Vol.2012-BIO-28 No.8
2012/3/28

RECOT: 次世代シーケンサの比較ゲノムに向けたゲノム座標の変換

伊澤　亜紀子[†1]　瀬々　潤[†2]

次世代シーケンサの台頭により多様な種のゲノムが読まれ，それらを利用したRNA-seqやChIP-seqなどの実験が行われている．この実験結果を種間比較するには，次世代シーケンサ用のブラウザは複数のゲノムが扱えず，比較ゲノムブラウザはリードを気軽に追加できない問題点がある．本研究ではこれらの問題点を解消するため，リードの位置や変異情報を異種のゲノム情報に対応させ，次世代シーケンサ用ブラウザで可視化可能になるよう変換するRECOTを開発した．

RECOT:A tool for the genome coordinate transformation of next-generation sequencing reads between related species

Akiko Izawa[†1] and Jun Sese[†2]

The whole-genome sequences of many non-model organisms have recently been observed. Based on the genome sequences, experiments using high-throughput sequencers such as RNA-seq and ChIP-seq have been performed, and the experiments have been compared between related species. Although these comparisons require changes of the genome coordinates of the reads between the species, current software tools are not suitable.In this paper, we introduce a set of programs, called REad COordinate Transformer (RECOT), which transform the coordinates of short reads from the genome of the originally observed species to that of another species after aligning the gene/genome sequences between the species.

†1 お茶の水女子大学
Ochanomizu University
†2 東京工業大学
Tokyo Institute of Technology

1. はじめに

DNAを読む技術が急速に発展し，新型のシーケンサ (以下NGS) が生まれたことで，多様な生物のゲノム配列が解読されている．これにより，今までゲノム既知の種限定で行われてきたRNA-seq[1)]やChIP-seq[2)]などのNGSを使った実験が多様な種で可能になった．これらの実験を比較することで，進化プロセスや生物メカニズムについての解明が期待できる．

一般にNGSを用いた実験では，ゲノムに配列断片 (以下リード) をアラインメントし，対応づける．近縁種同士の比較には，種間でリードをアラインメントし，対応づける必要がある．しかし現在，大量のリードを他種のゲノム配列に対応づけることに適したソフトウエアがない．そして，複数種のNGSによる実験結果を同時に可視化し，種間で比較することは，現在公開されているソフトウエアでは容易ではない．このような問題点に対し本手法では，2種で読まれたNGSのリードのアライメント結果を同時に表示し，比較可能にするためのツール群REad COordinate Transformer(RECOT)を開発した．また，RECOTはNGSの計算結果で最も一般的な形式であるSAM形式を出力する．これにより，ユーザが使い慣れた可視化ソフトを選ぶことが可能である．

2. 提案手法

本研究では，リードが読まれた種をoriginal speciesと呼び，リード位置の変換対象種をtarget speciesと呼ぶ．手法の概観を図1に示す．RECOTは4つの入力ファイルを入力とする．(1)NGSから読まれたoriginal speciesのリード (2)original speciesのゲノム配列 (3)遺伝子情報 (4)target speciesのゲノム配列である．出力は (1) の配列の (4) 上でのアラインメント結果である．

本手法では，2つのアラインメント結果が必要になる．1つは (1) と (2) のアラインメント結果．そして，(2) の遺伝子配列と制御領域の遺伝子周辺領域の配列を，(4) にアラインメントした結果である．前者はBWAやBowtie，後者はGMAPなどが利用可能である．遺伝子周辺領域の配列は提供されないことも多いので，ゲノム配列と遺伝子領域から，遺伝子周辺領域の配列を得るスクリプトも作成した．図1は，これらのアラインメントした例である．SAM形式は塩基の対応をCIGARで表す．図1(B)では3M3N12Mとなる．Mはマッチ又はミスマッチ (以下マッチ) を示し，Dは欠失，Nはギャップを指す．

しかし，相同性のある遺伝子により，1つのtarget speciesの領域に複数の遺伝子がアラ

インメントされ，遺伝子間の対応付けで混乱を招くかもしれない．このような問題を避けるため，同じ領域に複数の遺伝子がアラインメントされた場合，その複数の遺伝子から 1 つを選択するスクリプトを作成した．このスクリプトでは以下の 2 つの選択方法が用意されている．(1) ユーザ指定の遺伝子の対応表により，対応に優先順位をつける．(2) アラインメントスコアが最も高い遺伝子配列を優先する．

最後に 2 つのアラインメント結果から，リードを target species にアラインメントしていく．今までに計算した 2 つのアラインメント (1)-(2) と (2)-(3) の結果をもとに，(1)-(2) 間の CIGAR を (2)-(3) 間の CIGAR に対応づける．そのために (1) target species のゲノム配列にアラインメントされた original species の遺伝子と，その遺伝子の領域にアラインメントされたリードの対応関係を表す表を作成，(2) この対応表を使用し，リードの CIGAR の書き換えをおこなう．図 1(A) は，original species の遺伝子配列とリードの，図 1(B) は original species の遺伝子配列と target species のゲノム配列のアラインメントを示す．また，図 1(A)(B) の遺伝子配列は同一のものである．

3. 実 行 結 果

NGS の実データに本手法を適用し，有用性を検証するため，*D. simulans*(ERR020078) と

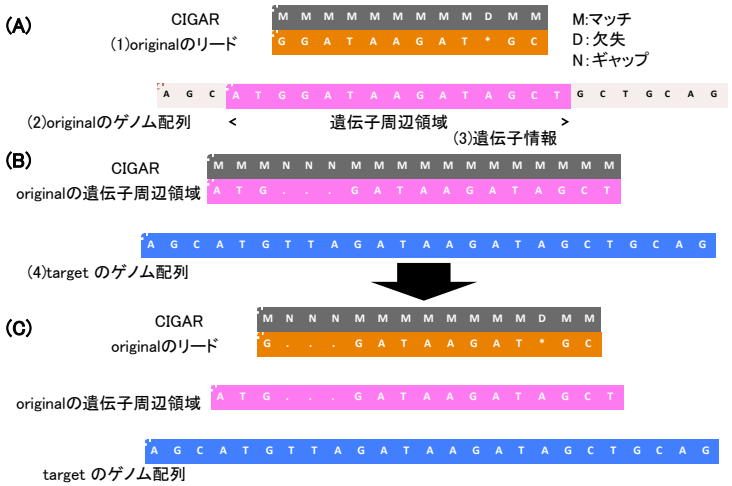

図 1　提案手法の概観
Fig. 1　Overview of Proposed Method

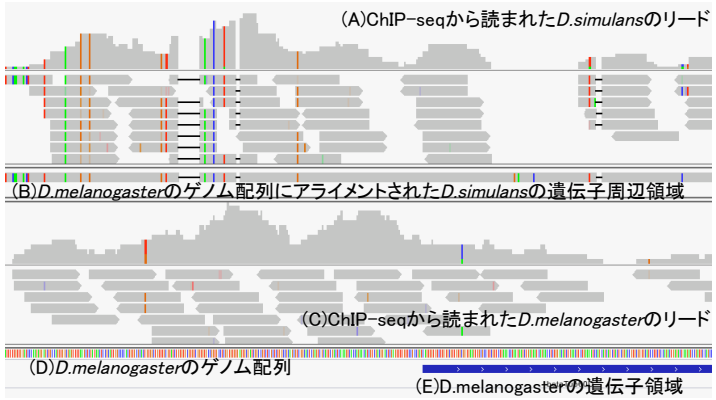

図 2　実データによる実行結果
Fig. 2　Result using a Real Dataset

キイロショウジョウバエ (*D. melanogaster*(ERR020066)) の ChIP-seq の比較をおこなった．これらは約 500 万年前に分岐した近縁種である．結果の可視化にはゲノムビューア Integrative Genomics Viewer(IGV)[3] を使用した．図 2(A) は，*D. simulans* のリードを RECOT により，*D. melanogaster* にアラインメントした結果である．図 2(B) は，*D. melanogaster* の遺伝子周辺領域を D. *melanogaster* のゲノムにアラインメントした結果である．図 2(C) は，*D. melanogaster* のリードを，*D. melanogaster* のゲノムにアラインメントした結果である．図 2(D) は *D. melanogaster* のゲノム，図 2(E) はその遺伝子である．2 種のヒストグラムを比較すると，発現制御部位の変化を読み取ることができる．このように本手法は種間を比較するために有用である．

参 考 文 献

1) Wang Z. *et al.*: RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, Vol.10, pp.57–63 (2009).
2) Jothi *et al.*: Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucl Acids Res*, Vol.36, pp.5221–5231 (2008).
3) Robinson *et al.*: Integrative Genomics Viewer. *Nature Biotechnology*, Vol.29, pp. 24–26 (2011).