

大規模ゲノム疫学研究の 統合情報基盤の構築

～ここまでは「ライフサイエンスデータベースの統合」～

シンポジウム2013 トーゴの日

京都大学医学研究科附属ゲノム医学センター

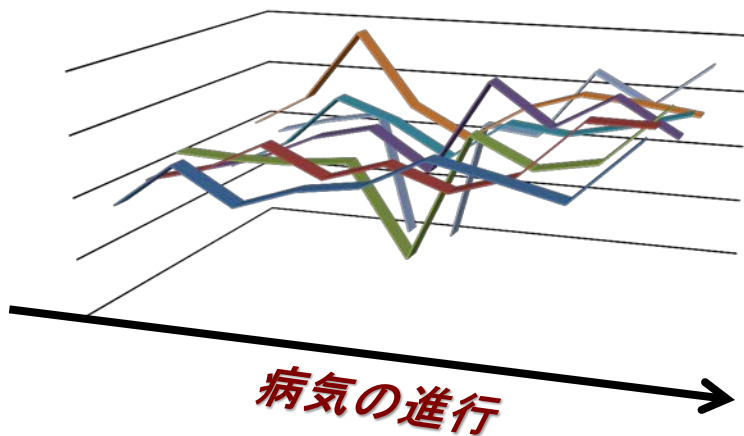
松田文彦

2013年10月4日(金)
9:45～17:05

2013年10月5日(土)
9:30～17:10

生命分子の動きを通して健康や病気をながめる

生物の身体の中では、多くの生命分子が常に変化しつつ生命活動を支えている



慢性疾患は**体内の微小な変化**が時間とともに蓄積され**緩徐に進行**するため、健常と病気の区別が極めて困難
なうえ表現型も異なる

ヒトの誕生から死までの時間経過の中で、体質の多様性や老化といった**正常の生命活動とともに病気を理解する縦断的アプローチ**が不可欠

大規模ゲノムコホート研究で得られる情報を正確、網羅的、かつ効率的に蓄積するための情報基盤を構築する

ゲノムとオミックス(転写物、タンパク質、代謝物など)の網羅的解析によるバイオマーカーの探索

- 疾患の予知、診断、予後予測、治療法の開発や創薬には、網羅的ゲノム解析による関連遺伝子の同定に加え、新たな疾患関連バイオマーカーの探索が必須
- 大規模コホート集団の網羅的ゲノム解析に加え、時系列を考慮した血液・尿中の生体分子の網羅的解析を実施
- 疾患関連遺伝子や新たなバイオマーカーを利用した疾患の分子機序の解明や、患者群内での亜集団解析による重症度・予後予測因子の同定

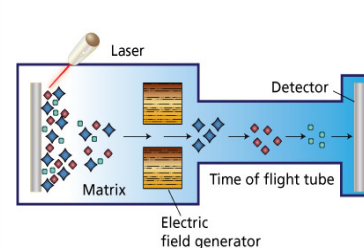
ゲノム



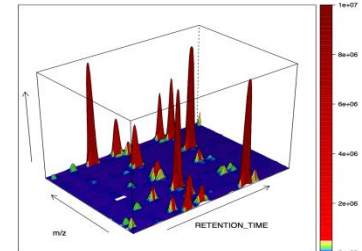
トランスクリプトーム



プロテオーム



メタボローム



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、ゲノム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。

ながはま0次予防コホート事業（ながはまコホート）

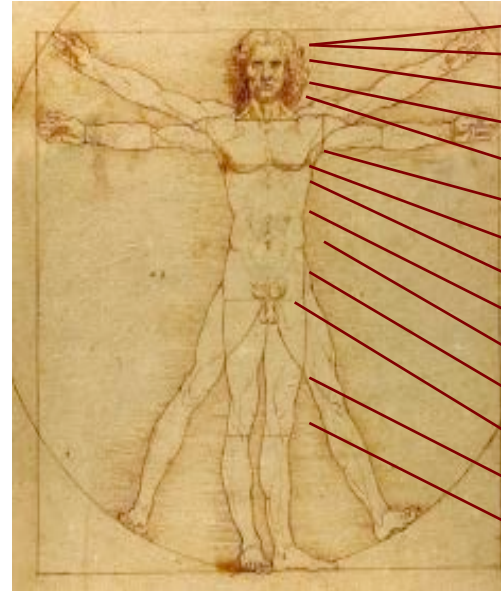


湖北地域の地方中堅都市
(人口 約120,000人)

- 65歳以上老年人口割合は18～20%
(我が国の平均値とほぼ一致)
- 住民の移動が少ない
(80%以上が、長浜市で一生を過ごす)
- 3つの基幹病院で長浜市を中心に20万人以上の診療圏

- 滋賀県長浜市の健康な住民（30歳～74歳）**1万人**を対象とし、5年ごとの詳細な健康診断と追跡をベースとした**20年以上の長期縦断ゲノムコホート研究**
- 京都大学医学研究科の総力を結集し、疾患のみならずヒトの持つ**健常体質の多様性や老化**にも目をむけた、多目的型の**ヒト生物学研究**

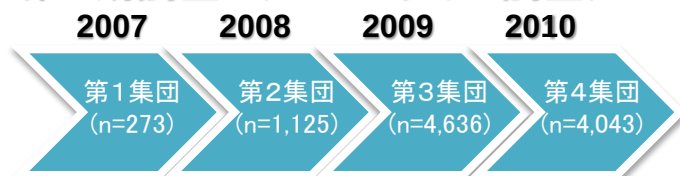
ながはまコホートの対象疾患



- 認知症・MCI
- 脳血管障害
- 高度近視
- 睡眠障害
- 歯周病・咀嚼機能・顎関節疾患
- ぜんそく・COPD
- 虚血性心疾患
- 高血圧・動脈硬化
- 2型糖尿病・肥満・NASH
- 自己免疫疾患
- 椎体圧迫骨折・脊椎疾患
- 前立腺がん・排尿障害
- ロコモティブシンドローム
- 骨粗鬆症・変形性膝関節症

