東北大学
TOHOKU UNIVERSITY

Press Release

2020 年 10 月 5 日

東北大学東北メディカル・メガバンク機構
日本医療研究開発機構

ゲノム・オミックス解析情報の公開データベース jMorp を大幅拡充

～日本人の主な変異を網羅、メタボローム解析情報も～

【発表のポイント】

- 疾患の比較対照として有効な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）の収載データを大幅に拡充しました。
- 8.3 千人の全ゲノム解析データをもとに全ゲノムリファレンスパネル*1 8.3KJPN を公開しました。8.3KJPN は、日本人のもつ 0.01%以上の頻度の変異をほぼ網羅しました。
- 3 セットのゲノム配列を統合することで構築した日本人基準ゲノム配列*2 を、6 セットに拡張することで、より日本人集団の遺伝的多様性を代表し、基準としての精度を格段に向上させた日本人基準ゲノム配列 JG2 を公開しました。
- jMorp のメタボローム解析*3 対象者数を 1.5 万人から約 2.5 万人に拡張するとともに、対象代謝物の種類を大幅に拡張しました。

【概要】

東北大学東北メディカル・メガバンク機構（ToMMo）は、jMorp の全ゲノム解析データを 8.3 千人に、メタボローム解析データを 2.5 万人にそれぞれ大幅拡充すると共に、日本人基準ゲノム配列を再構築し JG2 として公開しました。個別化医療・個別化予防の実現に向けた、疾患の原因や治療法の発見への貢献が期待されます。

東北メディカル・メガバンク計画（【参考】を参照）では、当初より 8 千人の全ゲノム解析を目指してきましたが、今回目標を超える人数の全ゲノムリファレンスパネルを構築したことで、日本人のゲノムデータを有するデータベースとして一番網羅性の高いゲノム情報となりました。

約 2.5 万人分のメタボローム情報は世界で最大級の収載数となり、また最新の解析手法による新規代謝物の定量値データのリファレンスパネル*4 の公開は世界初です。

日本人基準ゲノム配列を 3 セットから 6 セットのゲノム配列にし、精度を格段に向上させて JG2 として公開しました。

全国の研究者が jMorp のデータを用いることにより、疾患の原因や予防法、バイオマーカー*5 の発見が期待されます。本研究は、国立研究開発法人日本医療研究開発機構（AMED）による東北メディカル・メガバンク計画のもと東北大学東北メディカル・メガバンク機構によって行われています。

【詳細】

■jMorp の経緯と機能

jMorp は、東北メディカル・メガバンク計画のコホート調査*6 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たデータベースです。2015 年 7 月に世界で初めて 500 人以上の血漿に対する網羅的メタボローム及びプロテオーム*7 の統合解析結果を公開し、以後着々と収載データの量と種類を増やしてきました。2018 年にはゲノム解析情報を、2020 年には岩手医科大学いわて東北メディカル・メガバンク機構（IMM）による解析情報を追加し、現在ではゲノミクス、エピゲノミクス、トランスクリプトミクス、プロテオミクス、メタボロミクスというヒトに関わる生命科学の総合的な情報を網羅的に収載するリファレンスパネルとなっています。jMorp の情報は主に東北メディカル・メガバンク計画の長期健康調査に参加された方々のご協力のもとに収集されています。

表 1：jMorp の経緯

時期	アップデート内容
2014 年 8 月	1KJPN（1,000 人分の全ゲノムリファレンスパネル）公開。（別サイト）
2015 年 7 月	jMorp 初公開。500 人分の網羅的メタボローム及びプロテオーム統合解析結果を公開
2016 年 6 月	2KJPN 公開（別サイト）
2016 年 8 月	メタボロームの解析人数を 1,000 人に拡大。項目間相関情報・ペプチド情報を追加
2017 年 7 月	3.5KJPN 公開（別サイト）
2017 年 10 月	メタボロームの解析人数を 5,000 人に拡大。年齢別の代謝物濃度分布・ネットワーク解析結果を追加
2018 年 6 月	jMorp に統合した 3.5KJPNv2 公開。X 染色体とミトコンドリアゲノム情報を追加
2018 年 8 月	メタボロームの解析人数を 10,000 人に拡大
2019 年 2 月	日本人基準ゲノム配列 JG1 の公開
2019 年 9 月	メタボロームの解析人数を 15,000 人に拡大。全ゲノム解析情報を 4,700 人まで拡張（4.7KJPN）
2020 年 1 月	DNA メチル化情報、遺伝子発現情報(iMETHYL)掲載
2020 年 8 月（本発表）	メタボロームの解析人数を 25,000 人に拡大。全ゲノム解析情報を 8,300 人まで拡張（8.3KJPN），日本人基準ゲノム配列を JG2 にアップデート

