東北大学 TOHOKU UNIVERSITY

Press Release

平成 30 年 6 月 25 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
学校法人岩手医科大学いわて東北メディカル・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

東北メディカル・メガバンク計画による 3.5 千人分の日本人全ゲノムリファレンスパネルに X 染色体とミトコンドリアゲノム情報を追加

より広範な疾患に対応、国際比較も容易に

【発表のポイント】

・東北大学東北メディカル・メガバンク機構（ToMMo)はこれまで進めてきた全ゲノム解析*1 のデータ約 3.5 千人分をもとに、内容を大幅に拡充・更新した全ゲノムリファレンスパネル 3.5KJPNv2 を公開しました。

・今回の公開にあたっては、これまでのリファレンスパネルのデータに含まれていなかった性染色体*2 である X 染色体、そして、ミトコンドリア*3 のゲノム情報を新たに加えました。日本人の X 染色体やミトコンドリアの情報を含む 1,000 人規模以上のゲノム情報を発表した例はこれまでになく、初のデータとなります。

・また、データ解析手法を国際標準に準拠した手法に更新、海外の大規模ゲノム解析との比較がより容易なリファレンスパネルとして再構築しました。更に、データ公開サイトも刷新して、研究者の方々により使いやすいものとしました。

・X 染色体やミトコンドリアには、比較的発症頻度の高い遺伝性疾患の原因となる遺伝子が多数含まれており、新規リファレンスパネルの公開は特に小児科・産婦人科・遺伝カウンセリング領域の疾患について、研究者や医療関係者に利活用されることが想定され、我が国の個別化医療・個別化予防に向けた研究が促進されることが期待されます。

【概要】

ToMMo は全ゲノム解析のデータ約 3.5 千人分をもとに、日本人による標準的な全ゲノムリファレンスパネルを国際比較に利用しやすくするために再構築し、新たな情報を大幅に拡充した 3.5KJPNv2 を公開しました。本パネルの対象者構成は、2017 年 9 月に公開した全ゲノムリファレンスパネル（以下、3.5KJPN）とほぼ同様で、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*4 への協力者 3,342 人、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 181 人、ながはま 0 次予防コホート事業への協力者 29 人で構成される日本人一般住民合計 3,552 人分です。

今回のパネル更新の要点は以下の通りです。

- ●X 染色体の情報を追加したこと
  X 染色体上で、200 万個を超える一塩基バリアント*5（SNV）を発見して、パネルに収載しました。
- ●ミトコンドリアの情報を追加したこと
  ミトコンドリア DNA 上で、2 千個以上の SNV を発見して、パネルに収載しました。
- ●国際標準に準拠した解析手法を用いて国際比較に優れたデータとしたこと
  現在の国際的な標準手法とされる GATK Best Practices*6 に準拠した方法により再構築を行いました。これにより、海外のゲノムデータと定量的な比較がより容易となりました。
- ●認証されたユーザーに対して詳細な検索機能を導入したこと
  頻度 1%以上の SNV に限っていたブラウザ上での検索が、認証されたユーザーでは全ての SNV で可能になりました。また、検索画面上で国際 1000 人ゲノムデータ*7 におけるアレル頻度データも同時に表示する機能を実装しました。これにより、国際的なデータベースと比較することも簡便になりました。

また、従来の 3.5KJPN が有していた以下の特長は 3.5KJPNv2 にも引き継がれています。

- ●長崎や長浜（滋賀県）におけるプロジェクトの協力により、東北在住以外のデータを取り込んだ。また、岩手・宮城両県でのコホート参加者についても、調査票情報から由来が両県以外の方々を多く抽出し、北は北海道から南は沖縄までカバーする日本人として代表性の高いゲノム情報とした。
- ●延べ約 329 兆塩基もの高品質な全ゲノム断片配列情報を解読し、スーパーコンピュータによる情報解析技術と他の手法による実験結果による検証とを組み合わせた信頼度の高い配列情報であること。

3.5KJPNv2 は、東北メディカル・メガバンク計画を推進する国立研究開発法人日本医療研究開発機構（AMED）が掲げるデータシェアリングの推進方針を踏まえ、ToMMo の日本人多層オミックス*8 参照パネル（jMorp：Japanese Multi-Omics Reference Panel）のウェブサイトにおいて公開します（※）。

※jMorp は、東北メディカル・メガバンク計画によって得られた、オミックスデータの統合を目指して構築されたデータベースで、2015 年以降メタボローム情報やプロテオーム情報[*9] を公開しています。今回、データ利用者の利便性を高めるため、ゲノム情報も jMorp 上で公開することとしました。
URL：https://jmorp.megabank.tohoku.ac.jp/
なお、従来のリファレンスパネルのゲノム情報が掲載されている iJGVD（integrative Japanese Genome Variation Database）には、jMorp からもアクセスすることが可能です。

【背景】
ToMMo は岩手医科大学いわて東北メディカル・メガバンク機構と協力し、宮城県・岩手県の地域住民 15 万人規模のコホート調査を 2013 年から実施しています。ToMMo では、このコホート調査に参加された宮城県住民の 1,070 人分の全ゲノム解読が完了したことを 2013 年 11 月に発表し、2014 年 8 月には解読された全ゲノム配列に基づく全ゲノムリファレンスパネル 1KJPN のアレル頻度[*10]5%以上の SNP 頻度情報をウェブサイト上で公開、翌年 12 月には公開範囲を全ての SNV 頻度・位置情報に拡充しました。さらに 2016 年 6 月には 2,049 人分の全ゲノム配列に基づく全ゲノムリファレンスパネル 2KJPN を構築し公開、更に 2017 年 9 月に 3.5KJPN を公開してきました。

【今回のパネルの新規性・特徴の詳細】
今回公開した 3.5KJPNv2 構築のための解析にあたって、海外の大規模ゲノム解析との定量的な比較を行うことが容易となるよう、近年、国際的に標準的な解析手法となりつつある GATK Best Practices 法を用いました。
3.5KJPNv2 の主な特徴は以下の通りです。

●常染色体[*11] の再解析の結果
常染色体上に検出された SNV の総数は、51,168,347 個（基準をクリア（※）したものは 44,107,909 個）であり、当計画のこれまでの全ゲノム解析によって常染色体上に検出され、それ以外の国際的なデータベース等に収載されてこなかったものは 28,127,100 個（同 23,652,000 個）です。
※基準をクリア：解析によって検出されたバリアントに対して一定の基準でチェックを行い変異データの信頼性を高めるようにしています。今回公開した常染色体と X 染色体のパネルは、いくつかの国際プロジェクトでも採用されている GATK の Variant Quality Score Recalibration (VQSR) 法を用いた基準で判断をしています。

●X 染色体
X 染色体はヒトゲノム全体の約 5%に相当する鎖長[*12] を持つ比較的大きな染色体であり、1,000 以上の遺伝子を含み、特に血友病や筋ジストロフィー症などの疾患遺伝子の存在が知られる重要な染色体です。これまで難しかった一般住民集団での伴性疾患遺伝子頻度の正確な推定などが可能になり、疾患研究に対して貢献が期待されます。

今回、標準的な解析（※）で 2,005,093 個（PAR1+PAR2、基準をクリアしたものは 1,726,127 個）の SNV が検出され、また XTR（X-chromosome-transposed region）領域[*13] を考慮した解析方法（※）では 2,065,505 個（PAR1+PAR2＋XTR、基準をクリアしたものは 1,750,054 個）の SNV が検出されました。

※解析方法の特色：X 染色体を 2 本持つ女性では常染色体と同じように二倍体[*14] ゲノムとしてのバリアントコール[*15] を行います。一方、X と Y 染色体それぞれ一本ずつ持つ男性では、相同性の高い領域のみ二倍体ゲノムとしてバリアントコールを行います。これは、本解析で採用した手法では、相同性の高い領域で X と Y の情報を明確に区別できないためです。本解析では、PAR（pseudoautosomal region）領域[*16] である PAR1、PAR2 を二倍体ゲノムとして取り扱う標準的な解析と、PAR1、PAR2 に加え、同様に相同性の高い XTR 領域を二倍体ゲノムとして取り扱う解析で、バリアントコールを行っています。どちらの解析でも、該当する領域では、Y 染色体のバリアント情報の一部が、X 染色体の情報として含まれる可能性があります。

●ミトコンドリア

ミトコンドリア DNA は、核ゲノムとは別に、16kbp の長さを持つ重要なヒトゲノム構成要素で、特にミトコンドリア病と総称される一群の疾患遺伝子の存在も知られています。また、組換えを起こさずに、母系系統を追跡できることから、人類進化研究においても重要な貢献が期待できる DNA です。

本解析では 2,483 個の SNV を発見して、パネルに収載しました。このうち ToMMo で新規に発見したものは 1,555 個であり、これらのアレル頻度情報からミトコンドリア DNA における疾患関連バリアントの高精度な絞り込みなどが可能となります。

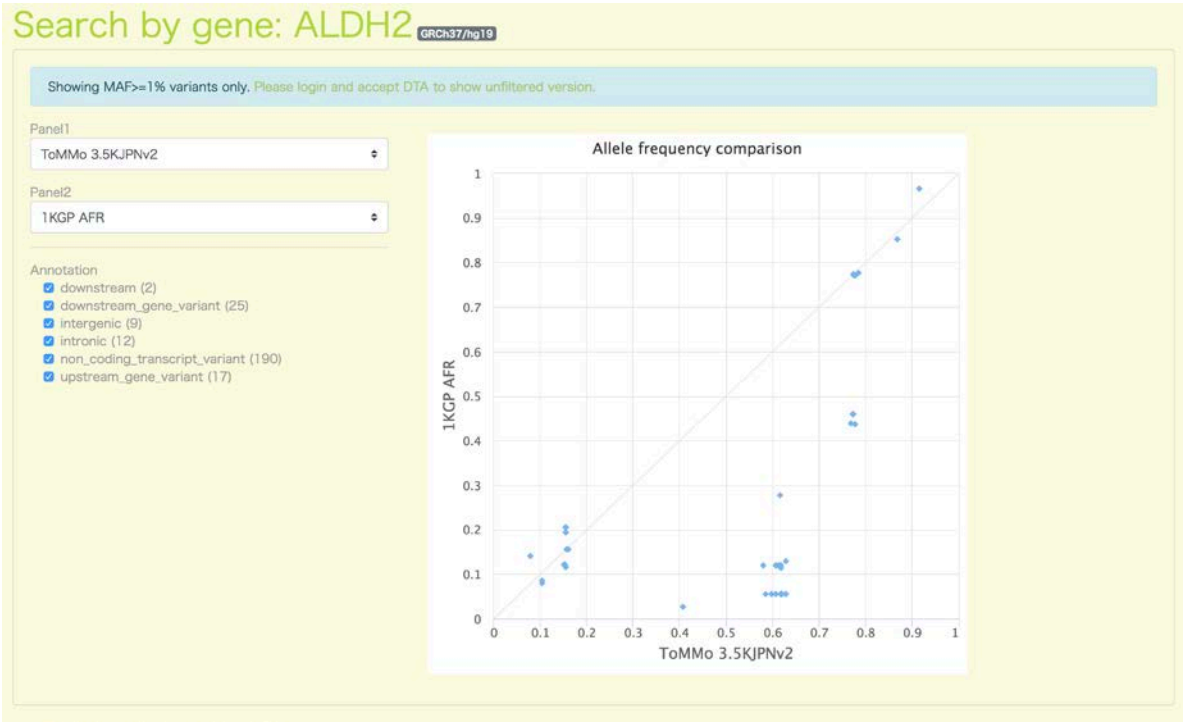
※解析方法の特色：ミトコンドリア DNA は環状ですが、ある 1 箇所を開始点・終点と考え 1 本の配列として解析します。この時切れ目部分におけるバイアスが懸念されますが、リファレンスゲノムを 10,000 塩基シフトした配列も考慮することでこの問題の解決を図りました。リファレンスゲノムには国際標準配列である revised Cambridge Reference Sequence を用いています。各検体の代表的なバリアントの検出を目的として一倍体ゲノムとみなしたバリアントコールを行い、公開データは VCF 形式による、解析結果からの SNV のアレル頻度情報となっています。

	バリアント数：収載数	バリアント数：新規発見数
常染色体	51,168,347	28,127,100
X 染色体	2,065,505 (PAR1+PAR2＋XTR)	1,281,108
ミトコンドリア	2,483	1,555

●パネルユーザーへの認証機能と詳細検索

今回の更新にあたり、ORCID[*17] と連携する認証機構を実装しました。認証されたユーザーは、ローカルの環境に大きなファイルをダウンロードすること

なしに検索機能を用いて、遺伝子ごとに各バリアントのアレル頻度を調べ、またそれを海外で公開されている代表的なデータベースサイト（国際 1000 人ゲノム）と比較できるようになりました。



jMorp 画面（二つの集団間のアレル頻度差を可視化）

【今後の展開】

今回 3.5KJPNv2 で初めて公開した X 染色体やミトコンドリアには、比較的発症頻度の高い遺伝性疾患に関連する遺伝子が多数含まれることから、全国の疾患研究者による利活用が想定されます。当計画では今後、より広範なデータシェアリングを進め、全国の研究者のゲノム研究を加速するとともに、さらなるゲノム情報の高精度化やより高度な情報解析を進めていきます。

