第 1 回 SPARC Japan セミナー2020
「研究データ公開:フルオープンと制限公開の境界線」

ヒトデータの共有のための取り組み

三橋 信孝
(科学技術振興機構 バイオサイエンスデータベースセンター)

講演要旨

2013 年に運用を開始した NBDC ヒトデータベースは、生命科学や医療の分野で生み出されたヒトに関する様々なデータの共有を推進しています。個人単位で決定されたゲノム配列が主な格納データです。ヒトデータの共有では、データの使いやすさと研究参加者のプライバシー保護のバランスが非常に重要です。
NBDC ヒトデータベースでは、データ提供者は、研究参加者との間のインフォームドコンセント、データのプライバシーレベル、研究の進捗状況などによって、適切な共有方法を選択することができます。個人情報の保護や倫理面にも配慮しながら、データの利用を促進するためのこれまでの取り組みを紹介したいと思います。

三橋 信孝

国立研究開発法人 科学技術振興機構（JST）バイオサイエンスデータベースセンター（NBDC）研究員。東京大学大学院理学系研究科修士課程修了後、民間企業にてライフサイエンス分野のソフトウェア開発に従事。ライフサイエンスデータベースセンター（DBCLS）研究員を経て、2011年より現職。NBDCヒトデータベースの情報セキュリティポリシー策定・情報システムを担当。NBDC、DBCLSの研究者と共同で、日本人ゲノム変異の統合データベース「TogoVar」の開発・運用に取り組んでいる。

機関紹介

バイオサイエンスデータベースセンター（NBDC）は科学技術振興機構（JST）の一組織です。ライフサイエンス研究の成果が最大限活用されるように、オープンサイエンスという方針の下で、ライフサイエンス分野のデータを統合して利活用を促進するセンターとして 2011 年に設立されました。私どもは以下の三つのテーマで活動を行っています。一つは「ひろく」ということで、データを共有するためのルールやガイドラインを作成してデータを集めることです。二つ目は「つなげて」で、RDF というフォーマットで集めたデータの統合を行っています。その他、辞書を作ったり、さまざまな統合をするための技術開発を行ったりしています。三つ目は「つかう」ということで、統合したデータを使いやすくするためのアプリの開発、データ駆動型研究に向けての研究などを行っています。

NBDC の Web サイトでは、統合データベースとそれを使うためのツールが誰でも無料で利用できるようになっています（図 1）。共同研究先のライフサイエンス統合データベースやファンディングプロジェクトで作られたサービス、データベースなども利用できます。TogoTV という動画による解説がありますので、ご覧いただければと思います。

NBDC のサービスとヒトデータベース

サービスの一覧のところに、今日紹介させていただく NBDC ヒトデータベースやバリアントのデータベースがあります（図 2）。なぜヒト由来のデータを集

めて共有する必要があるのか。人間のゲノムにはごくわずかな個人差があります。このゲノムの個人差が、病気のかかりやすさ、薬の効きやすさ、治療率などの個人差につながっていきます。このようなゲノムの個人差を調べることができるようになったのは、ヒトのゲノムの塩基配列、DNA の配列を決定するコストが非常に下がったからです。今、1 人当たり約数百ドルで配列決定ができるといわれています。

こうしてゲノムの個人差と疾患の関係等が分かると、ゲノムの違いによって治療方針を決めることができる個別化医療（プレシジョンメディスン）が実現するといわれています。そのためには、より多くの個人の医療情報とセットになったゲノム情報を蓄積したデータベースの構築が必要になってきます。

欧米では、米国国立衛生研究所（NIH）の The database of Genotypes and Phenotypes（dbGaP）や、ヨーロッパの European Genome-phenome Archive（EGA）という公共データベースが構築されました。そのため、ヨーロピアンのデータが圧倒的に多いです。ゲノムにももちろん人種差がありますので、日本人のデータベースを作ることは非常に重要です。日本でも疾患に関する知見を得るためのコホート研究を、バイオバンク・ジャパンや東北メディカル・メガバンクが数千人から十万人を集めて行っています。そういった研究から出たゲノムのデータや医療データなどを共有するための受け皿として、ヒトデータベースが作られました。

NBDC のヒトデータベースは、国立遺伝学研究所の DDBJ センターが運営する Japanese Genotype-phenotype Archive（JGA）と共同運営しています。NBDC が、制限公開をするためのガイドラインの作成や、制限公開時にはデータ提供や利用をするための審査を実施しています。一方、DDBJ は、データの管理やユーザーへの払い出しなどの業務を行っています。

