東北大学 TOHOKU UNIVERSITY

Press Release

2017 年 6 月 6 日

国立大学法人東北大学 東北メディカル・メガバンク機構
国立研究開発法人日本医療研究開発機構

日本人基準ゲノム配列、精度が向上した新版（JRGv2）を公開

一分子長鎖型シークエンサーを用いた複数名の高深度ゲノム情報を元に 日本人に特徴的なゲノム情報を 10 倍に拡充

【発表のポイント】

○ 一分子長鎖型シークエンサー*1 を用いて、3 名の日本人の全ゲノム解読を完了。従来の技術では精確な解読が困難だった約 9,600 箇所の挿入配列の解読に成功。

○ 国際ヒトゲノム参照配列に約 620 万塩基を拡充した、日本人の基準ゲノム配列（JRG）の新バージョンを公開。

○ 日本人の基準ゲノムとして現在多くの研究者が利用可能な唯一の配列であり、これを継続して改訂することで日本人ゲノム解析のクリニカルシークエンスなどでの精度向上、疾患関連遺伝的多様体の高精度検出・同定が期待できる。

【概要】

東北大学東北メディカル・メガバンク機構（以下、ToMMo）は、コホート調査*2 の参加者から提供された DNA をもとに、一分子長鎖型シークエンサー PacBio RS II（Pacific Biosciences 社製：以下、PacBio）を用いて日本人の基準ゲノム*3 の構築を進めています。2016 年 6 月に日本人基準ゲノム配列 JRGv1（Japanese Reference Genome version 1）を公開していましたが、今回、新たに 2 名分を加え、合計 3 名の全ゲノムを高精度に解読しました。本シークエンス解析の結果、国際ヒトゲノム参照配列*4 に対して、前回の公開分をあわせて、日本人が保有しこれまで報告されてこなかった約 9,600 箇所の新たな挿入配列、約 620 万塩基の同定に成功しました。この解読完了から、ToMMo では、新たに同定された配列群を挿入するなどして得られた日本人の基準ゲノム配列 JRGv2 （Japanese Reference Genome version 2）を作成、公開することとしました。今回の日本人に特徴的なゲノム構造の拡充は、前回の約 10 倍の規模になります。

また、あわせてデコイ配列*5 decoyJRGv2 を公開することとしました。デコイ配列とは、国際ヒトゲノム参照配列には含まれていない領域をまとめたもので、

decoyJRGv2 は日本人に高頻度でありながら国際ヒトゲノム参照配列には含まれていない配列で、ゲノム解読時の重要な行程であるアライメント*6の精度の向上に活用されます。

両配列の公開は、国際ヒトゲノム参照配列だけを用いている際には精確に読みとることのできなかった領域の研究に大きく寄与し、クリニカルシークエンスを含む、日本のゲノム研究全体を底上げ、加速させるものと期待されます。また今回の成果は、日本人に特徴的なゲノム構造を拡充する成果であり、今後、日本の医学研究の大きな基盤となる成果と考えられます。

【背景】

東北大学と岩手医科大学は、東日本大震災の被害からの復興事業として、2012年から東北メディカル・メガバンク計画に取り組み、東北大学は東北メディカル・メガバンク機構（以下、ToMMo）を、岩手医科大学はいわて東北メディカル・メガバンク機構をそれぞれ設立して事業を進めています。両機構は、宮城・岩手両県の住民15万人に対し健康調査（コホート調査）を実施しており、2017年5月現在で、約15万人のベースライン調査を完了し現在追跡調査を行っています。

また、日本人の個別化医療、個別化予防の実現に向け、ToMMo は、コホート参加者から提供された血液から DNA を抽出し、短鎖型次世代シークエンサーである HiSeq 2500（Illumina 社製）を用いて 1,070 人の全ゲノム解析を行いました。その結果、約 2,120 万個の一塩基多様体（SNV）*7を含む日本人の全ゲノムリファレンスパネル*8（1KJPN）の構築を実現しました。この成果は、2014年 8 月 29 日に当機構のポータルサイトの 1 つである iJGVD*9から先行して公開され、2016 年 4 月 1 日時点ですでに世界 100 カ国、1 万人以上の方から利用されています。成果は論文化され*10、さらに、2015 年 12 月 15 日には、SNV 約 2,120 万個すべての情報の公開を行いました。現在、2,049 人の全ゲノム解析をもとにした日本人の全ゲノムリファレンスパネル（2KJPN）に拡充し、得られた SNV 約 2,800 万個すべての情報の公開を行っています*11。

しかし、短鎖型次世代シークエンサーを用いた解析では、解読できる塩基の長さが数百塩基ずつしかないため、日本人のゲノム配列が持つ数千塩基以上の構造多型を詳細に解明することは困難でした。日本人を対象にしたゲノム医療研究を進めるにあたっては、日本人が独自に有するこれらの構造多型を明らかにしたゲノム配列情報が求められていました。

【日本人の基準ゲノム配列の意義】

ヒトゲノムは約 30 億塩基対から構成されています。現在は、ヒトゲノムを高速かつ安価に解析するために、短鎖型次世代シークエンサーが世界中で数多く使用されています。この機器は、ヒトゲノムを非常に短い長さ（数百塩基程度）

に細断し、それらを同時並行で解析するものですが、バラバラになったヒトゲノムを元の長さに効率よく復元するためには、「お手本」が必要となります。「お手本」を参照して、シークエンス解析した数百塩基程度の配列がどこにあてはまるのか並べていく（アライメントする）ことで、長いゲノム配列の解読が完了します。

その「お手本」として、国際的組織であるゲノムリファレンスコンソーシアム（Genome Reference Consortium）が管理する国際ヒトゲノム参照配列（以下、国際参照配列）が広く用いられています。この配列は、日本人の配列情報も一部用いて作成されていますが、欧米人の配列情報を主としており、日本人によく見られる構造多型の情報は含まれていないのではないかと考えられていました。そのことが短鎖型次世代シークエンサーを用いて日本人のゲノム配列を解読しアライメントする際の精度に影響するものと考えられていました。

そのため、東北メディカル・メガバンク計画においては、日本人ゲノム配列の高精度化を図るため、参照ゲノムについても日本人独自のもの（日本人の基準ゲノム配列）を作成することを進めています。

日本人の基準ゲノム配列が完成すれば、短鎖型次世代シークエンサーを用いたゲノム解読のアライメント精度の向上に活用されるとともに、国際参照配列だけを用いても精確に解読することのできなかった領域の解明に貢献します。それにより、日本人のゲノム情報に基づく疾患解析を行っている研究者が、それぞれの疾患の原因変異をより正確に同定することが可能となると考えています。

このため、日本人の基準ゲノム配列の公開および継続した拡充は、我が国のゲノム研究全体を底上げ、加速させるものと期待されます。また、今回の成果は日本人のゲノム配列上の特徴を明らかにする成果であり、今後、日本の医学研究の大きな基盤になると考えられます。

【事業の概要】

日本人の基準ゲノム配列構築のため、ToMMo は、コホート調査の参加者から提供された DNA をもとに、一分子長鎖型シークエンサーPacBio を用いて、全ゲノムを解読しました。

この機器は、平均で一度に 1 万塩基以上のゲノム配列情報を連続して読むことが可能です。しかし、この機器は読み取りエラーの割合が高いことが知られており、これまで、日本人ゲノムの大規模解析に用いられたことは、ほとんどありませんでした。ToMMo では、この問題を克服するため、日本人のゲノム DNA から平均 1.2 万塩基長という長い DNA ライブラリーを作成し、一気に読み取ることで大量の配列情報を取得することと、高精度の塩基配列の新規再構成（デノボアセンブル[*12]）という情報科学的手法を用いることにしました。

そのため、一分子長鎖型シークエンサーを用いて日本人の DNA を繰り返し

読み取り、一人あたり合計約 3,000 億塩基分（ヒトゲノム全体を 100 回以上繰り返して全て読んだ場合に相当）の配列情報を得ました。また、ToMMo のスーパーコンピュータシステム[*13] を数ヶ月間にわたって利用して情報解析を行いました。

昨年 6 月に日本人基準ゲノム配列 JRGv1 を公開していましたが、この度、新たに 2 名（合計 3 名）のシークエンスを行うことで、新たに同定された配列の一部を国際参照配列上に配した日本人の基準ゲノム配列 JRGv2、およびデコイ配列 decoyJRGv2 を構築し、両配列を公開することとしました。

デコイ配列とは、国際参照配列には含まれていない領域をまとめたもので、decoyJRGv2 は日本人に高頻度でありながら、国際参照配列には含まれていない配列ということになります。デコイ配列は主に短鎖型次世代シークエンサーを用いたゲノム解読時の精度の向上に活用されます。

【解読結果の概要】

ゲノムリファレンスコンソーシアムは、国際ヒトゲノム参照配列を管理するとともに、定期的に改定しています。2016 年 4 月時点の最新の国際参照配列は、2013 年 12 月にリリースされた GRCh38 です。GRCh38 と今回のデノボアセンブルの結果とを網羅的にスーパーコンピュータ上で比較することで、最終的に国際参照配列には収載されていない、約 9,600 箇所の新たな挿入配列（総塩基として約 620 万塩基分）の同定に成功しました。

今回の拡充は、JRGv2 においては、JRGv1 と比べて約 510 万塩基の拡充となります。また、decoyJRGv2 においては、decoyJRGv1 とくらべて、約 360 万塩基の日本人に特徴的な配列を拡充できたことになります。（表）

これまで検出できなかった挿入配列を同定可能にした要因は、PacBio が一度に 1 万塩基以上を連続して読むことが可能な一分子長鎖型シークエンサーであること、さらには、PacBio では DNA をクローニングや PCR 増幅することなく一分子ごとに直接観察する手法が用いられていることによります。（これまでの手法では、DNA を複製させる必要がありました）

【公開予定】

日本人基準ゲノム JRGv2 は、近日中に下記のサイトで公開する予定です。なお、今後の研究の進展により、より精細な配列情報が得られた場合は、バージョンアップする予定です。

日本人基準ゲノム公開 URL（図）http://jrg.megabank.tohoku.ac.jp/

1. 日本人基準ゲノム JRGv2　約 30 億塩基（国際参照配列に約 9,600 箇所の挿入配列、約 620 万の新規塩基を含む）

2. 日本人基準ゲノム用デコイ配列 decoyJRGv2　約 620 万塩基

【参考】

＜東北メディカル・メガバンク計画について＞

本計画は、東日本大震災を受け、被災地住民の健康不安の解消に貢献するとともに、個別化予防等の東北発の次世代医療を実現するため、被災地域の復興を推進する、国の復興事業として行われているものです。2015 年度より、国立研究開発法人 日本医療研究開発機構が本計画の研究支援担当機関の役割を果たしています。

被災地に医療関係人材を派遣して地域医療の復興に貢献するとともに、15 万人規模の地域住民コホートと三世代コホートを形成し、そこで得られる生体試料、健康情報、診療情報等を収集してバイオバンクを構築します。さらに、ゲノム情報、オミックス情報、診療情報等を解析することで、個別化医療等の次世代医療に結びつく成果を創出することを目指しています。また、得られた生体試料や解析成果を同意の内容等に十分留意し、個人情報保護のための匿名化等の適切な措置を施した上で、外部に提供することや、コホート調査や解析研究を行うための多様な人材の育成も行っています。

本計画の事業の実施は、東北大学東北メディカル・メガバンク機構と岩手医科大学いわて東北メディカル・メガバンク機構とが連携して行っています。

【用語解説】

＊1. **一分子長鎖型シークエンサー**：主に2000年代半ば以降に登場した、DNA配列を超並列に読み取る(シークエンス)機器は、次世代シークエンサーと一般に呼ばれている。主に、ランダムに切断されたDNA断片の塩基配列を1塩基ずつ決定する解析過程を、数百万から数十億ものDNA断片に対して同時並列的に処理することが可能。この技術はDNAを増幅する必要があり、読み取り長は最大でも数百塩基である。一方、DNAを増幅することなく、一分子のまま1万塩基以上読みとることのできるシークエンサーが登場し、一分子長鎖型シークエンサーと呼ばれている。増幅しないため、ゲノムをより均一に読みとることが可能である。

＊2. **コホート調査**：ある特定の人々の集団を一定期間にわたって追跡し、生活習慣などの環境要因・遺伝的要因などと疾病の関係を解明するための調査のこと。

＊3. **日本人基準ゲノム**：より日本人のゲノム情報を反映した参照配列。現在は、一般的に国際ヒトゲノム参照配列が用いられている。東北メディカル・メガバンク機構において、2016年4月23日に日本人基準ゲノム JRGv1（Japanese Reference Genome version 1）の構築のニュースリリースを行った。

参考：プレスリリース「日本人の基準ゲノム配列(JRG)を決定―長鎖読みとり型次世代シークエンサーを用いて日本人のもつゲノム構造を解明―」
URL: http://www.megabank.tohoku.ac.jp/news/16174

＊4. **国際ヒトゲノム参照配列**：国際的な学術組織Genome Reference Consortiumが継続的に改訂を行っているヒトゲノムの全染色体の塩基配列。同配列は主に欧米の複数のヒトゲノムを読むことで構築されている。事実上、ヒトゲノムのデファクトスタンダードの塩基配列として全世界のヒトゲノム研究に利用されている。2016年4月現在、最もよく使われている最新の国際ヒトゲノム参照配列はGRCh38である。

＊5. **デコイ配列**：繰り返し配列などの、短鎖型次世代シークエンサーでの難読領域等を仮想配列として統合した配列。難読配列は、短鎖型次世代シークエンサーによって読みとりはされるが、その読みとり結果を、国際ヒトゲノム参照配列等に照らして並べ（アライメントし）ようとすると、適合しなかったり、特定箇所に過度に集中するなどして、うまくアライメントすることができない。そうした領域を、（デコイ＝おとりのようにして）人為的に集める仮想配列をつくると、短鎖型次世代シークエンサーの結果の解析に対して有用である。

＊6. **アライメント**：ゲノム解読において、特に短鎖型のシークエンサーにおいてシークエンス解析された配列を、お手本となる参照配列のどこ由来の配列かを探索し並べること。

＊7. **一塩基多様体（SNV）**：個人間でゲノムの一塩基が異なる状態。なお、通常は一定以上の頻度(通常1%)で確認されるSNVを特に一塩基多型(Single Nucleotide Polymorphism)SNPと呼ぶ。

＊8. **全ゲノムリファレンスパネル**：大規模な人数の全ゲノム解読を行った結果を総合し、DNA 配列の多型の頻度などの情報をまとめることで、今後のゲノム研究の参照情報となるよう、東北大学東北メディカル・メガバンク機構が構築を進めている全ゲノムリファレンスパネルのこと。

＊9. **iJGVD**：東北大学東北メディカル・メガバンク機構が公開している、一塩基多様体についてのデータベースのポータルサイトIntegrative Japanese Genome Variation Databaseの略語。アレル頻度5%以上のSNP頻度情報について、一般に公開している。また、誓約事項に同意いただくことで、1KJPNに含まれる、全てのSNVの位置情報、アレル頻度情報およびアレル数情報についてダウンロード可能になっている。
URL: http://ijgvd.megabank.tohoku.ac.jp/
参考：プレスリリース「東北メディカル・メガバンク計画『全ゲノムリファレンスパネル』情報の部分的な一般公開を開始～アレル頻度5%以上のSNP頻度情報がウェブサイトにて検索可能に～」