ながはま0次予防コホート事業（ながはまコホート）

第1期調査（ベースライン調査）

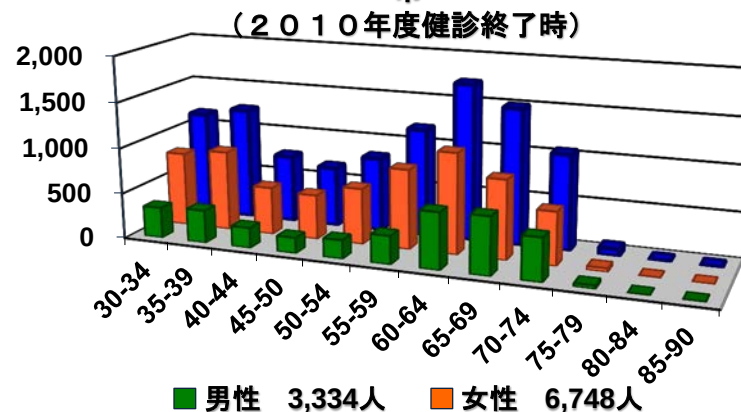


第2期調査（フォローアップ調査）



認知機能検査・MRIを導入

ながはまコホートの参加者数と年齢分布



大規模生体試料バンク

DNA (数10μg)	10,082 検体
血漿・血清 (各3ml)	10,082 検体
尿 (10ml)	10,082 検体
RNA	300 検体

(第1期調査分を集計)

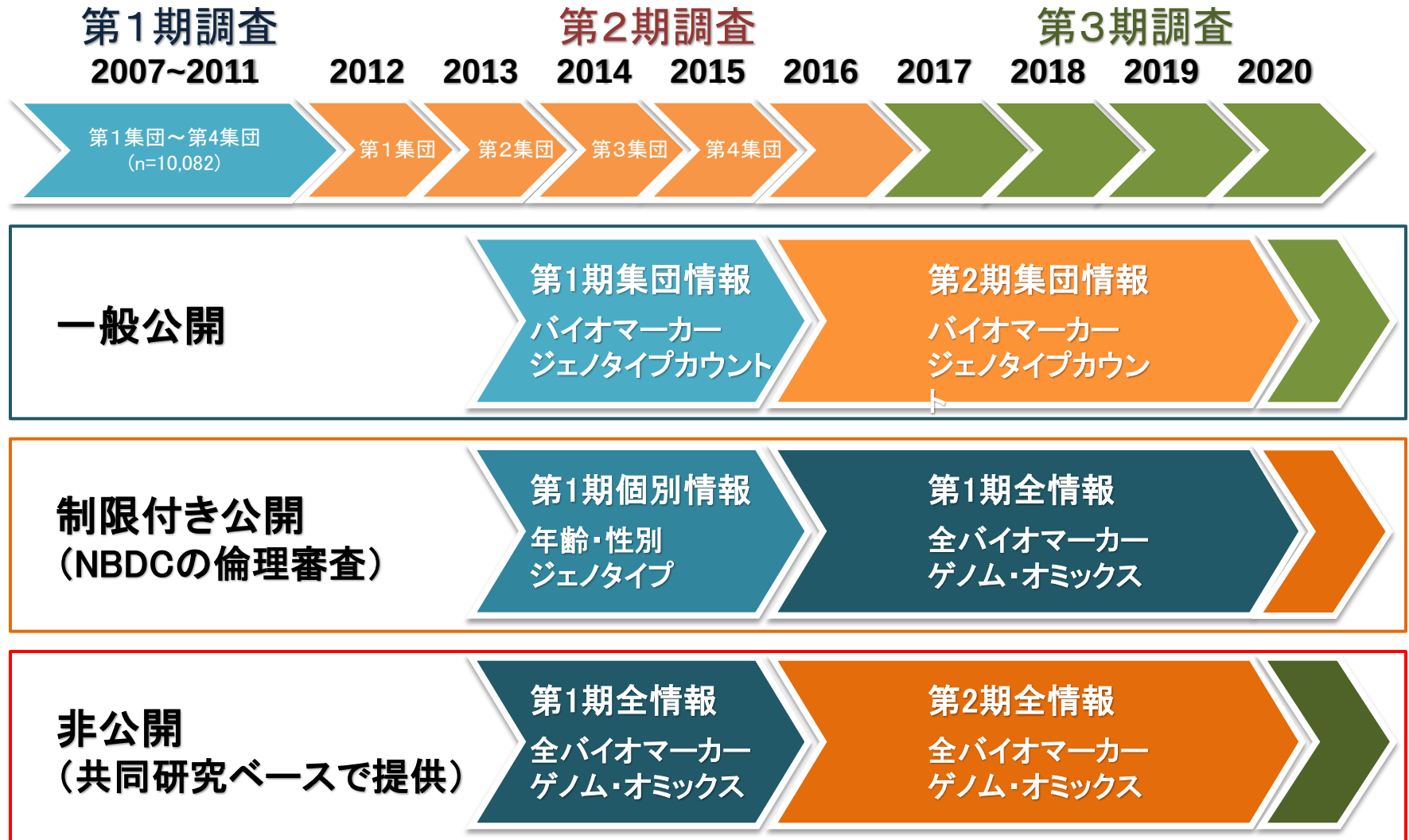
生命情報統合データベース

網羅的SNP解析	4,000 検体
エクソームシーケンス	300 検体
転写物	300 検体
低分子水溶性代謝物	2,000 検体
生活習慣調査・疾患スクリーニング	742 項目
生理学／生化学／血液学検査	145 項目

本事業でNBDCに蓄積するデータ（第1期分）

- 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
- ゲノムスキャン情報
 - Illumina610K 約1,800検体、2.5M 約1,800検体 （2013年9月）
- エクソームシーケンス情報
 - 300検体 （2013年9月）
- トランスクリプトーム情報
 - 300検体 （2013年9月）
- 網羅的メタボローム解析情報
 - 水溶性低分子化合物 119化合物分 約2,000検体 （2013年9月）

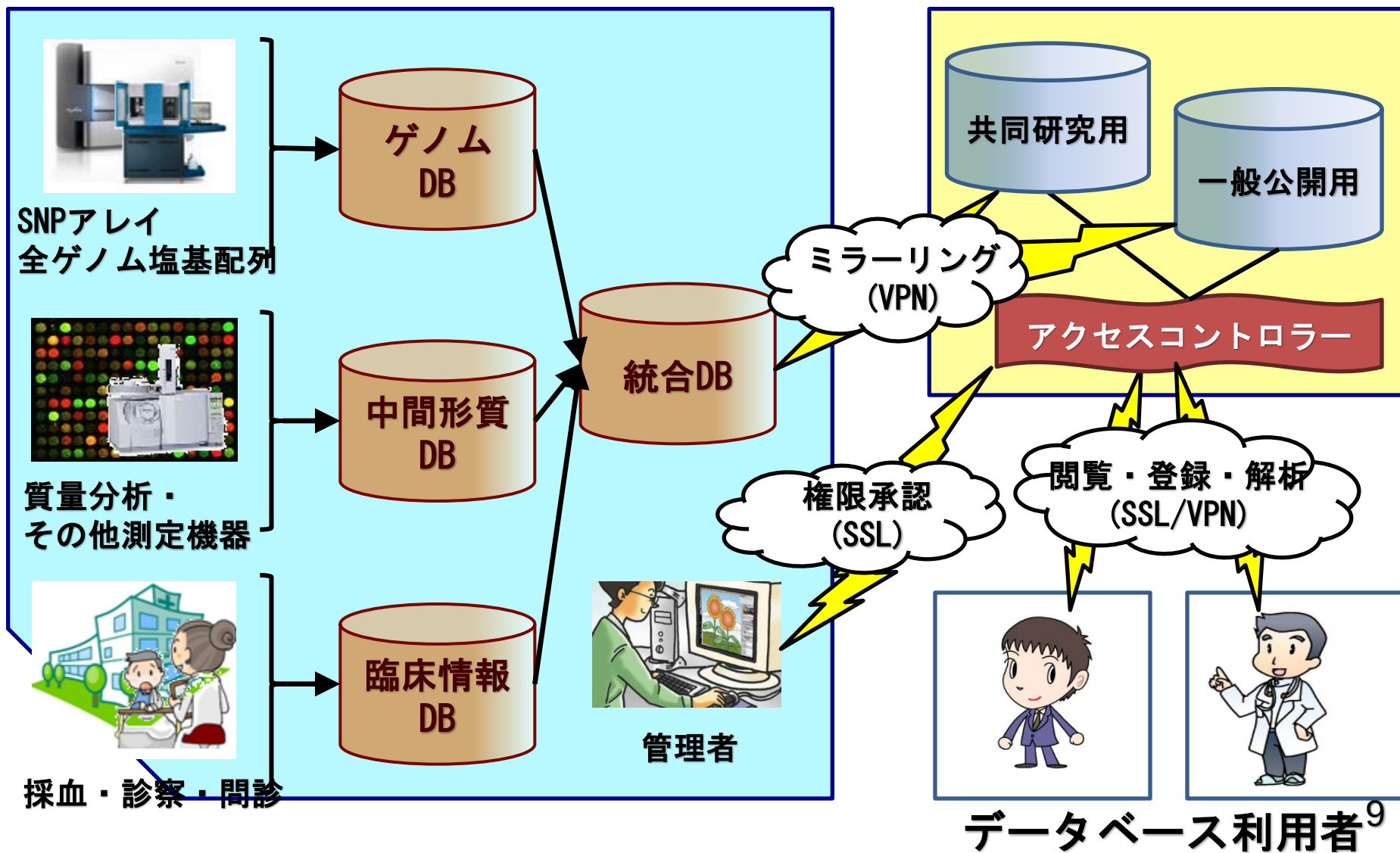
データ公開の計画



データベースシステム概念図

京 都 大 学

NBDC



本事業での開発項目

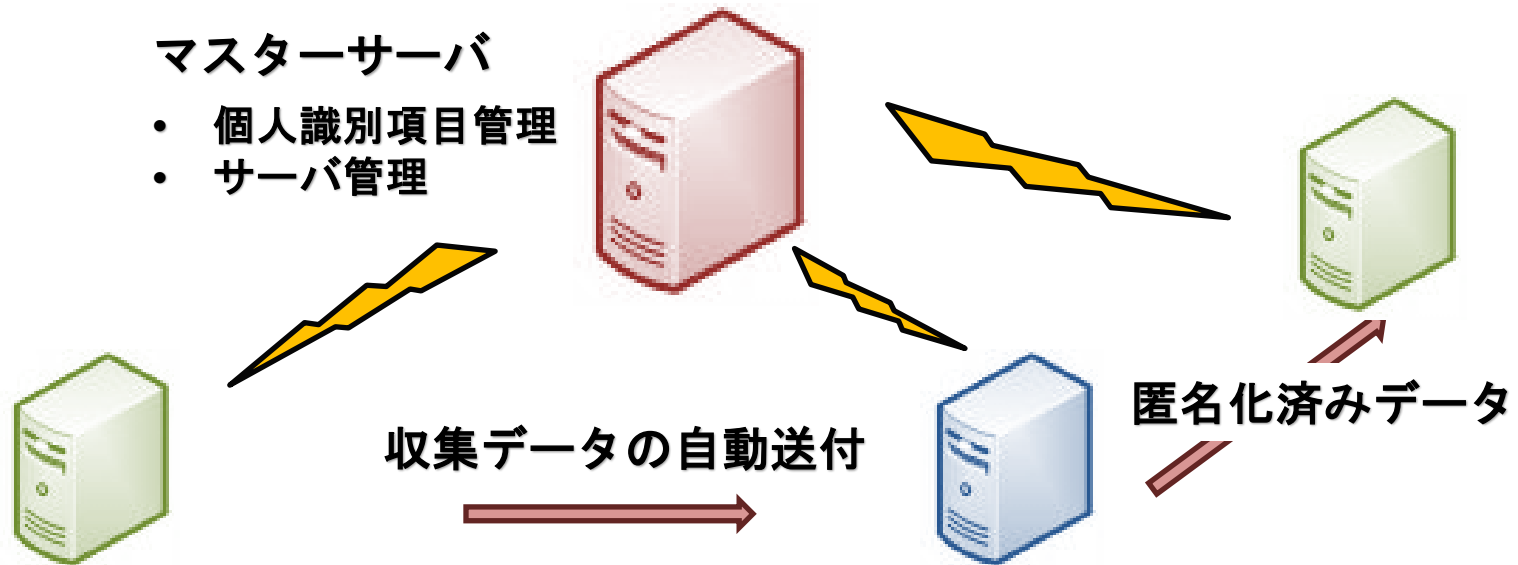
1. 個人情報保護・匿名化の枠組み
2. データ共有システムの実装
3. データ項目の標準化
4. データ収集省力化ツールの開発
5. EHRによる疾患関連情報の取得
