令和4年度 厚生労働科学研究費補助金
（政策科学総合研究事業（臨床研究等 ICT 基盤構築・人口知能実装研究事業））
分担研究報告書

次世代バイオデータ基盤の構築に向けたデータパッケージ整備

データセットおよび倫理申請・データ利用申請のひな形のパッケージの整備

研究分担者　荻島　創一　　東北大学東北ﾒﾃﾞｨｶﾙ・ﾒｶﾞﾊﾞﾝｸ機構医療情報 ICT 部門　教授

研究要旨

次世代バイオデータ基盤の構築に向けて、令和元年度に実施した調査研究で取りまとめた「コホート・バイオバンクの産業利活用促進策」の具体化のために、コホート・バイオバンクの産業利用促進のための調査研究を踏まえてデータの利活用を促進するデータパッケージを整備した。データパッケージは企業からニーズを受けて、わかりやすいデータセットに加え、倫理申請ひな形及びデータ利用申請ひな形をセットとしたものである。このデータパッケージを整備することで企業のニーズに応え、データ利活用のハードルを下げる。さらにこのデータパッケージによりコホートの負担も軽減し、円滑な産学連携を実現する。

A. 研究目的

バイオ戦略では、「医療とヘルスケアが連携した末永く社会参加できる社会」の実現を目指し、「大規模統合コホート・バイオバンクの構築」のため、「健常人コホート等の実施主体が連携し、データを統合・強化する大規模健常人コホート・バイオバンクの構築」、すなわち、次世代バイオデータ基盤を構築することとされた。

このような政策的位置付けの中で、次世代バイオデータ基盤の構築に向けて、令和元年度に実施した調査研究で取りまとめた「コホート・バイオバンクの産業利活用促進策」をとりまとめた。そのなかで、産業界からのコホートの利活用にあたってのハードルの一つとして、企業向けのわかりやすいデータパッケージがないという課題が指摘された。

そこで、企業からニーズを受けて、わかりやすいデータセットに加え、倫理申請ひな形及びデータ利用申請ひな形をセットにしたデータパッケージを整備する。

B. 研究方法

1. データパッケージとデータ説明書の整備

コホート横断検索システムカタログについて、大規模なゲノムコホートのカタログの整備を進めるのに併せて、ゲノムデータのあるコホートデータのパッケージの整備を行う。

東北メディカル・メガバンク計画のゲノムデータのあるコホートデータのパッケージの整備を行う。このパッケージには、ゲノムデータのみならず、コホートの調査票、検査のデータ、メタボローム、プロテオームのデータのパッケージを整備する。また、データのパッケージにデータの説明書を用意する。データの説明書は、東北メディカル・メガバンクデータを初めて解析する立場から、必要な説明事項を洗い出して、とりまとめる。

2. 倫理申請及びデータ利用申請のひな形の整備

上記のデータのパッケージの利用にあたって必要な倫理申請及びデータ利用申請のひな形を整

備する。これについても、東北メディカル・メガバンクデータを初めて解析する立場から、必要な説明事項を洗い出して、とりまとめる。

（倫理面への配慮）

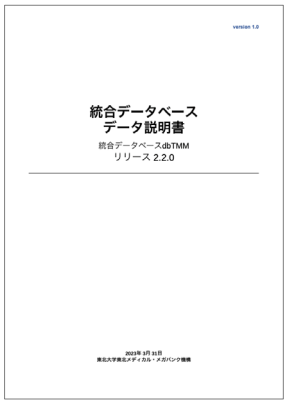
本研究はヒトゲノム・遺伝子解析、臨床研究、ヒトを対象とする医学系研究、動物実験等の実施はない。したがって倫理面の問題はないと判断した。

