研究開発課題別事後評価結果

１．研究開発課題名

ヒトゲノムバリエーションデータベースの開発

２．代表研究者名

東京大学大学院医学系研究科　教授　徳永勝士

３．研究実施概要

本課題では、変異・疾患・臨床情報を整理・体系化し、成果・情報を俯瞰可能とすると共に、健常者のゲノム多様性情報を提供することを目的とし、平成 18～22 年度の統合データベースプロジェクトにおいて構築したゲノムワイドな関連解析（Genome wide association study：GWAS）「GWAS-DB」を引き続き運用して新たな成果を受け入れると共に、データベースの高機能化を実施した。次世代シークエンサ（NGS）を含む様々な実験手法によって発見される疾患関連変異の登録・管理を行うための「Human genome variation DB」を構築し、パスウェイデータなどのオミックス情報とも連携させ変異と疾患の関係を体系化させた。

１）次世代型シークエンサー用多型・変異データベースの構築と関連計算手法の開発

次世代シークエンサー用バリエーションデータベースを構築し、Single Nucleotide Polymorphism (SNP)のみならず、構造多型も登録できるようにし、HLA のように、ハプロタイプでないと意義が乏しい領域については、ハプロタイプとして登録できるように工夫し、新たな「HLA DB」を構築した。各種プロジェクトで産出されたデータの収集と整理、登録に当たり、各種データの変異をコールする条件について詳細な検討を行い、また、NGS による解析手法の差によるクオリティの差がないように品質管理を行う解析手法の検討を行った。さらに、集団によって標準となる参照配列が異なることから、複数の集団の参照配列を登録できるようにするとともに、日本人健常者の標準ゲノム構築を行うための計算手法の開発を行い、1000 人ゲノムなどの公開データを用いて標準ゲノムを構築した。

２）次世代型シークエンサー用以外の実験手法に関する多型・変異データベースの構築と関連計算手法の開発

次世代シークエンサー以外で発見した変異データを登録するためのバリエーションデータベースを構築し、文献情報を含めデータを収集するとともに、次世代シークエンサー用データベース,「GWAS-DB」と共に同時検索可能とした。具体的には、30,000 件以上の変異と疾患の情報、およびそれに付随する情報（人種、検体数、予後など）を文献から抽出して登録した。データは基本的には、疾患オントロジーなど世の中で広く使われている疾患

名称などに標準化している標記に沿って登録した。

３）知識型データベースの開発

「Human genome variation DB」に対して、本プロジェクトに搭載されるその他のオミックスデータとの連携を深め、特定の locus 上で変異が起きた時のオッズ比、特定のパスウェイ上での変異が起きた時のオッズ比などを計算できるようにするとともに、同一遺伝子でも特定の位置での変異によって予後が大きく異なる要因が予測できるような知識型データベースへと発展させるなど、パスウェイ情報や種間での保存性を鑑みた形で変異を解釈できるような知識型データベースとしての機能追加を行った。

４）GWAS 関連データベースの開発

GWAS関連データベースであるSNP、GWAS、CNVについては、前プロジェクトからの継続開発として、データの受け入れと再配布の運用、メタ解析やインピュテーション（予測）機能の追加、及び、新プラットフォームデータへの対応などや、ユーザーフレンドリーで汎用性のあるメタ解析プログラムを開発した。

４．事後評価結果

４－１．当初計画の達成度

本課題は、１）次世代型シークエンサー用多型・変異データベースの構築と関連計算手法の開発、２）次世代型シークエンサー用以外の実験手法に関する多型・変異データベースの構築と関連計算手法の開発、３）知識型データベースの開発、４）GWAS 関連データベースの開発という当初の研究計画を達成した。対象がヒトデータであるため、データの RDF 化は難しい状況にあるが、健常者のゲノム多様性情報を提供のため、パスウェイデータなどのオミックス情報とも連携させ変異と疾患の関係などのデータの統合化を進めた。

４－２．研究開発成果の公開および利用の状況等

開発した「Human genome Variation DB」は公開されており、月間ユニーク IP アクセス数は平均 700 件程度で、利用者数がやや少ないといえる。

４－３．研究開発成果によるライフサイエンス分野のデータ活用への波及効果

本課題は、日本人の SNP、ゲノム配列などゲノムバリエーションと関連の疾患情報が統合されており、疾患解析から医療応用研究まで利用可能である。今後更なるデータ集積と共に疾患や薬剤感受性とゲノムバリエーションとの相関解析が一層進捗することが期待される。

４－４．広報・アウトリーチ活動等

論文発表、学会講演・発表などで研究成果を積極的に周知するとともに、複数の展示会において成果データベースのデモを実施するなど積極的に活動したことが評価できる。

５．総合評価

本課題は従来のデータベースへのデータ蓄積だけでなく、新規に NGS のバリエーションデータベース、「HLA DB」 を構築し、集団データの登録も実現し、利用者視点に沿ったユーザービリティの向上等も図っている。本データベースがさらに充実したものになるには、確実に研究コミュニティからの研究者にデータサブミッションをしてもらう方策を講じることが急務と思われるが、現在国内外の研究コミュニティとの連携において各種方策が講じされつつある。また、日本人データに留まらず他人種のデータ等との連携も含め、データベースの統合化をさらに推進することが望まれる。今後、データが十分な蓄積となれば、ヒトの疾患遺伝子の探索に有効であるとともに、医学を含むライフサイエンス研究開発の重要な研究基盤になると考えられ、一層の推進が望まれる。