■全ゲノムリファレンスパネルの拡充内容

これまで公開を行なってきた約 4,700 人からなるリファレンスパネル（4.7KJPN）の更新版として、8.3KJPN の構築を行いました（表1）。8.3KJPN 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や、その他外部コホート事業における協力者、合計 8,300 人から構成されています（表2）。本拡充により日本人が持つアレル頻度*8 0.01%以上の SNV*9 をほぼ収集できたと考えられます。

表 2:8.3KJPN を構成する協力者のコホート別人数

コホート名	解析プラットフォーム（シークエンサー）	人数
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	Illumina HiSeq2500	3,619
	Illumina HiSeq X Five	71
	Illumina NovaSeq 6000	3,691
	MGI DNBSeq G400	577
独立行政法人国立病院機構長崎医療センターにおける協力者	Illumina HiSeq2500	191
ながはま0次予防コホート事業における協力者	Illumina HiSeq2500	65
国立がん研究センターにおける協力者	Illumina HiSeq2500	47
J-MICC Study における協力者	Illumina HiSeq2500	60
大阪大学眼科における協力者	Illumina HiSeq2500	30
大阪大学ツインリサーチセンターにおける協力者	Illumina HiSeq2500	29
合計		8,380

8.3KJPN はこれまでの 4.7KJPN に含まれていた検体を引き継ぎつつ、また、外部のコホート研究との連携により東日本・西日本両方からの協力者が含まれるため、より日本人として代表性の高いゲノム情報となっています。また、4.7KJPN までは Illumina 社のシークエンサーによるゲノム解析をもとにしたパネルとなっておりましたが、8.3KJPN では新たに MGI 社 DNBSeq G400 によるデータも含まれるようになりました（表 2）。

8.3KJPN に収録される SNV（一塩基バリアント）・INDEL*10（挿入・欠失）の数は以下の通りです（表 3）。

表 3:8.3KJPN に収録された SNV・INDEL 数

	SNV		INDEL	
	総数	新規数	総数	新規数
常染色体	76,324,188	35,911,274	10,017,023	5,646,786
X 染色体 (PAR1+PAR2)	2,994,485	1,506,206	411,478	234,386
X 染色体 (PAR1+PAR2+XTR)	3,031,683	1,521,465	415,785	237,476
ミトコンドリア	3,357	1,715		

- （dbSNP release 153 に記載のないものを新規として計算）
- （X 染色体は 2 種類の解析方法で解析された結果を公開。解析の詳細は Tadaka et al, 2019, *Human Genome Variation* をご参照ください）

また、アレル頻度情報以外に、ジェノタイプ頻度*11 情報の公開も同時に行いました。これにより、より詳しい精度での解析が可能になります。両データとも jMorp ウェブサイト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*12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。)

■日本人基準ゲノム配列解析の拡充内容

日本人基準ゲノム JG2 は、JG1 と同じ 3 人の検体を用いながら、新たに Hi-C 法*13 やナノポア長鎖リードシークエンサー *14 などの新規手法を追加して解析を行い、6 ハプロイドゲノム*15(3 人×2 セットのゲノム配列)を構築し統合しました。この手法により JG2 は JG1 はもとより他の人類集団固有の基準ゲノムより高い精度でゲノム解析を行うことが可能になります。特に JG2 ではアセンブリの誤りと推定される領域の数が JG1 から 1,800 程度減りました(表 4)。このことは、誤って構造多型として検出される数が最大で 1,800 程度減り、さらにこの 1,800 の領域における SNV や短い INDEL の解析精度が向上することを示唆します。

表 4:JG2、JG1、および他の人類集団固有の基準ゲノムとの比較

	推定ミスアセンブリ数 *16(1kb 超)	推定ミスアセンブリ数 (1 kb 以下)	遺伝子領域におけるエラー数の推定値
JG2	1,055	2,323	780
JG1	1,654	3,613	781
ZF1 (チベット)	2,066	3,043	1,174
AK1 (韓国)	2,138	5,233	1,129

■メタボローム解析の拡充内容

解析人数、対象とする代謝物の種類を拡充、さらに同じ対象者集団についての経時的な情報を追加し、大規模・高精度・多彩な代謝物の濃度分布情報を含むデータベースとなりました。

