大規模ゲノム疫学研究の統合情報基盤の構築

京都大学医学研究科附属ゲノム医学センター
松　田　文　彦

JSTバイオサイエンスデータベースセンター
「基盤技術開発プログラム」および「統合化推進プログラム」
平成24年度　進捗報告会
2013年1月21日　JST東京本部

研究開発の目標・ねらい

- **ゲノム疫学研究の情報基盤の構築と公開**

  「ながはま0次コホート研究」の一万人の生活習慣・環境情報、臨床情報、ゲノム・オミックス情報を標準化し、データベースを構築する。

  集積した情報を、個人情報保護のもと、医学・生命科学研究者に提供する。

- **データベースの枠組みの提供と情報の連結**

  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、分子疫学研究者にデータベースの枠組みを提供する。

  他の研究で蓄積された遺伝型・表現型データを連結、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。

- **ゲノム情報科学の若手研究者の育成**

  バイオインフォマティクス、遺伝統計学の若手研究者に教育訓練（OJT）をおこない、これらの分野の将来における中心的研究者の育成をはかる。

「パーソナルヘルスレコード」の情報提供先として機能できる汎用性の高い健康情報管理システムを提案

データベースシステム概念図

京　都　大　学

NBDC

SNPアレイ
全ゲノム塩基配列

ゲノム
DB

質量分析・
その他測定機器

中間形質
DB

採血・診察・問診

臨床情報
DB

分子疫学
DB

管理者

ミラーリング
(VPN)

共同研究用

一般公開用

アクセスコントローラー

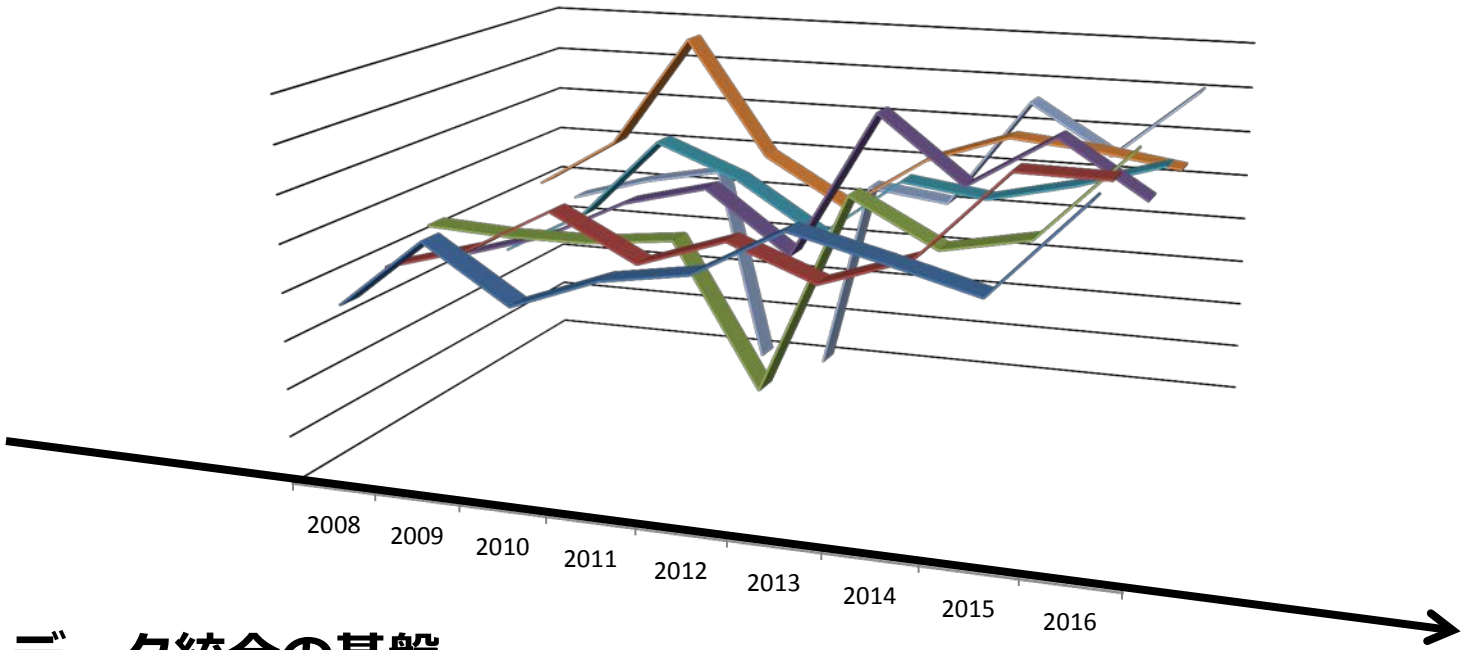
権限承認
(SSL)

閲覧・登録・解析
(SSL/VPN)

