

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  
研究開発課題「大規模ゲノム疫学研究の統合情報  
基盤の構築」

研究開発終了報告書

研究開発期間：平成23年4月～平成26年3月  
研究代表者：松田 文彦  
(京都大学大学院医学研究科)



## § 1 研究開発実施の概要

### (1) 実施概要

現在、日本国内で多数の疫学研究が実施されているが、各研究が独自にデータを収集・解析しており、標準的な研究実施基盤が存在しない。異なる研究では同じ目的のデータが異なる水準で収集されていることも多く、また、データが死蔵されている場合もあり、その統合は容易ではなかった。本課題は、こういった疫学研究における大きな問題を解決し、また、今後予防医学研究の主流となる多施設共同のゲノム疫学研究の情報基盤を構築することを目的とし、5つの課題を設定し、順次実施した。

#### 1) ゲノム疫学研究の統合基盤構築

複数拠点にわたる様々なゲノム疫学研究を統合するためには、データディクショナリの定義とその統合的な管理が重要であり、マスターサーバを中心とする分散系システムとして構築した。具体的な設計・開発にあたっては、京都大学が実施するながはま0次コホート研究をモデルケースに、データフローや匿名化、試料のロジスティック等の手順を洗い出し、一般化したのち、パフォーマンスやセキュリティに配慮しつつシステムとして実装した。主要な課題として、1)メタデータの開発・保守、2)統計解析手法の研究開発、3)データハンドリング効率化、4)セキュリティポリシー確立、5)パフォーマンスチューニングを検討し実装した。また、オミックス解析のうち特に網羅的メタボローム解析に注力し、測定データの標準化とデータベースへの蓄積を試みた。

#### 2) データ公開

本システムで得られたながはま0次コホートで得られたデータを公開した。実施にあたっては、まず、NBDC内のデータ共有部会において、データ共有のルールを確立し、それに従って構築されたNBDC内のデータ共有サイトへデータを提供した(2014年3月末予定)。さらに、ゲノムブラウザ様のインタラクティブなWebサイトを独自に構築し、各SNPのジェノタイプ頻度やマイクロアレイとの関連解析(eQTL)の結果を公開した(2013年11月)。

#### 3) データベースシステムの提供

本情報基盤を用いて複数の疾患ゲノム解析の多施設共同研究を支援するとともに、システムそのものを国内のゲノム疫学研究グループに提供した。

#### 4) 人材育成

3~4日間の次世代シーケンサー解析トレーニングコースと疾患ゲノム解析シンポジウムの実施を開催し、若手のインフォマティシヤンの育成に務めた。

#### 5) PHR展開

将来の、地域コホートでのPHR展開の布石として、疾患コホートをモデルケースにEHR連携を実施した。

本基盤が複数の疫学研究を統合し、より効果的で強固な日本の疫学研究の基礎となれば幸いである。

## (2) 研究開発成果のデータベース等

名称	概要	URL	公開日	レコード数	レコード数の定義、説明等	NBDCアーカイブへのデータ提供	備考
大規模ゲノム疫学研究統合DB	大規模ゲノムコホートから他施設共同臨床研究まで、コホート研究を統合的に管理するデータベース。本研究課題の最大の成果物であり、システムそのものの提供に加え、京都大学内で新たなプロジェクトのデータ管理が出来る機能を有する。	<a href="https://genome.cohort.med.kyoto-u.ac.jp/">https://genome.cohort.med.kyoto-u.ac.jp/</a>	2013年度より要望に応じてデータベースシステムの使用(データ登録と解析のための利用)、システム自体の移設を行ってきた。	1,200万件の情報の蓄積が可能	約12,000人について、最大1,000項目の臨床情報	利用を希望する研究者には随時提供を行っている。	一般公開して内容を閲覧してもらう性質のものではなく、データベースへのデータの登録や、システム自体の提供を目的として構築された。
Human Genetic Variation Browser	ヒト遺伝子多型と様々なオミックスとの関連解析の結果をゲノムブラウザ形式で提供する。現時点では、特にマイクロアレイによる発現データとSNPとの関連を収録	<a href="http://www.genome.med.kyoto-u.ac.jp/SnpDB/develop/index.html">http://www.genome.med.kyoto-u.ac.jp/SnpDB/develop/index.html</a>	2013/11/12	22億	約150万の遺伝子多型と1.5万の遺伝子との網羅的関連解析	約3,600人の網羅的ゲノム解析情報を制限付き公開する予定(2013年度末を予定)	

## § 2. 研究開発構想(および構想計画に対する達成状況)

### (1) 当初の研究開発構想

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、本提案では以下の 5 点を目標として設定した。

#### 1. ゲノム疫学研究の統合基盤構築

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。

この目的のために、以下に示す研究開発項目を実施する。

##### 1) メタデータの開発・保守

データ型の定義、オントロジーによる類似(同義)項目間の関連付けなどを行う。

##### 2) 統計解析手法の研究開発

データ種別や実験手法に応じたクオリティコントロールと統計解析手法の開発を行う。

##### 3) データハンドリング効率化

膨大な種類の臨床情報を以下に扱うかを検討し、また、各種登録・閲覧が容易に実施可能な Web インタフェースを提供する。

##### 4) セキュリティポリシー確立

アクセス権コントロールのモデル化、および、最適な IT セキュリティを検討し、実装する。

##### 5) パフォーマンスチューニング

利便性やセキュリティに配慮しつつ、パフォーマンスを向上させる。

##### 6) 定量的メタボローム解析

質量分析計(MS)による血漿の低分子水溶性化合物の網羅的測定値のデータの標準化とデータベースへの蓄積方法を確立する。

#### 2. データ公開

統合情報基盤で得た情報を、セキュリティ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. データベースシステムの提供

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. 人材育成

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じた教育訓練(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. PHR 展開

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。

(2)新たに追加・修正など変更した研究開発構想

1. ゲノム疫学研究の統合基盤構築

3)データハンドリング効率化において、以下の機能を追加した。

○健診支援機能の開発

ながはま0次コホート研究での第二期健診業務が2012年度から開始されたが、健診項目の大幅な増加もあり、健診補助やデータ登録ツールの必要性が高まった。そのため、以下の機能を新たに実装した。

