G-31

次世代シーケンサーによるゲノム解析のマッピング情報を用いた精度向上法の提案

Proposal of accuracy improvement method using mapping information of genome analysis by Next-generation sequencer

東 銀史 *　柿花 優成 †　大沢 勇統 *　高橋 篤 ‡　大星 直樹 *

AZUMA Ginji　KAKIHANA Yusei　OHZAWA Yuto　TAKAHASHI Atsushi　OHBOSHI Naoki

1　序論

次世代シーケンサーなどのゲノム解読技術の発展により, ゲノム解析研究が急速に進展し, ゲノム情報を医療や創薬等に応用しようとする動きが高まっている. 次世代シーケンサーによるゲノム解析では, リードと呼ばれる短い塩基配列断片のゲノム上の位置を, 配列が決定されているリファレンス配列と照らし合わせ決定するマッピング処理が行われる. 生物はゲノム塩基配列に基づいてタンパク質を生成しているが, 塩基がわずか 1 つ異なるだけで別のタンパク質が生成される原因となることがある. 不正確なマッピング処理によって得られた塩基配列は本来の塩基配列と異なる結果を示す可能性があり, その後の解析において, 本来とは異なるタンパク質が生成されるなどの誤った解釈を誘発する. よって, 正確なマッピング結果を得ることが後の遺伝子等の解析の精度向上に繋がると考えられる. 正確なマッピング結果を得るためのアプローチの一つとして, リードが誤った位置にマッピングされた確率を定量化した値である MAPQ (MAPping Quality score) によるフィルタリング (以下, MQ フィルタリング) が利用されている. マッピングされたリードが正しい位置でない確率を P とすると, MAPQ は数式 (1) で求められる [1].

$$Q = -10 \log_{10} P \tag{1}$$

2　目的

マッピング後のリードには, マッピング時の状況を示す様々な付加情報が記録されている. 付加情報の一つとして, 対象のリードが実際にマッピングされた位置を除く, 他のマッピング候補位置が記された XA (Alternative hits) タグが存在する. マッピングを行う際に利用されるリファレンス配列には, ある人種特有の特徴や混入した微生物の DNA を取り除くことを目的としたデコイ配列が含まれる. リードが正確にマッピングされていても XA タグにデコイ配列が挙げられている場合, そのリードは微生物由来のものである可能性が考慮され, MAPQ が低くなる傾向にある.

MQ フィルタリングでは MAPQ のみに着目してフィルタリングを行っており, 正確にマッピングされているリードまで取り除いてしまっている可能性が考えられる. 正確にマッピングされているリードが取り除かれてしまうことで, マッピング結果の精度が低下し得られた塩基配列が本来の塩基配列と異なる結果を示す可能性が高くなることが懸念される. 上記の XA タグの例では, リードが正確にマッピングされていても XA タグにデコイ配列が挙げられている為に MAPQ が低くなり, MQ フィルタリングによって取り除かれてしまう. 本研究ではマッピング処理後のデータに含まれるマッピング情報に着目したフィルタリング精度向上の手法について提案し, 提案手法によってフィルタリングの精度が向上したか検定を行う.

3　方法

3.1　提案手法

本研究ではマッピング処理後のデータに対して MQ フィルタリングを行い, 取り除かれたリードのうち, MAPQ が一定値以上のリードの XA タグに着目する. XA タグに第一候補箇所としてデコイ配列が挙げられている場合, 正しくマッピングされたリードであると判断する. これにより, 正常にマッピングされたリードが MQ フィルタリングによって取り除かれてしまう量を減少させ, フィルタリングの精度向上を図る.

3.2　利用データ

人工ゲノムデータを利用する. リードには図 1 に示すように Single–End read と Paired–End read が存在する. Single–End read は塩基配列の断片の片側からのみ読み取るのに対し, Paired–End read は反対側からも読み取る. これにより, Paired–End read の情報量は Singel–End read の 2 倍となり, より高精度な解析が可能となる.

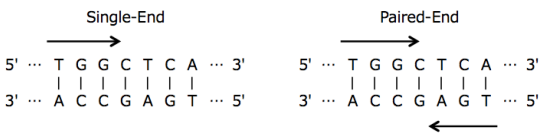

図 1　リードの種類

本研究で利用する人工ゲノムデータは, ヒトの常染色体中で最も短い染色体 21 番の Paired–End read とする. 各リードの長さは EMBL-EBI [2] にて公開されているゴールドスタンダードな全ゲノムデータセットである CEPH1463

* 近畿大学大学院 総合理工学研究科
Graduate School of Science and Engineering, Kindai University
† 近畿大学 理工学部
Faculty of Science and Engineering, Kindai University
‡ 国立循環器病研究センター
National Cerebral and Cardiovascular Center

家系データ同様に 101 塩基とする．各塩基は読み取りエラーを含み, エラー率はリードの先頭塩基で 0.1%, リードの末尾塩基で 1.0% となるように線形に遷移している．

3.3 マッピングツール

マッピングには広く用いられているツールである BWA (Burrows–Wheeler Aligner) [3] を使用する．BWA は比較的短いリード向きのマッピングツールであるが, 1,000 塩基程度の長いリードにも対応している．BWT (Burrows–Wheeler Transform) [4] をベースとした BWA–backtrack, BWA–SW, BWA–MEM の 3 つのアルゴリズムが利用可能である．本稿では, 他の 2 つよりも比較的新しく, 処理が速い BWA–MEM を使用した．使用バージョンは bwa–0.7.12 である．

3.4 評価方法

人工データのマッピング後のデータに対して閾値を 30〜40 の間で変化させながら MQ フィルタリングを行う．MQ フィルタリングによって取り除かれたリードから, XA に第一候補としてデコイ配列が挙げられており, かつ MAPQ が 20(マッピング結果の信頼度が 99%) 以上のものも正しくマッピングされたリードとする (図 2 参照)．

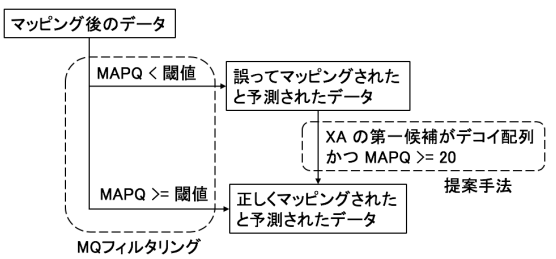

図 2　提案手法によるフィルタリング

提案手法の効果は以下 5 つの評価指標を MQ フィルタリングのみの結果と比較し, 対応のある t 検定によって評価する．本稿において, 陽性とは正しくマッピングされていること, 陰性とは誤ってマッピングされていることを表す．

1. 感度 (sensitivity)
   陽性のデータを正しく陽性と予測した割合．正しくマッピングされているが MQ フィルタリングでは取り除かれてしまうリードを, 提案手法ではどれだけ取得できるかを確認する．
2. 精度 (accuracy)
   データを正しく予測できた割合．提案手法により, どれだけフィルタリングの精度が向上したか確認する．
3. 偽陽性 (false positive rate)
   陰性のデータを誤って陽性と予測した割合．提案手法により増加することが懸念される為, どの程度増加するか確認する．
4. 適合率 (precision)
   陽性と予測したデータが本当に陽性である割合．フィルタリングの結果, 陽性と判断されたリードの信頼性が向上していると言えるかを確認する．
5. F1 値 (F1-measure)
   トレードオフの関係にある感度と適合率を同時に評価できる指標．感度と適合率の調和平均．

4 結果・考察

MQ フィルタリングを行った結果とその後に提案手法を実行した結果を図 3〜6 に示す．True Positive, False Positive, False Negative, True Negative の意味は以下の通りである．

True Positive (TP)	陽性のリードを陽性と判断した．
False Positive (FP)	陰性のリードを陽性と判断した．
False Negative (FN)	陽性のリードを陰性と判断した．
True Negative (TN)	陰性のリードを陰性と判断した．