データベース利用者

大規模ゲノム疫学情報

「個人」、「項目」、「時点」で特定される多次元の情報



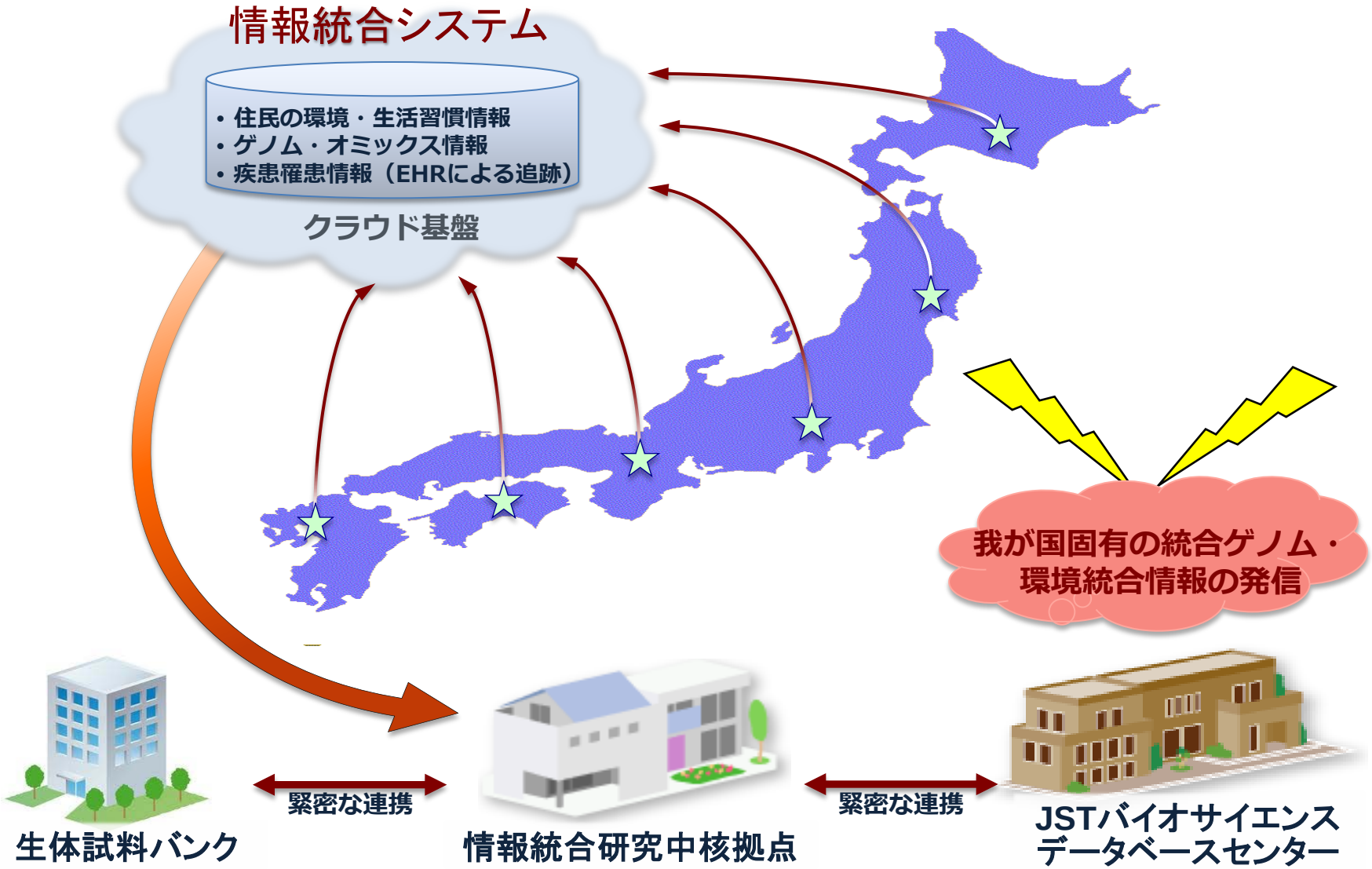
データ統合の基盤

<u>個人</u>：　倫理指針を遵守し、厳格なセキュリティのもとユニークな匿名化IDのもに利用

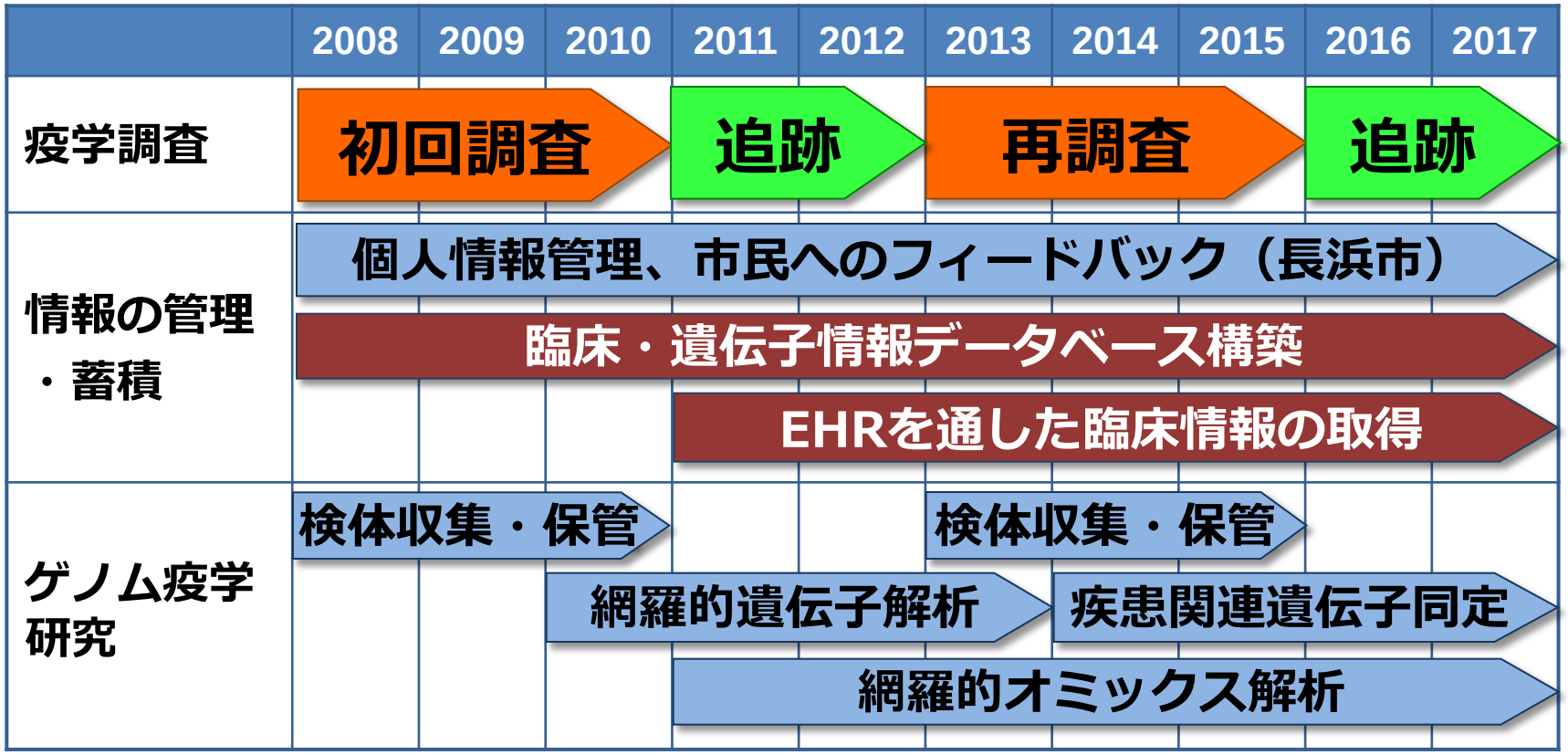
<u>項目</u>：　厳密な定義による測定・解析情報の収集と、ユニークID化による管理

正確、網羅的、かつ、効率的に収集するための情報基盤を構築する。

地域研究拠点の連携による大規模ゲノム疫学研究基盤構築と世界へ向けた情報の発信



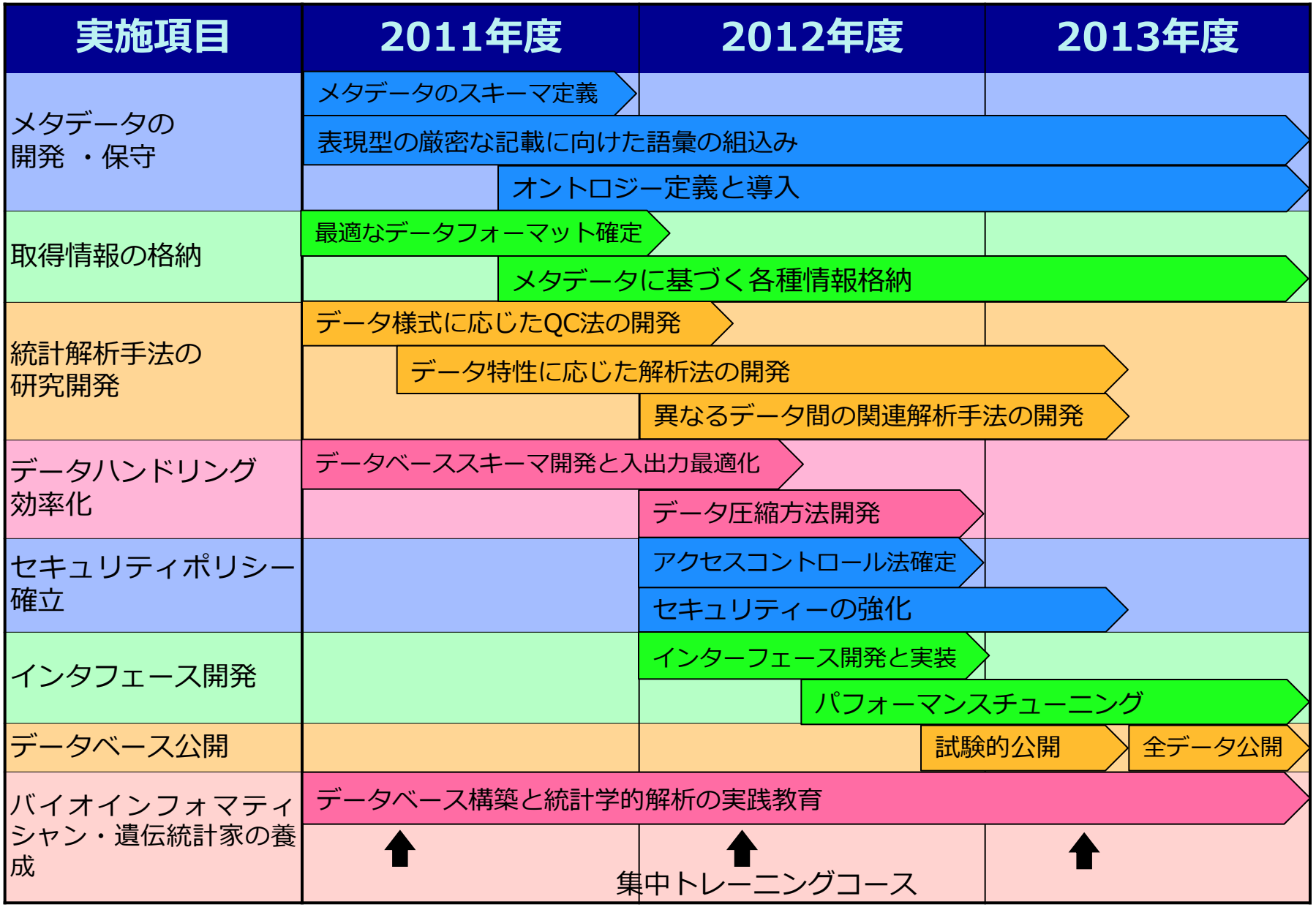
ながはま0次コホート事業のロードマップ



本事業でNBDCに蓄積するデータ項目

- 質問票による環境・生活習慣情報
  - 742項目　約748万件
- 生化学・血液学・生理学的測定値
  - 145項目　約146万件
- 参加者のゲノムスキャン情報
  - Illumina610K（約1,800検体）、2.5Mアレイ（約3,200検体）
- エクソームシークエンス情報
  - 500検体を目標とする
- 網羅的メタボローム解析情報
  - 島津製作所GC-MSを利用　全検体解析　200～250ピーク分
- EHRによる疾患罹患情報
  - ITネットワークが整備されることが前提

