールの代表的なものに，SAMtools が挙げられる [11]．SAMtools はマッピングの結果として得られたファイルの編集・解析を行うツールであり，その機能のひとつに変異コールが含まれている．マッピング結果のファイルにはアライメントの際に考慮したミスマッチやギャップの情報が含まれており，それらの位置にアライメントされたリードの数などの情報から変異発生箇所の抽出を行う．変異コールの結果はマッピング結果の影響をうけるため，使用したマッピングツールの精度によって変異コール結果が大きく変化するという特徴がある．たとえば，マッピングに Bowtie を用いた場合では，Bowtie はギャップアライメントを行わないため，挿入や欠失が起こっていても検出できないという結果になる．

次章で，Bowtie, BWA, SAMtools を用いた変異コールの精度の検証と比較を行う．

3. 真正細菌を対象とした変異コールの検証

3.1 実験対象

本検証は，真正細菌 Escherichia coli str. K-12 substr. MG1655 chromosome を対象として実験を行った．以下，E.coli と記載する．今回の検証には，置換が 2 箇所，挿入が 1 箇所，計 3 箇所の変異が観察されるものを利用した．

3.2 検証に用いた解析手順

Bowtie と BWA それぞれを用いて，ミスマッチを 3 箇所まで許容するマッピングを行った．マッピング結果の編集と変異コールには SAMtools を用いた．変異コール結果のうち，次の要素に注目する．

ポジション (POS)
　レファレンスの何番目の塩基かを示す数値.

変異内容 ($TYPE$)
　発生している変異が置換/欠失/挿入のいずれかを示す.

正確性のスコア ($QUAL$)
　0〜255 の数値で出力される．検出結果が $1 - 10^{(-QUAL/100)}$ の精度であることを示す.

リードの本数 (DP)
　そのポジションに張り付いたリードの本数.

3.3 検証結果

3.3.1 変異コールの結果

マッピングに Bowtie を使用した場合は全部で 468 箇所の変異が検出された．検出結果から低い精度のものを除外するため $QUAL \geq 200$ の変異箇所のみに注目し，表 1 に示す.

表 1 Bowtie+SAMtools による変異コール結果

POS	$type$	$QUAL$	DP	正誤
547694	SNP	212	164	○
3422257	SNP	222	1265	×
3422258	SNP	222	1281	×
3422259	SNP	222	1282	×

また，マッピングに BWA を使用した場合は，置換が 34 箇所，挿入が 2 箇所検出された．同様に低い精度のものを除外するために $QUAL \geq 200$ の変異箇所のみに注目し，出力結果を表 2 に示す.

表 2 BWA+SAMtools による変異コール結果

POS	$type$	$QUAL$	DP	正誤
547694	SNP	222	187	○
547831	INDEL	214	228	○
3957957	SNP	222	128	○

$QUAL \geq 200$ のものに注目すると，マッピングに Bowtie を用いた場合では置換が 4 箇所，BWA を用いた場合では置換が 2 箇所と挿入が 1 箇所検出された．Bowtie を用いて検出された 4 箇所の置換のうち，実際に変異が起こっている箇所は 547694 のみであり，残りの 3 箇所については誤りであった．一方，BWA を用いた場合で検出された変異は 3 箇所であった．これらは変異発生箇所を正確に検出したものであった.

3.3.2 検証結果の考察

マッピングに bowtie を用いた場合に検出されなかった 1 箇所の置換 3957957 は，$QUAL = 178$, $DP = 109$ という結果になっていた．しかし，誤った検出のうちに $QUAL = 177, 176, 175$ となる 3 点が含まれていたため，マッピングに Bowtie を使用した場合，全置換箇所の検出には厳密な $QUAL$ の閾値決定が必要である．今回の検証ではあらかじめ変異の個数が判明しているデータを用いたので設定すべき $QUAL$ の閾値を知ることができたが，変異の個数や箇所が不明なデータを対象に変異コールを行う場合，閾値の設定を正確にできないことが予測される．また，全置換箇所の検出に成功しても，変異が発生していない箇所を誤って検出した場合のほうが高いスコアとなることもあった．一方，BWA による変異コールは $QUAL \geq 200$ の箇所に注目した場合は正確であった．しかし，Bowtie を用いた場合の $QUAL$ の閾値を BWA を用いた場合の結果に適用すると，誤った検出結果 ($QUAL = 183$) が混入した．この結果から，$QUAL$ を基準とした変異コール結果の評価ができないということがわかる.