- 1) 健診ワークフローによる、健診時の健診順序制御と受診状況のモニタリング
- 2) タブレット端末を利用した Web アンケート・健診結果入力インタフェース
- 3) 論理矛盾排除の仕組み
- 4) 試料管理機能とインタフェース

2. データ公開

当初想定していた NBDC データベースへのデータ提供は、2013 年度下半期を予定していたが、長浜市が主催する「ながはま 0 次コホート事業審査会」から、ゲノム情報を公開するにあたり一度オプトアウト式で同意撤回の意思確認をおこなうほうがよいのでは、との慎重な意見が出されたため、今後 NBDC へ提供する予定の様々な計測・測定値、質問票による環境・生活習慣情報なども含め、オプトアウト式の再同意を実施した。そのため、データ提供が2013年度末にずれ込んだ。

また、NBDC のデータベースは生データのリポジトリであり、一般ユーザーが容易にデータを閲覧できる状況にはない。そのため、当初予定していなかったインタラクティブな Web インタフェースツールを京都大学内に構築し、ゲノム情報に関してゲノムスキャン済みの全検体の集約情報を先行して公開した。

3. データベースシステムの提供

特記事項なし

4. 人材育成

特記事項なし

5. PHR 展開

特記事項なし

(3) 達成状況

研究項目	2011年度	2012年度	2013年度	変更点
<b>1 ゲノム疫学研究の統合基盤構築</b>				
1) メタデータの開発・保守				
・メタデータのスキーマ定義	←→			
・語彙の組み込み	←→			
・オントロジーの定義と導入	←→			
2) 統計解析手法の研究開発				
・QC法の開発	←→			
・解析法の開発	←→			
3) データハンドリングの効率化				
・データベーススキーマ開発	←→			
・入出力最適化	←→			
・データ圧縮方法の開発		←→		
・インタフェース開発	←→			
4) セキュリティポリシー確立				
・アクセスコントロール		←→		
・セキュリティの強化	←→	←→		
5) パフォーマンスチューニング	←→			
6) 網羅的メタボローム解析				
・定量化手法の開発	←→			
・解析法の開発		←→		
				実装を継続中
<b>2 データ公開</b>				
1) 試験的公開			←→	
2) 一般公開			←→	
				2014年3月末を予定
<b>3 データベースシステムの提供</b>				
1) データの格納と管理	←→			
<b>4 人材育成</b>				
1) 統計学的解析の実践教育	←→			
2) 集中トレーニングコース		↑	↑	
<b>5 PHR展開</b>	←→			

#### (4) 研究開発の今後の展開について

大規模ゲノム疫学研究に関わるゲノム・オミックス情報、環境・生活習慣情報、臨床情報などの詳細かつ多様な情報を統合・一元管理するデータ標準化形式の策定および利便性と汎用性の高い統合データベースの構築は、今後我が国で実施・展開される大規模ゲノムコホートをを用いたヒト生命情報統合解析の情報基盤としてのデータベース構築のパイロット研究として極めて重要なものである。特に、全国の複数の研究拠点のデータを一元化しつつも、個々の研究者がリアルタイムでデータを共有できる情報基盤の構築はプロジェクトを成功させるために不可欠である。

本研究課題では、多施設共同研究で実施されるゲノムコホート研究のデータを同時に蓄積する枠組みを完成させ、またデータディクショナリの一元管理により、データの統合や共有が容易おこなえるデータベースシステムを完成した。

今後のゲノムコホート研究は、大規模かつ多施設での共同研究として実施されることは明らかであり、また生活習慣や環境に関わる情報や分析・測定データもより高感度・高精度になることが予想される。したがって、これまで疫学研究者がおこなっていた、質問票をベースとした情報の収集や手作業の解析では、今後の大型プロジェクトを支えることはとうてい不可能である。また、今後はゲノムコホート研究で蓄積した新知見を、個人に最適な予防・健康づくりや治療に生かす仕組みへと展開する必要がある。具体的には、診断支援や投薬の効果予測、最新のライフログ技術や、個人の健康履歴である PHR との連携等による健康情報サービスであり、そこで得たデータによる新たな医学知識の探索と体系化である。本研究で開発した情報基盤がそういった将来の展開へつながるものとして利活用されるよう、さらなる改良を加え、発展させる予定である。

### § 3 研究開発実施体制

(1) 研究チームの体制について

① 「研究代表者: 松田」グループ

研究参加者

氏名	所属	役職	研究開発項目	参加時期
松田文彦	京都大学医学研究科 附属ゲノム医学センター	センター長 教授	研究の総括	H23.4～H26.3
山田 亮	同上	教授	統計手法の開発	H23.4～H26.3
日笠幸一郎	同上	特定講師	多型データベース構築	H24.4～H26.3
寺尾知可史	同上	特定助教	検体臨床情報の収集	H23.7～H26.3
川口喬久	同上	研究員	データベース構築	H23.4～H26.3
SARAVANAM UTHU, Gunalini	同上	研究員	代謝物の網羅的解析	H24.3～H26.3
岡部みどり	同上	教務補佐員	トレーニングコース開催 に関する業務および研 究資料の作成など	H24.9～H26.3

② 「研究分担者: 佐藤」グループ

研究参加者

氏名	所属	役職	研究開発項目	参加時期
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	グループの総括	H23.4～H26.3
九山浩樹	株式会社 島津製作所 基盤研究所 臨床プロ テオミクス G	研究員	網羅的代謝物解析のデ ータ解析	H25.4～H26.3
園村和弘	株式会社 島津製作所 基盤研究所 臨床プロ テオミクス G	研究員	網羅的代謝物解析のデ ータ解析	H24.4～H26.3