本研究開発のロードマップ



H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

ながはま0次コホートにおける二段階匿名化

長浜市役所（個人情報管理者）

個人識別ID：住基コードより連結不可能な方法で生成
収集情報：個人情報（住所・氏名・電話番号など）
個人識別ID　⇔　一次匿名化ID対応表

研究事務局（匿名化ID管理者）

個人識別ID：一次匿名化ID ⇒ 健診時一時ID
収集情報：健診情報（血液検査、問診等）、臨床情報（EHR連携）

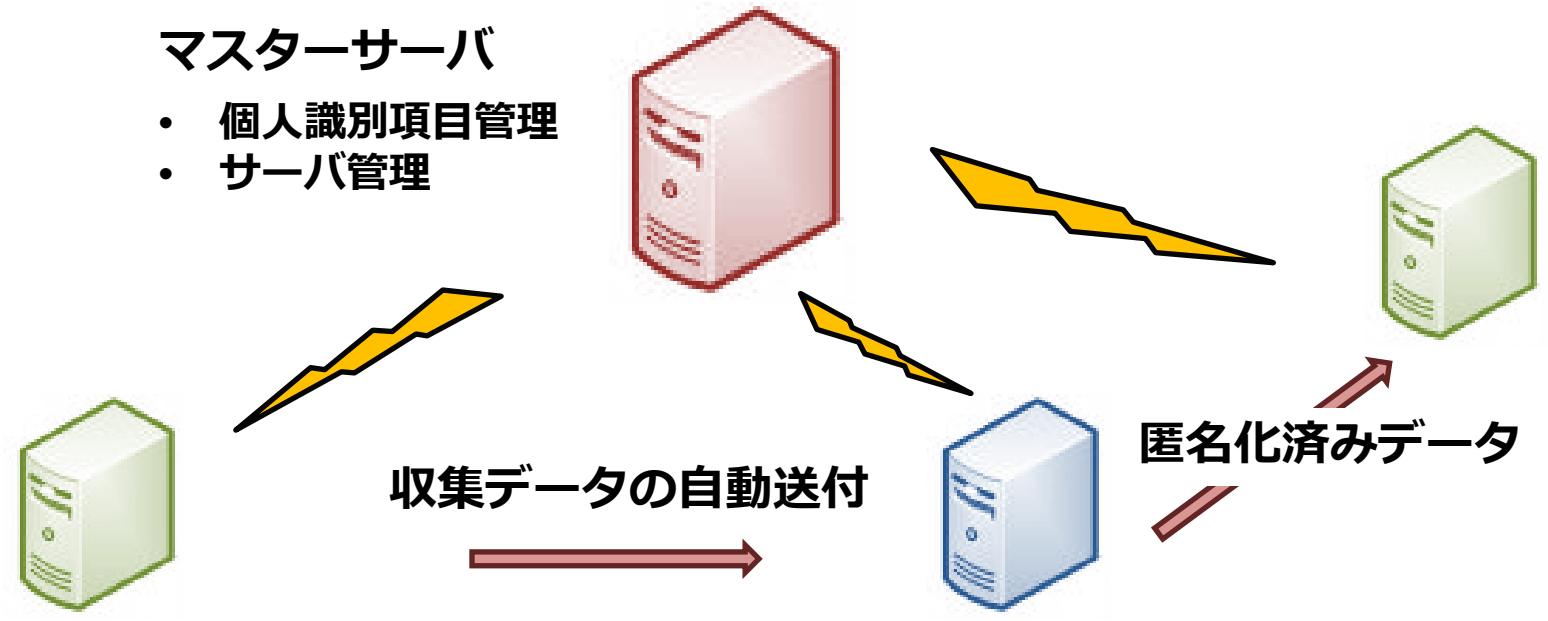
一次匿名化ID　⇔　再匿名化ID対応表

ゲノム情報解析部門（ゲノム医学センター）

個人識別ID：二次匿名化ID ⇒ サンプルID
収集情報：ゲノム情報

長浜市条例に基づいた
二段階匿名化による
セキュリティーの担保

個人ID管理と情報匿名化の一般化



情報収集管理サーバ

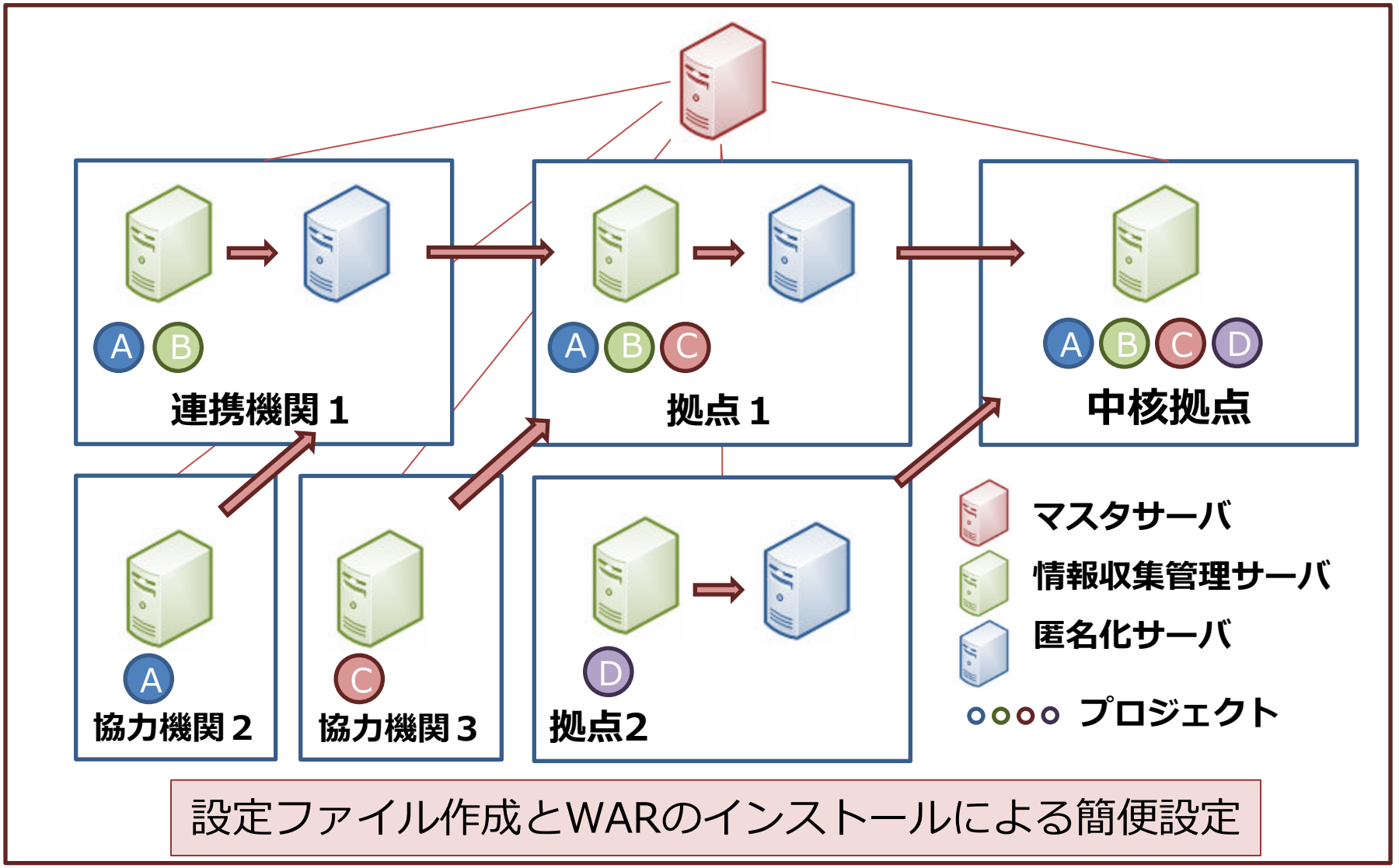
- 新規個人ID採番
- 個人情報を含む収集データ管理
- プロジェクト内で一意化された別名の管理

