

ヒトゲノムマッピングにおけるツールの違いが SNP 解析結果の信頼性に及ぼす影響

大沢 勇統[†] 東 銀史^{††} 村上 潤一郎^{††} 高橋 篤^{†††} 大星 直樹[†]
[†] 近畿大学大学院総合理工学研究科 ^{††} 近畿大学理工学部 ^{†††} 国立循環器病研究センター

1. はじめに

ヒトの全遺伝情報(ゲノム)を解析する際、同一データに対して解析を行っている場合でも、使用するツールが異なると解析結果に差異が生じる。より信頼性の高い結果を得るためには、この差異の起因やツール結果の信頼性を検証し、様々な解析指針を検討する必要がある。

本稿では信頼性検証の1つとして、マッピング処理にて使用するツールが異なる場合、SNP 解析結果の信頼性に影響を及ぼすのかを探るため、遺伝法則に基づき SNP の信頼性を評価した試みについて報告する。

2. SNP 解析

ヒトゲノムを構成する DNA の塩基配列中には、1つの塩基が別の塩基に置き換わっている SNP(Single Nucleotide Polymorphism:一塩基多型)と呼ばれる箇所が存在する(図 1)。個人の SNP 情報を解析することで、個人差に対応した医療の実現などが期待されている。

3. マッピング

ヒトゲノムの解析では、まず長大な DNA を適当な長さ に切断し解読する。マッピングとは、得られた DNA 断片の塩基配列情報を、ヒトの基準ゲノム配列に貼りつける形で解読元の配列を再構築する処理である。この処理に用いるツールが異なると、ツール間にてマッピング結果の一部に差異が生じ、解析結果に差異を生じさせる。

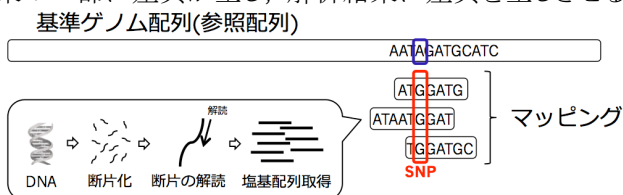


図1. マッピングのイメージ図

4. 評価方法

本稿ではマッピング処理にて広く用いられているツールである BWA[1], Bowtie2[2]を検証対象とした。同じ解析データを各ツールを用いてマッピングし、2つのマッピング結果に対して SNP 解析を行い結果の差異を比較した。解析データは EMBL-EBI[3]が公開する CEPH1463 家系(図 2)を使用し、家族集団全体の SNP を解析した。

本稿では、SNP 解析結果の信頼性評価を行う指標として、SNP 型の遺伝法則(図 2)を適用した。SNP の結果ファイルには 17 個体における SNP 型の情報が格納されており、各個体の SNP 型が家族構成に準じている SNP を高信頼性のものとして評価した。この高信頼性の

SNP の検出率を対象ツール別に調査した。

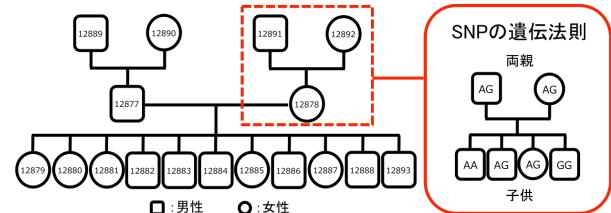


図 2. CEPH1463 家系の家族構成と SNP の遺伝法則パターン

5. 結果

各ツールの評価結果を図 3 に示す。BWA では 6054515 箇所、Bowtie2 では 5259807 箇所の SNP が検出された。BWA 使用時では Bowtie2 に比べ SNP 検出数が多く、遺伝法則が成り立っている SNP も多く検出できた。しかし、低信頼性の結果も Bowtie2 より多く検出された。一方、Bowtie2 使用時では BWA に比べ、SNP 検出数は少なかったが、Bowtie2 の SNP 検出数に対する高信頼性の SNP 存在割合が高い傾向が見られた。

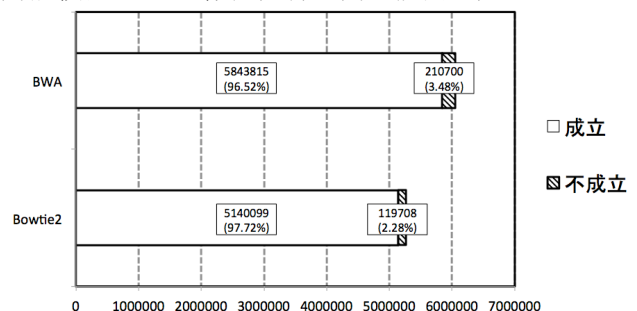


図 3. 遺伝法則に基づく信頼性評価結果

6. まとめ・今後の展開

本稿の結果から、マッピングツールが異なると、SNP 検出率や高信頼性の結果の存在割合に差異が生じることがわかった。このような各ツールの結果の傾向や差異を理解しておくことは解析結果の質の向上に繋がる。

今後は、より詳細な差異の傾向や起因を探るべく、本稿にて信頼性を評価した SNP が検出された座位周辺のマッピング状況を分析していく予定である。

参考文献

- [1] Li H and Durbin R, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, vol.25, no.14, pp.1754-1760, 2009.
- [2] B.Langmead and S.L.Salzberg, Fast gapped-read alignment with Bowtie2, *Nature Methods*, Vol.9, No.4, 357-359, 2012.
- [3] The European Bioinformatics Institute (EMBL-EBI) <http://www.ebi.ac.uk/> (Accessed 23 August 2016)