7ZF-04

マッピングの特性を考慮した次世代シークエンサーを用いた一塩基多型解析の精度向上の試み

大沢 勇統† 東 銀史† 高橋 篤‡ 大星 直樹†
近畿大学† 国立循環器病研究センター‡

1. はじめに

ヒトゲノムで最頻度の個体差である一塩基多型(Single Nucleotide Polymorphism:以下 SNP)は，個別化医療や疾患原因究明への応用が期待されている．次世代シークエンサーによる SNP 解析では，前処理として得られた塩基配列断片を，既存の参照ゲノム配列と比較し位置を特定するマッピングが行われる．しかしゲノム配列の個体差や類似配列の影響で僅かなエラーが含まれることがあり，SNP の誤検出への影響が懸念される．現状，マッピングに利用するツールの性能評価は行われているが，ツールのマッピング結果の特性が，実際の解析における SNP の誤検出にもたらす具体的な影響や，それらを考慮した解析指針に関する分析は十分であるとはいえない．

本研究では，マッピング結果の特性と SNP の誤検出との関連を確認し，特性の考慮が精度向上に如何に寄与するかを検証することを目的とし，SNP 周辺のマッピング特性を考慮したエラーSNP 判別実験を通して，誤検出に関連するマッピング特性の抽出と判別精度評価を行った．

2. 方法

2.1. 家系 SNP 解析によるエラーの評価

SNP 解析結果を分析するため，図 1 に示す EMBL-EBI[1] で公開されている CEPH1463 家系に対して SNP 解析推奨フロー[2]に基づき解析を行った．マッピング処理には，近年広く用いられている BWA[3], Bowtie2[4]を使用した．解析では膨大な数の SNP が得られるため，最も短い染色体 21 番に存在する SNP を分析に用いた．また，得られたエラーSNP とそうでない SNP との検出数の差が極めて大きく，判別実験にてエラーでない SNP の傾向に偏った結果が得られる可能性がある．そこで，検出数が少ないうえエラーが多く，エラーSNP とそうでない SNP との検出数の差が比較的大きくなり過ぎないと考えられる，NCBI の公共変異データベースである dbSNP に未登録の SNP を分析対象とした．次に分析対象 SNP に対して，図 1 に示す遺伝法則に基づき信頼性を評価し，遺伝子型エラーが疑われる SNP を抽出した．評価では，家系の子供全個体のいずれか 1 個体のみが家系の遺伝法則に反する SNP を，その個体のエラーSNP とし評価した．

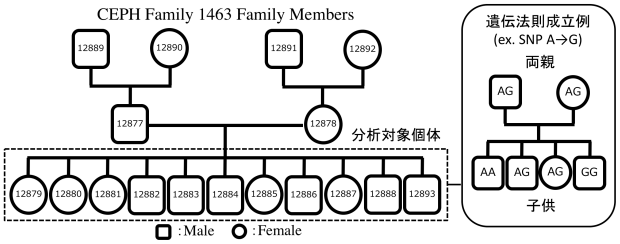

図 1 解析データと遺伝法則成立パターン

2.2. マッピング特性の抽出とエラー判別実験

検出された SNP 周辺のマッピング特性を得るため，図 2 のように分析対象個体のマッピング結果に対して，SNP 座位周辺の塩基配列を抽出した．抽出した配列に対して，SNP に影響を及ぼす可能性が考えられるマッピング特性を抽出した．本研究では，図 2 に示す f_1〜f_{11} のマッピング特性を検討し分析を試みた．次にエラーSNP と関係が深いマッピング特性を探るため，マッピング特性を特徴量とした Random Forest による重要度評価を試みた．また，マッピング特性が精度向上に寄与できるかを検証するため，Random Forest によるエラーSNP 判別実験を行い，エラーSNP の判別結果の評価を試みた．

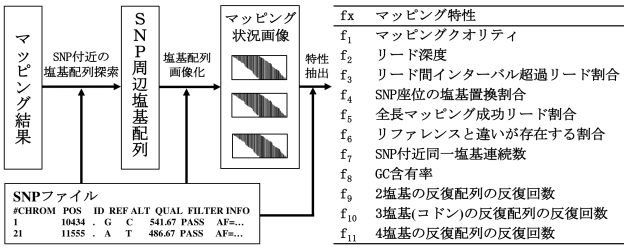

図 2 マッピング特性抽出フロー

Improvement of Accuracy for Single Nucleotide Polymorphism Analysis using Next Generation Sequencer considering Genome Mapping Characteristics
Yuto Ohzawa† Ginji Azuma† Atsushi Takahashi‡ Naoki Ohboshi†
†Kindai University ‡National Cerebral and Cardiovascular Center

