東北大学
TOHOKU UNIVERSITY

Press Release

2021 年 12 月 8 日

東北大学東北メディカル・メガバンク機構
東北大学高等研究機構未来型医療創成センター
日本医療研究開発機構

# ゲノムリファレンスパネルとゲノム解析情報を拡充

## ～日本人全ゲノムリファレンスパネルが 14KJPN に構造多型データベースなど公開～

【発表のポイント】

- 1 万 4 千人の全ゲノム解析データをもとに日本人全ゲノムリファレンスパネル*1 14KJPN を公開しました。
- 日本人にみられる約 6 万 8 千個のゲノム構造多型*2 とその頻度情報をデータベースとして公開しました。
- 次世代シークエンス解析の参照配列として実用性を高めた日本人基準ゲノム配列*3 の新バージョン JG2.1 を公開しました。
- 連鎖不平衡*4 情報をもとにした遺伝地図*5 の対象人数を 96 人から 150 人に拡大し解像度が向上しました。

【概要】

東北メディカル・メガバンク計画は、公開データベース日本人多層オミックス参照パネル（jMorp：Japanese Multi Omics Reference Panel）を大幅に更新し、1 万 4 千人分の全ゲノム解析情報に基づく日本人全ゲノムリファレンスパネル「14KJPN」を新たに公開しました。同データベースは均一性の高い民族集団における世界最大規模のもので、約 1 億個のバリアント*6 を収載しています。また、主に最新の長鎖リードシークエンサー*7 を用いた 333 人分の全ゲノム解析に基づく構造多型データベース、「JSV1：Japanese Structural Variation」を作成・公開しました。さらに、日本人基準ゲノム配列の実用性を高めた新しいバージョンである「JG2.1」を公開し、遺伝地図は対象人数を 150 人に拡大し精度の向上をはかりました。

jMorp はこれまでもゲノム解析情報の公開により、ゲノム医学・医療研究の発展に貢献してきました。今回の拡充により、ますます精緻なゲノム解析が可能となります。

【詳細】

＜背景＞

東北メディカル・メガバンク計画（【参考 1】）では長期健康調査によって得られた試料を解析した結果を、個人識別性のない頻度情報等にして jMorp（【参考 2】）として公開しています。jMorp は 2015 年 7 月、代謝物やタンパク質の解析結果の公開からスタートし、2018 年 6 月に別サイト（iJGVD）で公開していた全ゲノムリファレンスパネルを統合しました。全ゲノムリファレンスパネルは 1,000 人分（1KJPN）からスタートし、着々と解析人数や情報を拡大し、2020 年 8 月には 8.3 千人分のゲノム解析の頻度情報（8.3KJPN）を公開しました。全ゲノムリファレンスパネルでは、最初は SNV[*8] の頻度情報、2018 年 11 月には 50 塩基対未満の INDEL[*9] 頻度情報を公開しました。ゲノムにはさらに長い挿入や欠失が存在し、疾患等の表現型に大きな影響を持つと考えられていますが、短鎖リードシークエンシング解析では長い配列の解析は困難であるため、日本人における構造多型の個人ごとの違いや頻度はこれまで不明でした。

現在ヒトゲノム解析の主流である短鎖リードシークエンシング解析にひな型として必要なのが「基準ゲノム配列」です。2019 年 2 月に日本人のゲノム解析を行うためのひな型となる JG1 を発表しました。JG1 およびその後発表した JG2 は、民族集団の違いを考慮したゲノム解析に有効ですが、長鎖リードシークエンシング技術では解読が難しい配列部分を中心に未決定領域が残されていました。

また、2019 年 11 月には連鎖不平衡情報をもとにした遺伝地図を発表しました。これは日本人民族集団の家系から算出した初めての遺伝地図です。対象は三世代コホート調査参加者の 96 人でありアジアでは最大級の遺伝地図でした。

なお、jMorp はゲノム解析情報以外にもヒトに関わる生命科学の総合的な情報を公開しており、リファレンスパネルとして多くの研究者に利用されています。

＜内容＞

◆14KJPN

これまで公開を行ってきた約 8.3 千人からなるリファレンスパネル（8.3KJPN）の更新版として、「14KJPN」の構築を行いました。8.3KJPN と 14KJPN の違いは大きく分けて 2 点あり、(1)検体数の拡充と(2)参照配列の変更になります。