C. 研究結果

1. データパッケージとデータ説明書の整備

東北メディカル・メガバンク計画のゲノムデータのあるコホートデータとして、調査票、検査のデータ、メタボローム、プロテオームのデータのパッケージを整備した。

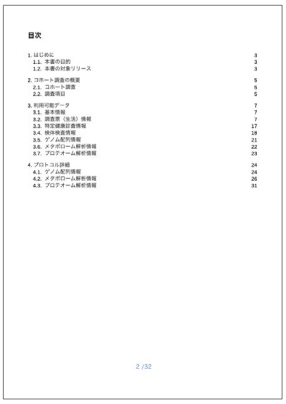
また、データのパッケージにデータの説明書を用意した。


version 1.0

統合データベース
データ説明書

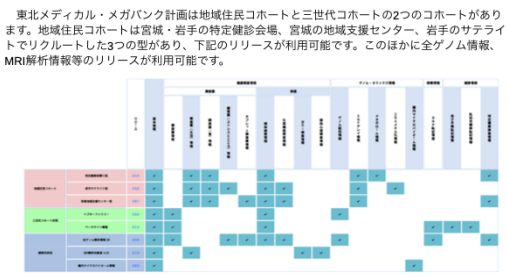
統合データベース dbTMM
リリース 2.2.0

2023年 3月 31日
東北大学東北メディカル・メガバンク機構


目次

2 /32

東北メディカル・メガバンクデータを初めて解析する立場から、東北メディカル・メガバンク計画で収集しているデータの全体像の説明から入って説明書を作成した。


東北メディカル・メガバンク計画は地域住民コホートと三世代コホートの2つのコホートがあります。地域住民コホートは宮城・岩手の特定健診会場、宮城の地域支援センター、岩手のサテライトでリクルートした3つの型があり、下記のリリースが利用可能です。このほかに全ゲノム情報、MRI解析情報等のリリースが利用可能です。

各リリースの掲載内容は上記表の通りです。リリースごとに項目などの詳細に違いがあります。
各データ項目の詳細については、以下をご参照下さい。

- リリース 2.3.3
  - 地域住民コホート特定健診相乗り型ベースライン調査 TMM 67K (2022年3月22日リリース)
- リリース 2.5.2
  - 地域住民コホート岩手サテライト型ベースライン調査 9.3K (2021年9月1日リリース)
- リリース 2.8.1
  - 地域住民コホート 宮城 地域支援センター型 ベースライン調査 18K (2020年6月30日リリース)
- リリース 2.4.0
  - 三世代コホートヘプタファミリー (TMM BirThree HeptaFamily) (2020年1月10日リリース)
- リリース 2.7.3
  - 三世代コホート ベースライン調査 73K (2021年11月25日リリース)

- 基本情報
  - 性別
  - 年齢
- 調査票 (生活) 情報　全員 / 男性 / 女性
  - 基本情報 **(13項目)**
  - 運動について **(19項目)**
  - 飲酒について **(24項目)**
  - 喫煙について **(12項目)**
  - ストレスについて **(7項目)**
  - 家族構成および健康状態について
    - 家族構成 **(12項目)**
    - 現在の健康状態 **(6項目)**
    - 輸血経験、生活習慣病の治療 **(8項目)**
    - 罹患歴 **(131項目)**
    - 家族の出生地などについて **(25項目)**
  - 体質について **(13項目)**
  - 仕事の状況について **(15項目)**
  - 睡眠について **(13項目)**
  - 人とのつながりについて **(11項目)**
  - 東日本大震災の記憶について **(5項目)**
  - 東日本大震災での被災と体験について **(9項目)**
  - うつについて **(20項目)**
  - 女性の健康について **(36項目)**
- 特定健康診査情報　全員 / 男性 / 女性
  - 検査データ **(22項目)**
- 検体検査情報　全員 / 男性 / 女性
  - 血液学検査 (血球計算、白血球分画等) **(13項目)**
  - 生化学的検査 **(13項目)**
  - 免疫学的検査 (アレルギー検査等) **(66項目)**
  - 尿検査 **(5項目)**
- ゲノム配列情報 **(1項目)**
- メタボローム解析情報 **(2項目)**

調査票については、解析する立場からの要望として、それぞれの設問の出典の説明を記載した。

3.2. 調査票（生活）情報

- 基本情報
  - 性別　出典：JPHC　ベースライン　I & II.
  - 年齢　出典：JPHC-NEXT　ベースライン
  - あなたの血液型は何型ですか　出典：JPHC-NEXT　ベースライン
  - 最終学歴　出典：JPHC-NEXT　ベースライン
  - 現在の身長　出典：JPHC　ベースライン　I & II.
  - 現在の体重　出典：JPHC　ベースライン　I & II.
  - 20歳ころのおよその体重　出典：JPHC　ベースライン　I & II.
  - 1年前の体重　出典：JPHC-NEXT　ベースライン
  - 出生時体重　出典：JPHC-NEXT　ベースライン
  - 震災時の被害状況　出典：ToMMo
  - 主に居住している場所　出典：ToMMo
  - 損壊した家屋やがれきを日常的に見る　出典：ToMMo
    現在も震災により損壊した家屋やがれきを、日常的に見ることがありますか。
  - 何回住居が変わりましたか　出典：ToMMo
    震災後、避難所を含めて何回住居が変わりましたか。