加えて，変異コールの精度が低いため，結果が正確であるかの評価はマッピング結果を目視で確認する必要がある．今回の検証に用いた E.coli のゲノムの全長は約 4.6Mb で

あったため，変異が検出された箇所をすべて目視で確認することができたが，対象が真核生物であった場合，たとえばヒトゲノムの場合は全長 3,100Mb となり，SNP は 500 万〜1,100 万種あると試算されているため，目視での確認は非現実的であるといえる [12].

次章では，これらの問題を解決するべく，置換・挿入・欠失を同時に高精度で検出するための手法を提案する.

4. SNP と INDEL 同時検出手法の提案

提案する手法には，マッピングに mpsmap を用いる．これまでの研究ではリードあたりのミスマッチ許容数は通常 3，最大でも 5 程度であったが，本手法ではリード長さの約半分までのミスマッチを許容する設定とする．本研究に用いたリードデータの長さは 35bp であるため，ミスマッチの許容数は 17 とした．また，マッピング結果のうち注目する要素として，レファレンス位置に張り付いたリードの本数 DP とミスマッチの数 $mismatch$，その割合 $score$ の 3 要素に注目する．$score$ は下の式によって算出される.

$$score = 100 * mismatch/DP \quad (1)$$

また，各変異について簡潔にするためにモデル化を行う．モデル化の際には

- シーケンスとベースコールは正確であり，変異箇所以外にミスマッチは発生しない
- DP はレファレンス全体で均一である

と仮定する.

4.1 モデルを用いた変異コールのアルゴリズム

mpsmap により，E.coli の変異発生箇所におけるマッピング結果を可視化した．ミスマッチ許容数は一般的な設定である 3 と，提案手法に用いる 17 の二通りで可視化を行った.

4.1.1 置換検出手法の提案

置換発生箇所周辺のマッピング結果を抽出し，観察を行った．マッピング結果を図 1 に示す.

図 1 より，変異位置でのミスマッチは縦一直線に集中していることがわかる.

次に，置換発生箇所についてモデル化を行い (図 2)，1 ベース単位で DP, $mismatch$, $score$ を算出した．置換発生箇所を中心とした $score$ の分布を図 3 に示す.

図 3 より，置換発生箇所では特異的な $score$ が観察される．この $score$ の値を検出することで，置換発生位置の特定を行う.

4.1.2 挿入検出手法の提案

次に，挿入発生箇所周辺のマッピング結果を抽出し，観察を行った (図 4).

図 4 では変異箇所周辺に特徴的な形状でミスマッチが集中している様子が確認できる．この結果より，ミスマッチの許容数を多く設定することでギャップアライメントを行わなくても INDEL が検出できるようになると考えられる.

次に，挿入発生箇所についてモデル化を行い (図 5)，1 ベース単位で DP, $mismatch$, $score$ を算出した．変異箇所を中心とした $score$ の分布を図 6 に示す.