匿名化サーバ

- 新たな匿名化IDの発行
- 既存の匿名化IDとの連結

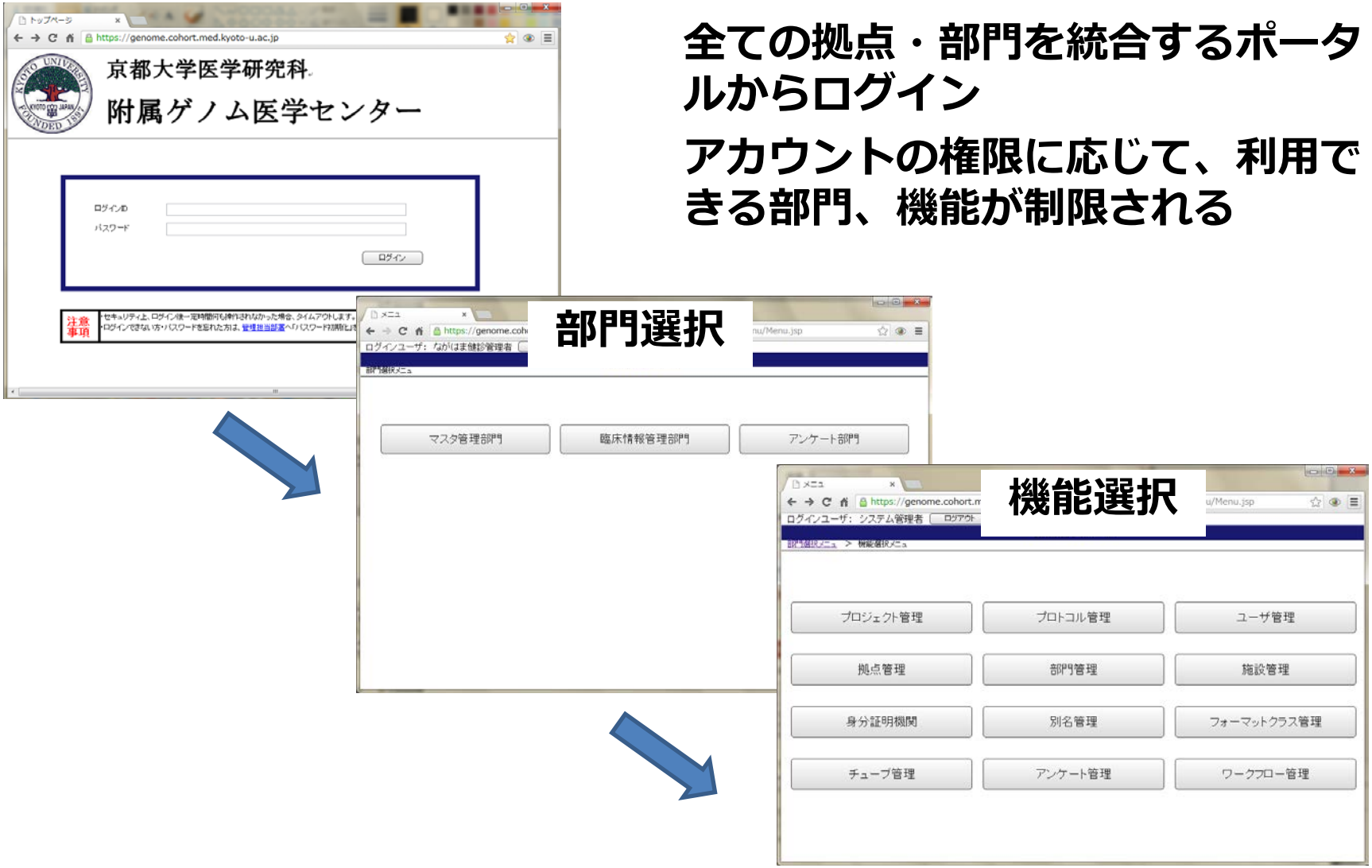
個人ID（サーバ毎に一意な接頭辞を有す）を自動採番
⇒異なるプロジェクトを統合的に管理
⇒ｎ段階匿名化・拠点の自由な追加が可能

システムのパッケージ化により複数拠点からの情報を容易に統合可能



設定ファイル作成とWARのインストールによる簡便設定

統合インタフェース



全ての拠点・部門を統合するポータルからログイン

アカウントの権限に応じて、利用できる部門、機能が制限される

H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

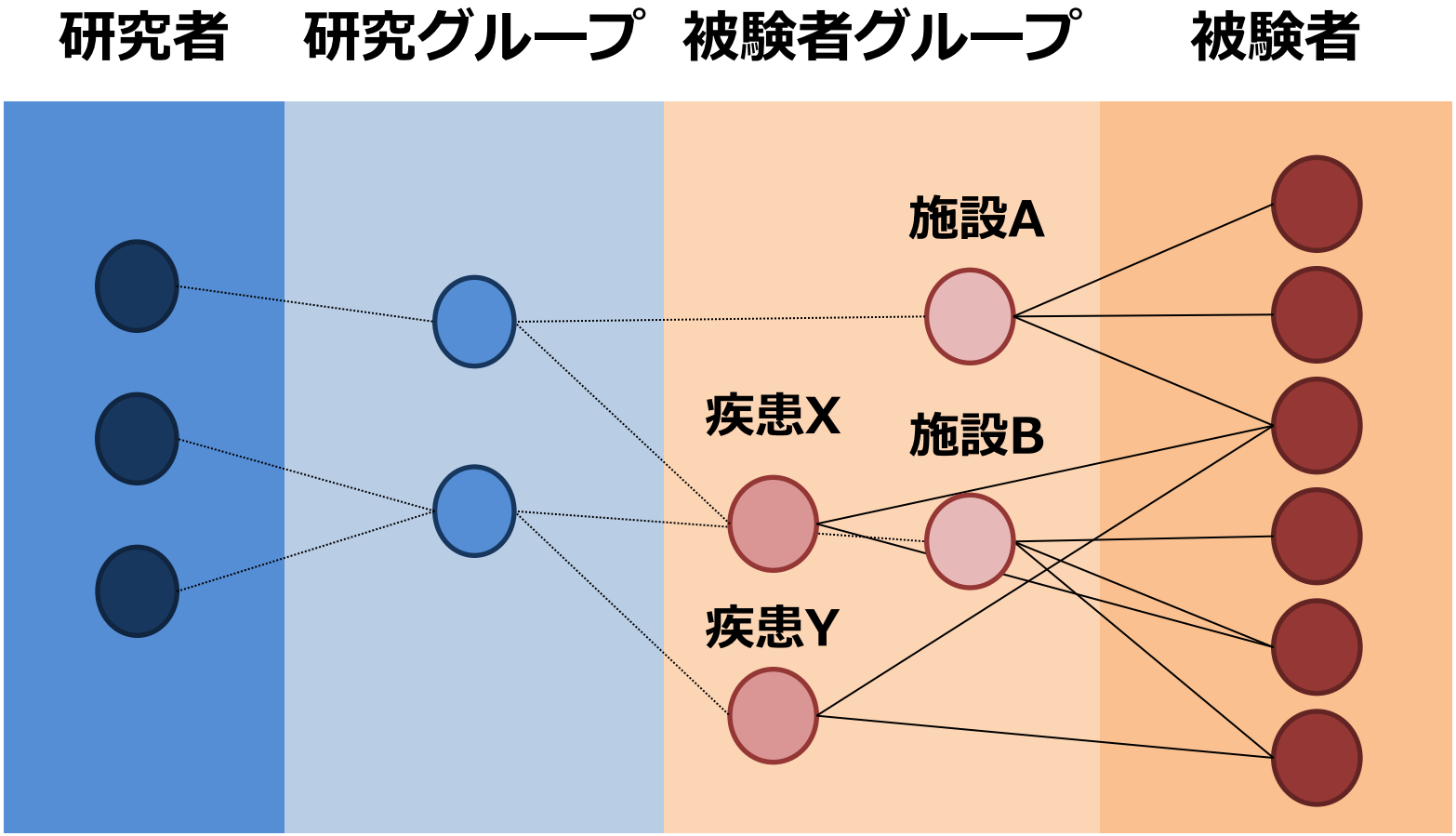
機能・データの種類によるアクセス制限

	データ項目	ユーザ情報	被験者個人情報	匿名対応表	個人別疫学情報	集団情報
共通マスタ管理者	○	○				△
個別プロジェクト＊1						
プロジェクト管理者	○	○			○	○
個人情報管理者			○			
匿名化情報管理者				○		
研究参加者	○				○	△
データ利用者						
一般研究者（制限付き公開）	△				△	△
不特定の閲覧者（一般公開）	△					△