公開からちょうど 7 年たちますが、300 件程度の申請があり、追加更新なども行われています。約 32 万人のデータが格納されていて、そのうち約 150 件が公開されており、約 140 件の利用申請があります。

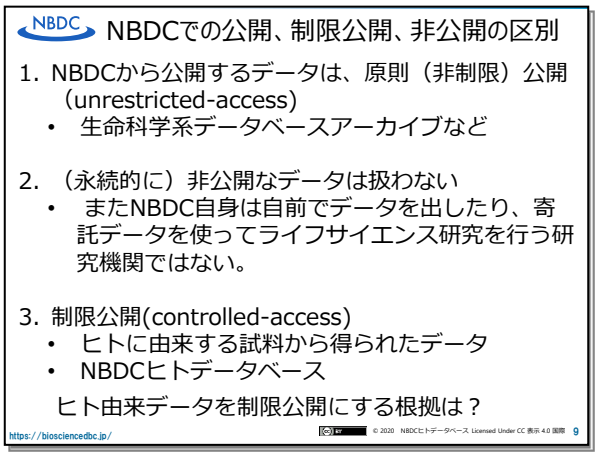
（図 1）

（図 2）

公開、制限公開、非公開の区別

ここからは、セミナー概要で問い掛けられていた四つのテーマに沿ってヒトデータベースの内容を紹介させていただきます。

まず、データの区分です（図 3）。NBDC では、ヒ



（図 3）

ト以外のデータも扱っていますが、基本的には非制限公開（フルオープン）です。生命科学系データベースアーカイブなどを、クリエイティブコモンズなどのライセンスで公開しています。NBDC 自体は、論文発表前のデータなど一時的に非公開のデータを預かることはありますが、永続的に非公開なデータを扱うことはありません。また、NBDC 自体が NBDC データベースのデータを使ってライフサイエンス研究そのものをすることもありません。中立的な立場でデータベースを運営しています。

その中で、やはりヒト由来のデータだけはどうしても制限公開（controlled-access）で扱わなければいけません。その根拠が、ヒト由来試料を対象にした医学研究を実施する上での医学指針やゲノム指針など、各種の倫理指針です（図 4）。これに沿って研究なりデータ共有なりをしなければなりません。また、これらの指針は個人情報保護法に裏付けられています。

ゲノム指針は 2001 年に策定されました（図 5）。基本方針の 5 番目、「個人の人権の保障の科学的又は社会的利益に対する優先」というところが一番重要ではないかと思います。そういった精神に沿って制度設計がされています。

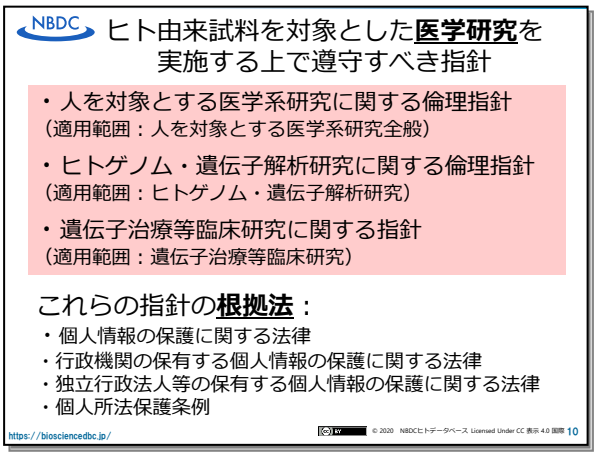
また、個人情報保護法も、完全施行されたのが 2005 年ですが、2017 年に改正されました。個人情報の定義が明確化され、ヒト由来データに関しても、その枠組みで個人情報として扱われることになりました。

では、具体的に何が個人情報になるのかというと、まず、ゲノム研究で扱うレベルの DNA の配列情報は、ほぼ全て個人識別符号という個人情報に位置付けられました。また、病歴や投薬情報といった医療関係データは、さらに厳格な管理が求められる要配慮個人情報に該当することになりました。これらの定義は法律や政令にきちんと定義されています。要配慮個人情報になると何が大変かというと、データを第三者に提供するためにあらかじめ同意を得なければいけない、いわゆるオプトインの手続きを経ないといけないことです。