(2) 国内外の研究者や産業界等との連携によるネットワーク形成の状況について

京都大学は革新的イノベーション創出プログラム(COI Stream)拠点として採択されており、本年度から9年間にわたって産学連携で予防・先制医療のプログラムを開始した。このプログラムでは、京都大学にある最先端の研究手法やながはまコホートという研究フィールドと収集された検体・情報と企業の持つ最先端技術を融合し、大規模ゲノムコホート研究で集積した情報と生体試料に、最先端の測定・分析技術と統計学、計算科学を駆使したビッグデータ解析を組み合わせ、新たな生涯健康モデルにもとづく健康科学研究を推進する。この研究開発事業の要となるのが本研究課題で構築されたゲノム疫学データベースであり、今後このデータベースに個々人の運動、環境被曝、生活パターンや睡眠などのライフログ情報を重層し、個人の健康関連情報を本人、医療者、医学研究者が共有し有効活用する仕組みを構築する予定である。また、本課題の実施中に、



UKBiobank、China Kadoorie Study、Taiwan Biobankなどの先行研究との情報交換を密にし、データ項目、データ形式などの統一を図る努力を続けてきた。今後は、データの相互利用や統合解析を目標とした共同研究を進めていきたいと考えている。

## § 4 研究実施内容及び成果

### 4.1 研究課題名:ゲノム疫学研究の統合基盤構築

#### 実施方法

大規模ゲノム疫学研究の情報基盤に最適な、複数の研究拠点で得られる情報を統合し、また、データ項目を統一しリポジトリサーバで集中管理が行えるような、標準化されたシステムを実現することを要件とし、ながはま0次コホート研究をモデルケースとして設計と実装をすすめた。

#### 実施内容・成果

##### 1) メタデータの開発・保守

最初に、データの各項目のデータ型を定義する枠組みを構築した。型は、連続数値型、カテゴリ型、文字列型などである。これらの型は後からも追加可能である。各データ項目はデータ型とともにその型に応じた制約を設定することができる。さらに、項目間の論理矛盾を排除するための制約を導入可能である。また蓄積されたデータのうち同義のデータ項目を拠点間で共有できるよう、検索機能の実装やテンプレートセットから流用も可能となるよう実装した。なお、オントロジーについては、既存オントロジーの調査とRDF化のための実装方法の検討にとどまり、実装には至っていない。

##### 2) 統計解析手法の研究開発

データの種類、状況に応じて最適な品質管理(QC)を実施し、一連の手続きをルーチン化した。SNPジェノタイプ、次世代シーケンサーによるDNAシーケンシング、遺伝子発現アレイ、ガスクロマトグラム質量分析計(GC-MS)、フローサイトメトリー等で実績がある。解析項目のうち特にオミックスデータに関しては、古典的なマニュアルアナレーションによるものではなく、クラスタリング等を用いた情報学的理論に基づく解析手法を開発した。

##### 3) データハンドリング効率化

連結可能匿名化の自動化、データの一括登録・ダウンロード、フォームでの逐次キュレーション機能、ユーザー登録・プロジェクト登録等のWebインタフェースにより、データハンドリングの利便性向上を図った。また、健診支援機能として、Webアンケート、健診ワークフロー、試料管理等の機能を追加した。なお、アンケートやキュレーション用Webインタフェースは、データ項目のセットから自動生成され、都度の画面実装が不要であり、またプログラム言語様の演算式により複雑な論理矛盾チェックが実施できるようにした。

##### 4) アクセスコントロール、セキュリティ強化

研究者ごとに閲覧・解析できる症例や情報および解析の機能を制限するアクセスコントロール機能、匿名化対応表の暗号化、匿名化対応表サーバとのネットワーク断続的接続(夜間の自動匿名化バッチの際にのみ接続する)により、データおよびシステムセキュリティの向上を図った。

##### 5) パフォーマンスチューニング

健診会場での複数の一般ユーザーの同時利用に耐えられるよう、一つのアンケートの最大項目数を1000、同時ログイン数100での速度遅延がないよう、パフォーマンスチューニングを実施した。

##### 6) 網羅的メタボローム解析

血漿中の低分子化合物のGC-MS測定データのデータ形式を標準化し、データベースに蓄積できる仕組みを開発した。そのために、ながはまプロジェクトで測定した3,286検体の測定値をデータセットとして利用した。測定に用いる方法論の限界から、絶対定量は困難で、

標準化合物に対する比率で化合物の量を表現した。測定ピークと化合物の対応付けをおこない、達成できた化合物数は徐々に増加し、60(2012年)、119(2013年)、139(2014年)となった。統計解析にあたっては、測定値の生データからシグナルを取得し、三次元展開することで、数千のピークを同定しており、現在化合物を特定せず、全シグナルのプロファイルで異なる測定を比較できるような統計学的手法を開発している。

以下、実装したシステムの構成・機能の一覧を示す。

#### ○システムの構成

システムの構成は、各拠点内の部門(個人情報管理する部門、一次匿名化された ID で臨床情報等を収集する部門、それらを二次匿名化し遺伝情報等を追加して解析する部門。すべての拠点がこれら三部門を有する必要はない)の独立した複数のサーバと、プロジェクトやデータ項目など、拠点間で共通する情報を管理する唯一のマスターサーバからなる分散系とした。これにより、拠点や部門の追加を柔軟に行うことができ、今後展開される大規模複数拠点の疫学研究を実施する基盤とした。システムは、<https://genome.cohort.med.kyoto-u.ac.jp/>として提供しており、権限を有するユーザーは外部からもアクセス可能である。また、システムはパッケージ化しており、複数拠点への展開が容易である。