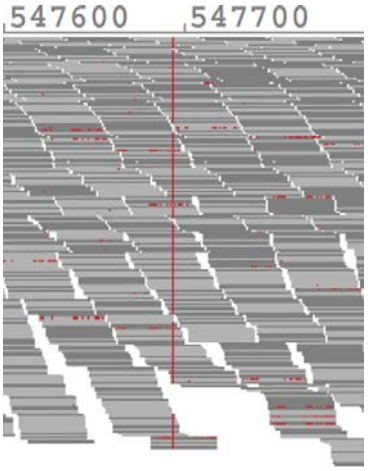


図 1 置換発生箇所周辺のマッピング結果

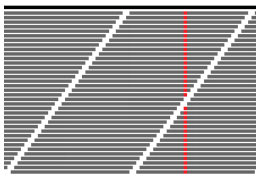


図 2 置換発生箇所のモデル

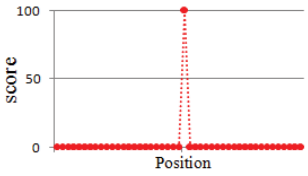


図 3 置換発生箇所の $score$

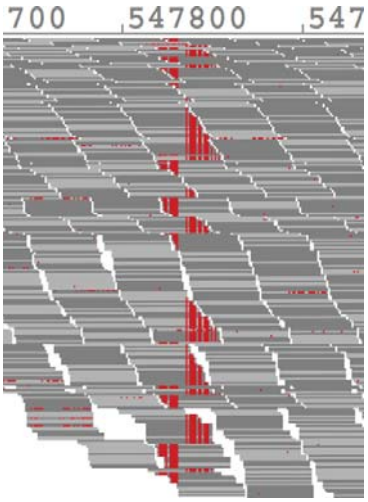


図 4 挿入発生箇所周辺のマッピング結果

図 5 挿入発生箇所のモデル

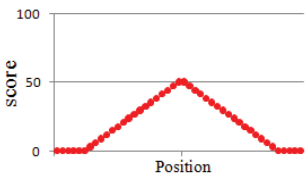


図 6 挿入発生箇所の $score$

図 6 より，*score* の分布は変異箇所の前後 1 ベースを中心とした山型となることがわかる．ミスマッチが集中する領域の幅は，ミスマッチ許容数の倍かリード長のいずれか少ないほうの値をとる．また，*score* のピーク値は 50 となる．この山型の形状を抽出することにより，挿入発生箇所を特定する．

score のグラフから傾きを算出することで山型の形状を抽出する．*score* のピーク値が 50，ミスマッチが集中する領域の幅はリード長 (*read*) と同一なため，*score* のグラフの傾きは $\pm 50/(read/2) = \pm 100/read$ となる．以上より，リード長と同じ長さの領域に傾きが $\pm 100/read$ となる *score* が集中している領域を抽出することで，挿入発生箇所を特定できると考えられる．

4.1.3 欠失検出手法の提案

検証の対象とした E.coli には欠失は観察されなかったため，モデルによる考察を行う．欠失発生箇所のモデル化を行うと図 7 に示す結果となる．モデルから *score* を算出し，変異箇所を中心とした分布を図 8 に示す．

図 7 欠失発生箇所のモデル

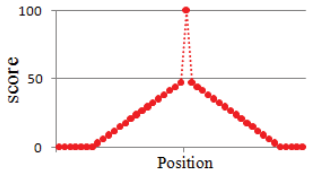


図 8 欠失発生箇所の *score*

図 7 は図 5 とほぼ同様であるため，ミスマッチ許容数を多く設定することで欠失の発生箇所を観察できると予測できる．また，図 8 では変異箇所を中心とした山型の *score* が観察される．ミスマッチが集中する領域の形状はピーク値が 100 となること以外は挿入の場合とほぼ同様である．

以上より，挿入と欠失はいずれも変異箇所周辺の *score* の分布が山型の形状をとることがわかる．この形状を抽出することで挿入と欠失を同時に検出でき，*score* のピーク値によって挿入と欠失を分別できると予想できる．

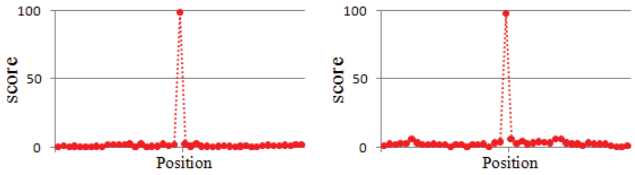
4.2 提案手法の検証

各提案手法について，*score* の算出実験を行った．実験の対象は 3 章で行った実験の対象と同一の E.coli とした．マッピングに用いるツールは mpsmap とし，ミスマッチ許容数は 17 とした．全レファレンス位置について *DP*, *mismatch*, *score* を算出した後，変異が発生している箇所を中心とした 45bp の範囲について *score* をグラフ化した．

