G-004

人工ゲノムデータを用いたゲノム解析におけるマッピングの精度検証
Accuracy evaluation of mapping on genome analysis using artificial genome data

東 銀史† 大沢 勇統† 高橋 篤‡ 大星 直樹†
Azuma Ginji Ohzawa Yuto Takahashi Atsushi Ohboshi Naoki

1. 背景

近年，次世代シーケンサーなどのゲノムを読み取る技術が発展し，ヒトをはじめとする様々な生物の全ゲノム塩基情報を取得することが可能になった．これにより，ゲノム解析研究が急速に発展し，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．ゲノム解析では，次世代シークエンサーによって読み取られたリードと呼ばれる短い塩基配列の断片が得られ，配列が決定されているリファレンス配列上における断片配列の位置を決定するマッピングと呼ばれる処理が行われる．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は解析された塩基配列が本来の塩基配列と異なる結果を生じ，後の遺伝子等の解析において，本来生成されるタンパク質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．そのため，高精度なマッピング処理が，後の遺伝子等の解析における精度向上につながる．マッピング処理を行うソフトウェアはマッピングツールと呼ばれ，現在，数種類のツールが存在する．本来マッピングを行う際，数種類ある中のマッピングツールのいずれを用いても，入力データが同一であれば処理結果も同一であることが望ましい．しかし，実際にはマッピング処理のアルゴリズムなどの違いにより，同一の入力データを用いても，ツール間でマッピング処理結果は一致しないことが知られている．

2. 目的

現在のマッピングツールによるマッピング処理結果が異なる結果を含む原因として，マッピング処理が不正確になる可能性の高い状況が存在するのではないかと考えた．そこで，本稿ではリファレンス配列を用いて，どこにマッピングされるべきであるのかを明確にした人工ゲノムデータを作成し，この人工ゲノムデータに対して，マッピングに広く用いられている Bowite2 および BWA によってマッピング処理を行う．マッピング処理を行った結果，不正確な場所にマッピングされた塩基配列を見つけ出し，それらの特徴を調べることで正確なマッピング処理が難しい状況が存在するか検討する．

3. 方法

3.1 マッピングツール

3.1.1 Bowtie2[1]

Bowtie2 は比較的短いリード向きのマッピングツールであり，BWT をベースとしたアルゴリズムが利用されている．Bowtie2 はプログラムが OS ごとに用意されている．本稿では Linux 用のものを使用し，バージョンは bowtie2-2.2.5 を使用した．

3.1.2 BWA （Burrow-Wheeler-Alignment Tool）[3]

BWA も Bowtie2 同様に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 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．バージョンは bwa-0.7.12 を使用した．

3.2 人工データの作成

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．これによって Paired-End read の情報量は Single-End read の 2 倍となり，より高精度な解析が可能となる．本稿では染色体 21 番について Paired-End read の人工データを作成し，マッピング処理を行う．リファレンス配列には human_g1k_v37_decoy.fasta（hs37d5）を使用し，人工データにはシークエンシング時のエラーや，個人差である変異を含まないものとする．次世代シークエンサーによるシークエンシングは，理論的にはゲノム全体を一様に読み取る．従って，人工データを作成するにあたってリファレンス配列から読み取る位置は，Mersennne Twister[4]を用いて一様にランダムに決定した．

図 1 リードの種類

3.2.1 リード長

本稿で作成する人工ゲノムデータのリードの長さは，EMBL-EBI[2]にて公開されている CEPH1463 家系データ同様に 101 塩基とした．

3.2.2 リード数

ゲノム上のある位置が読み取られた回数を表す DP（read Depth at this Position）という値がある（図 2）．

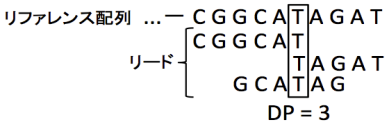

図 2 DP の例

CEPH1463 家系に含まれる NA12877 個体に対し，Bowtie2 によってマッピング処理を行うと DP は 47 付近であることが確認できた．この値を基準とし，人工ゲノムデータにおける DP が平均 47 となるようにリード数を

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

決定した．染色体 21 番の，配列は 35,908,807 塩基であり，リード長は 101 とするため，DP が平均 47 となるようにするには$(35{,}908{,}807 \div 101) \times 47 \cong 16{,}710{,}039$より，必要なリード数は 16,710,039 となる．本稿では Paired-End read を作成するため，片側および反対側から読み取るリード数は$16{,}710{,}039 \div 2 \cong 8{,}355{,}020$より，それぞれ 8,355,020 とする．