#### ○主要機能

- 1) 被験者情報管理  
個人情報の管理、健診予約や健診受診券の発行、健診結果返却時の匿名化番号の実名への変換などを可能とする機能。
- 2) データディクショナリ  
疫学データで収集する情報に型付け(continuous, ordinal, categorical, binomial, date, time, string, biallelic)し、それぞれの型に応じた制約を導入するための機能。型を後から追加することも可能である。一般的な項目セットをテンプレートとして登録しており、それを流用することで、異なるプロジェクト間での項目の整合性を高めた。
- 3) Web フォーム自動生成  
収集したいデータ項目を選択し、画面での表示順やラベル名等を Web から登録するだけで、キュレーションやデータ登録用のインタフェースを自動的に生成する機能。
- 4) Web アンケート自動生成  
Web フォーム自動生成同様、項目セットの選択により Web アンケートを自動生成する機能。Web フォームとは異なり、本機能は一问一画面でページ遷移し、また、パスワードログインによって、被験者が自身のコンピュータからデータを登録することができる。
- 5) 自動匿名化  
被験者の個人識別情報を、自動的に一次ないしは二次匿名化する機能。利用の際は、プロジェクトごとに検体 ID のフォーマット(接頭辞や連番部分)を事前に決めておくことで、新たに登録された被験者個人識別情報を自動で匿名化し、その被験者のデータを夜間バッチにて、次の部門へ自動で転送する。
- 6) 試料(チューブ)管理  
分注や貼り替えの際のバーコード自動発行、匿名化 ID と検体チューブにふられたバーコードとの対応表を一括でアップロードする機能など、試料の取り違えをなくするための支援機能。
- 7) アクセスコントロール  
ユーザーごとに、利用できる機能や参照できる被験者情報を制限する機能。被験者情報への閲覧権限の制御では、健診実施施設や受診病院等で被験者をグループ化し、ユーザーの関与するプロジェクトごとに、その被験者グループへの閲覧権限を付与する。データ一括登録

キュレーション前もしくは後のデータをテキストファイルを用いて、一括でデータベースに登録するための機能。Web フォームからの逐次登録やキュレーションが必要ない場合に有効。

- 8) データ一括ダウンロード  
登録済みの臨床データとチューブバーコードを一括でダウンロードするための機能。
- 9) 論理矛盾チェック  
データ間の論理矛盾を排するための機能。項目セット内において、項目の ID を用い、論理演算式をプログラム言語同様に記述することで、データ登録・変更時に、関連する項目間で値のチェックを行う。
- 10) 健診ワークフロー  
健診会場において、各ブースでの受診歴の登録、バーコード発行、問診データの登録、受診順序の制御を実施するための機能。
- 11) EHR 連携  
データ取得の労力をおさえるため、EHR から臨床情報を自動的に取得し、それをキュレートするためのインタフェースを提供する機能。具体的には、電子カルテシステムから Medical Markup Language(MML)形式で臨床情報を抽出し、データベースへ「キュレーション前」として登録する。それに対し、データ管理者等が被験者単位で Web フォームからキュレーション(文字列データから、カテゴリ型のコードへの変換など)を実施し、「キュレーション済」にステータスを変更する。



### 成果の位置づけ・類似研究との比較

地域ベースで行われる健常者のゲノムコホート研究から疾患に着目した疾患ゲノム解析までを網羅し、データリポジトリとしてのみならず、データ登録や匿名化などのツールを提供しつつ、データ収集を支援する本システムと類似する国内外の大規模な研究は例がない。類似のシステムとしては、欧米では NCBI の dbGaP (<http://www.ncbi.nlm.nih.gov/gap>)、国内では臨床試験登録システムの UMIN-CTR (<http://www.umin.ac.jp/ctr/index-j.htm>)などがあげられるが、いずれもデータと研究結果のリポジトリであり、研究実施を支援する機能はまったく持ち合わせていない。研究を実施できる基盤を共有することで、データ項目の標準化・統合が進み、各研究のデータインフラや諸手続き、それに続く解析が簡素化され、研究コストの低減が期待できるとともに、データ公開(共有)までがシームレスに進めることができ、データの死蔵や隠ぺい等を減少させることができる。

#### 4. 2 研究課題名: データ公開

##### 実施方法

NBDC のデータ共有部会において、データ共有のルールを策定し、それに基づいて構築された NBDC 内のデータベースにデータを提供する。また、独自に公開サイトを構築し、集約情報を公開する。

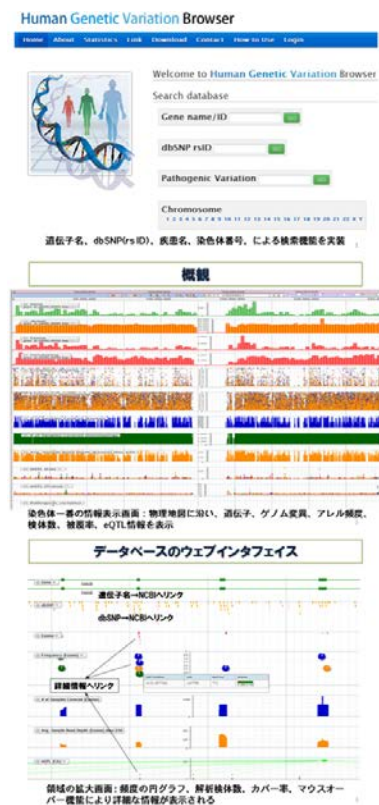
##### 実施内容・成果

2013 年 4 月公開の NBDC ヒトデータ共有ガイドライン、NBDC ヒトデータ取扱いセキュリティガイドラインの策定に深く関わった。ながはま 0 次コホート研究のデータは本ガイドラインに沿って公開する。

まずは、ながはまコホート研究で得られた網羅的ゲノム解析情報(個人のジェノタイプ)を制限付き公開する。質問票情報、生化学・血液学・生理学測定値、ゲノムスキャン情報、エクソームシーケンス情報、トランスクリプトーム情報、網羅的メタボローム解析情報は、当面、当事業に参加した研究者の優先的解析権を担保するため、共同研究ベースでの制限付き公開とするが、2015 年度末を目処に通常の制限付き公開とし、データを広くコミュニティで共有する。また、ゲノムスキャンによるジェノタイプ情報、網羅的発現解析情報などを集計した情報は一般公開を終えており、臨床検査情報、生理学的測定値情報などは順次公開していく。

なお、NBDC のデータ公開サーバは生データのアーカイブと共有が主目的であるため、ユーザーはその内容を容易に閲覧することができない。そこで、独自にゲノムブラウザ様のインタラクティブな Web ツールを構築・公開し、幅広く情報を配信した。