4.2.1 置換発生箇所の score

置換が発生している 2 箇所の *score* のグラフを図 9 に示す．

変異発生箇所の *score* は，*score* = 98.1±0.3 となった．また，グラフの形状について，予想とほぼ一致する結果が



[a] 547694　　[b] 3957957

図 9 置換発生箇所の *score*

得られた．この実験結果から，*score* が 97 以上の箇所を選択的に抽出すれば，置換の検出は可能であると考えられる．

4.2.2 挿入発生箇所の score

挿入が発生している 1 箇所の *score* のグラフを図 10 に示す．

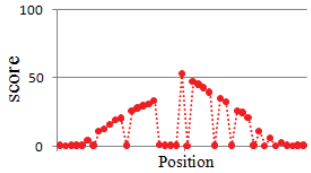


図 10 挿入発生箇所 (547831) の *score*

変異発生箇所の *score* のピーク値は 52.9 となった．グラフの形状について，*score* の起伏が予想よりも激しいという結果が得られた．起伏を無視すれば，予想通りの山型であると言える結果となった．また，2 ベース分のピーク値が観察できると予測したが，ピーク値は 1 ベース分しか観察されなかった．

4.2.3 考察

置換発生箇所の *score* については全箇所で予想と一致する結果が得られたため，提案した手法により問題なく検出できると考えられる．

挿入発生箇所周辺の *score* の値に起伏が発生する原因は，塩基は 4 種類しかないため，数塩基のずれが生じても全ての位置で 1/4 の確率で正解となるためであった．また，ピーク値が 1 ベース分しか観察されなかった原因は，変異後に 2 塩基以上のホモポリマーが形成されていたためであった．

以上より，置換の検出はモデルを用いた場合と同様の手法で行えると予測できる．一方，挿入と欠失をグラフの傾きから検出するには起伏を無視することが必要である．また，変異後にホモポリマーが形成されることによって *score* に長いブランクが発生している場合も考慮する必要がある．ブランクの対策には，*score* のグラフの傾きを算出したのち，傾きの正負からミスマッチが集中する領域を前半と後半に分割し，前半部分の終点と後半部分の始点のギャップが数塩基以内であることを検出の条件とすればよいと考えられる．

E.coli を対象とした検証ではホモポリマーは 5 塩基の長さであったため，作成した変異コールツールではさらに長

いギャップに対応できるようギャップを 7 塩基までとした．この数値を大きく設定することでより長いホモポリマーに対応することができる．

4.3 提案するアルゴリズムのフロー

提案するアルゴリズムのフローを以下に示す．例として，E.coli における挿入発生箇所を中心とした 45 ベースの領域での処理を図示する．

(1) mpsmap でのマッピングで得られたファイルを読み込み，レファレンスの全位置について *score* の算出を行う (図 11)．

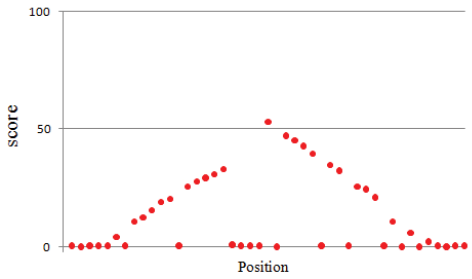


図 11 *score* を算出

(2) INDEL 発生位置における *score* の起伏を無視するため，$score \geq 5$ となった点のみを出力する (図 12)．同時に，$score \geq 80$ 以上の位置を取得する．この位置は置換または欠失が発生している位置として記憶する [出力 2]．

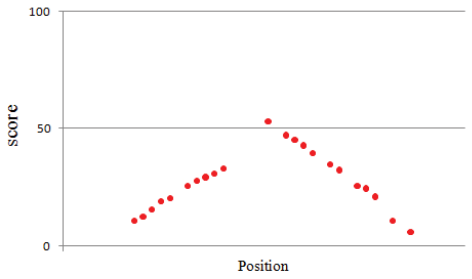


図 12 $score \geq 5$ の箇所を抽出

(3) 出力 1 より，リード長 [bp] 以内に 3 箇所以上抽出された領域を抽出し，*score* の傾きを算出する (図 13)．
リード長/2 [bp] 以内に傾き 2 以上の点が 3 箇所以上ある領域を抽出する [前半部分]．
リード長/2 [bp] 以内に傾き-2 以下の点が 3 箇所以上ある領域を抽出する [後半部分](図 14)．

(4) 前半部分の終点と後半部分の始点が 7[bp] 以内になる組み合わせを検索する (図 15)．成立した組み合わせを INDEL 発生位置として記憶する．[出力 3]