6. データ解析手法の開発
7. データ公開用インタフェースの開発

本事業での開発項目

1. 個人情報保護・匿名化の枠組み
2. データ共有システムの実装
3. データ項目の標準化
4. データ収集省力化ツールの開発
5. EHRによる疾患関連情報の取得
6. データ解析手法の開発
7. データ公開用インタフェースの開発

1. 個人情報保護・匿名化の枠組み

～マスターサーバによる情報の一元管理～



マスターサーバ

- ・ 個人識別項目管理
- ・ サーバ管理

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

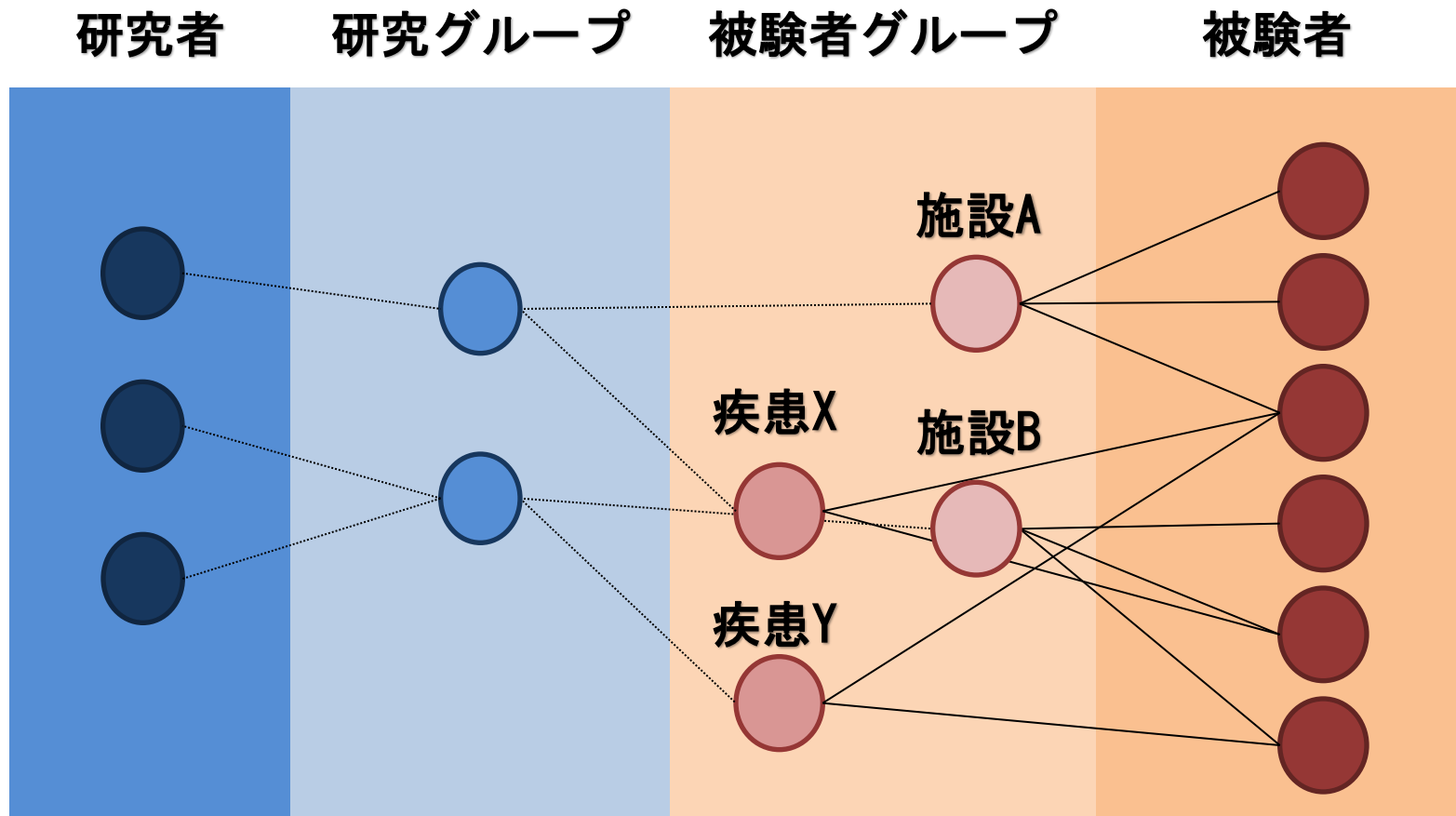
2. データ共有システムの実装 ～機能・データの種類によるアクセス制限～

	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 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

2. データ共有システムの実装

～アクセス可能な被験者情報のコントロール～

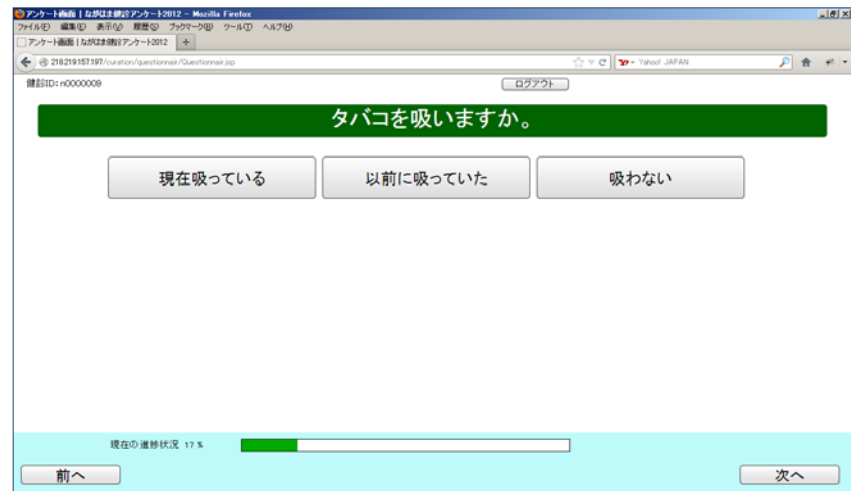


いずれも、GUIから設定可能

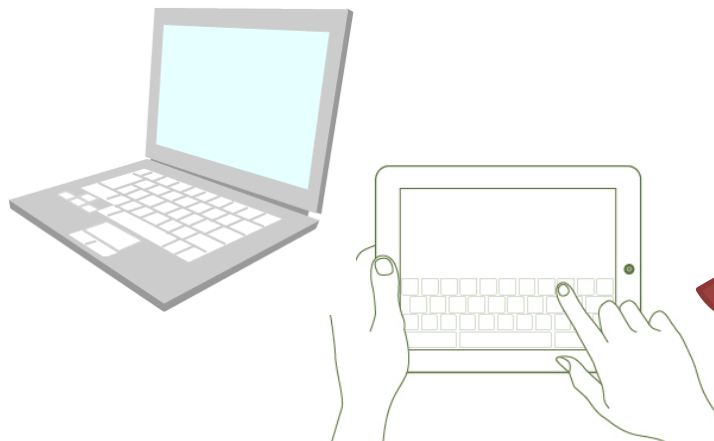
4. データ収集省力化ツールの開発 ～WebQ&Aの開発と実装～



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%
-------	-----

4. データ収集省力化ツールの開発 ～健診ワークフロー一元化管理機能の実装～

ログインユーザ: 京大 太郎 (ログアウト)

健診結果入力画面

健診ID: 検索

血液検査

8 枚 印刷

コメント/再入力理由

戻る 印刷

ログインユーザ: 京大 太郎 (ログアウト)

健診結果入力

健診ID: 検索

片足保持時間

00:00.0 開始 リセット

戻る 印刷 印刷

ブラウザ: http://172.22.254.1/workflow/workflow/Wor_ID = 〇 X

おすすめサイト: Web スライス ギャラリー Web スライス 確認 (全体)

ログインユーザ: 京大 太郎 (ログアウト)