3. 実験・考察

3.1. 家系 SNP のエラー評価結果

表 1 にて，分析対象データの解析で得られた全個体の遺伝法則に基づく SNP(True)および，エラーが疑われる SNP(False)の検出数に関して，1 個体あたりの平均と全子供個体の総数を示す．各個体において，BWA では検出数の約 5%，Bowtie2 では約 11%のエラーSNP の存在が確認できた．これらの SNP を判別実験にて用いた．

表 1　遺伝法則の信頼性別 SNP 分布

Children's SNPs	BWA		Bowtie2	
	True	False	True	False
Average	685.73	34.27	194.36	23.18
Total	7,543	377	2,138	255

3.2. マッピング特性による判別実験結果

図 3, 4 にて，Random Forest を用いた各ツールのエラーSNP に対するマッピング特性の重要度評価の結果を示す．重要度評価では，Random Forest を用いて，全子供個体データを学習データとしたモデルについてエラーSNP の予測可能性に関する各特性の相対的重要度を算出した．学習データは不均衡であるため，Over Sampling 手法の SMOTE を用いてエラーSNP 数を同数とした．パラメータは木の数を 500，ランダムサンプリングする特性の数を正の平方根に設定し学習を行った．結果として，各ツールにおいて SNP 座位の塩基置換割合が共に高い重要度を示した．また，BWA では Bowtie2 と比べ，他の特性についても，判別時の重要度が高いことが確認できた．

次にエラーSNP 判別実験を行った．全子供個体に関して，1 個体の SNP データをテストデータ，残りを学習データとしたモデルを構成し，全 11 個体分の判別実験を行った．表 2 に，各ツールの各子供個体に行った，Random Forest を用いたエラーSNP 判別実験結果の平均を，再現率，適合率，正答率の平均で示す．結果から，エラーSNP の約半数を判別でき，判別に成功したエラーSNP 数と同数程度のエラーではない SNP を誤判別したことが確認できる．しかし，SNP 総数に対するエラーではない SNP の割合が高いため，誤判別による正答率の大きな低下がなかったことが確認された．

以上の結果から，具体的なマッピング特性として SNP 座位の塩基配列置換割合が各ツールの解析結果のエラーに影響を与えることが確認できた．また，SNP 解析にてマッピング特性を考慮することにより，エラー判別に大きく寄与することないが，エラーではない SNP を大幅に減らすことなく，結果の偽陽性率をわずかに下げることが示唆された．

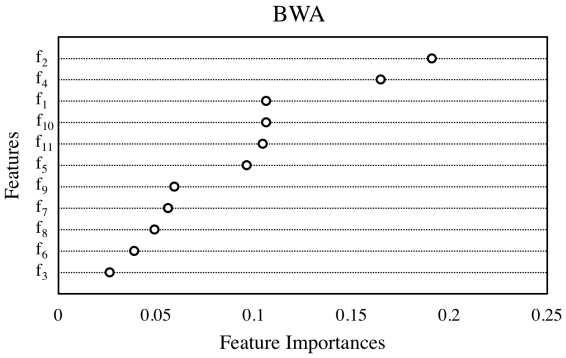

図 3 BWA におけるマッピング特性の重要度評価

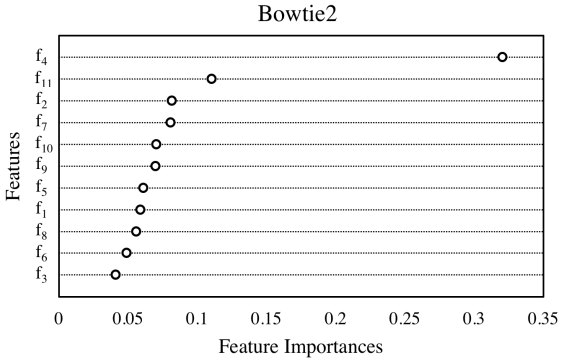

図 4 Bowtie2 におけるマッピング特性の重要度評価

表 2　Random Forest を用いた判別結果

Tool	Recall	Precision	Accuracy
BWA	47.27%	68.65%	96.62%
Bowtie2	57.48%	66.64%	92.83%

4. まとめ

本研究ではマッピング特性と誤検出との関連を分析し，SNP 座位の塩基配列置換割合が各ツールのエラーに影響を与えることを確認した．また，マッピング特性を考慮することにより，如何に精度向上に寄与するかを検証し，解析結果の偽陽性率に対する効果を確認した．

参考文献

[1] The European Bioinformatics Institute (EMBL-EBI). Retrieved August 23, 2016 from http://www.ebi.ac.uk/
[2] GATK Best Practices. Retrieved August 23, 2016 from https://software.broadinstitute.org/gatk/best-practices/.
[3] Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. vol.25. no.14. 1754-1760. Jul 2009.
[4] B.Langmead and S.L.Salzberg, Fast gapped-read alignment with Bowtie2, Nature Methods, Vol.9, No.4, 357-359, 2012.
[5] National Center for Biotechnology Information (NCBI) dbSNP Short Genetic Variations. Retrieved August 23, 2016 from https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi