

(別紙1)

総 括 研 究 報 告 書

課題番号：2019C-15

課 題 名：エクソームデータを用いた大規模レトロトランスポジション検出法の開発

主任研究者 (所属施設) 国立成育医療研究センター
(所属・職名 氏名) メディカルゲノムセンター 青砥早希

ヒトゲノム中には機能をもつ遺伝子と相同性の高い偽遺伝子領域が約2万箇所存在する。このような配列類似の高い非機能領域はショートリードシーケンシングの特性からミスマッピングを誘発しやすい。特に親遺伝子と配列類似度の高い偽遺伝子の一つであるレトロコピー配列のリードが親遺伝子にミスマッピングされた場合、レトロコピーの多型は親遺伝子の多型として検出されてしまい、さらにその頻度は理論上33%となる。これは生殖細胞変異の多型頻度としてもおかしくない値であり、体細胞変異やモザイク変異と区別するのは困難である。そこで本研究では、レトロコピー検出方法を確認し、クリニカルシーケンスを目的として変異検出を行う際に既知遺伝子と混同されうるレトロコピー配列の同定を行った。

1. 研究目的

レトロコピーは、LINE1等のレトロトランスポゾンの逆転写活性を利用してプロセシングを受けた mRNA がゲノム中に再挿入(レトロトランスポジション)することで生じる配列である。分子進化においてはパラガスな新規遺伝子獲得機構の一つとしても知られるが、そのほとんどは多型が蓄積し遺伝子としては壊れており機能を持たない。そのため、人類集団に対して頻度情報などを網羅的に調べた例は非常に少ない。

本申請では全ゲノム配列データに比べ、容易に入手できるエクソームデータを用いて効率的にレトロコピーの検出をおこなう手法を確認することをめざした。エクソームシーケンスはそのシーケンシングデータが国内外で既に蓄積されており、そのデータが入手しやすいという側面だけではなく、4万円以下と比較的安価でありながら遺伝子配列上の多型のみを同定するというクリニカルシーケンス、つまり診断の面で非常に効率的な側面を有する。

一方でエクソーム解析に主に利用されるショートリードシーケンサーやその解析ソフトウェア(主にマッピングソフトウェア)は非常に似た配列を区別するのは得意ではない。短い塩基配列を個人が有する多型を許容して国際参照配列に貼り付けるようにしてゲノム配列を組み上げていくため、レトロコピーが相同性の高い親遺伝子にミス

マッピングされうる。

レトロコピー配列のリードが親遺伝子にミスマッピングされた場合、レトロコピー由来の多型は親遺伝子の多型として検出されてしまい、さらにその頻度は理論上33%となる。これは生殖細胞変異の多型頻度としてもおかしくない値であり、ましてや体細胞変異やモザイク変異と区別するのは困難である。そこで本研究では、レトロコピー検出方法を確認し、エクソームシーケンシングにおいて既知遺伝子と混同されうるレトロコピー配列の同定を行い、特に日本人集団で考慮すべきレトロコピー配列の同定を目指した。

2. 研究組織

研究者	所属施設
青砥早希	国立成育医療研究センター
緒方広子	国立成育医療研究センター

3. 研究成果

本年度の研究では、以下にの項目について実施し、その結果、本研究成果を利用して得た成果を文献として発表することができた。

(1) エクソームデータを用いたレトロコピー配列の検出方法の開発

レトロコピー配列検出プログラムの作成はペアエンドシーケンスによるエクソームデータを対象として行った。レトロコピー

はイントロンを欠く構造を有するため、そのリードはコーディング領域にしかマッピングされない。この特性を利用し、隣り合うエクソンを2つ以上含むキメラリードや、ペアとなるリードどうしの間隔が 1000 bp を超えるテンプレートリードを形成しているリードをマッピングデータから抽出することでイントロン配列の欠如が疑われるリードをレトロコピー候補として抽出した。それに加え、リードのクオリティやコピー数を手掛かりとしてレトロコピー配列を特定する方法を確立し、解析ツールとして作成・パイプライン化を行った。本パイプラインは、研究所内クラスターサーバにおいて利用可能な状態にしている。

(2) レトロコピー検出ツールの精度設定

本手法で実施するレトロコピー検出法の検出精度は、平均リード深度に依存する傾向があることが予備解析で明らかになっている。そこでバリデーションデータとして平均リード深度が75リード以上ある正常経産婦のエクソームデータ8検体を利用する。シーケンス深度の高いデータを用い、データ削減を行う事で様々なリードデプスのデータを仮想的に作成し、様々なリードデプスに対する本ツールの精度との関係性を算出した。これにより実用的なデプスで検出できるパラメータの値を取得した。本センターでシーケンスしたデータからは1検体あたり平均20個のレトロコピー候補が検出できることを確認した。

(3) 日本人に特徴的なレトロコピーの導出

特定の遺伝的背景を持つ集団に特異的なレトロコピーを検出するために1000 Genomes Projectから日本人、漢民族、白色人種、ヨルバ人（アフリカ系集団）の4つの遺伝的背景を持つ集団のエクソームデータを取得し、本申請で作成したツールを用いてレトロコピーの検出を行った。

その結果、既報告の通りヨルバ人でレトロコピー多型が多いことが本パイプラインの実施においても確認された。また、白色人種は大きく2パターンに分かれ、日本人集団特有の配列として3つの遺伝子のレトロコピーを特定した。引き続き、データ量を増やしてこれらの成果の検証に当たる予定である。

成果報告等

1, 関連学会での成果報告

本研究成果は第42回日本分子生物学会年会において以下の題目で発表した。

「色素性乾皮症患者由来 ERCC2 欠損細胞における内在性遺伝子のレトロトランスポジション」

2, 査読付雑誌

Aoto S, Katagiri S, Wang Y, Pagnamenta AT, Sakamoto-Abutani R, Toyoda M, Umezawa A, and Okamura K. “Frequent retrotransposition of endogenous genes in ERCC2-deficient cells derived from a patient with xeroderma pigmentosum”

Stem Cell Research and Therapy.

2019;10:273.

4. 研究内容の倫理面への配慮

本提案では文献やデータベースなどで公開されているヒトエクソームデータを用いて遺伝子進化及びゲノム不安定性の解明を進めた。つまり、すでにデータの利用と公開の同意を取得し、匿名化された既存のゲノムデータを二次的に利用し、新たな知見を得ようとするものであったため、データを利用するに際し新規の研究対象者に対する人権擁護上の配慮は不要であった。公開情報であるゲノムデータの再利用についても当該ウェブサイトの方針に従って適切に行った。