(5) 出力 2 のうち，出力 3 の領域内にある点を検索する．検索された点は，DEL 発生位置である．

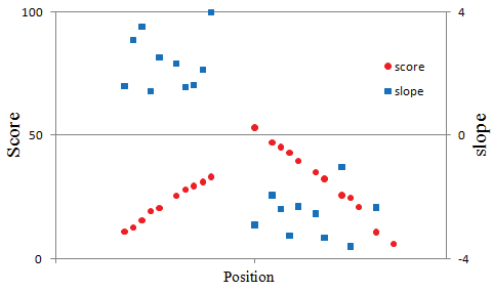


図 13 図 12 から傾きを算出

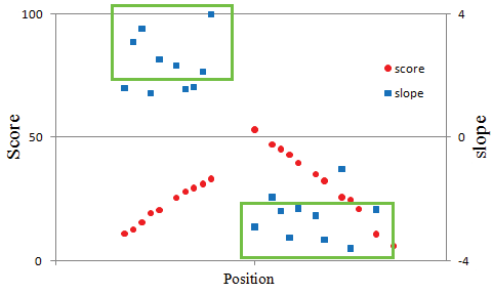


図 14 傾きから領域抽出

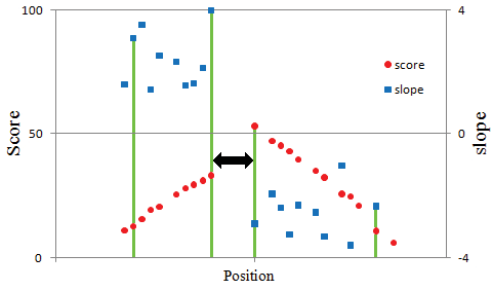


図 15 領域の間隔を抽出

出力 2 のうち，検索されなかった点は SNP 発生位置である．
出力 3 のうち，検索されなかった点は INDEL 発生位置である．

このフローに従って変異コールツールを作成した．

5. 提案手法を用いた変異コール実験

5.1 実験対象

本実験の対象は，2 章に用いた E.coli に加え，Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293 を対象とした．

それぞれのリードデータに対し，Bowtie, BWA, mpsmap によるマッピングを行い，SAMtools と独自アルゴリズムによる変異コールを行った．リード長は全て 35[bp] であるため，ミスマッチ許容数の設定は，Bowtie と BWA では 3，mpsmap では 17 とした．

5.2 変異コール実験の結果

5.2.1 E.coli を対象とした変異コールの結果

E.coli を対象とした実験を行うにあたって，第 3 章に記載した既存のツールによる変異コール結果から，マッピングに Bowtie を用いた変異コールでは検出されなかった置換箇所の検出と誤ったコール箇所の除外ができ，BWA を用いた結果と同様の検出結果が得られると予測した．

E.coli を対象として独自のアルゴリズムで変異コールを行った結果，全部で 3 箇所の変異が検出された．検出された変異箇所と変異内容を表 3 に示す．

表 3 独自アルゴリズムによる E.coli を対象とした変異コールの結果

POS	*type*	*DP*	正誤
547694	SNP	192	○
547831 - 547836	INDEL	227	○
3957957	SNP	138	○

この結果から，すべての変異位置が予想の通り正確にコールされていることがわかる．

5.2.2 Leuconostoc を対象とした変異コールの結果

1. Bowtie + SAMtools による変異コール

合計で 102 箇所の変異が検出された．そのうち $QUAL \geq 200$ の変異箇所について，出力結果を表 4 に示す．

表 4 Bowtie+SAMtools による変異コール結果から $QUAL \geq 200$ の箇所を抽出した結果

POS	*type*	*QUAL*	*DP*	正誤
338699	SNP	222	507	○
410044	SNP	222	401	○
559188	SNP	222	384	○
755527	SNP	222	326	○
953160	SNP	222	262	○
1094250	SNP	222	335	○
1236979	SNP	222	400	○
1693283	SNP	222	552	○
1993032	SNP	222	406	○

出力結果として検出された 102 箇所について目視による確認を行ったところ，SNP であると断定できる箇所は 9 箇所あり，そのすべての箇所が正確に検出されていた．

2. BWA + SAMtools による変異コール

置換が 21 箇所，欠失が 1 箇所，合計で 22 箇所の変異が検出された．そのうち $QUAL \geq 200$ の変異箇所について，出力結果を個数を表 5 に示す．

検出された 22 箇所について目視による確認を行ったところ，変異箇所と断定できる箇所は Bowtie を用いた実験で検出された置換 9 箇所と，今回検出された欠失が 1 箇所であり，$QUAL \geq 200$ として抽出された他の箇所は誤りであった．しかし，誤りであった箇所には正解であった箇所よりも *QUAL* が高い箇所があり，*QUAL* を基準とした変異箇所の検出を行えないことが明らかになった．

表 5 BWA+SAMtools による変異コール結果から $QUAL \geq 200$ の箇所を抽出した結果

POS	*type*	*QUAL*	*DP*	正誤
115659	SNP	225	484	×
197593	SNP	222	67	×
269842	SNP	225	135	×
338699	SNP	222	511	○
410044	SNP	222	404	○
472524	SNP	225	535	×
559188	SNP	222	385	○
615573	SNP	225	152	×
755527	SNP	222	330	○
796684	SNP	225	71	×
953160	SNP	222	268	○
1094250	SNP	222	339	○
1236979	SNP	222	409	○
1237017	SNP	225	73	×
1600218	INDEL	214	224	○
1693283	SNP	222	561	○
1993032	SNP	222	407	○

3. mpsmap + 独自ツール による変異コール

合計で 21 箇所の変異が検出された．検出された変異箇所と変異内容を表 6 に示す．

表 6 独自アルゴリズムによる Leuconostoc を対象とした変異コールの結果

POS	*type*	*DP*	正誤
197592 - 197593	INDEL	330	○
269842 - 269845	INDEL	414	○
338699	SNP	515	○
410044	SNP	405	○
558511 - 558515	INDEL	299	○
559188	SNP	386	○
615573 - 615576	INDEL	555	○
755527	SNP	332	○
796684 - 796686	INDEL	306	○
953160	SNP	269	○
1094250	SNP	341	○
1236979	SNP	410	○
1237015 - 1237020	INDEL	300	○
1291049 - 1291051	INDEL	349	○
1600219 - 1600224	INDEL	224	○
1624087 - 1624088	INDEL	490	○
1693283	SNP	566	○
1993032	SNP	411	○
(2038394 - 2038395	INDEL	424	×)

出力されたすべての箇所について目視で確認を行ったところ，置換 9 箇所と INDEL9 箇所について変異発生個所であると断定できた．2038394 - 2038395 は本来 INDEL ではない箇所だが，検出結果に混入した．これ

はLeuconostocのレファレンスの全長が2038396[bp]であり，次の塩基(2038396)以降はプラスミドのレファレンスであったため，異なったレファレンスの間に生じたミスマッチの集中である．つまり，Leuconostocの配列のみに注目するとこの検出結果は無視できる．

また，確認のため，マッピング結果全体の目視による確認を行った．その結果，今回の実験に用いたLeuconostocのデータには表6に示した18箇所以外の変異はが確認されなかった．

これらより，Leuconostocを対象とした変異コールでは，mpsmapと独自ツールを組み合わせた場合ですべての変異位置が正確にコールされていることがわかる．BowtieとSAMtoolsを用いた場合は，置換の検出は正確であったが，INDELを検出できない仕様のため，完璧であるとは言えない．BWAとSAMtoolsを用いた場合は，置換は全箇所検出できていたが検出の誤りが目立つほか，2箇所の挿入と5箇所の欠失を検出できていなかった．

以上の結果から，本研究で作成した変異コールツールはSNPとINDELの同時検出を可能としており，変異箇所の見逃しもなく，SAMtoolsを用いた場合よりも精度が高いといえる．