(1)検体数の拡充

14KJPN は、東北メディカル・メガバンク計画による宮城県と岩手県でのコホート調査への協力者、合計 14,129 人から構成されており、より低頻度なバリアントを収載しました。14KJPN に収録される SNV および INDEL の数は以下の通りです。

| | SNV | INDEL |
|---|---|---|
| 常染色体 | 106,705,823 | 13,130,321 |
| X 染色体<br>(PAR1+PAR2)*10 | 4,015,929 | 518,977 |
| X 染色体<br>(PAR1+XTR +PAR2) *10 | 4,074,917 | 526,406 |
| ミトコンドリア | 3,832 | - |

- X 染色体は 2 種類の解析方法で解析された結果を公開
  （解析方法の詳細: Tadaka *et al.*, 2019, *Human Genome Variation*）

(2)参照配列の変更

8.3KJPN は GRCh37/hg19 と呼ばれる国際的に使われるヒトゲノム配列を参照配列として用いて解析を行ってきましたが、14KJPN では現時点の最新版である GRCh38/hg38 を用いて解析を行いました。

14KJPN のアレル頻度*11 情報・ジェノタイプ頻度*12 情報は jMorp ウェブサイト（https://jmorp.megabank.tohoku.ac.jp/）からダウンロード可能です。（ジェノタイプ頻度情報のダウンロードにあたっては、ORCID*13 と連携する認証を行い、データ移転契約（DTA：Data Transfer Agreement）をご確認いただく必要があります。）

◆構造多型データベース

これまで公開を行っていたリファレンスパネル（8.3KJPN, 14KJPN など）は主に短鎖リードシークエンス技術を用いた全ゲノム解析データに基づいており、その解析対象を SNV や小規模の INDEL に限定していました。一方で、ヒトゲノム中には、構造多型と呼ばれる大規模な塩基の挿入や欠失が存在することが知られています。そこで今回、一部の検体で長鎖リードシークエンス解析を実施し、日本人集団にみられる構造多型を網羅したデータベース「JSV1」を作成しました。解析には三世代コホート調査への協力者（111 トリオ（両親、子の組み合わせ）、333 人）の培養細胞試料から取得した高品質 DNA を用いており、家族間の遺伝子型一致性を利用することで解析結果の精度検証を実施しました。JSV1 では、ゲノム中の構造多型の位置情報のみならず、解析集団におけるアレル頻度、精度検証結果も合わせて公開しています。

JSV1 とは別に短鎖リードシークエンス解析を用いた構造多型解析結果も公開しました。大規模な挿入やリピート部分の欠失の精度については長鎖リードシークエンス解析に及びませんが、より人数の多い約 8.3 千人からなる構造多型リファレンスパネルです。低頻度でみられる構造多型の調査など JSV1 を補完する用途に利用可能です。

◆JG2.1

基準ゲノム配列はゲノム解析を行う上で遺伝子や突然変異のある位置を表すための重要な情報基盤です。東北大学東北メディカル・メガバンク機構（以下ToMMo）ではこれまで2019年に日本人基準ゲノム配列JG1を、2020年にその後継となるJG2を構築・公開してきました。JG1は3組のゲノム配列を構築し統合したもので、JG2は6組のゲノム配列を統合し、より日本人の代表性を高めたものです。しかし、ヘテロクロマチン領域*14周辺や大規模な重複が知られている領域など難読領域は未解読の状態でした。今回我々は、次世代シークエンス解析における実用性を高めるために、JG2でも未解読であった領域に対し、国際基準ゲノム配列GRCh38/hg38を参照して更新し、「JG2.1」を構築しました。JG2.1では、JG2で未解読だった常染色体領域2.1億塩基対を1.27億塩基対まで減らしました。またタンパク質をコードする遺伝子がJG2では19,429個検出されていましたが、JG2.1では19,743個検出することに成功しました。JG2.1は日本人以外に由来するゲノム配列領域があるため使用には注意が必要ですが、配列の由来の情報も同時に公開しています。また次世代シーケンス解析のデファクトスタンダードであるGATKベストプラクティス*15の実行に必要なGATKリソースバンドル*15のJG2.1座標バージョンも合わせて公開しています。これにより、より実用性の高い基準ゲノム配列となっています。

