誰がためにTogoVarはある

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 研究員

本シンポジウムのテーマである「つないで使う」は、データベース統合プロジェクトにとって数年来の課題であり、前回のトーゴーの日でのバイオサイエンスデータベースセンター（NBDC）高木センター長の発表においても、データを利活用して新たな発見につなげるために、データベース利用者の意見をこれまで以上に収集・反映していく方針が示されている。この方針を具体化する一環として、TogoVar（日本人ゲノム多様性統合データベース）をライフサイエンス統合データベースセンター（DBCLS）と共同で開発し、2018年6月に公開した。多種多様なデータベースに散在して収録されてきたgenotypeやphenotypeに関連する情報を整理・統合し、バリアントを解釈するための情報をワンストップでわかりやすく提供することを目指している。

開発の背景には、NBDCが2013年から国立遺伝学研究所DDBJセンターと連携して運用している「NBDCヒトデータベース」に登録されているデータの利用を促進するために、個人由来のゲノムデータを個人特定の問題がないバリアント頻度情報に加工して制限なく公開することでデータ概要を提示し、データセットにたどり着き易くしたいという要望があった。また、バリアントデータベースや関連文献が散在しているために、それらのデータを研究者が各々収集・整理せざるを得ない状況も目の当たりにしていた。NGSデータから大量に産出される新規バリアントにdbSNP rs番号のようなIDが付与されていない状況もデータ統合が進まない要因であった。

開発段階から第三者の意見を収集しているが、複数のバリアントデータベースのアレル頻度を一度に比較可能であることなどは一定の評価を得ている。一方で、課題も多く寄せられ、エンドユーザーの痒いところに手が届くためには、まだまだ改良する必要がある。そのひとつとしてphenotypeでの検索が可能なPubCaseFinderとの連携も検討を開始している。ワークショップ「日本人ゲノム多様性統合データベースTogoVarを使ってみる」では、エンドユーザーだけでなくデータベース開発者からも忌憚のない意見をいただくことで、今後のTogoVarの改良につなげていきたいと考えている。

それでもこの時期にTogoVarを公開したことは大変意義があったと感じている。日本人NGSデータは臨床データなどと共に、バイオバンク・ジャパンや東北メディカル・メガバンク機構などにおいて数千人規模で集められているが、これらのプロジェクトの関係者からTogoVarに注目していただき、NBDCも参画してデータベース間の連携やデータの集約が具体化しようとしているからである。国外ではプロジェクト間の連携が進んでおり、例えば、米国ブロード研究所が主導するGenome Aggregation Database（gnomAD）に、40以上のプロジェクトから収集した13万人を超える個人由来のNGSデータを再解析して集計した頻度情報が公開されている。プロジェクトの垣根を超えて1万人を超えるデータを集約する日本版のgnomADの実現は待ったなしである。

そして何よりも、開発計画から1年以内での公開が実現したのは、NBDCやDBCLSを中心にデータベース統合プロジェクトで整備してきた基盤技術を「つないで使う」ことで、アプリケーション固有の開発に集中できたからである。例えば、主要なバリアントデータベースであるExome Aggregation Consortium（ExAC）やClinVarは開発開始時点でRDFデータが利用可能であったし、可視化に関してもTogoStanzaを利用することができた。

今後もTogoVarの開発を通じて、データ産出者、基盤技術開発者、研究プロジェクト運営管理者が「つながり」、エンドユーザーのために「使える」データを提供できるように改良を重ねていきたい。