ステータス確認 (全体) 画面

予約人数: 40 受付人数: 40 実検数: 0 検診日: 2012 年 12 月 01 日 09:10:00 印刷

表示件数: 10 全 40 件 1 2 3 4

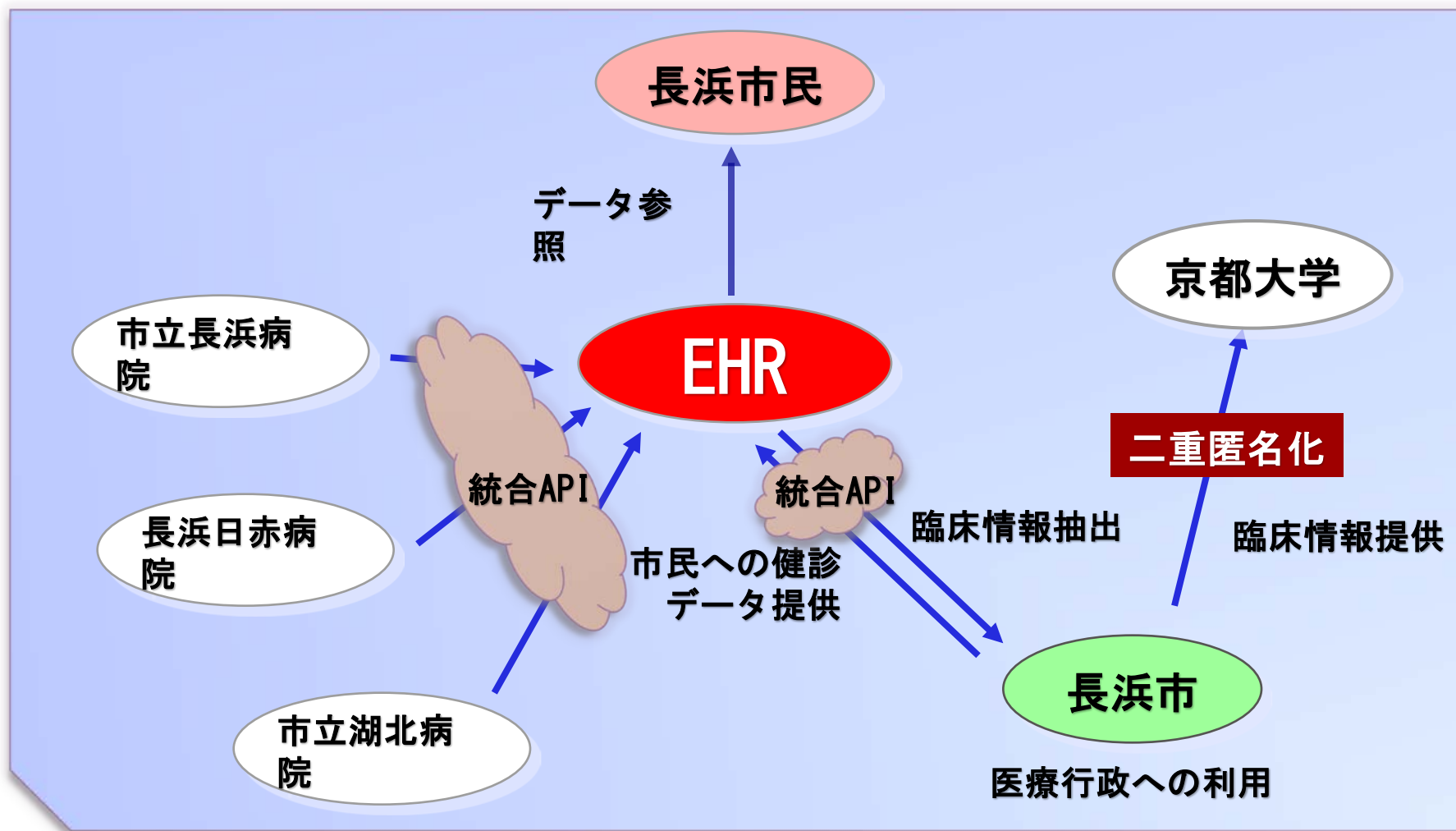
健診ID	受付	血液検査	片足保持時間	それ以外	検診、検査し	コンタクト	身長	体重	算出BMI
2007552	○	○	○	○	○	○	○	○	○
2007571	○	○	○	○	○	○	○	○	○
2007100	○	○	○	○	○	○	○	○	○
2047054	○	○	○	○	○	○	○	○	○
2007026	○	○	○	○	○	○	○	○	○
2140381	○	○	○	○	○	○	○	○	○
2071933	○	○	○	○	○	○	○	○	○
2081077	○	○	○	○	○	○	○	○	○
2064153	○	○	○	○	○	○	○	○	○
2140450	○	○	○	○	○	○	○	○	○

1 ~ 10 / 全 40 件 1 2 3 4

健診順序のコントロール
健診の進捗の一元的な管理

健診会場におけるバーコード
印刷
ストップウォッチ入力

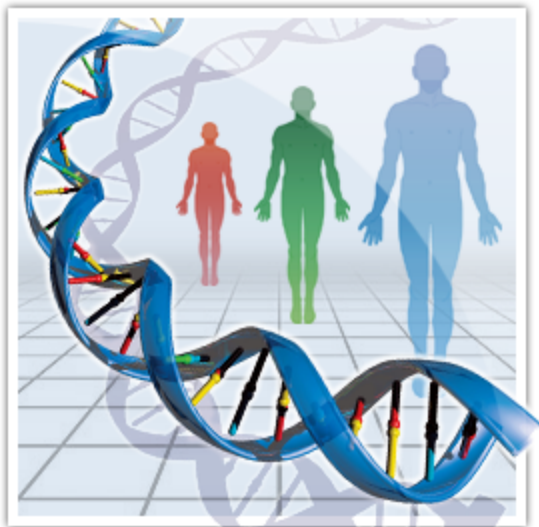
5. EHRによる疾患関連情報の取得 ～ 京都大学「まいこネット」の例 ～



7. データ公開用インターフェースの開発

Human Genetic Variation Browser

[Home](#) [About](#) [Statistics](#) [Link](#) [Contact](#) [Message](#) [Login](#)