◆遺伝地図

ヒトは、両親からそれぞれ一組のゲノムDNAを受け取る二倍体の生き物です。個体が次世代にゲノムDNAを伝える際には、精子や卵子を作って一倍体ゲノムにして伝達します。この過程で、両親から受け取ったゲノムDNAの一部を交換する「組換え」と呼ばれる現象が起こります。組換えにより多様な遺伝情報を持った個体を生み出すことが可能になると考えられます。ゲノムDNA上の遺伝子や多型を示すマーカー間の距離は、物理的な距離だけでなく、この組換えの起こりやすさを用いても表すことができます。これを遺伝地図と呼びます。組換えはゲノムDNA上で均一に生じるのではなく、むしろ組換えホットスポットと呼ばれる限られた範囲でよく生じることが知られており、物理的な距離と関連するものの、異なった描像が得られます。

ToMMoでは、2019年11月に96検体の全ゲノムシークエンス結果から、ゲノム全体にわたって組換えが起こりやすかったか（組換え頻度）を推定し、その結果である「遺伝地図」を構築・公開してきました。2019年版では8,735,371マーカーとその間の距離を推定しましたが、今回、150検体に解析対象を広げることで、10,092,551マーカー間の距離を推定しました。これにより高解像度な遺伝地図を構築することに成功しました。

遺伝地図はハプロタイプフェージング*16、遺伝子型インピュテーション*17、連鎖分析等、様々な遺伝統計学的解析の基盤情報となります。この遺伝地図は日本人集団を対象とした遺伝統計解析の高精度化に貢献するものと期待されます。

【今後の展望】

14KJPN と構造多型データベースの公開により、jMorp が網羅する遺伝的バリアントの数および種類が大幅に拡張しました。今後は一層、SNV、構造多型ともに解析対象人数を増やすとともに、INDEL 以外の重複や逆位の構造多型解析情報の搭載を検討しています。

JG2.1 への更新による精緻化でこれまで不可能であった難読領域におけるバリアントコール*18 が可能になりました。今後は性染色体など一層未解読領域を減らしていくとともに、日本人検体のゲノム情報を用いた難読領域の解読に挑みます。

連鎖不平衡情報に基づく遺伝地図は三世代コホート調査をもとにした世界最大規模のもので連鎖分析*19・関連解析*20 等、様々な遺伝統計学的解析の情報基盤としての利用が期待されます。

今後も研究基盤の強化を継続し、常に最前線で日本のゲノム医学・医療の発展を牽引していきます。

【用語説明】

*1　全ゲノムリファレンスパネル ： 東北メディカル・メガバンク計画で実施された、日本人の一般住民数千人の全ゲノム次世代シークエンシング解析により、検出されたゲノム DNA バリアントから構築された日本人ゲノム配列のパネル。

*2　ゲノム構造多型 ： ゲノム配列において、SNV（後述）や短鎖リードシークエンサーで検出できるような INDEL（後述）などの短い長さの多型ではなく、数十から数千、あるいはそれ以上の塩基が個人間で異なる多様性のこと。

*3　基準ゲノム配列 ： 次世代シークエンシング解析を行う際、ひな型となるゲノム配列。参照配列ともいう。次世代シークエンシング解析ではリードと呼ばれる小さな単位で大量に配列解読を行い、リードを基準ゲノム配列に当てはめて検体の元のゲノム配列を推定する。そのため基準ゲノム配列の品質がゲノム解析の精度を左右する。

*4　連鎖不平衡 ： ゲノム上の連鎖している（座位）の各 2 種類の塩基（アレル）について、ランダムに生じる以上の偏った組み合わせ（ハプロタイプ）の存在。

*5　遺伝地図 ： 染色体上のマーカー間の距離を、組換えの頻度で表したもの。100 回の減数分裂で 1 回の組換えが生じる距離を 1 センチモルガン（cM）と呼ぶ。

*6　バリアント ： 標準となるゲノム配列とは異なる箇所のこと。

*7　長鎖（短鎖）リードシークエンサー ： 大量のゲノム情報を同時並行で高速に解析可能な装置が次世代シークエンサーであり、数百塩基単位で解析しその後情報を基準ゲノム配列に当てはめるのが短鎖リードシークエンサー、数千から万単位の塩基を解析可能なのが長鎖リードシークエンサーである。短鎖リードシークエンサーは解