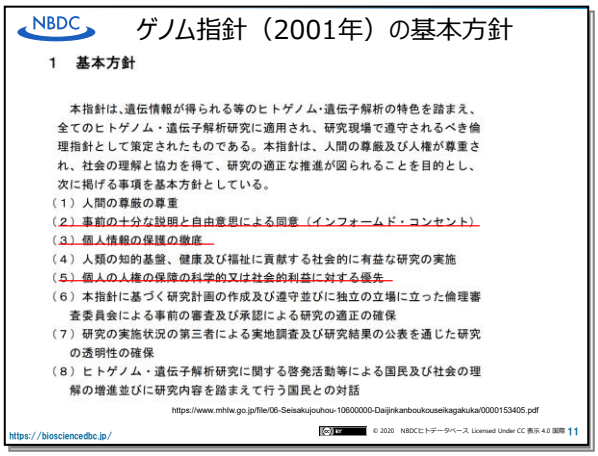
ステークホルダーとの取り決め

2 番目に、データ所持者、データ利用者等のステークホルダーとどのような取り決めを行っているか。ヒトのデータですので、データ提供者、データ利用者の他に、研究参加者（試料提供者）もステークホルダーになります。データ提供者が研究のために研究参加者から試料をもらうときには、どういう研究に使うかという目的を説明した同意説明文書が必ずあり、研究参加者はその文書にサインをしているはずであるという取り決めがあります。また、データベースセンターである NBDC とデータ提供者や利用者との間の取り決めとしては、NBDC が作ったヒトデータ共有ガイドラインがあり、これにのっとって運営しています。

ガイドラインの中には、データ提供時に提供者が守らなければいけない必要事項が書かれています。まず、データ提供者が実施した研究がその組織の倫理審査委員会で承認されているということを、研究計画書、承



（図 4）



（図 5）

認書とともに示していただく必要があります。そして、もちろん研究参加者からインフォームドコンセントを取っていて、なおかつ、そのインフォームドコンセントの中にデータベースでデータを共有するということが明記されていることを必ず確認しています。これに同意がないものはヒトデータベースで扱うことができません。最後に、例えば特定の疾患に関する研究にしかこのデータを利用できないなど、インフォームドコンセントに基づいた制限事項があれば、それも申請していただくことになっています。

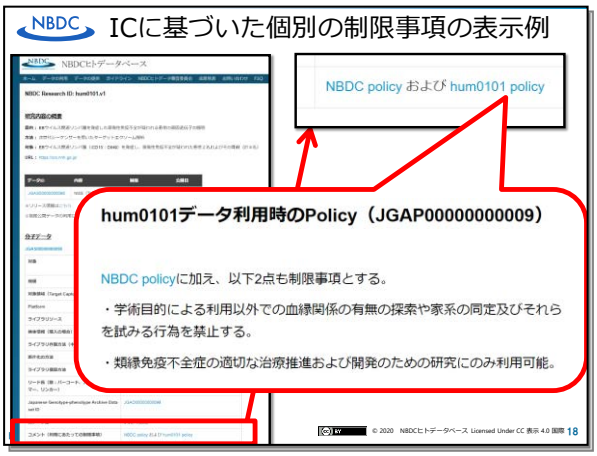
同じように、データ利用時の必要条項もガイドラインで決めています。データを利用する研究も倫理審査委員会の承認を得なければいけません。また、研究を行う方はそのデータを使った分野での研究経験があるということで、論文等の書誌情報を求めています。あとは、利用するインフラ、IT 環境できちんとセキュリティ対策が行われていることをガイドラインのチェックリストを使って確認しています。

図 6 は、インフォームドコンセントに基づいた個別の制限事項の表示例です。左にあるのが各データの内容を表示する概要ページです。どのようなデータに関してもこういった概要データは公開しています。その中にポリシーのリンクがあります。NBDC ポリシーはガイドライン標準のポリシーです。隣に hum0101policy というリンクがあるのは、このデータセットに特有の制限事項を付けることができるという例です。

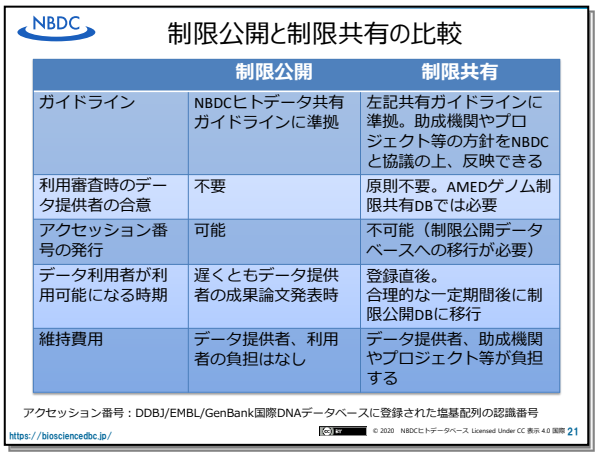
制限公開と制限共有の違い

では、制限公開と、特定の研究者・研究機関内のみに限定して研究データを提供する制限共有との違いは何でしょうか。NBDC ヒトデータベースにも制限共有という仕組みがあります。これはデータの公開に先駆けてグループの中で共有する仕組みです。将来的には制限公開系のデータベースに入れる、もしくは非制限公開することが条件になったデータを、一時的に制限共有する枠組みになっています。現在、NBDC グループ共有データベースを運用しており、国立研究開発法人日本医療研究開発機構（AMED）が実施する三つのプロジェクトから出てきたデータが制限共有として格納されています。もちろん、AMED 以外のグループもグループ共有データベースを作る枠組みがあります。

図 7 の表は、制限公開と制限共有の違いをまとめたものです。まずガイドラインは、制限共有の場合も NBDC ヒトデータ共有ガイドラインに準拠していただきますが、AMED ゲノム制限共有データベース（AGD）の場合は AMED の方針でガイドラインを多少変更することができ、データ提供者が合意した場合のみデータが利用できるというように変更されています。ですので、AGD のデータを利用したいユーザーは、あらかじめデータ提供者の合意を得て NBDC に利用申請していただかないといけません。その代わり、制限共有の場合は、国際 DNA データベースに登録する際に発行されるアクセッション番号が発行されませ



（図 6）



（図 7）

ん。論文発表するためには、データを使った場合には基本的にこのアクセッション番号を書かないといけないので、きちんとした論文発表ができないことがあります。その場合は制限公開に移行していただく必要があります。これはデータ提供者だけではなく利用者も同じ条件になりますので、利用者にとってもやや制限の多い共有となっています。

制限公開や制限共有を行う機関の実態

次に、データ提供者との関係性についてです。NBDC 自体は、データを作り出す研究の助成はしていません。米国の dbGaP というデータベースを運営する NIH は巨大なファンディング機関ですので、制度上、データ公開を義務化することができますが、われわれはそういうことができません。

その代わり、データ公開を促進する工夫として、まず日本の助成機関である AMED と連携しています。具体的には、AMED 研究データ利活用に係るガイドラインの中に、AMED の研究では必ずデータマネジメントプランを作るようにして、その中にデータ公開予定を記載することを義務化しました。その受け入れ先として AGD やヒトデータベースが含まれています。

2 番目に、アクセッション番号を取ることが論文受理の条件となっています。これが制限公開を促進する最も大きなドライビングフォースだと思っています。

3 番目に、インフォームドコンセントに基づいた個別の制限事項を受け付けます。これによって、再同意を取らなければデータベースに寄託できないものに関しても、なるべく制限事項も一緒に取り込むことで公開のハードルを下げる工夫をしています。

図 8 の図が制限公開共有までの流れです。データ審査に 2 週間ほどかかり、専用サイトを作り、JGA にデータを登録するとアクセッション番号が発行されるという仕組みになっています。

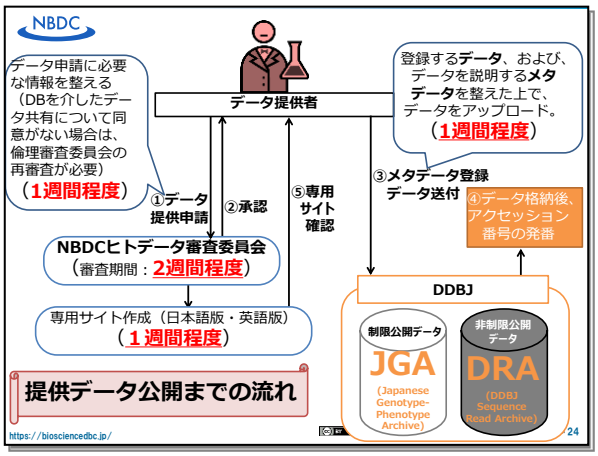
一方、データ利用者に関しては、ひたすらヒトデータを使いやすい環境を整備してデータ利用を促進しています。これによって制限公開等の必要性も高まるのではないかと考えています。方法としては、まず申請手続きの効率化のために Web システム化を行っており、その運用が始まっています。

