平成 23 年度　研究開発実施報告書

ライフサイエンスデータベース統合推進事業「統合化推進プログラム」
平成 23 年度採択　研究代表者

松田文彦
京都大学大学院医学研究科　附属ゲノム医学センター　センター長・教授

大規模ゲノム疫学研究の統合情報基盤の構築

§1. 研究実施体制

(1)「松田」グループ
① 研究代表者:松田　文彦 （京都大学大学院医学研究科附属ゲノム医学センター 教授）
② 研究項目
・大規模ゲノム疫学研究の統合情報基盤の構築
・メタデータの開発保守
・情報格納とハンドリング効率化
・EHR システムとの連携
・統計手法の開発研究

(2)「佐藤」グループ
① 主たる共同研究者:佐藤　孝明 （(株)島津製作所　基盤技術研究所　ライフサイエンス研究所　所長）
② 研究項目
・網羅的メタボローム解析データの定量化手法の確立

§2. 研究実施内容

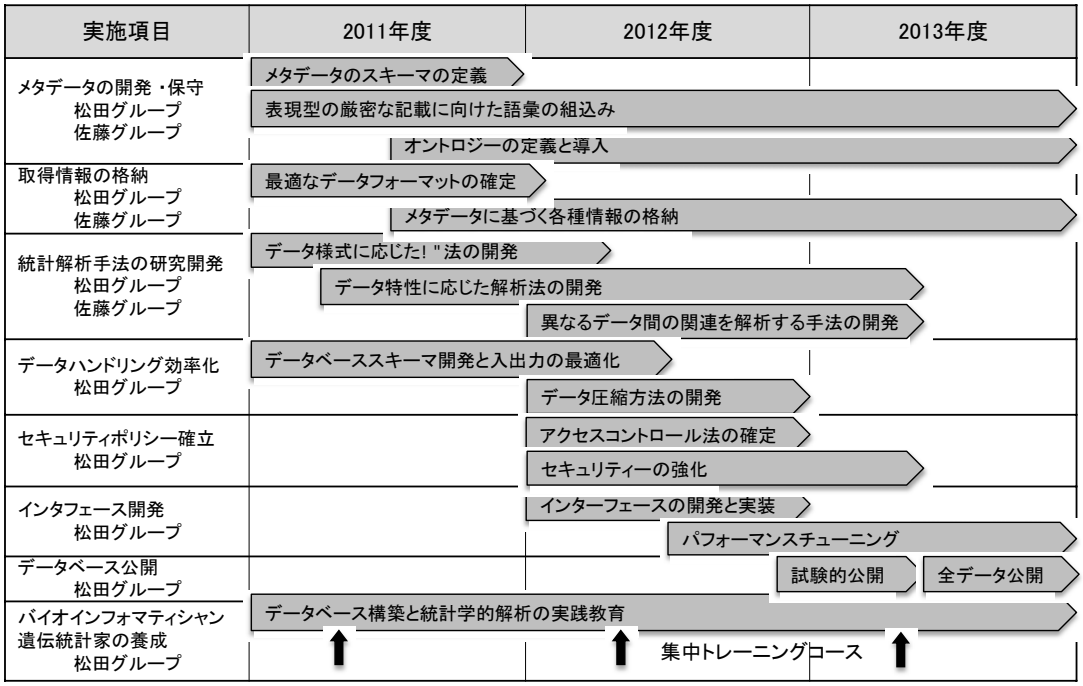
予防医学に関わるヒト疾患研究に供される日本人の詳細かつ網羅的なデータは、統合データベースセンターが提供する情報基盤の中でもきわめて大きな価値を持つものである。日本人集団を用いたゲノム疫学解析で得られた情報を集約し、標準化・一元管理のうえ研究者に公開し、それらを用いた解析で得られるオリジナリティーの高い研究成果を加えて世界へ発信することで、統合データベースセンターは次世代の予防医学研究において世界をリードするデータセンターとしての確固たる地位を占めることが可能となる。そこで、本研究開発では以下の 5 点を目標として設定した。