Welcome to **Human Genetic Variation** Browser

Search database

Gene name/ID

GO

dbSNP rsID

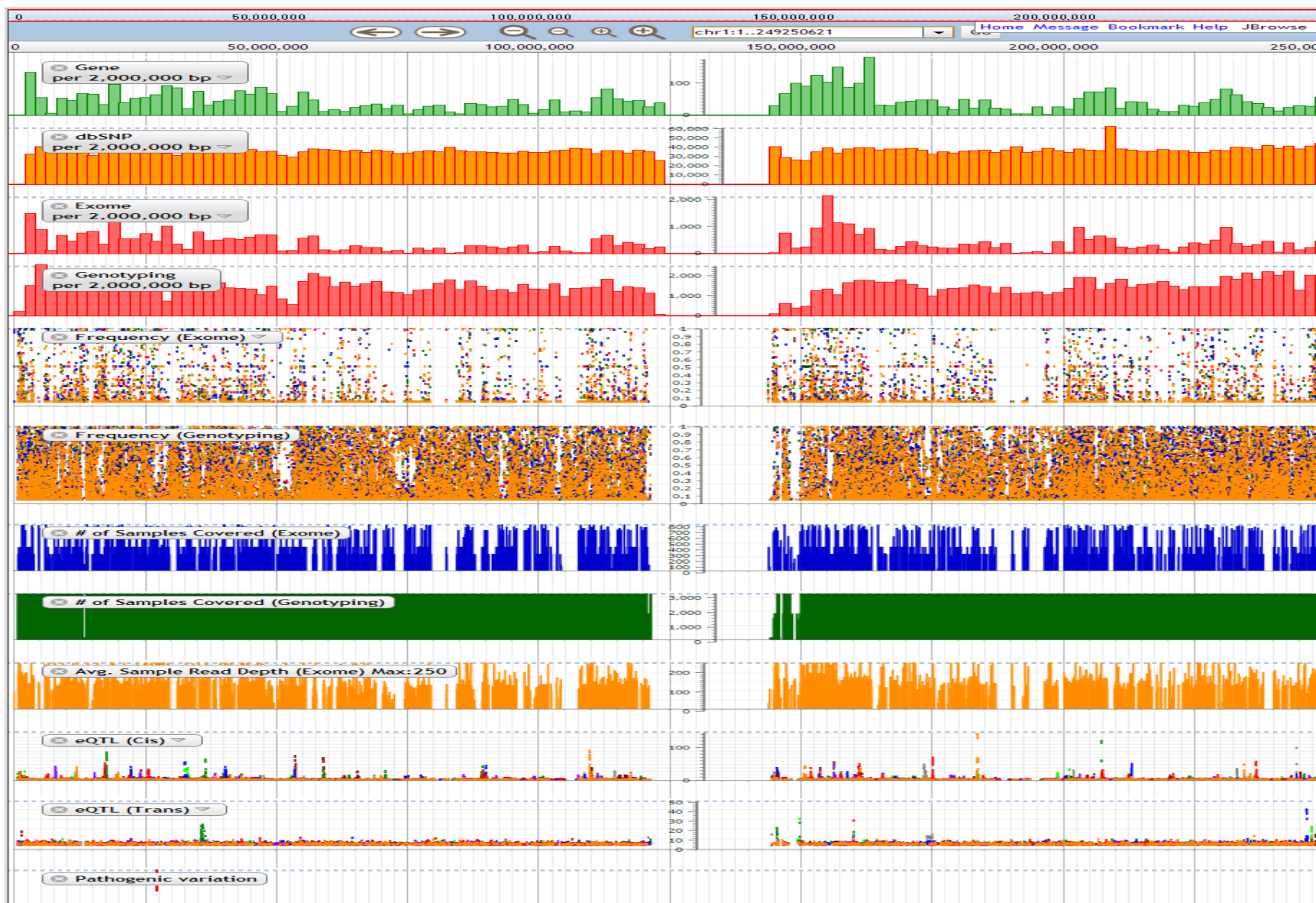
GO

Pathogenic Variation

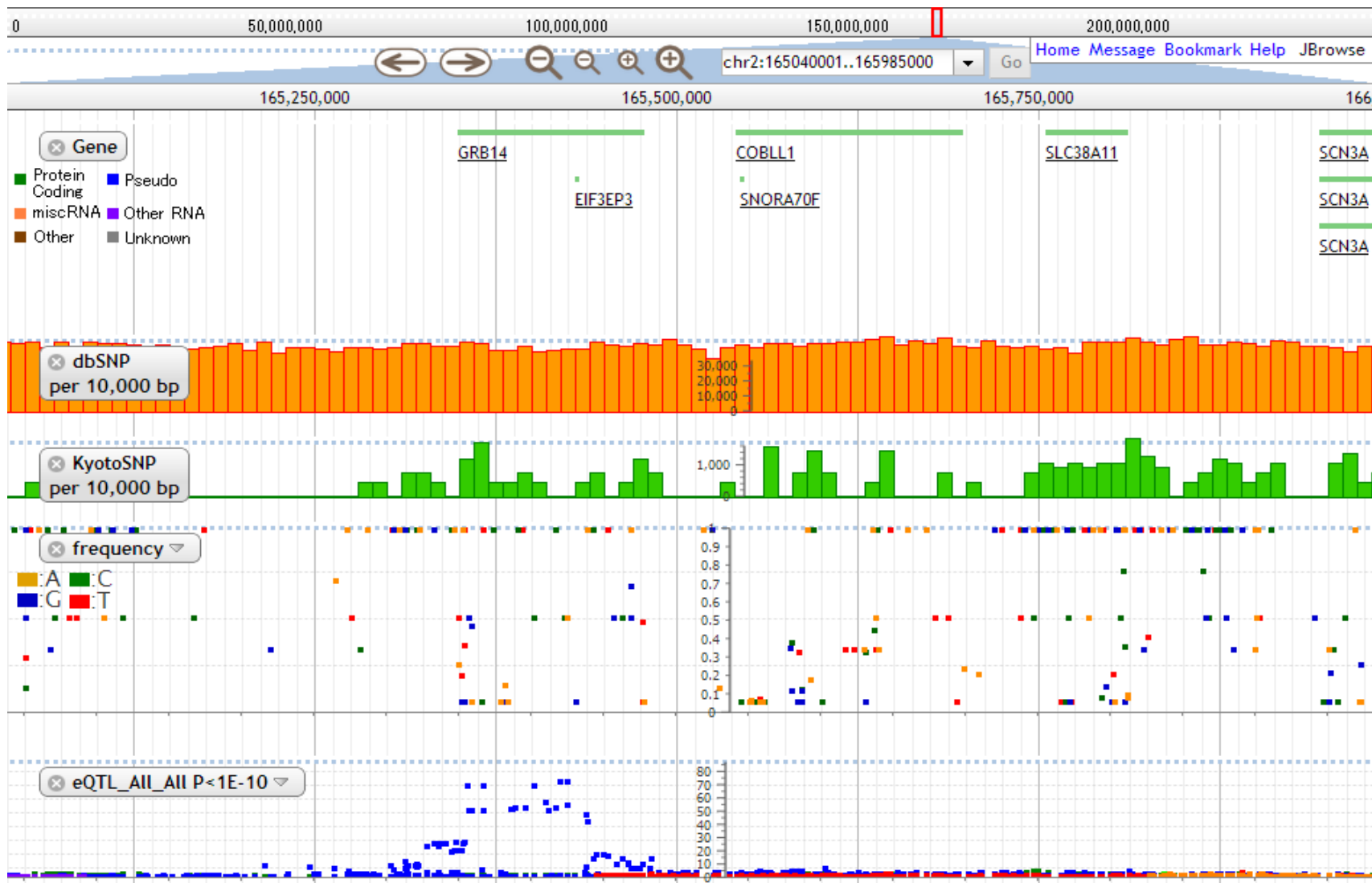
GO

Chromosome

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