2 番目に、ダウンロードしたデータを解析する必要がありますが、そのためにセキュアな計算環境を整えるにはかなりの手間がかかります。そこで利用者が所属する機関内に自前で計算機を用意することを求めていたガイドラインを改訂し、自組織以外に利用可能な解析用の計算環境を提供することにしました。現時点では JGA を運用している DDBJ と東北メディカル・メガバンクという、どちらもスパコンを持っているところで、非常に潤沢な解析環境が利用可能です。これを機関外サーバと呼んでいます。

3 番目としては、制限公開データの概要を把握できるように、ゲノムの個人差をあらかじめ検出して、ヒトデータベース内での個人差の頻度を非制限公開として出すことによって、データの中身を少しでも把握してもらおうとしています。

図 9 が利用の流れです。申請していただいて、2 週間ほどの審査の後、JGA からダウンロードできるようになります。図 10 は機関外サーバの仕組みです。以前は自組織のサーバ以外は使えませんでしたが、現在は自組織のサーバも、DDBJ や東北メディカルのスパコンも使えるようになっています。

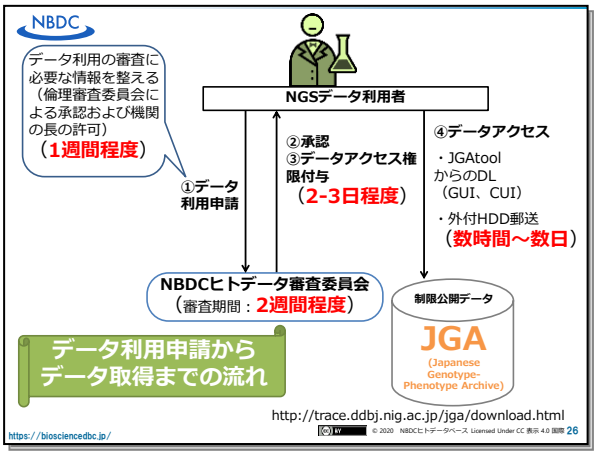
図 11 はヒトデータベースのデータセットの一覧ですが、データの中身がこれだけだと把握しにくいということで、重要なバリアントの部分だけを取り出し、



（図 8）

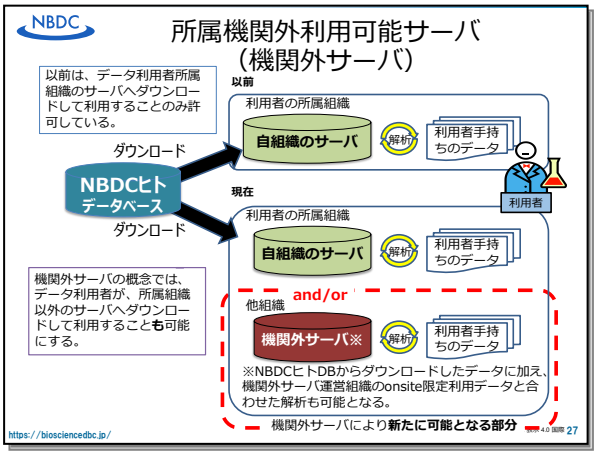
頻度という統計情報に加工して非制限公開で閲覧できるようにしたものが TogoVar というデータベースです（図 12)。左側にあるのが、ヒトデータベースに格納されている、それぞれのプロジェクトから出てきたゲノムのデータです。これを同じ手法で再解析して、日本人という大きな集団でのバリアント頻度を計算します。これは非制限公開できますので、ユーザーが見て、ヒトデータベースに利用申請するかどうかを決めることができます。

同じく、研究に必要なさまざまな他の集団での頻度データや文献データなども、ワンストップで解析できるようにして利便性を上げています。

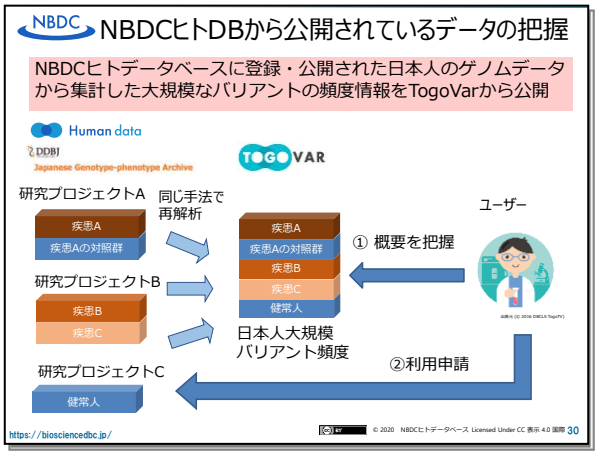


（図 9）

（図 11）



（図 10）



（図 12）