3.2.3 インサート長

片側から読み取った塩基配列と，反対側から読み取った塩基配列の距離をインサート長と呼ぶ（図 3）．

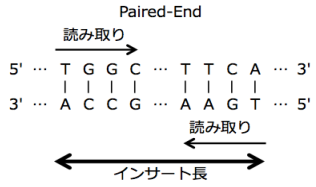

図 3　インサート長

CEPH1463 家系に含まれる NA12877 個体に対し，Bowtie2 によるマッピング処理では，インサート長の中央値は 316 であった．しかし，標準偏差が 30 万を超えていたため，明らかに解析エラーであると思われる結果を除去して再度集計を行うと，平均 318，中央値 316，標準偏差 68 となる結果が得られた．本稿ではこの値を用いて，インサート長は平均 318，標準偏差 68 の正規分布に従うと仮定した．

3.2.4 クオリティスコア

リードにはシークエンシング時のエラーが発生している確率を示すクオリティスコアが塩基ごとに付与されている．エラーの発生率を p とすると，クオリティスコア Q は以下の式で求められる．

$$Q = -10\log_{10} p$$

従って，クオリティスコアが高いほどエラー率が低いと言える．今回はリファレンス配列から直接取得したデータであり，エラーは含まないようにするため，クオリティスコアを 40（エラー生起確率 0.01%）で固定した．

3.3 評価方法

作成したリードは，それぞれが対応するリファレンス配列の位置が明確に判明している．このことを利用し，リファレンス上の元の位置へマッピングされたものを成功とし，それ以外の場所へマッピングされたものを失敗とした（図 4）．

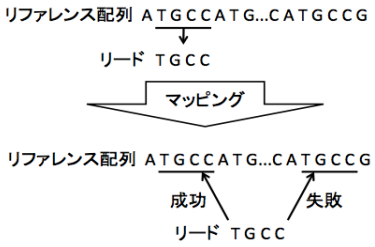

図 4　マッピング結果の評価

4. 結果・考察

マッピングに成功したリード数と，失敗したリード数は表 1 の通りである．

表 1　マッピング結果

	Bowtie2		BWA	
成功	16,529,351	(98.9%)	16,534,038	(98.9%)
失敗	180,689	(1.1%)	176,002	(1.1%)

マッピングに失敗したリードについて，ゲノム上の位置を 50 万区切りで描画したヒストグラムが図 5 である．

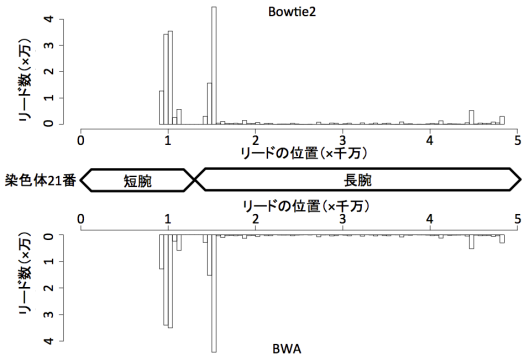

図 5　マッピングに失敗したリード

図 5 より，Bowtie2，BWA ともに染色体 21 番上の位置が 1,000 万および 1,500 万あたりのリードは間違いやすいことが確認できる．これは染色体の短腕と長腕の境界付近であり，これらのリードが間違ってマッピングされているマッピング先は別の染色体の短腕と長腕の境界付近に多く見られた．このことから，染色体の短腕と長腕の境界付近の塩基配列は染色体によらず似た構造を持ち，それがマッピングの失敗を引き起こす原因になっているのではないかと考えられる．そのため，これらの位置にマッピングされたリードに対し，何らかの修正を加えることでマッピングの精度を向上させられる可能性が示唆される．

5. 結論

染色体 21 番の人工ゲノムデータを作成し，Bowtie2 および BWA を用いてマッピングした結果，約 1%程度マッピングに失敗することを確認した．マッピングに失敗したリードは大部分が短腕と長腕の境界付近であったため，実際の解析においてこれらの結果が得られた場合は解析エラーの可能性が高いことを考慮する必要がある．今後はこれらのエラーとなりやすい領域に対して結果を修正するような手法について追求したい．

参考文献

[1] 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, vol.10, no.3, p.10:R25 (2009).
[2] EMBL-EBI. http://www.ebi.ac.uk/ena/data/view/PRJEB3381 (5 June 2017).
[3] Li H and Durbin R. “Fast and accurate short read alignment with Burrows-Wheeler transform”. Bioinformatiocs, vol.25, no.14, July pp.1754-1760 (2009).
[4] M.Matsumoto and T.Nishimura, “Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator”, ACM Trans. on Modeling and Computer Simulation, Vol.8, No.1, January pp.3-30 (1998).