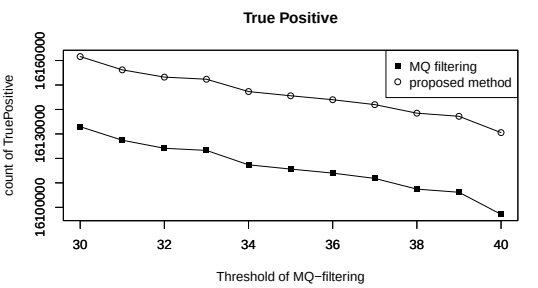

図 3　各閾値における True Positive のリード数

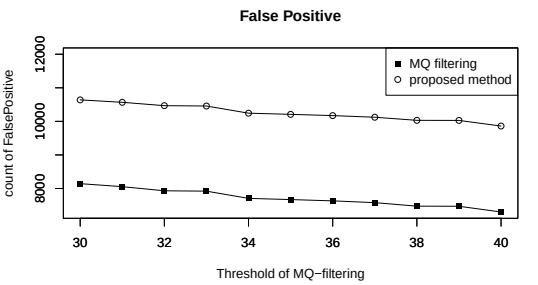

図 4　各閾値における False Positive のリード数

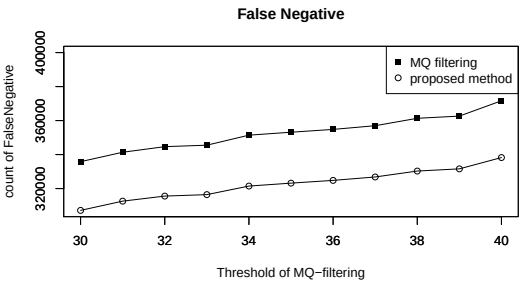

図 5　各閾値における False Negative のリード数

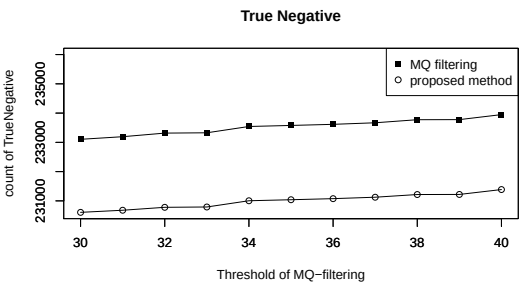

図 6　各閾値における True Negative のリード数

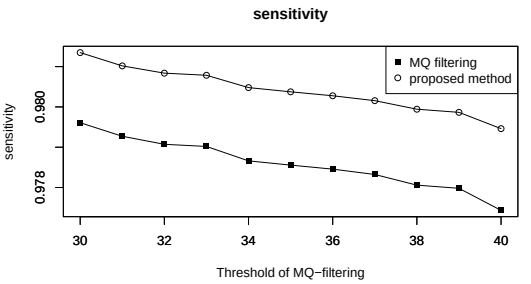

図 7　各閾値における感度

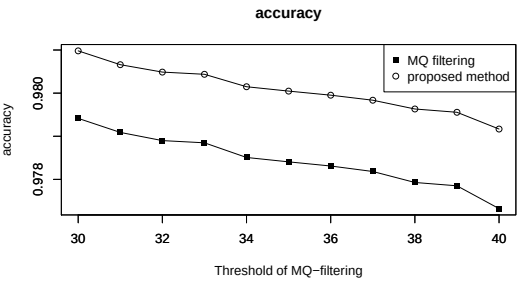

図 8　各閾値における精度

図 3〜6 より, MQ フィルタリングと比較して TP のリード数が増加し, FN のリード数は減少していることが確認できる. しかし, 同時に FP のリード数も増加し, TN のリード数も減少している. このことから, 正確にマッピングされたにも関わらず MQ フィルタリングによって取り除かれたリードを取得することは可能となったが, 同時にマッピングを誤ったリードも取得してしまっていることが分かる. FP のリードには XA タグに複数の候補箇所が記載されているものが多くみられ, XA の第一候補のみでなく候補の挙げられた数も考慮することで FP のリード数を減少させられる可能性が考えられる.

MQ フィルタリングと提案手法の各評価指標の比較結果を図 7〜11 に示す. また, 対応のある t 検定を行った結果を表 1 に示す. t 検定における帰無仮説は「MQ フィルタリングと提案手法に差は無い」, 対立仮説は「MQ フィルタリングと提案手法には差がある」である.

図 7〜11 および表 1 より, 感度, 精度, 偽陽性, F1 値は有意に向上しており, 適合率は有意に低下していることが確認できる. 従って提案手法を適用することで, 正確にマッピングされているにも関わらず MQ フィルタリングでは取り除かれてしまうリードを取得することが可能であり, その結果フィルタリング後のデータの精度が向上することが示された. しかし, 誤ってマッピングされたリードも数多く取得してきてしまい, 適合率は低下してしまう. そのため, マッピ

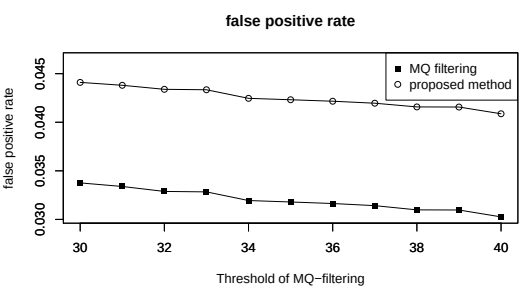

図 9　各閾値における偽陽性

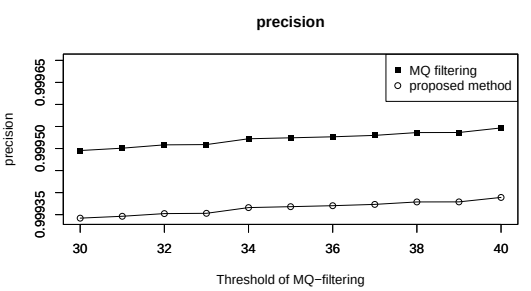

図 10　各閾値における適合率

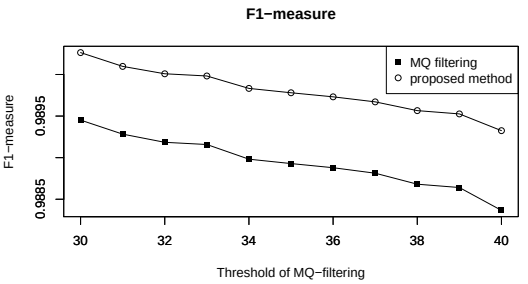

図 11　各閾値における F1 値

表 1　各評価指標の t 検定の結果

評価指標	統計量 t	p 値	
感度	-73.073	$5.621e^{-15}$	***
精度	-67.67	$1.21e^{-14}$	***
偽陽性	-438.25	$2.2e^{-16}$	***
適合率	407.62	$2.2e^{-16}$	***
F1 値	-66.916	$1.353e^{-14}$	***

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$

ング処理の後に行われる遺伝子変異解析等において, 本提案手法を適用した際のデータの精度向上による影響と適合率の低下による影響のどちらがより大きな問題となるかを検証する必要がある.

5　結論

人工ゲノムデータに対してマッピングを行い, マッピング結果に対してフィルタリングを行った. 従来の MAPQ のみを考慮したフィルタリングに加え, マッピング後に付与されるデータである XA に着目したフィルタリングを行うことでフィルタリングの精度が向上することが確認された. 今後は他のゲノムデータに対しても提案手法のフィルタリングが有用であるかの検証を行う必要がある. また, フィルタリング後のデータを用いることで遺伝子変異解析の精度がどの程度向上するかを検証することが重要と考えられる.

参考文献

[1] SourceForge (2018). `http://maq.sourceforge.net/qual.shtml`.

[2] EMBL-EBI (2018). `https://www.ebi.ac.uk/`.

[3] H. Li and R. Durbin: “Fast and accurate short read alignment with burrows–wheeler transform”, Bioinformatics, **25**, 14, pp. 1754–1760 (2009).

[4] B. Langmead, C. Trapnell, M. Pop and S. L. Salzberg: “Ultrafast and memory-efficient alignment of short dna sequences to the human genome”, Genome Biology, **10:R25**, 3 (2009).