＊1　該当プロジェクトに関わる情報のみ

○　登録、△　閲覧・利用のみ

アクセス可能な被験者情報のコントロール



いずれも、GUIから設定可能

H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

多様なデータ項目の定義

- データ型と制約の厳密な定義を実施

データ型	制約	例
連続値	最大最小、打ち切りの有無	バイオマーカー
順序ありカテゴリ	カテゴリ値とコード、その順	質問票
順序なしカテゴリ	カテゴリ値とコード	質問票
文字列・日付	文字数、日付形式	質問票（自由記載）
遺伝子多型	ゲノム上の位置とアレル	SNP, CNV

- データの特性に応じた複数の標準フォーマットの活用
  - JCAMP、FASTQ、BAM等

収集する情報間の論理矛盾の解消

- **回答権の制約**
  - **複数の質問の間での論理矛盾**
    - **例）出産に関連する質問**
- **値の制約**
  - **値が取りうる範囲内であることを保証**
  - **入力前選択肢絞り込み・入力後のチェックを実施**
    - **例）０ ＜ 喫煙期間 ≦ 現年齢-喫煙開始年齢**

項目IDと演算子を用い、論理式（Java様式）で記述
複合演算・基本的な集合演算が可能

H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

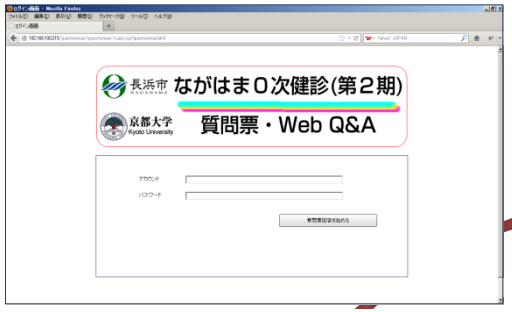
項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

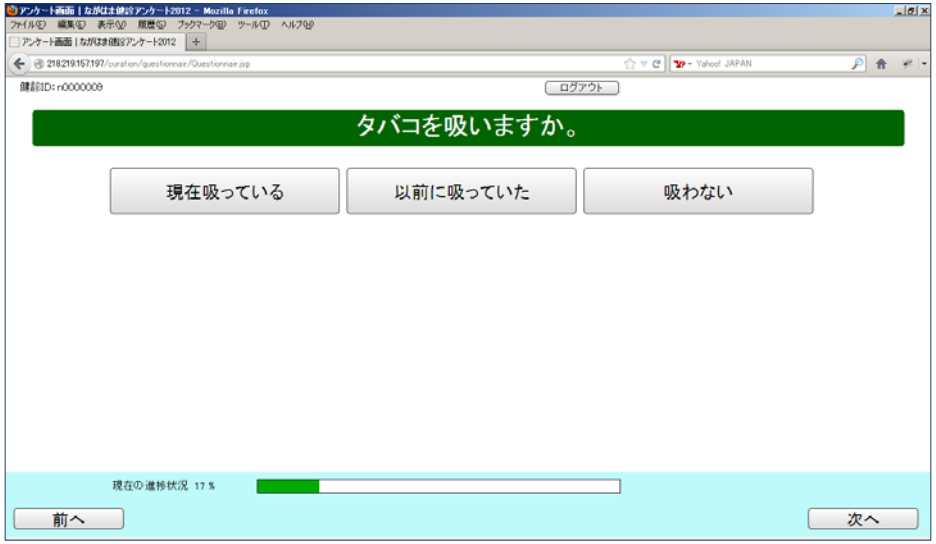
データ収集省力化ツールの開発

1. WebQ&Aの開発と実装
2. 健診ワークフロー一元化管理機能の実装
3. データ登録・キュレーションフォームの改良
4. 実験データアップロード機能の開発

1. WebQ&Aの開発と実装



自宅から事前回答



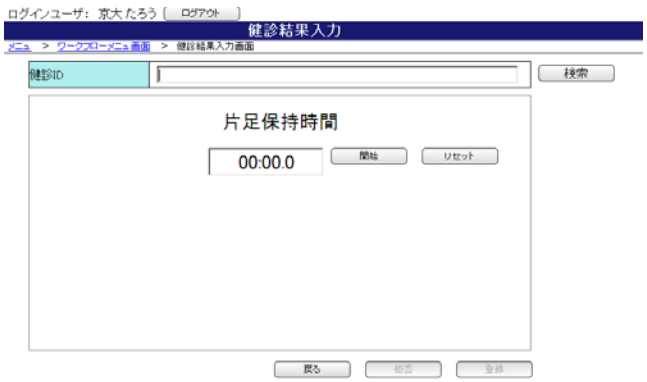
一問一答形式のWebQ&A

健診会場にて
iPadで回答

2012年度ながはま０次健診
自宅回答率

60歳以上	26%
60歳未満	46%

2. 健診ワークフロー一元化管理機能の実装



健診会場におけるバーコード印刷やストップウォッチ入力

健診順序のコントロールと健診の進捗の一元的な管理

3. データ登録・キュレーションフォームの改良

入力フォームの自動作成

ID	枝番	プロトコル項目名	カテゴリ	表示順	見出し	表示方法
10	1	施設名	患者基本情報／その他	10		セレクトボックス
11	1	性別	患者基本情報／その他	20		ラジオボタン(横)
12	1	診断時の年齢	患者基本情報／その他	30		
13	1	現在の喫煙	患者基本情報／その他	40		ラジオボタン(横)
14	1	喫煙中の場合 1日の本数	患者基本情報／その他	50		
15	1	喫煙中の場合 何年前から喫煙…	患者基本情報／その他	60		
16	1	禁煙者の場合 何年前に禁煙し…	患者基本情報／その他	70		
17	1	禁煙者の場合 それまで1日何…	患者基本情報／その他	80		
18	1	禁煙者の場合 それまで何年間…	患者基本情報／その他	90		
19	1	アルコール摂取	患者基本情報／その他	100		ラジオボタン(横)
20	1	アルコール摂取量(毎日の場合)	患者基本情報／その他	110		
21	1	既往歴	患者基本情報／その他	120		
22	1	悪性腫瘍の家族歴(第一親等)	患者基本情報／その他	130		ラジオボタン(横)
23	1	悪性腫瘍の家族歴ありの場合…	患者基本情報／その他	140		
24	1	IgG4値(診断時)	患者基本情報／その他	150	IgG4関連疾患の診断	
25	1	IgG4関連疾患の臓器	患者基本情報／その他	160		ラジオボタン(横)
26	1	特徴的なびまん性あるいは限…	患者基本情報／その他	170		
27	1	IgG4関連疾患の組織	患者基本情報／その他	180		ラジオボタン(横)
28	1	リンパ球・形質細胞の著明な浸…	患者基本情報／その他	190		
29	1	臓器特異的診断基準を利用し…	患者基本情報／その他	200		ラジオボタン(縦)
30	1	IgG4関連疾患 or 自己免疫性…	患者基本情報／その他	210		
31	1	膵画像	患者基本情報／その他	220	国際膵臓学会の診断	ラジオボタン(縦)
32	1	ERP像	患者基本情報／その他	230		ラジオボタン(縦)