1)【データ拡充】NMR によるメタボローム解析の代謝物の種類と解析人数を拡充:

NMR メタボローム解析*17 では 2019 年に 43 代謝物について、15,403 人分の定量値・分布情報を公開しています。この 43 代謝物についての定量解析法を 2019 年度 jMorp 公開データに適用し再解析した結果に加え、さらに今回新たに 10,000 人分の解析結果を追加し、約 25,000 人という大規模なものへと拡張するとともに新たに 2 代謝物を追加しました。また、この中には妊娠中の代謝物の変動データ約 2,000 人分も含まれます。

2)【新規】最新の手法による MS 標的メタボローム定量情報解析:

MS メタボローム解析*18 では、超高速液体クロマトグラフ三連四重極型質量分析装置(LC-MS/MS)による標的メタボローム解析に、測定可能な代謝物が大幅に増えた新たな代謝物

キット解析[*19]を導入し、血漿 10μL から検出された 421 代謝物を新規に追加しました。なお解析対象者数は、約 2,400 人分です。この 421 代謝物の新規データは，他の研究機関での解析結果と比較可能な定量値情報として提供します。また、ガスクロマトグラフ三連四重極型質量分析装置（GC-MS/MS）による代謝物についても、約 600 人分の分布・頻度情報を追加し、約 3,000 人としました。

3)【データ拡充】経時変化情報を拡充：
詳細二次調査[*20]参加者からご提供いただいた血漿試料に対して NMR/MS メタボローム解析を実施し、NMR メタボローム解析では 700 人分、44 代謝物の定量値情報、MS メタボローム解析では GC-MS/MS による約 600 人分、165 代謝物の測定値情報を追加致しました。この拡張により加齢によるメタボロームの経時変化について、最大 1,600 人規模の比較解析を行うことが可能になります。
また、約 400 人分の妊娠中の代謝物の変動データを追加し、約 2,000 人分とすることにより、妊娠中及び産後 1 ヶ月の比較解析がより高精度に可能となりました。

なお、いずれの代謝物についても、2019 年拡張版 jMorp の公開情報と同様に、代謝物間の相関情報、年齢層別の分布情報等を公開します。

今回の jMorp 拡充は、宮城県で実施中のコホート調査の協力者の方々からご提供いただいた生体試料をもとに行われました。

【今後の展望】

東北メディカル・メガバンク計画では最終的に 8,000 人の解析データをもとにした全ゲノムリファレンスパネルの構築を目指して解析を行ってきました。今回、8.3KJPN を構築し、目標の 8,000 人を達成したことで、日本人が持つアレル頻度 0.01%以上の SNV データをほぼ収集できていると考えています。そのため、かなり稀な疾患まで含めた遺伝疾患の原因解明における、より信頼度の高い頻度フィルタ用途など、当計画のリファレンスパネルが更に広く有効に使われることが期待されます。また、非常に解析が難しい Y 染色体の情報や、これまで収載している、SNV・INDEL 頻度のほかに、構造多型[*21]の頻度情報を収載することも検討しており、「日本人全ゲノムリファレンスパネル」としてブラッシュアップし続けます。

JG2 は、JG1 に比べ正確性がさらに向上しました。しかし、未決定領域は依然 250 Mb 程度残されており、特にセントロメアやテロメア領域と呼ばれる解読の難しい領域は未解読のままです。今後、正確性の高い長鎖リードシークエンサーなど新たな技術を適用することで、これらの領域の解読を進め、さらなる高精度化を目指します。

メタボローム解析データは、リリース当初 500 人だった対象者数が約 25,000 人となりました。今後も順次対象者数を増やしリファレンスとしての精度を高めるとともに、追跡調査データを加え経時情報についても厚みを持たせたいと考えます。さらに、最新の解析手法を絶えず取り入れ、代謝物の種類を増やしていきます。このように解析を加速していくことで、男女、年代のバリエーションを持ち、かつ数万人単位の経時変化を含む、世界でも類を見ないメタボローム情報のコレクションが構築され、ヒトの加齢変化を分子的、かつ網羅的に捉えることができる貴重な基盤となります。

【jMorp】
サイト名：Japanese Multi Omics Reference Panel (jMorp)
言語：英語
URL：https://jmorp.megabank.tohoku.ac.jp/

【参考】
＜東北メディカル・メガバンク計画について＞
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と、個別化予防・医療の実現を目指しています。ToMMo と IMM 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、平成 25 年より合計 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等を実施して、試料・情報を収集したバイオバンク*22 を整備しています。東北メディカル・メガバンク計画は、平成 27 年度より、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。