(<http://www.genome.med.kyoto-u.ac.jp/SnpDB/index.html>)。



##### 成果の位置づけ・類似研究との比較

公開されるながはま 0 次コホート研究のデータは、今後発展が予想される網羅的オミックス研究の重要なリファレンス情報として大いに活用されることが見込まれる。ゲノム網羅的なジェノタイプデータのデータベースとしては、国際 Hapmap プロジェクト (<http://hapmap.ncbi.nlm.nih.gov/index.html.ja>)、NCBI が運営する dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>) がよく知られている。本研究課題で構築したデータベースは、日本人 1,208 人のエクソーム解析由来のジェノタイプカウント、また、300 人の網羅的 eQTL 解析(mRNA 発現量関連遺伝子)の結果を公開しており、日本国内では最大規模のデータベースである。なお、網羅的 eQTL 解析のデータベースは、Sanger 研究センターの genevar (<http://www.sanger.ac.uk/resources/software/genevar/>) が知られているが、日本人検体数は、113 人に限られる。

#### 4. 3 研究課題名: データベースシステム提供

##### 実施方法

共同研究として実施するゲノムコホート研究の情報管理については、松田グループで運用するシステムにデータ全体を格納し、複数の研究から得られるデータの統合を行う。また、構築したデ

ータベースシステムをパッケージ化して提供し、新たなゲノムコホートデータベースの構築を支援する。

更に、本データベースの最大の特徴は、健常者集団を対象としたゲノムコホート研究のみならず、多施設共同での病院ベースでの疾患ゲノム解析にも大きな変更を加えることなく利用できる点である。将来このような多施設横断型の疾患研究は数を増していくことが予想されるが、そのような研究における情報基盤として利用可能である。

#### 実施内容・成果

複数の疫学研究を受け入れ、実施中である。以下、受け入れた実施中の主要な研究(20を超えるプロジェクトのうち4プロジェクト)を以下に紹介する。

プロジェクト	介入の有無、関連施設数、被験者数、データ点数
IgG4 関連疾患	介入研究、50 施設以上、約 800 人、時系列情報(4 点)・延べ約 100 項目
肺高血圧症	介入研究、1 施設、約 150 人、時系列情報(4 点)・延べ約 100 項目
がんコホート	症例対象研究、複数施設、5000 人、約 80 項目
好酸球性食道炎	介入研究、約 20 施設、150～200 人、時系列情報(4 点)

また、健常者の前向きゲノムコホート研究である岩手東北メディカル・メガバンクに対しては、データベース全体のシステムパッケージを提供し、また、その構築を支援した。

#### 成果の位置づけ

システムの実証実験と、システム活用による実利を目的として、複数の研究を受け入れた。問題点が解決され、有効性が確認されれば、徐々に受け入れ研究を増やし、またパッケージを提供したいと考えている。

#### 4. 4 研究課題名:人材育成

##### 研究開発実施内容及び成果

##### 実施方法

バイオインフォマティクス・統計遺伝学の若手人材の育成を目的として、トレーニングコース、シンポジウムを開催した。

##### 実施内容・成果

2013年1月に4日間、2014年3月に3日間のトレーニングコースを開催し(詳細は§6)、それぞれ海外からその分野でトップレベルの講師を招きチュートリアルと演習を行った。2013年、2014年ともに国内外の若手研究者約40人が受講した。また、それに続いて講師を招待講演者に迎えた国際シンポジウムを開催した。

#### 成果の位置づけ

海外の著名な研究者を招いて実施したトレーニングコースは、実践的で高度な解析手法を、体系的・効率的に学ぶよい機会であったと受講者から高い評価を得た。さらに、講師陣とのつながりから、留学や共同研究に発展する例もあり、その点でも期待以上の成果を得た。今後も、可能な限り継続して開催していきたいと考えている。

#### 4. 5 研究課題名:PHR 展開

##### 研究開発実施内容及び成果

## 実施方法

データベースに蓄積された生命分子情報、環境・生活習慣情報と用いた PHR を展開するには、これらに加えて複数の医療機関に蓄積されている複数の疾患関連の臨床情報を迅速かつ効率よくデータベースに取り込む必要がある。しかしながら、最初からすべての疾患に関わる疾患関連情報を網羅的に蓄積するようなシステムの構築は極めて困難である。そこで、まずは1医療機関から1疾患の情報を確実に抽出し、データベースへ登録することから始めて、1疾患・複数医療機関、多疾患・複数医療施設に拡張する方針で臨むこととし、まずは京都大学病院の医療情報を、EHR を介して収集することを試みた。EHR には、京都大学附属病院を中心とした「まいこねっと」を採用し、一施設一疾患によるシンプルなモデルで、必要な項目の利用可能性や連携方法を検討し、それに従ってデータ入出力やキュレーション機能を実装した。

## 実施内容・成果

ある自己免疫疾患をモデル疾患とし、疾患ゲノム解析に必要なデータ 131 項目のうちどの程度が EHR を通して取得可能かを確認したところ、74 項目について何らかの情報が得られることが分かった。取得不能の 57 項目のうち 42 項目は病院システムに電子情報として存在するが、EHR に移行されていない情報であり、残りの 15 項目は電子化されていない情報であった。取得可能な項目については、メディアによるバッチ型で電子情報としてデータを受け取った。データは文字列で記載された情報が多く、データ登録用の Web フォームに、そういった文字列情報から数値やカテゴリなどへの変換を支援する機能を実装した。

## 成果の位置づけ

疫学研究では、その地域の病院等医療機関からの臨床情報の収集は非常に重要である。地域コホートにおいては、病院からの情報収集はその地域特有の事情があり容易に他の地域の結果を適用できないが、この結果をもとに効率的な PHR へと発展していきたいと考えている。

## § 5 成果発表等

(1)原著論文発表 (国内(和文)誌 0件、国際(欧文)誌 12件)