URL: http://www.megabank.tohoku.ac.jp/news/5696

プレスリリース「integrative Japanese Genome Variation Database～全ゲノムリファレンスパネルの公開データベース～」

URL: http://www.megabank.tohoku.ac.jp/news/13171

＊10. **日本人1,070人の全ゲノム解読に関する論文**：2015年8月21日、東北メディカル・メガバンク計画のコホート調査事業に参加した宮城県在住の健常な日本人1,070人分 の全ゲノムを解析した成果が英国科学誌「Nature Communications（ネイチャー・コミュニケーションズ）」に掲載された。

参考：プレスリリース「日本人1,070人の高精度全ゲノムデータの統合的な解析に成功～お米の消化の遺伝子の個人差やHLAの詳細などが統合解析からみえてくる～」

URL: http://www.megabank.tohoku.ac.jp/news/11873

＊11. **日本人2,049人の全ゲノム解読と公開に関するお知らせ**：2016年6月15日、東北メディカル・メガバンク計画のコホート事業に参加した宮城県在住の健常な日本人2,049人分の全ゲノムに拡充した成果をiJGVDに公開を行った。

参考：「日本人ヒト全ゲノム解析に基づく高精度の住民ゲノム参照パネル（2,049人）から全SNV頻度情報等を公開します」

URL: http://www.megabank.tohoku.ac.jp/news/15894

＊12. **デノボアセンブル**：断片化されてよみとられた塩基配列の、重複した部分を見つけ出してつなぎ合わせることで、元の染色体の塩基配列での並び順に再構築する情報科学的な手法を指す。今回の場合、概ね1万塩基配列でよみとられた配列から数百万塩基程度の配列（コンティグ）につなぎ合わせることをスーパーコンピュータシステム上で行った。

＊13. **スーパーコンピュータシステム**：東北メディカル・メガバンク機構は複合バイオバンクとしてデータバンクおよび解析の機能も併せ持っており、ライフサイエンス分野では日本最大級のスーパーコンピュータシステムの本格運用を有している。

URL: http://sc.megabank.tohoku.ac.jp/

図　日本人基準ゲノムのロゴ

(a) JRGv1 と JRGv2 の比較表

	国際ヒトゲノム参照配列に対し拡充できた塩基数	挿入配列の数
JRGv1	1,086,173	903
JRGv2	6,199,419	9,612
拡充率	約 5.7 倍	約 10.6 倍

(b) decoyJRGv1 と decoyJRGv2 の比較表

	総塩基数（単位：塩基数）	拡充できた挿入配列数
decoyJRGv1	2,536,870	3,559
decoyJRGv2	6,199,419	9,612
拡充率	約 2.4 倍	約 2.7 倍

表　前バージョンと新バージョンの比較表

【お問い合わせ先】
（研究に関すること）
東北大学東北メディカル・メガバンク機構
シークエンス解析室　室長
准教授　勝岡 史城（かつおか ふみき）
電話番号：022-273-6214
E メール：kfumiki@megabank.tohoku.ac.jp

東北大学東北メディカル・メガバンク機構
ゲノム情報解析室　室長
教授　長﨑 正朗（ながさき まさお）
電話番号：022-273-6051
E メール：nagasaki@megabank.tohoku.ac.jp

（報道に関すること）
東北大学東北メディカル・メガバンク機構
広報戦略室　室長
長神 風二 （ながみ ふうじ）
電話番号：022-717-7908
ファックス：022-717-7923
E メール：f-nagami@med.tohoku.ac.jp

（AMED 事業に関すること）
日本医療研究開発機構（AMED）
基盤研究事業部　バイオバンク課
電話番号：03-6870-2228
E メール：tohoku-mm@amed.go.jp