1. ゲノムコホート研究で網羅的に収集された一万人の生活習慣・環境情報、臨床情報、ゲノム情報を標準化し、データベースを構築することで、ゲノム疫学研究の情報基盤の整備をおこなう。
2. それらの情報を、セキュリティーの強化により個人情報の漏洩に最大限の注意をはらい、医学・生命科学研究者の研究に供するかたちで公開する
3. これをモデルケースとして、同様の研究をおこなう際に即時活用可能なかたちでデータベースの枠組みを提供し、他のゲノム疫学研究で蓄積された遺伝型・表現型データを標準化した後に連結、共有することで、個別の研究で得られた情報の一元化によるそれらの再利用を促す。
4. バイオインフォマティクス、遺伝統計学の若手研究者に研究現場での実務を通した教育訓練（OJT）をおこなうことで、我が国で手薄なこれらの分野の、研究の中心となりうる専門家の育成をはかる。
5. 将来的に、国民一人ひとりが自身の医療情報を持つ「パーソナルヘルスレコード（PHR）」の情報提供先として機能できるような、汎用性の高い健康情報管理システムのプロトタイプを提案する。

本年度は、開始時に設定したロードマップ（図 1）に従って、以下の項目を実施した。

1） メタデータの開発保守
2） 情報格納とハンドリングの効率化
3） EHR システムとの連携
4） 統計手法の開発研究
5） バイオインフォマティシャン・遺伝統計家の養成

図 1　本研究開発のロードマップ



1)　メタデータの開発保守

データベースに集約する各項目（データ項目）は、通常、各研究において任意に定義されるため、異なる研究間のデータ項目を統合することは難しい。例えば、単位や桁の違い、カテゴリ値のコードの違い、各カテゴリのクライテリアの違い、検査における実験プロトコルの違いなどである。そこで、異なる研究を統合することを目的に、統一的な基準でデータを収集するための基盤として、まずは、各データ項目のデータ型を決定し、それに付随する制限（制約）を定義するための枠組みと Web インタフェースを構築した（表1）。

表 1. データ型とその制約

データ型	制約	例
連続値	最大最小、打切り値の可否	バイオマーカー
順序有カテゴリ	カテゴリの値とコード、その順	質問票
順序なしカテゴリ	カテゴリの値とコード、取りうる値の個数	質問票
SNP 情報(diplotype)	ゲノム上の位置とアレル	SNP
その他の多型情報	ゲノム上の始終点	CNV
他次元	次元数、各次元の最大最小	中間形質
文字列・日付	最大文字数・日付のフォーマット	自由記載情報

データベースは、研究者が独自に定義した、統合を念頭にいれないデータ項目を受け入れることも可能であるが、プロジェクト横断的なデータの統合は、特定のキュレーターが“データ項目の門番”としての役割を担い、登録の際に標準化を進め、さらに、複数のプロジェクトでの再利用を推進することによって、はじめて可能となる。現在、ICD10 や MedDRA のような既存のオントロジーを活用して、データ項目間の意味的な関連付けを行うことで、類似のデータ項目の統合を進めてい

る。また、各項目のデータ数を増やすことで、それらの事実上の標準化を図りたいと考えている。

2) 情報格納とハンドリング効率化

実験を伴う分析・解析結果から得られる生データを、最小限の加工でデータベース登録可能なパイプラインを構築した [1]。例えば、次世代シークエンサーのパイプラインではマッピングから多型同定、データベースへの格納と解析可能なデータ形式での出力を一貫して行った。また、臨床情報に関しては、「1」メタデータの開発保守」で述べた方法に基づき、厳密な定義のもと収集された情報を、その情報の生成時点(もしくは期間)とともに登録するためのフォーマットを策定し、Web 経由でデータベースへ格納できるようにした。臨床情報は文字列として記録されている項目が非常に多く解析には適さないが、変換前後の値対応表(文字列においてはディクショナリ、数値やカテゴリ値においてはルールと呼ぶ)によって、それらをカテゴリ型や連続値型のような容易に解析できるデータ型に変換する機能を実装した。このほかのオミックスデータ、すなわち、質量分析やマイクロアレイによるオミックスデータは、正規化や加工前の実験データをバイナリ、および、テキスト形式でデータベースに格納できるよう、各実験プラットフォーム別にデータ形式や、そのプラットフォームを定義する特有な情報について洗い出しを行った。佐藤グループは GCMS を用いた網羅的メタボローム解析における定量化の方法を開発し、1,000 検体を超える質量分析を実施した。本データに基づき、クオリティコントロールやデータフォーマットを策定し、パイプラインを実装した。

3) EHR システムとの連携

ある希少性難病をモデルケースに、京都地域連携医療推進協議会が提供する EHR システム、「まいこねっと」との連携を実施し、病院のカルテ情報から追跡情報を収集する仕組みを構築した。定義された 131 項目について、EHR システムからの取得可能性を検討した(表 2)。