- 運動について
  - 身体の動かし方が変わる忙しい期間
    昨年１年間の「身体の動かし方」についておたずねします。農繁期など、１年のほかの時期に比べて、仕事時の「身体の動かし方」が大きく変わる特に忙しい期間がありましたか。　出典：JPHC　10年後　I & II.
  - ふだん1日の体を動かす時間の内訳をおたずねします。(計6項目)
    - 力作業の時間　出典：JPHC　5年後　I & II.
    - 歩いている時間　出典：JPHC　5年後　I & II.
    - 立っている時間　出典：JPHC　5年後　I & II.
    - すわっている時間　出典：JPHC　5年後　I & II.
    - 散歩　頻度　出典：JPHC　10年後　I & II.
    - 散歩　一回あたりの時間　出典：JPHC　10年後　I & II.
  - 余暇での「身体の動かし方」についておたずねします。昨年、次のことを行う頻度と1回当たりの時間はどのくらいでしたか。(計6項目)　出典：JPHC　10年後　I & II.
    - ウォーキング　頻度
    - ウォーキング　一回当たりの時間
    - 軽・中度の運動　頻度
    - 軽・中度の運動　一回当たりの時間
    - 激しい運動　頻度
    - 激しい運動　一回当たりの時間

これにより企業の利用者は他コホートと連携して環境曝露の要因の利活用が可能となり、東北メディカル・メガバンク計画固有の設問については震災の影響についての環境曝露の要因の利活用が可能となる。

検査については単位や最小値、最大値、データ

数を示して、企業が研究利用するにあたって必要な説明を記載した。

- 血液学的検査 (血球計算、白血球分画等)

項目名	単位	最小値	中央値	最大値	最頻値	データ数
白血球数	/μL	2,000	8,700	15,400	5,100	1,894
赤血球数	万/μL	270	465	660	455	1,894
血色素量	g/dL	8	13.6	19.2	13.4	1,894
ヘマトクリット値	%	28.5	41.75	55	42.25	1,894
平均赤血球容積	fL	66	91.5	117	95.5	1,894
平均赤血球色素量	pg	18	28	38	30.75	1,894
平均赤血球色素濃度	%	27.2	31.7	36.2	32.7	1,894
血小板数	万/μL	6	36.5	67	22.5	1,892
リンパ球	%	7	37.5	68	31.5	1,893
単球	%	1	7.5	14	4.9	1,893
好酸球	%	0-0.5	-	29	1.25	1,893
好塩基球	%	0-0.1	-	3	55	1,893
好中球	%	28	-	NA	56.5	1,823

ゲノムデータ、メタボローム、プロテオームデータについてはデータの取得方法、プラットホームを記載した。

3.5. ゲノム配列情報

- 概要
  - 日本人3,552人の全ゲノム解析の結果を元に構築した全ゲノムリファレンスパネル“3.5KJPNv2”です。
  - 3.5KJPNv2は常染色体・X染色体・ミトコンドリアDNAにおける一塩基変異及び挿入・欠失のアレル頻度情報を含むデータベースです。
  - データ解析には国際標準に準拠した手法を用いたため、海外の大規模ゲノム解析との比較がより容易なパネルです。
  - 3,315人分のデータを分譲可能です。
- データ取得方法
  - 日本国内の3,552人から取得したゲノム配列情報です。
  - 参加者のバフィーコートからゲノムDNAを抽出し、超音波処理でライブラリ調製した後、HiSeq 2500（Illumina）を用いて配列を決定しました。
  - 取得した FASTQファイルは BWA-MEM（ver. 0.7.12）を用いてヒト参照ゲノム配列（GRCh37）に対してアライメントをおこないました。
  - その後、GATK（ver. 3.7）のHaplotypeCallerを用いて変異の検出をおこないました。
  - 解析の詳細は 4. 技術詳細の項に記載します。
  - 参考論文
    - Tadaka S, Katsuoka F, Ueki M, *et al.* 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. Hum Genome Var. 2019; 6: 28. https://doi.org/10.1038/s41439-019-0059-5
- プラットフォーム
  - HiSeq 2500