1. Yoshimura, K., Nakayama, T., Sekine, A., Matsuda, F., Kosugi, S., Yamada, R., Shimizu, Y., Kanematsu, A., Yoshimura, K., Ogawa, O., the Nagahama Cohort Research Group. (2012) B-type natriuretic peptide as an independent correlate of nocturnal voiding in Japanese women. *Neurourol. Urodyn.* 31, 1266-1271
2. Kawaguchi, T., Yoshio Sumida, T., Umemura, A., Matsuo, M., Takahashi, M., Takamura, T., Yasui, M., Saibara, T., Hashimoto, E., Kawanaka, M., Watanabe, S., Kawata, S., Imai, Y., Kokubo, M., Shima, T., Park, H., Tanaka, H., Tajima, K., Yamada, R., Matsuda, F. and Okanoue, T. for Japan Study Group of Nonalcoholic Fatty Liver Disease (JSG-NAFLD) (2012) Genetic Polymorphisms of the Human PNPLA3 Gene are Strongly Associated with Severity of Non-Alcoholic Fatty Liver Disease in Japanese. *PLoS One* 7, e38322.
3. Yoshimura, K., Nakayama, T., Sekine, A., Matsuda, F., Kosugi, S., Sugino, Y., Yoshimura, K. and Ogawa, O. Nagahama Cohort Research Group. (2013) Prevalence of postmicturition urinary incontinence in Japanese men: Comparison with other types of incontinence. *Int. J. Urol.* 20, 911-916.
4. Tabara, Y., Takahashi, Y., Setoh, K., Muro, S., Kawaguchi, T., Terao, C., Kosugi, S., Sekine, A., Yamada, R., Mishima, M., Nakayama, T. and Matsuda, F. on behalf of the Nagahama Study Group. (2013) Increased aortic wave reflection and smaller pulse pressure amplification in smokers and passive smokers confirmed by urinary cotinine levels: The Nagahama Study. *Int. J. Cardiol.* 168, 2673-2677.

5. Tabara, Y., Takahashi, Y., Kohara, K., Setoh, K., Kawaguchi, T., Terao, C., Igase, M., Yamada, R., Kosugi, S., Sekine, A., Miki, T., Nakayama, T. and Matsuda, F. on behalf of the Nagahama Study Group. (2013) Association of longer QT interval with arterial waveform and lower pulse pressure amplification: The Nagahama Study. *Am. J. Hypertens.* 26, 973-980.
6. Miyake, M., Yamashiro, K., Nakanishi, H., Nakata, I., Akagi-Kurashige, Y., Tsujikawa, A., Moriyama, M., Ohno-Matsui, K., Mochizuki, M., Yamada, R., Matsuda, F. and Yoshimura, N. (2013) Insulin-like growth factor 1 is not associated with high myopia in a large Japanese cohort. *Mol. Vis.* 19,1074-1081.
7. Yamazaki, T., Yamori, M., Asai, K., Nakano-Araki, I., Yamaguchi, A., Takahashi, K., Sekine, A., Matsuda, F., Kosugi, S., Nakayama, T., Inagaki, N. and Bessho, K.; Nagahama Study Collaboration Group. (2013) Mastication and risk for diabetes in a Japanese population: a cross-sectional study. *PLoS One.* 8, e64113.
8. Khor, C. C., Miyake, M., Chen, L. J., Shi, Y., Barathi, V. A., Qiao, F., Nakata, I., Yamashiro, K., Zhou, X., Tam, P. O., Cheng, C. Y., Tai, E. S., Vithana, E. N., Aung, T., Teo, Y. Y., Wong, T. Y., Moriyama, M., Ohno-Matsui, K., Mochizuki, M., Matsuda, F.; Nagahama Study Group, Yong, R. Y., Yap, E. P., Yang, Z., Pang, C. P., Saw, S. M. and Yoshimura, N. (2013) Genome-wide association study identifies ZFHX1B as a susceptibility locus for severe myopia. *Hum. Mol. Genet.* 22, 5288-5294.
9. Terao, C., Bayoumi, N., McKenzie, C. A., Zelenika, D., Muro, S., Mishima, M.; The Nagahama Cohort Research Group, Connell, J. M., Vickers, M. A., Lathrop, G. M., Farrall, M., Matsuda, F. and Keavney, B. D. (2013) Quantitative variation in plasma angiotensin-I converting enzyme activity shows allelic heterogeneity in the ABO blood group locus. *Ann. Hum. Genet.* 77, 465-471.
10. Nakata, I., Yamashiro, K., Nakanishi, H., Akagi-Kurashige, Y., Miyake, M., Tsujikawa, A., the Nagahama Cohort Research Group, Matsuda, F. and Yoshimura, N. (2013) Prevalence and characteristics of age-related macular degeneration in the Japanese population: The Nagahama Study. *Am. J. Ophthalmol.* 156, 1002-1009.
11. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, D. A., Jiang, L., Yin, J., Ye, L., Su, D. F., Yang, J., Xie, G., Keystone, E., Westra, H. J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E. A., Diogo, D., Cui, J., Liao, K., Guo, M. H., Myouzen, K., Kawaguchi, T., Coenen, M. J. H., van Riel, P. L. C. M., van de Laar, M. A. F. J., Guchelaar, H. J., Huizinga, T. W. J., Dieude, P., Mariette, X., Bridges Jr, S. L., Zhernakova, A., Toes, R. E. M., Tak, P. P., Miceli-Richard, C., Bang, S. Y., Lee, H. S., Martin, J., Gonzalez-Gay, M. A., Rodriguez-RodriguezL., Rantapaa-Dahlqvist, S., Arlestig, L., Choi, H. K., Kamatani, Y., Galan, P., Lathrop, M., the RACI consortium, the GARNET consortium, Eyre, S., Bowes, J., Barton, A., de Vries, N., Moreland, L. W., Criswell, L. A., Karlson, E. W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J. S., Bae, S. C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Raychaudhuri, S., Stranger, B. E., De Jager, P. L., Franke, L., Visscher, P. M., Brown, M. A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T. W., Siminovitch, K. A., Momohara, S., Matsuda, F., Yamamoto, K. and Plenge, R. M. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376-381.
12. Tabara, Y., Muro, S., Takahashi, Y., Setoh, K., Kawaguchi, T., Terao, C., Kosugi, S., Sekine, A., Yamada, R., Nakayama, T., Mishima, M. and Matsuda F.;

Nagahama Study Group (2014) Airflow limitation in smokers is associated with arterial stiffness: The Nagahama Study. *Atherosclerosis*. 232, 59-64.