いずれも、用いるデータ項目のセットを決めれば、Webフォームは動的に作成される。また、体裁の微調整もWeb画面から可能

H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

いかにして疾患罹患情報を取得するか

- 質問票による調査

  自己申告情報の確度に問題

- 特定健診の情報

  受診率の低さに問題

- 地域の疾病登録制度の利用（がん登録など）

  情報が一部の疾患に限定される

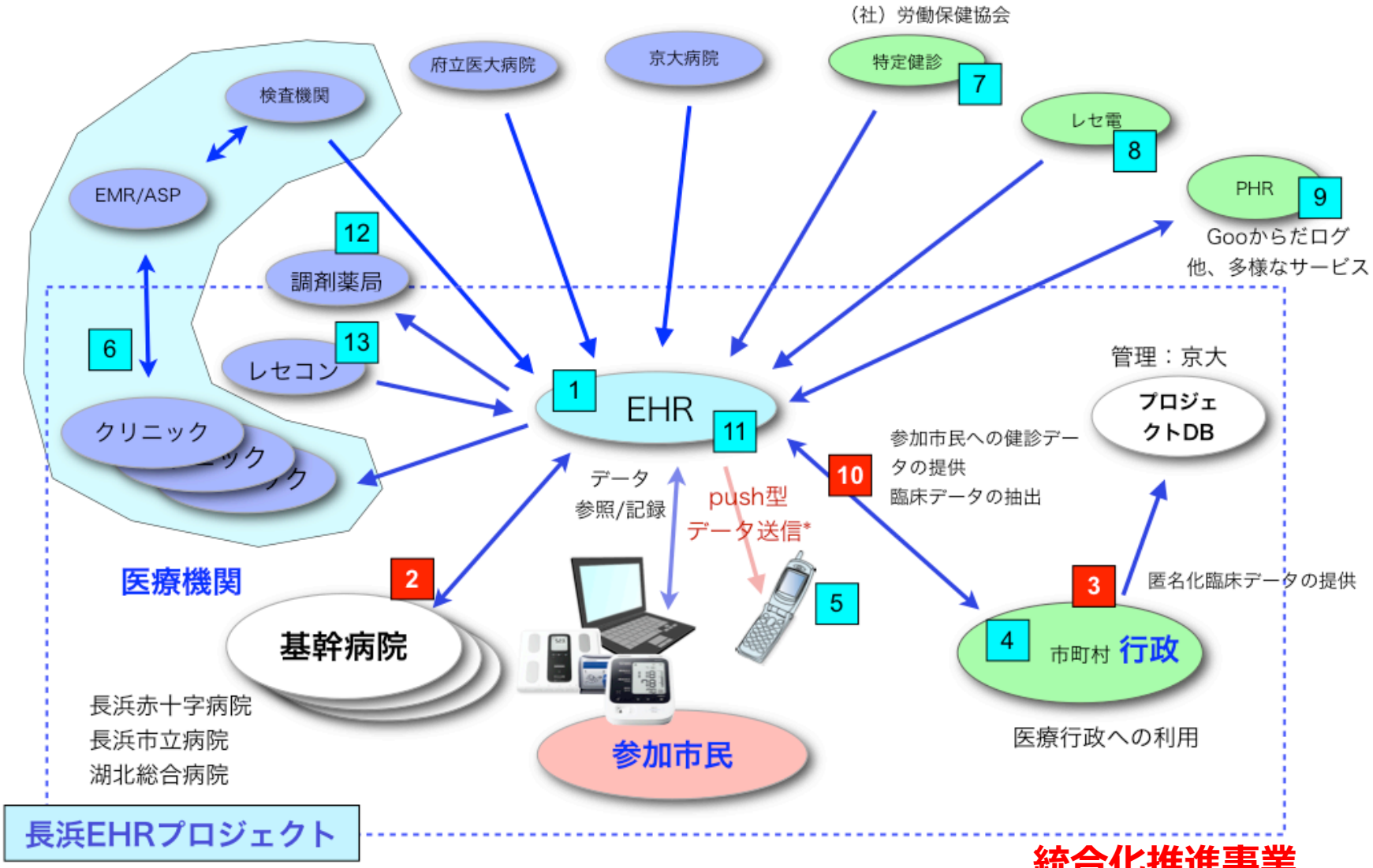
- 移動・死亡情報（住民基本台帳、死亡小票の閲覧）

  死亡者の情報のみしか集まらない

EHRを利用した疾患関連情報取得の検討を開始

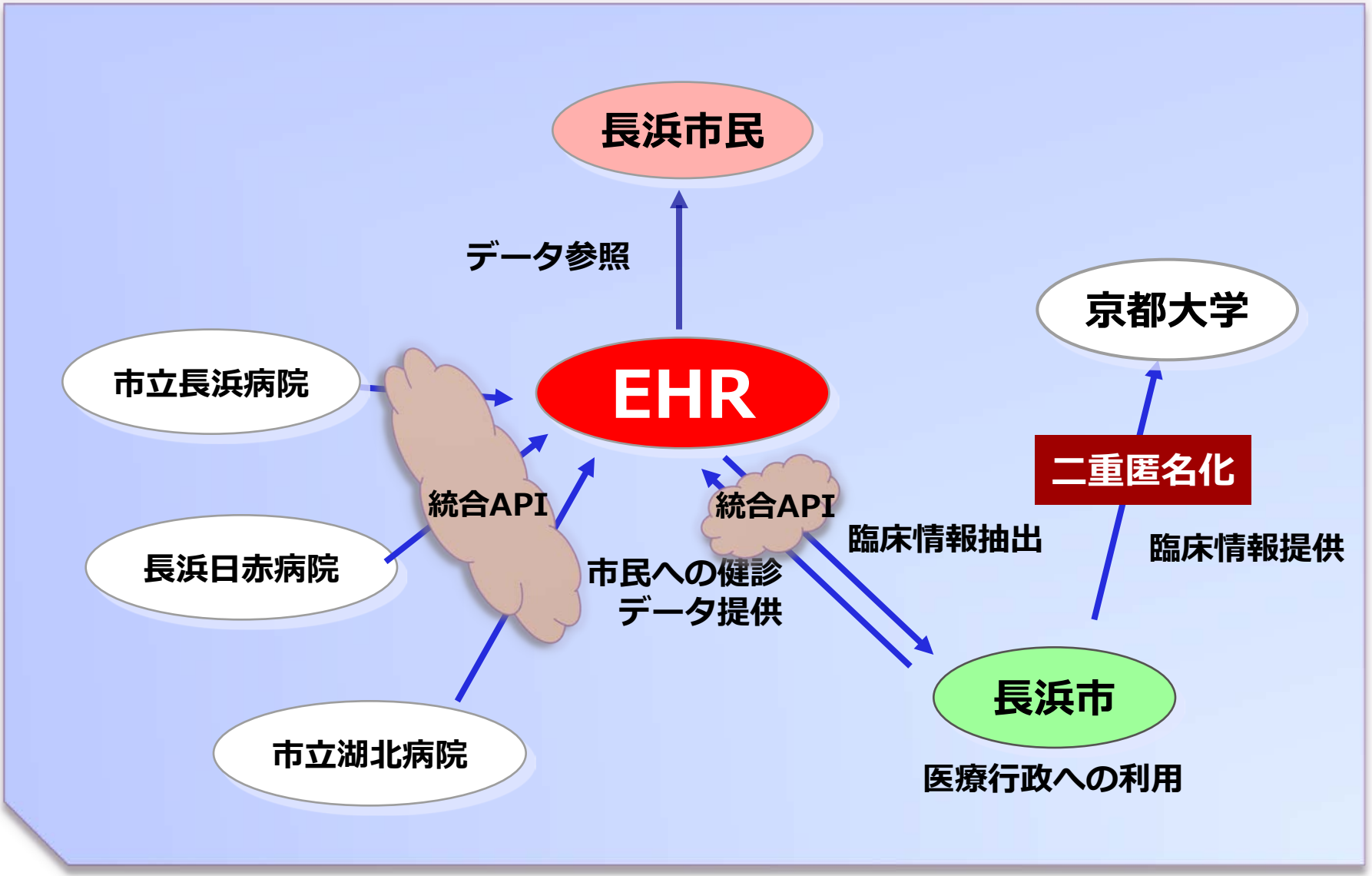
EHRを利用した医療ネットワークの構築

― 京都大学「まいこネット」の例 ―



統合化推進事業

まいこネットの疫学研究利用



H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

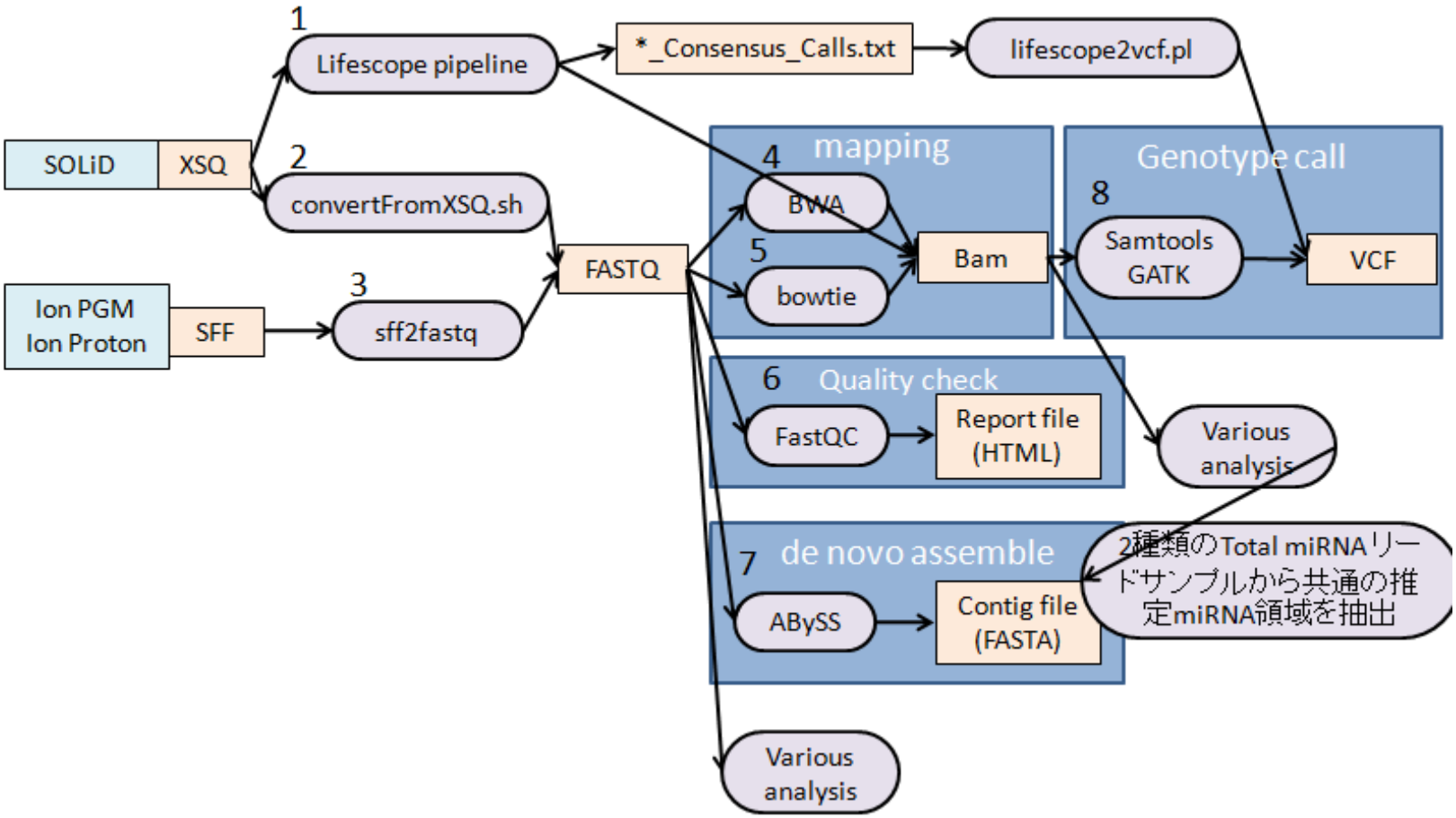
項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

次世代シークエンサー解析パイプライン