また、それぞれのプロトコル詳細の説明について、東北メディカル・メガバンクデータを初めて解析する立場から、必要な説明をまとめた。

4.1. ゲノム配列情報

- サンプル
  - 日本国内の3,552 名
- 全ゲノムシーケンス
  - バフィーコートから抽出したゲノムDNAを超音波処理により平均550 bpに断片化しました。
  - シーケンスにはHiSeq 2500 system（Illumina）を用いました。
  - TruSeq Rapid PE Cluster V1 and SBS Kits（1サンプル/フローセル）およびTruSeq Rapid PE Cluster Kit V2 and SBS Kit（2サンプル/フローセル）をそれぞれ162bpペアエンド（162PE）および259bpペアエンド（259PE）プロトコルで使用しました。
- 全ゲノムre-sequencing
  - GATK Best Practices ワークフローに基づいています。
  - 参照ゲノムの選択、BQSR（Base Quality Score Recalibration）の使用、ジョイントジェノタイピングの3ステップをカスタマイズしました。
- アライメント
  - リファレンスゲノムには、GRCh37のhs37d5.fa（1000ゲノムプロジェクトPhase214で使用されたものと同じ参照配列）を用いました。
  - アライメントにはBWA-MEM v0.7.12 を用いました。
  - Picard v2.10.6のSortSamを用いて座標でソートしています。
  - 重複リードの削除にはMarkDuplicatesを用いました。
  - BQSRを組み込んだジェノタイピング結果とBQSRを組み込まない場合の結果の一致を確認しました。
- 常染色体におけるバリアントコールとジョイントジェノタイピング
  - バリアントコールにはGATK HaplotypeCaller（output: GVCF files）を用いました。
  - 複数サンプルのジョイントジェノタイピングにはGATK GenotypeGVCFsを用いました。
  - SNVとINDELのクオリティコントロールでは、Variant Quality Score Recalibration (VQSR) をおこないました。
- X染色体におけるバリアントコールとジョイントジェノタイピング
  - 女性のX染色体は、常染色体と同様の手順で実施しました。
  - 男性のX染色体とY染色体は、偽常染色体領域（PAR）と非PARリードの変異体判定を、異なる倍率の設定で実施しました。2種類のSNVコール手順を使用しました。
- ミトコンドリアDNAにおけるバリアント同定とジョイントジェノタイピング
  - 円形のミトコンドリア DNA 配列を、その中にブレークポイントを挿入することで、2 つの直線的な DNA 配列に変換しました。
  - ミトコンドリアゲノムにアライメントされたリードとマッピングされていないリードを抽出しました。
  - BWA-MEMを用いて、2 つの線形ミトコンドリアゲノム上に再アライメントしました。
  - バリアントコールにはGATK HaplotypeCallerを用いました。
- バリアントのアノテーション
  - snpEff v.4.3tを用いて、GENCODE release 28に基づくアノテーションをおこないました。
  - 各変異の rs番号はSnpSift v4.3tを用いて、dbSNP23 release 150に基づいて決定しました。

この説明書により、非常に詳細に解析データについて解析にあたって必要最小限度の情報を得ることができる。

2. 倫理申請及びデータ利用申請のひな形の整備

倫理申請のひな形、分譲申請のひな形についても整備した。

様式 004

東北メディカル・メガバンク機構 試料・情報分譲 研究計画書

東北大学 東北メディカル・メガバンク機構 機構長 殿
岩手医科大学 いわて東北メディカル・メガバンク機構 機構長 殿

西暦 20〇〇 年 〇 月 〇 日

【研究番号:20〇〇-〇〇〇〇】

	(ふりがな)	〇〇〇 〇〇〇
	申請者	〇〇 〇〇
	(ふりがな)	〇〇〇 〇〇〇
	研究責任者	〇〇 〇〇
申請者所属機関・法人	機関・法人名	ABC 大学 生命〇〇研究科
	部局・部署名	〇〇専攻 〇〇研究室
	職名	教授
	住所	宮城県〇〇市〇〇町〇丁目〇番〇号
	電話番号	〇〇-〇〇〇-8078
	FAX 番号	〇〇-〇〇〇-8079

利用者の視点でわかりにくいところについて、東北メディカル・メガバンクデータを初めて解析する立場からコメントを受けて、注釈をつけた。

D. 考察

企業向けのわかりやすいデータパッケージとして、データパッケージ、データ説明書、倫理申請ひな形及びデータ利用申請ひな形をセットにしたデータパッケージを整備した。

データ説明書は、今回は、東北メディカル・メガバンクデータを初めて解析する立場から、解析に必要な情報を洗い出して、それをもとに必要な記載をまとめた。利用者のバックグラウンドや利用目的などによって、必要な情報に幅があることが想定

された。

倫理申請ひな形及びデータ利用申請ひな形は、研究開発の体制や内容に大きく依存するため、作成は非常に難しかった。このひな形の作成を通じて、倫理申請及びデータ利用申請の利用者にとってわかりにくいところがわかってきたため、今後、改善に取り組んでいきたい。

E. 結論

企業向けのわかりやすいデータパッケージとして、東北メディカル・メガバンク計画のゲノムデータのあるコホートデータとして、調査票、検査のデータ、メタボローム、プロテオームのデータのパッケージを整備した。データ説明書、倫理申請ひな形及びデータ利用申請ひな形をセットにしたデータパッケージとして整備した。

F. 健康危険情報　なし

G. 研究発表

1. 論文発表　なし
2. 学会発表　なし

H. 知的財産権の出願・登録状況

1. 特許取得　なし
2. 実用新案登録　なし
3. その他　なし