【用語解説】
＊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読により、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. 基準ゲノム配列：次世代シークエンシング解析（ヒトゲノム解析でもっともよく利用されている手法）を行う際、ひな型となるゲノム配列。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。ToMMo では 2019 年 2 月に「日本人基準ゲノム配列」初版 JG1 を発表している。
＊3. メタボローム解析：生体内に含まれる代謝物（メタボライト）の網羅的な解析。
＊4. リファレンスパネル：ここでは疾患など健常時と異なる状態に対して、比較対照可能な参照用のデータ群を指す。
＊5. バイオマーカー：疾患の発症や進行を反映する生体内分子。
＊6. コホート調査：特定の集団を一定期間追跡することによって、環境要因や遺伝的要因と疾病発生の関連を調べる調査。当計画の調査は「前向きコホート調査」である。
＊7. プロテオーム：生体内に含まれるタンパク質.。
＊8. 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 8,300 人中の頻度なので、最大で約 16,000 アレルのうちにどれだけ検出されたか計算される。
＊9. 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。
＊10. 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。

＊11. 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 8,300 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。

＊12. 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つくられた英数字コード。

＊13. Hi-C 法：細胞の核の中における染色体 DNA 領域間の物理的な近さを次世代シークエンサーにより推定する方法。同一染色体は物理的にも近接しているため、この情報をゲノム配列の再構築に応用できる。

＊14. ナノポア長鎖リードシークエンサー：ナノポアと呼ばれる微細な穴に、解読したい DNA を通過させ、その際に生じる電流の変化を元に配列を解読する手法。短鎖リードシークエンサーに比べ正確性は劣るものの、より長い（10 kb 程度）単位で解読することが可能である。

＊15. ハプロイドゲノム：精子や卵子が持つゲノム情報 1 セットをハプロイドゲノムと呼ぶ。したがって一人のヒトが有するゲノム情報全体はディプロイド（二倍体）である。

＊16. 推定ミスアセンブリ数：DNA を解読したリードをつなげて元の染色体配列を再構築することをアセンブリと呼ぶ。アセンブリの際に異なる染色体を一つに繋げてしまったり、特定の領域を省略してしまったりするなどして生じた誤りの推定値をミスアセンブリ数と呼び、アセンブリ性能の指標として広く使われる。

＊17. NMR メタボローム解析：NMR 解析は、生体分子を含む様々な分子を強力な磁場の中において、分子中の各原子が持つ核磁気モーメントを計測することにより、分子の構造や量を測定する解析方法。当計画では、NMR 解析を用いて生体中の代謝物を網羅的に解析（メタボローム解析）している。

＊18. MS メタボローム解析：MS（マススペクトロメトリー）は、物質を荷電粒子に変え、質量電荷比（*m/z*）にて分離されたスペクトルとして検出する解析方法。生体内、食品及び環境に含有される様々な物質の存在量を測定することができる。当計画では、MS を用いて生体中の代謝物の定量的な解析を行っている。

＊19. 代謝物キット解析：当計画では、従来代謝物を UHPLC-MS/MS を用いて一斉に定量分析できる試薬キットである AbsoluteIDQ® p180 Kit (Biocrates 社製)による解析を行っていたが、今回新たに Biocrates MxP Quant 500 と呼ばれる最新のキットを導入し解析を行った。こちらは従来の p180 と比較して脂質代謝関連分子を中心に測定可能な代謝物の種類が大幅に増加しており、多数の代謝物の検出が可能である。

＊20. 詳細二次調査：当計画のコホート調査（*6 参照）において、参加者に対して最初に行った健康調査から約 4 年後に実施している 2 回目の詳細調査。数年おきに同じ人に対して健康調査を行うことにより調査開始時点からの前向きな経時的変化がわかる。

＊21. 構造多型：ゲノム配列において、SNV や短鎖リードシークエンサーで検出できるような INDEL 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。

＊22. バイオバンク：生体試料を収集・保管し、研究利用のために提供を行う。東北メディカル・メガバンク計画のバイオバンクは、コホート調査の参加者から集めた血液・尿などの生体試料だけではなく、それらを検査・解析した情報、調査票等から得られた情報も含む。

東北大学東北メディカル・メガバンク機構 https://www.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
生命情報システム科学分野 教授 木下賢吾
電話番号：022- 274-5952
生体分子解析分野 教授 小柴生造
電話番号：022-274-6016
ゲノム遺伝統計学分野 教授 田宮元
電話番号：022-274-5996

（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
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