また、日本の他の地域のコホート事業などとも連携を進め、東北メディカル・メガバンク計画の第二段階の目標である 8,000 人の全ゲノムリファレンスパネルの構築に向け、より日本人として代表性の高い網羅的なパネルとなるよう拡張をしていく予定です。

【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。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。

【用語等説明】

＊1. 全ゲノム解析：生物がもつ遺伝情報のセットをゲノムと呼ぶが、その全てを解析することを全ゲノム解析と呼ぶ。ヒトにおいては、一人あたり約 30 億塩基分 x2 セット持ち、その全ての情報を解析すること。

＊2. 性染色体：ヒトにおいてはヒトが持つ 23 対 46 本の染色体のうち、X 染色体と Y 染色体を性染色体と呼ぶ。原則として男性は X 染色体と Y 染色体を 1 本ずつ持ち、女性は X 染色体を 2 本持ち、性別により構成が異なる。

＊3. ミトコンドリア：細胞内において、エネルギー生産を担う小器官。独自の DNA を持つ。世代間においては、卵子を通じて継代される。

＊4. 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発症の関係を解明するための調査のこと。

＊5. バリアント：標準となる配列とは異なること。置換（塩基が置き換わっていること）、欠失（塩基がなくなっていること）、挿入（塩基が挿入され、増えていること）などがあるが、本リファレンスパネルにおいては現状では置換のみを搭載している。

＊6. GATK Best Practices：近年、国際的に標準的な解析手法となりつつある解析手法。米国の Broad Institute による全ゲノム解析をはじめ、最近の大規模なゲノム解析では、この手法を基にした解析を行う例が多々見られる。

＊7. 国際 1000 人ゲノムデータ：人類集団の広範かつ詳細な遺伝的多様性の収集を目指して、世界各地の約 1,000 人の全ゲノムシーケンスを行った国際研究計画のデータ。現在解析人数を 2,504 人にまで拡張した phase3 が完了している。

＊8. オミックス： 遺伝情報全てを指すゲノム（gen**ome**）、ゲノムから転写された情報全てを指すトランスクリプトーム（transcript**ome**）、タンパク質全てを指す（prote**ome**）など、それぞれの網羅的な情報を全体として指す言葉。東北メディカル・メガバンク計画では、ゲノムのみならず、メタボロームやプロテオーム、トランスクリプトーム、更にメチロームなどの解析を行っている。

＊9. メタボローム、プロテオーム：代謝産物全てを指すのがメタボロームで、タンパク質全てを指すのがプロテオームである。代謝産物とは、例えばアミノ酸、糖、脂質などで、比較的小さな分子群である。

＊10. アレル頻度：ある集団の DNA バリアントの塩基（A,T,G,C）の頻度をいう。今回は対象となった日本人 約 3,500 人中の頻度となる。

＊11. 常染色体：ヒトにおいてはヒトが持つ 23 対 46 本の染色体のうち、性別によらずに全てのヒトが持つ 1 番から 22 番染色体までの染色体を常染色体という。

＊12. 鎖長：DNA において塩基がつながっている個数（長さ）を指す。

＊13. XTR 領域：X-chromosome-transposed region、X 染色体転位性領域。X と Y 間で高い相同性を持つ長鎖長配列という点で PAR 領域に準じる領域であり、確立された観察ではないものの XY 間の組換えも報告されていることから、3 番目の PAR 領域と呼ばれることもある。
＊14. 二倍体：ヒトをはじめとするほ乳類などの脊椎動物や植物等で通常の個体において、同じ染色体は二つずつ存在するが、この二つずつ存在することを指す。
＊15. バリアントコール：ゲノム解析において解析対象の塩基配列中のどの位置にどのようなバリアントが発生しているかを特定すること。
＊16. PAR 領域：pseudoautosomal region、偽常染色体領域。性染色体において、X、Y で相同な配列をもつ領域。ここにある遺伝子は常染色体のように振る舞い X 染色体の不活化を受けないとされる。２つの PAR 領域では組換えが報告されている（PAR2 での報告はまだ確立していない）。XTR 領域も同じく XY 間で相同な領域であるが、組換えの報告が確立されておらず PAR 領域とは呼ばれないことがある。
＊17. ORCID：ORCID は、研究者はじめ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である。ORCID を用いることにより、個別の著者を同姓同名等であってもきちんと特定することができる。

【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
ゲノムプラットフォーム連携センター
センター長 木下賢吾（きのした けんご）
電話番号：022-274-5952
E メール：kengo@ecei.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
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：f-nagami@med.tohoku.ac.jp

（AMED 事業に関すること）
国立研究開発法人日本医療研究開発機構
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