6. おわりに

本研究では，既存の手順による変異コール精度の検証と，独自のアルゴリズムによる変異コールツールの作成および精度の検証を行った．既存の手順による変異コールは精度に問題があり，複数回の塩基配列決定と変異コールを繰り返す必要があるため，解析に必要な時間やコストが増大していた．また，用いられるマッピングツールのミスマッチ許容数の設定に限界があり，ギャップアライメントを行わない場合，ミスマッチの多いリードはアライメントされずに破棄されるという特徴があった．そこで，ミスマッチ許容数を任意に決定できるマッピングツールを用い，精度が高くSNPとINDELを同時に同定できるアルゴリズムを提案し，実装した．このアルゴリズムでは従来法ではアライメントされないリードを活用したINDEL発生箇所の同定が可能であり，シーケンスによって得られたデータを既存の手法よりも効率よく活用できる．さらに複数回の解析が不必要なため，解析に必要な時間とコストの削減ができる．

本研究で提案した手法の限界として，INとDELの区別ができない箇所があることが挙げられる．INまたはDELによってホモポリマーが構成された箇所では*score*のピークが特徴的な値をとらないため，それらの区別ができない．また，ホモポリマーの一部の欠失，または挿入によってホモポリマーが形成された場合は変異位置が厳密に特定できず，変異箇所の前後数塩基までしか特定できない．次に，変異後の塩基配列を特定できないという問題がある．*score*の算出を行う際にレファレンスとの正誤のみに注目したため，変異後にはA, C, G, Tのいずれに置換しているのかの特定ができない．

今後の課題として，挿入と欠失の区別を行う機能の実装，INDELの結果ホモポリマーが形成されていた場合の変異箇所の特定手法の考案，各変異発生箇所における塩基の変異の仕方を特定する手法の考案が挙げられる．また，真核生物のSNPには対立遺伝子が3〜4個あるものも存在するため，そのようなSNPに対する検出法を考案する必要がある．

以上より，本研究における成果と将来性は，ゲノム解析に貢献できると期待している．

謝辞 本研究を進めるにあたって実験対象の真正細菌のシーケンスデータを提供して頂いた奈良先端科学技術大学院大学 バイオサイエンス研究科 大島拓助教に深謝する．

参考文献

[1] 谷原正夫 : ゲノム情報による医療材料の設計と開発，シーエムシー出版, 2006.
[2] Steve R Bischoff, Shengdar Tsai, Nicholas E Hardison, Abby M York, Brad A Freking, Dan Nonneman, Gary Rohrer and Jorge A Piedrahita.: Identification of SNPs and INDELS in swine transcribed sequences using short oligonucleotide microarrays, *BMC Genomics*, Vol. 252, R14, 2008.
[3] Andrea Sboner, Xinmeng J Mu, Dov Greenbaum, Raymond K Auerbach and Mark B Gerstein. : The real cost of sequencing: higher than you think!, *Genome Biology*, Vol. 12, 2009.
[4] Christian Ledergerber and Christophe Dessimoz.: Base-calling for next-generation sequencing platforms, *Briefings in Bioinformatics*, Vol. 12, pp. 489–497, 2011.
[5] Martin Kircher, Udo Stenzel and Janet Kelso.: Improved base calling for the Illumina Genome Analyzer using machine learning strategies, *Genome Biology*, Vol. 10, R83, 2009.
[6] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C. Linak, Aki Hirai, Hiroki Takahashi, Md. Altaf-Ul-Amin, Naotake Ogasawara and Shigehiko Kanaya: Sequence-specific error profile of Illumina sequencers, *Nucleic Acids Research*, pp. 1–13, 2011.
[7] Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg. : Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, Vol. 10, R25, 2009.
[8] Heng Li and Richard Durbin. : Fast and accurate short read alignment with Burrows Wheeler transform, *Bioinformatics*, Vol. 14, pp. 1754–1760, 2009.
[9] Jotun Hein : An Algorithm Combining DNA and Protein Alignment, *J. theor. Biol.*, Vol. 167, pp. 169–174, 1994.
[10] 香川靖男，笠月健彦 : 遺伝と疾患, 岩波書店, 2000.
[11] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis and Richard Durbin : The Sequence Alignment/Map format and SAMtools., *BIOINFORMATICS*, Vol. 25, pp. 2078–2079, 2009.
[12] 西田奈央，徳永勝士 : 大規模SNPタイピングによる多因子疾患遺伝子の探, 実験医学, Vol. 25, pp. 178–184, 2007.