表 2. EHR システムからの情報取得可能性の例

	取得可能性	%
a)	1 対 1 で取得可能	16%
b)	ディクショナリやルールによって、システム的に取得可能	18%
c)	自由記載情報から、データキュレーターが人力で値を定義することが可能	22%
d)	EHR システムに移行されていない情報	32%
e)	電子カルテ上にも記録されていない情報	12%

c)に対しては、特定のキーワードで該当期間内のカルテを全文検索し、抽出された記述から、医師が総合的に値を判定するための Web インタフェースを実装した。今後は、c)を b)へ転換するため、上述のディクショナリや項目間のマッピングの機能を強化する。また、d)に関しては、電子カルテから直接情報を取得することも検討する。さらに、電子カルテや EHR システムへはたらきかけ、d)、e)の比率を下げることを試みる。

4) 統計手法の開発

各データ型に対応した、ゲノムワイド関連解析を実施するための基盤を開発し、公開した。また、

本システムを利用して、種々のゲノムワイド関連解析を実施した [2-8]。現在、各オミックスデータを統合する方法や、時系列解析の方法の検討を進めている。

5) バイオインフォマティシャン・遺伝統計家の養成

京都大学内外の5～10 人の若手研究者、および、学生らに対し、臨床研究手法の勉強会（毎週）、臨床情報の取り扱い手法の検討会（毎月）、R 等の統計ツールやプログラミングの講義（不定期）で実施した。

§3. 成果発表等

(3—1)データベースおよびウェブツールの構築と公開

① 公開中のデータベース・ウェブツール等

データベース名： KGWEB

概要：GWAS 実施用パイプライン。コマンドラインインタフェース、Web インタフェースを備える。

公開日：H23 年 3 月 10 日

URL： http://www.genome.med.kyoto-u.ac.jp/webpage/html/kgwast_eng.html

(3－2) 原著論文発表

① 発行済論文数(国内(和文) 0 件、国際(欧文) 7件)：

② 未発行論文数("accepted"、"in press"等)(国内(和文) 0 件、国際 (欧文)1 件)

③ 論文詳細情報

1. Kato L, Begum NA, Burroughs AM, Doi T, Kawai J, Daub CO, et al. Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes. Proc. Natl. Acad. Sci. U.S.A. 2012 Feb;109(7):2479–84. (doi:10.1073/pnas.1120791109)
2. Toyoda H, Kumada T, Tada T, Hayashi K, Honda T, Katano Y, et al. Predictive value of early viral dynamics during peginterferon and ribavirin combination therapy based on genetic polymorphisms near the IL28B gene in patients infected with HCV genotype 1b. J. Med. Virol. 2012 Jan;84(1):61–70. (doi:10.1002/jmv.22272)
3. Toyoda H, Kumada T, Hayashi K, Honda T, Katano Y, Goto H, et al. Antiviral combination therapy with peginterferon and ribavirin does not induce a therapeutically resistant mutation in the HCV core region regardless of genetic polymorphism near the IL28B gene. J. Med. Virol. (doi:10.1002/jmv.22145)
4. Matsuse M, Takahashi M, Mitsutake N, Nishihara E, Hirokawa M, Kawaguchi T, et al. The FOXE1 and NKX2-1 loci are associated with susceptibility to papillary thyroid carcinoma in the Japanese population. J. Med. Genet. 2011 Sep;48(9):645–8. (doi:10.1136/jmedgenet-2011-100063)
5. Ratanajaraya C, Nishiyama H, Takahashi M, Kawaguchi T, Saito R, Mikami Y, et al. A polymorphism of the POLG2 gene is genetically associated with the invasiveness of urinary bladder cancer in Japanese males. J. Hum. Genet. 2011 Aug;56(8):572–6. (doi:10.1038/jhg.2011.60)

6. Toyoda H, Kumada T, Tada T, Kawaguchi T, Murakami Y, Matsuda F. Impact of genetic polymorphisms near the IL28B gene and amino acid substitutions in the hepatitis C virus core region on interferon sensitivity/resistance in patients with chronic hepatitis C. J. Med. Virol. 2011 Jul;83(7):1203–11. (doi:10.1002/jmv.22092)
7. Terao C, Yamada R, Ohmura K, Takahashi M, Kawaguchi T, Kochi Y, et al. The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. Hum. Mol. Genet. 2011 Jul;20(13):2680–5. (doi:10.1093/hmg/ddr161.)
8. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, et al. (2012) Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nature Genetics.
Available:http://www.ncbi.nlm.nih.gov/pubmed/22446963. Accessed 18 April 2012. (in press).