(2)その他の著作物(総説、書籍など)

1. 川口 喬久、松田 文彦 個人ゲノム時代のゲノムコホート研究 *医学の歩み* 2011, 236 巻, 6 号, 607-615
2. 川口 喬久、松田 文彦 ゲノムコホート研究 *分子精神医学* 2011, 11 巻, 3 号, 209-210
3. 川口 喬久、松田 文彦 我が国のデータベース構築・統合戦略 疫学データベースの統合 *細胞工学* 2012, 31 巻, 4 号, 490-492

(3)国際学会発表及び主要な国内学会発表

① 招待講演 (国内会議 15 件、国際会議 3 件)

演題:松田 文彦 「先制医療の実現を目指した 21 世紀型ゲノムコホート研究」  
会議名:JST イノベーションサテライト高知 研究成果報告会 in 愛媛  
日時:2012 年 1 月 21 日  
場所:愛媛大学 南加記念ホール

演題:「地域に根ざした未来型健康づくりの試み～ながはま一万人コホートの例～」  
会議名:静岡ゲノムコホート講演会  
日時:2012 年 7 月 11 日  
場所:静岡県産業経済会館

演題:松田 文彦 「Human Biology とゲノム情報」  
会議名:日本 DNA 多型学会第21回学術集会シンポジウム  
日時:2012 年 11 月 7 日  
場所:京都教育文化センター

演題:松田 文彦 「ヒト生命情報統合研究とそのモデルケースとしてのながはまゲノムコホート事業」  
会議名:第 59 回日本臨床検査医学会学術集会シンポジウム「個別化医療と臨床検査」  
日時:2012 年 11 月 30 日  
場所:国立京都国際会館

演題:松田 文彦 「分子を通して自分を知る次世代の予防医療～大規模コホート研究とゲノム、タンパク、代謝物～」  
会議名:疾患メタボロミクスシンポジウム  
日時:2012 年 12 月 18 日  
場所:大阪大学吹田キャンパス・银杏会館

演題:松田 文彦 「ヒト生命情報統合研究に向けた大規模ゲノムコホート事業の推進～国内外の例を中心に～」  
会議名:日本学術会議公開シンポジウム「ヒト生命情報統合研究」  
日時:2013 年 1 月 23 日  
場所:日本学術会議ホール

演題:松田 文彦 「分子を通して自分の健康をながめる～ながはま0次コホート研究と次世代の予防医療～」  
会議名:いわて東北メディカル・メガバンク機構 発足記念シンポジウム  
日時:2013 年 2 月 2 日



場所:岩手医科大学矢巾キャンパス

演題:松田 文彦 「ヒト生命情報統合研究とその情報基盤」  
会議名:大阪大学蛋白研究所セミナー「ビッグデータ時代に向けた医療データベース」  
日時:2013年3月8日  
場所:大阪大学中之島センター

演題:松田 文彦 「ヒト生命情報統合研究とゲノムコホート」  
会議名:第20回日本遺伝子診療学会大会 教育講演  
日時:2013年7月20日  
場所:アクティシティ浜松コンgresセンター

演題:松田 文彦 「地域に根ざした新時代の予防医学の試み」  
会議名:2013年度ソニー／医科歯科臨床サミット  
日時:2013年8月20日  
場所:東京医科歯科大学鈴木記念講堂

演題:Fumihiko Matsuda ‘The Nagahama Study as a Model for the Comprehensive Human Bioscience’  
会議名:BioJapan 2013  
日時:2013年10月11日  
場所:横浜パシフィコ

\*演題:Fumihiko Matsuda ‘Large-Scale Genome Cohort Study and Multi-Omics Analysis’  
会議名:International Conference “Personalized Medicine & Global Health”  
日時:2013年10月17日～18日  
場所:Center for Life Sciences, Nazarbayev University, Astana, Kazakhstan

演題:松田 文彦 「網羅的ゲノム解析と疾患の分子機構の解明」  
会議名:第4回JBCRG 学術集会  
日時:2013年10月27日  
場所:京都市サーチパーク

演題:松田 文彦 「ヒト疾患と生命情報統合解析」  
会議名:第15回日本神経消化器病学会  
日時:2013年11月8日  
場所:ビッグハート出雲

演題:松田 文彦 「ヒト疾患と生命情報統合解析」  
会議名:平成25年度 第2回JPLSG 全体会議・合同班会議  
日時:2013年11月15日  
場所:国立病院機構名古屋医療センター

演題:松田 文彦 「大規模ゲノムコホート研究とヒト生命情報の統合」  
会議名:第33回医療情報学連合大会  
日時:2013年11月21日  
場所:神戸ファッションマート

演題:松田 文彦 「ながはまコホートとヒトの多様性に基づいた診断・医療」

会議名:第36回日本分子生物学会年会ワークショップ「未病社会の診断技術開発について」

日時:2013年12月3日

場所:神戸ポートアイランド

\*演題:Fumihiko Matsuda, 'Genome-wide association studies in IgG4-RD'

会議名:International Symposium on IgG4-RD and associated conditions

日時:2014年2月16日~19日

場所:Sheraton Waikiki Hotel

② 口頭発表 (国内会議 0件、国際会議 0件)

なし

③ ポスター発表 (国内会議 0件、国際会議 0件)

なし

#### (4)知財出願

① 国内出願 (0件)

② 海外出願 (0件)

③ その他の知的財産権

特になし

#### (5)受賞・報道等

≪受賞や新聞報道等について、具体的に記入してください。≫

① 受賞(顕著な受賞の前に\*を付記してください)