CUIベースのワークフロー

H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの改良

項目２　データ共有の仕組みの実装

項目３　データ項目の標準化

項目４　データ収集省力化ツールの開発

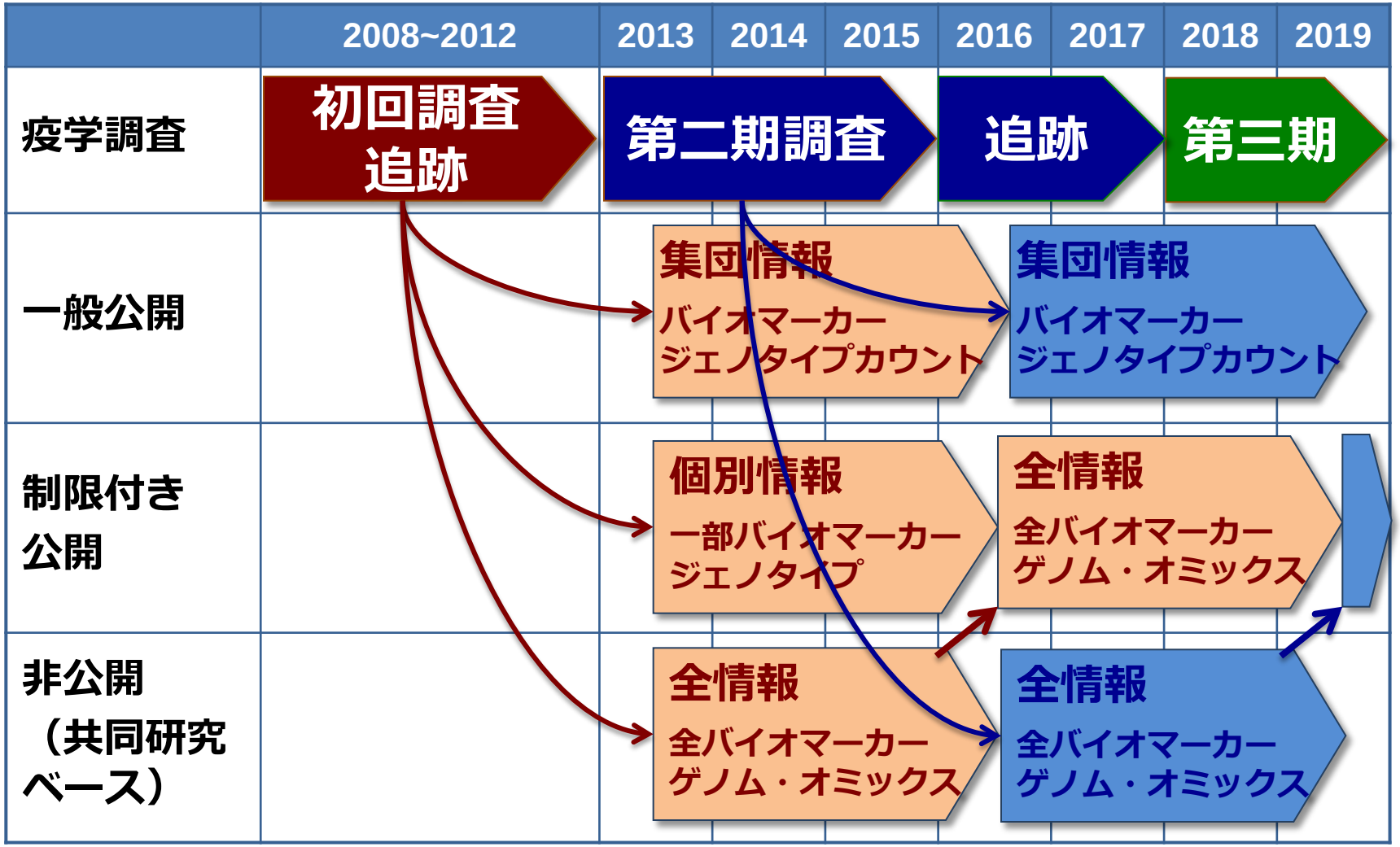
項目５　EHRによる疾患関連情報の取得

項目６　データ解析手法の開発

項目７　データ公開用インタフェースの開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

データの公開計画



SNP DB Browser

Home | About | Statistics | Link | Contact | Message

Welcome to SNP DB Browser

SNP DB Browser is created by Kyoto university. The database includes more than *** snp data. **(more...)**

Search

You can search the SNP database. Please input the keyword(e.g.:gene name,gene_id,chromosome number etc) and press the Search button.

Keyword Seach GO

Viewer

You can see the viewer directly by clicking the chromosome number below.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

What's New?

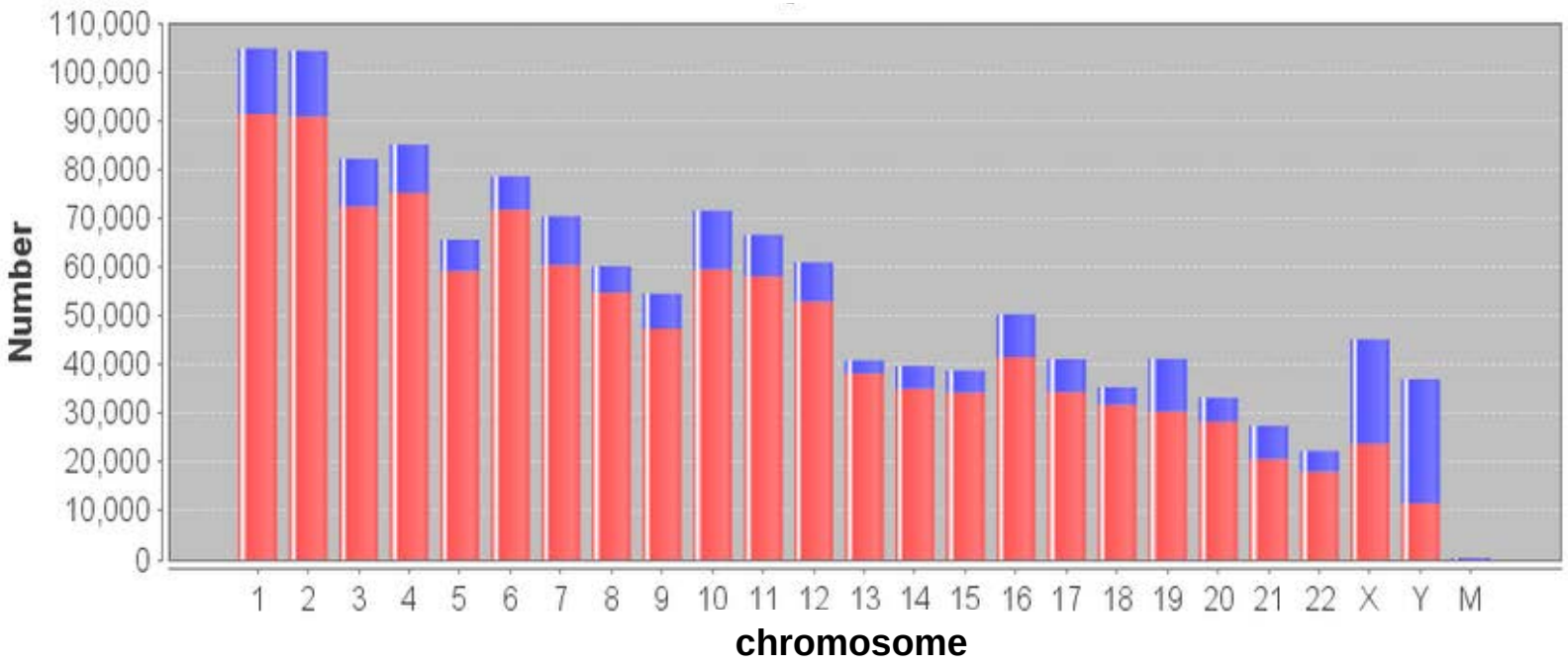
- **31.10.2012** Site is opened.

SNP DB Browser

Home | About | Statistics | Link | Contact | Message

Statistics

Number of SNPs in each chromosome



Known SNPs with rs number newly identified SNP

H24年度の開発項目

項目１　個人情報保護・匿名化枠組みの洗練

項目２　データ共有の仕組みの実装

項目３　EHRによる疾患関連情報の取得

項目４　データ項目の標準化

項目５　データキュレーション用インタフェース

項目６　質問紙調査のWeb化

項目７　解析手法の開発

項目８　バイオインフォマテイシャン・遺伝統計家の養成

FRONTIER MEDICINE
G-COE KYOTO

Shiran Kaikan
Kyoto University Faculty of Medicine
January 15th ~ 18th, 2013

NBDC
National Bioscience Database Center

Kyoto Course on Bioinformatics for Next Generation Sequencing with Applications in Human Genetics

A four-day course consisting of lectures and hands-on training for the analysis of NGS data

Co-organized by
Center for Genomic Medicine, Kyoto University
National Bioscience Database Center
Kyoto University Global COE Program 'Center for Frontier Medicine'

Supported by
The Molecular Biology Society of Japan
The Japan Society of Human Genetics

Day 1

EXOME AND WHOLE GENOME SEQUENCING USING MASSIVELY- PARALLEL SEQUENCERS

10:15 - 11:00	Lecture:	Introduction to next generation sequencing technologies and applications Mark Lathrop (CEPH, France)
11:30 - 12:30	Lecture:	Understanding the high-throughput sequencing data analysis workflow Guillaume Bourque (McGill, Canada)
13:30 - 14:30	Lecture:	Quality assessment and filtering, mapping strategies for sequence reads Guillaume Bourque (McGill, Canada)
15:00 - 16:00	Lecture and Practical: Rare variants association analyses Daniel E. Weeks (Pittsburgh, U.S.A.)	

Day 2

APPLICATIONS TO RNA ANALYSIS
Mar Gonzàlez-Porta (EBI, Cambridge U.K.)

09:00 - 10:00	Lecture: Introduction to RNA-seq: expression estimation and normalization
10:10 - 11:00	Lecture: Differential expression and differential exon usage
11:15 - 12:30	Practical: Dealing with raw RNA-seq data: quality assessment, filtering and mapping

DATA SUBMISSION
Rajesh Radhakrishnan (EBI, Cambridge U.K.)

13:30 - 14:30	Lecture: ENA introduction
14:30 - 15:30	Lecture: SRA introduction, submission, browsing, and data retrieval
16:00 - 16:25	Lecture: Data compression
16:25 - 17:55	Practical: ENA submission

Day 3

PRACTICALS FOR EXOME AND RNA SEQ

09:00 - 12:30 Practical: Exome Sequencing Data Analysis
Guillaume Bourque (McGill, Canada)

13:30 - 17:00 Practical: Dealing with aligned data
(including differential expression analysis)
Mar Gonzàlez-Porta (EBI, Cambridge, U.K.)

Day 4

HUMAN GENETICS

09:00 - 10:00	Lecture: Introduction to disease gene mapping
10:10 - 11:00	Practical: Disease gene mapping Jurg Ott (Rockefeller, U.S.A.)
11:30 - 12:30	Lecture: Imputation and testing imputed SNPs
13:30 - 14:30	Practical: Imputation and testing imputed SNPs Marcella Devoto (Pennsylvania, U.S.A.)
14:30 - 15:30	Lecture: Statistical tests for NGS data and heterogeneity statistics for association analysis
16:00 - 17:00	Practical: Statistical tests for NGS data and heterogeneity statistics for association analysis Derek Gordon (Rutgers, U.S.A.)
16:20 - 16:55	Lecture: Cancer genetics and genomics: Of Mice and Men Makoto Mark Taketo (Kyoto, Japan)

H25年度の研究開発予定

- 疫学データの収集
  - 第二期健診実施（約千名が対象）
  - EHRによる臨床情報の収集、死亡小票による死因調査
  - 第一期血漿のGC-MS解析（約一万サンプル）
- 疫学システムの改良
  - 被験者の複数ID処理の自動化
  - EHR連携機能の強化
  - データ項目検索機能の向上、パフォーマンスチューニング
  - システムのパッケージ化
- データ公開
  - 個人のジェノタイプ情報、一部疫学情報の制限付き公開
- バイオインフォマティシャン・遺伝統計家の養成
  - 第二回トレーニングコースの開催