Vol.2018-BIO-53 No.12
2018/3/10

# ビット列のソートに基づくリード直接比較による高速省メモリゲノム構造変異解析

新村 啓介[1]　加藤 有己[1,2,a)]　河原 行郎[1,2]

**概要**：体細胞のゲノム配列上に生じる構造変異はがんの発生や進行に関係することが知られている．そのため，同一個人のがん細胞と正常細胞のゲノムを採取し，次世代シークエンサーから得られるリード塩基配列をもとに変異を同定することが盛んに行われている．従来の変異検出法は，これらリード群の参照ゲノム配列へのマッピングに基づいているが，最初に実施するアラインメントの結果に強く依存してしまう欠点を持つ．本稿では，腫瘍細胞由来リードと正常細胞由来リードを直接比較することで，高精度にゲノム構造変異を推定する手法を提案する．人為的にヒトゲノム配列に変異を挿入しシークエンシングしたシミュレーションデータを用いた結果，提案手法の 1 塩基置換検出精度が高いことを確認した．

## 1. はじめに

ゲノム配列上に生じる 1 塩基置換（SNV）や挿入欠失（indel）などの構造変異（SV）の中には，がんなどの重篤な疾病を表すものがあり，これらを早期に発見することは医療において重要な課題である．そのための第一歩として，同一個人のがん細胞と正常細胞のゲノムを採取し，次世代シークエンサーでリード塩基配列を得た後，ゲノム上の変異箇所を調査することが考えられる．

これまでに広く使われている変異解析手法は，腫瘍細胞および正常細胞由来の全てのペアエンドのリードを参照ゲノム配列へマッピングした後，正しくマップされたリードと正しくマップされなかったリードの対応関係を取ることで，変異の位置および種類を特定する [1], [2], [6], [7]（図 1a）．ただし，この方法はマッピング手法に強く依存しており，マップされないリードに含まれる変異の検出が困難で，必ずしも感度が高いとは言えない．

一方，腫瘍細胞と正常細胞からのリード群を相互に直接比較することで，感度を向上させた手法が提案されている [4]（図 1b）．この手法はリード配列よりも長い規模の大きな構造変異を同定できる代わりに，大量のメモリを確保したコンピュータークラスターで多大な時間をかけなければ計算できず，汎用性が高いとは言えない．したがって，

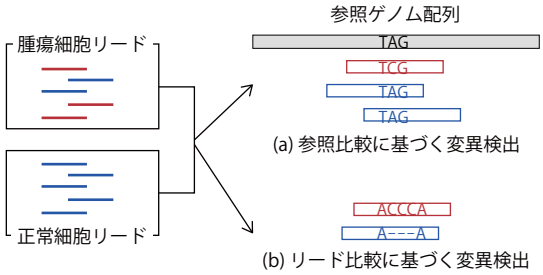


図 1　ゲノム構造変異検出法

変異検出感度が高く，デスクトップ PC などで簡便に使用できる省メモリの新たな変異同定法が必要である．

本稿では，リードを直接比較する方針は上記と同一だが，リードのビット列に基づくデータ構造を巧妙に利用することで，同一個人のゲノム構造変異を高精度かつ少ないメモリ使用量で検出できるツール Bivartect を提案する．

## 2. 手法

### 2.1 概要

まず，次世代シークエンサーの出力ファイル（FASTQ 形式）からリードを逐次読み出し，接尾辞配列としてメモリ上に展開する．この際，塩基情報を文字列ではなくビット列に変換することで，情報の圧縮によるメモリ使用量の削減のみならず，解析各工程での高速化を達成可能である．さらに，データ構造を工夫して，配列データを可変長ではなく固定長とすることで，一層の高速化を図る．

続いて，腫瘍細胞，正常細胞からの全てのリードを読むことで得られた接尾辞配列群を辞書順序でソートする．ビット列で表される接尾辞配列は単純な数値として扱われ

[1] 大阪大学 医学部
Faculty of Medicine, Osaka University, 2–2 Yamadaoka, Suita, Osaka 565–0871, Japan
[2] 大阪大学 大学院医学系研究科
Graduate School of Medicine, Osaka University, 2–2 Yamadaoka, Suita, Osaka 565–0871, Japan
[a)] ykato@rna.med.osaka-u.ac.jp

表 1　Bivartect による変異検出結果の精度評価

| SNV | | | SV | | |
|---|---|---|---|---|---|
| SEN | PPV | F | SEN | PPV | F |
| 0.859 | 0.902 | 0.880 | 0.643 | 0.701 | 0.671 |

るため，文字列に比し，非常に高速なソートがなされる．

ここで，ソートされた接尾辞配列群の接頭辞に着目すると，ある一定の読み枠で見たときに，同じ接頭辞を持つものが複数存在する．これらの配列はソートによって連続して整列しているので，容易にグルーピングが可能である．グルーピングの後，グループに含まれる接尾辞配列を並行して先頭から 1 塩基ずつ読み進め，接尾辞同士で塩基の比較を行う．多くの場合はいずれの接尾辞も同じ塩基を指しているが，稀に異なる塩基を指している場合があり，この塩基を breakpoint と推定する．

この手法で breakpoint を推定し続けると，全く同じものを何度も報告してしまう．そこで，最後に breakpoint 固有の塩基配列を生成し比較することで，発見した breakpoint の統合を行う．

### 2.2　ベンチマークデータ

ヒトゲノム（GRCh37）の 21 番染色体配列に対し，1000 ゲノムデータベース [5] からランダムに選んだヒトの 21 番染色体の変異データをダウンサンプリングし，5,160 個の SNV および 840 個の indel を反映させることで，人工的に変異型の染色体配列を作成した．ここで，indel の長さは 30 nt 以下までのものをランダムに選んだ．次に，次世代シークエンサーによるリードの生成をシミュレート可能な ART [3] を利用して，正常および変異の入ったリード群を作成した．ここで，読み深度を 30，リードの長さを 50bp とし，ペアエンドリードを生成した．

## 3.　結果

Bivartect の変異検出精度を評価するため，正解の変異集合に対してアルゴリズムがどの程度それらを検出できたかを示す sensitivity (SEN) と，アルゴリズムの予測集合の中でどれだけ正解の変異を含むかを示す positive predictive value (PPV) を用いた．さらに，SEN と PPV の調和平均で定義される F スコアを用いた．表 1 に結果を示すが，特に SNV に関しては高い変異検出精度を達成していることが見て取れる．

## 4.　おわりに

本稿ではゲノム構造変異解析ツール Bivartect を提案し，シミュレーションデータ上で高い 1 塩基置換検出精度が達成されることを示した．今後，実際のゲノム編集データやがんゲノムデータを用いて，多角的に提案手法の評価を行う予定である．

**謝辞**　本研究では，情報・システム研究機構国立遺伝学研究所が有する遺伝研スーパーコンピュータシステムを利用した．

### 参考文献

[1] Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).

[2] Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

[3] Huang, W., Li, L., Myers, J.R. & Marth, G.T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28,** 593–594 (2012).

[4] Moncunill, V., Gonzalez, S., Beà, S., Andrieux, L.O., Salaverria, I. *et al.* Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. Biotechnol.* **32,** 1106–1112 (2014).

[5] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

[6] Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Heatley, S.L. *et al.* CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8,** 652–654 (2011).

[7] Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).