特になし

② マスコミ(新聞・TV等)報道(プレス発表をした場合にはその概要もお書き下さい。)

2013年11月12日 HGVBの一般公開の共同記者会見

翌日の朝日新聞、日刊工業新聞などに掲載され、また共同通信から各地方紙に配信された。

③ その他

特になし

## §6 研究開発期間中に主催した会議等

主なワークショップ、シンポジウム、アウトリーチ等の活動

年月日	名称	場所	参加人数	概要
2013/1/15~ 2013/1/19	The First Kyoto Course on Bioinformatics for Next Generation Sequencing with Applications in Human Genetics	京都大学	40人	次世代シーケンサー(NGS)を用いた解析手法を基礎から、国外の一流研究者を招聘して4日間の集中トレーニングコースを実施。またそれに続いてヒト疾患関連遺伝子探索のシンポジウムを1日実施

2014/3/10~ 2014/3/14	The Second Kyoto Course on Bioinformatics for Next Generation Sequencing with Applications in Human Genetics	京都大学	40 人	同上。今期はトレーニングを 3 日間、シンポジウムを 2 日間実施
-------------------------	--	------	------	-----------------------------------

## § 7 ユーザー評価結果への対応

《平成 25 年 7 月に実施した「NBDC における事業活動のユーザー評価」(<http://biosciencedbc.jp/user-hyouka-2013/result-summary>)で得られたユーザーの意見、提案等（詳細は別紙 2 を参照）に対し、実施済み若しくは実施予定の対応策等を具体的に記載してください。》

### ① 実施済み

データ公開用 Web インタフェースを 2013/11/12 に公開した。

### ② 実施予定

2013 年 3 月中に、ゲノムスキャンによる約 3,600 人の網羅的遺伝子多型情報(SNP 情報)を、NBDC を通じて制限付きで公開する。

### ③ 実施予定にない

ユーザー評価結果への対応

本研究課題では、「大規模ゲノム疫学研究の統合情報基盤の構築」という課題名にある通り、多施設共同研究での大規模ゲノムコホート研究における情報基盤の構築を目標としている。すなわち、公開情報を統合して利用者が閲覧しやすいかたちにまとめるといった性質のプロジェクトではなく、今後得られる情報を標準化し、多施設共同研究を統合して管理するデータベースシステムの構築と提供であり、そのモデルケースとして「ながはまコホート」のデータを使用した。更に、そういったデータベースシステムは大規模ゲノムコホート研究のみならず、病院ベースでの疾患ゲノム解析研究にも十分利用可能であり、複数の疾患プロジェクトでの利用実績も上がってきた。

また、公開に関しては、ながはまコホート研究関係者に対する試験的公開を予定より3ヶ月遅れたが2013年度の初頭より実施し、データベース機能や利便性などの改善を年度末まで継続して行ってきた。また、ながはまコホートの集計情報に関しては、2013年11月より一般公開している。

以上の点を踏まえて、以下に対応を回答したい。

- 使用できないため、どの程度の情報が得られるのか判断できない。
- 公開されないなら評価できない。

(回答)本研究課題の当初よりの行程表では、2012 年度第四四半期より数ヶ月間の試験的公開を経て、一般公開を実施する予定であった。したがって、2013 年 7 月の一般公開(ユーザー評価とは一般公開であると認識している)は当初より想定しておらず、このような評価に当惑している。

また、上に述べたとおり、公開に関しては、データベースの枠組みを構築し、それにながはまコホートのデータを格納したものを公開することであり、当然の事ながら、評価する側はブラウザでゲノム情報などのデータを閲覧し研究利用するのではなく、データベースに格納するだけのデータを所有しており、実データを格納してデータベースの性能を評価しなければならないことを認識していただきたい。

- 必要性があるのか？
- 同様の内容について、より網羅的かつ信頼性の高いものが欧米を中心に公開されており、本プログラムの必要性が見出せない。

(回答) 諸外国に大規模ゲノム疫学研究において、多施設間のデータの標準化と一元管理が行え、ゲノム・バイオマーカー情報、時系列の量的形質、質問紙による環境・生活習慣情報などを網羅的に管理できるデータベースが存在するのなら、ぜひ具体的にどのデータベースか教えていただきたい。

評価時に公開できないものは、本プログラムの趣旨に合わないと考える。この取り組みはデータの独自性が本質であるので、データ獲得の予算の範囲でなされることでは？ ツールやデータベースとしての独自性を主張するなら、すでに公開されているデータを利用して公開版をつくることも可能であったらう。

(回答) 評価時に公開云々は、上の回答を参照していただきたい。当初からの予定でユーザー評価期間(2013年7月)に一般公開することは計画されていない。また、後半の「すでに公開されているデータ…」は、本研究課題の目的を正しく認識されていないのではないかと考える。繰り返しになるが、本提案の最終目標は、今後行われる大規模ゲノム疫学研究の情報基盤の構築にあり、他所で入手可したデータを見やすく表示するブラウザを作成するものではないことを申し上げたい。

- 海外も含め既存のデータと進行中プロジェクトのデータとの速やかな連携がカギとなると思われる。

(回答) 貴重かつ極めて重要な意見をいただき、心より感謝している。海外を含めたデータの共有と連携は極めて重要な課題であり、そのためにデータ項目の標準化、汎用性の高いデータベース設計などに注意して構築を続けてきた。諸外国でも大規模ゲノムコホート研究が先行して行われているが、得られたデータを統合して管理するデータベースはまだ構築されていず、各項目が別々に管理されていたり、実験者がデータを持つといった問題が生じていると聞いている。本課題の実施中に、UK Biobank、China Kadoorie Study、Taiwan Biobankなどの先行研究との情報交換を密にし、データ項目、データ形式などの統一を図る努力を続けてきた。今後は、本課題で構築されたデータベースにさらに改善を加え、諸外国の研究のデータと統合して研究者コミュニティに提供していきたい。

## §8 その他

(1) 研究代表者として、研究開発、プロジェクト運営等について、上記以外に報告したいことがあれば、自由に記載してください。

特になし

以上