析速度やコスト面で優位性があり、長鎖リードシークエンサーは基準ゲノム配列から外れた配列も解析可能であるため構造多型の解析に適している。

*8　SNV ： 一塩基バリアント。ゲノム配列において、ある領域で DNA の塩基配列が個人間で一塩基のみ異なる多様性のこと。

*9　INDEL ： ゲノム配列における塩基配列の挿入（insertion）または欠失（deletion）のどちらかあるいは両方。

*10　PAR1（+XTR） +PAR2 ： X 染色体の解析において、X と Y 染色体それぞれ一本ずつ持つ男性では、X 染色体を 2 本持つ女性や常染色体とバリアントコール（後述）の方法が異なり、解析方法は確立されていない。jMorp では、比較的有効と見られている 2 種類の方法の両方を用いて解析を行っている。

*11　アレル頻度: ある集団における DNA バリアントの塩基（A,T,G,C）の頻度で、アレル（同じ座位上で対立して存在する塩基）ごとに算出したもの。今回は対象となった日本人約 1 万 4 千人中の頻度となり、最大で約 2 万 8 千アレルのうちにどれだけ検出されたか計算される。

*12　ジェノタイプ頻度: 遺伝子型頻度。父母から由来する二つのアレルの組み合わせの頻度。今回の発表では、対象となる約 1 万 4 千人の中で、ホモで持つ（父母由来の情報が双方ともある）、ヘテロで持つ（父母いずれかからのみ持つ）などを分けて算出している。

*13　ORCID ： 研究者等学術的な著作の著者を一意的に識別するために作られた英数字コード。

*14　ヘテロクロマチン領域 ： 染色体上の常に凝集した領域であり配列決定が困難である。

*15　GATK ベストプラクティス／GATK リソースバンドル ：米国 Broad Institute が開発・提唱している次世代シークエンシング情報解析の標準的な手法（ベストプラクティス）と、その実行に必要な情報資源（リソースバンドル）のこと。

*16　ハプロタイプフェージング ： バリアントが 2 本の染色体のどちらに属するものか決定すること。

*17　遺伝子型インピュテーション ： 観測された遺伝子型の組み合わせから、観測されていない座位の遺伝子型を推定すること。他の検体で大規模に全ゲノム解析を行った結果であるリファレンスパネルを参照することで推定を行う。

*18　バリアントコール ： 参照配列との差異（バリアント）を検出すること。

*19　連鎖分析 ： 主に大規模な家系を用いて継承性のある形質と関連する遺伝子座を同定する方法の一つ。

*20　関連解析 ： 主に非血縁者の集団のアレル頻度情報を利用して特定の疾患や身長などの形質と関連する遺伝子座を同定するための解析手法。

【参考 1】＜東北メディカル・メガバンク計画＞

東北メディカル・メガバンク計画は、東日本大震災からの復興事業として平成 23 年

度から始められ、被災地の健康復興と、個別化予防・医療の実現を目指しています。
　ToMMo と岩手医科大学いわて東北メディカル・メガバンク機構を実施機関として、東日本大震災被災地の医療の創造的復興および被災者の健康増進に役立てるために、合計 15 万人規模の地域住民コホート調査および三世代コホート調査を平成 25 年より実施し、収集した試料・情報をもとにバイオバンクを整備しています。
　平成 27 年度より、日本医療研究開発機構（AMED）が本計画の研究支援担当機関の役割を果たしています。

【参考 2】＜jMorp＞
　公開データベース日本人多層オミックス参照パネル。東北メディカル・メガバンク計画のコホート調査によって得られた試料を解析した結果を、個人識別性のない頻度情報等にして公開している。
サイト名：Japanese Multi Omics Reference Panel (jMorp)
言語：英語
URL：https://jmorp.megabank.tohoku.ac.jp/

【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
基盤情報創成センター長
木下　賢吾（きのした　けんご）
電話番号：022- 274-5952

（報道担当）
東北大学東北メディカル・メガバンク機構
長神　風二（ながみ　ふうじ）
電話番号：022-717-7908
ファクス：022-717-7923
E メール：pr@megabank.tohoku.ac.jp

（AMED 事業に関すること）
日本医療研究開発機構（AMED）
ゲノム・データ基盤事業部　ゲノム医療基盤研究開発課
電話番号：03-6870-2228
E メール：tohoku-mm＠amed.go.jp