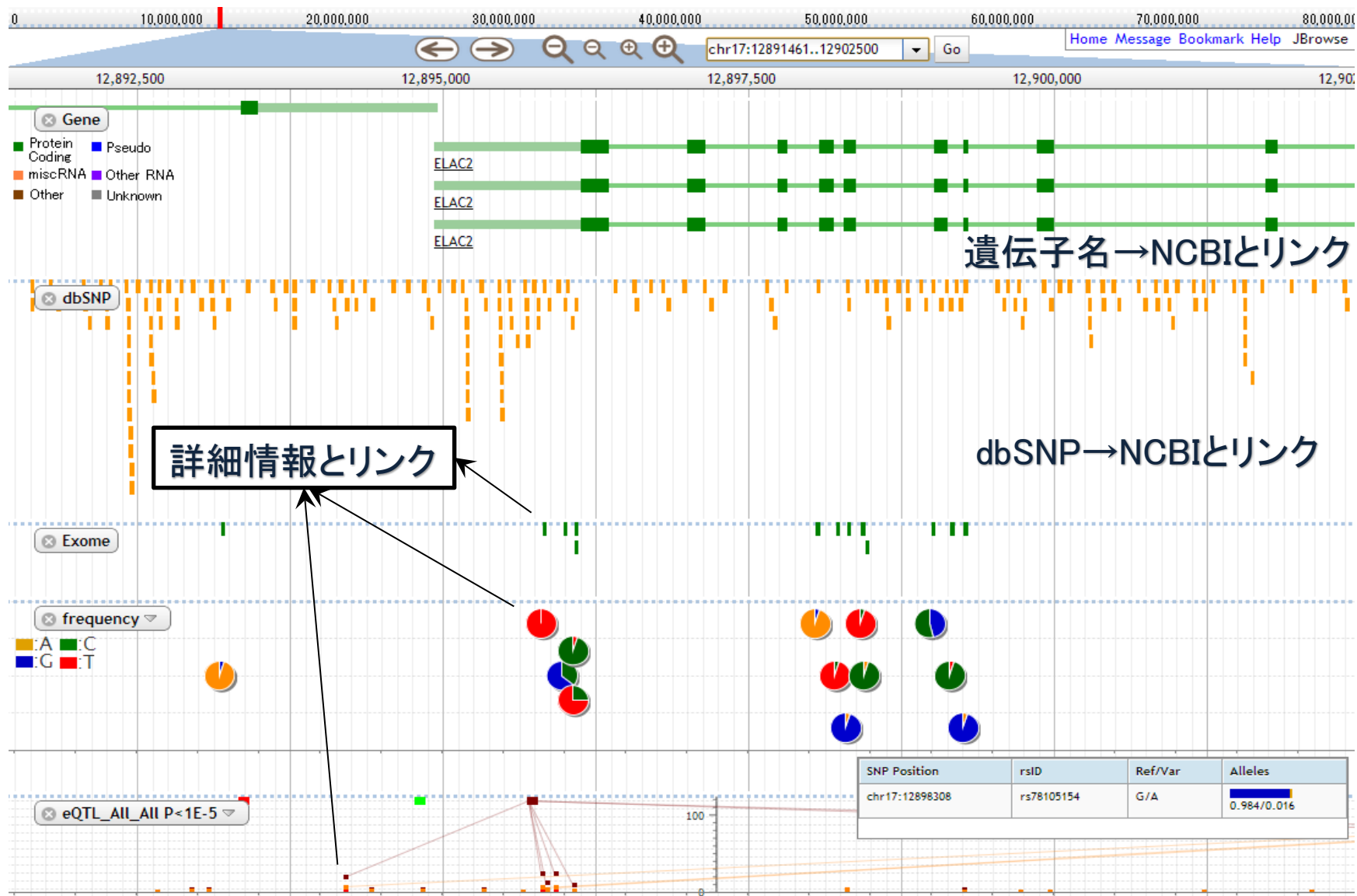
染色体 1 番の情報表示画面：物理地図に沿い、遺伝子、ゲノム変異、アレル頻度、検体数、被覆率、eQTL情報を表示



ズーム機能による拡大図

Frequency : non-reference側のアレル頻度を0～1の範囲で縦軸に表示

eQTL : $-\log_{10}P$ の値を縦軸に表示



ズーム機能による頻度の円グラフ表示、マウスオーバー機能による詳細情報表示

eQTL解析におけるプローブとSNPの関係を線で連結表示

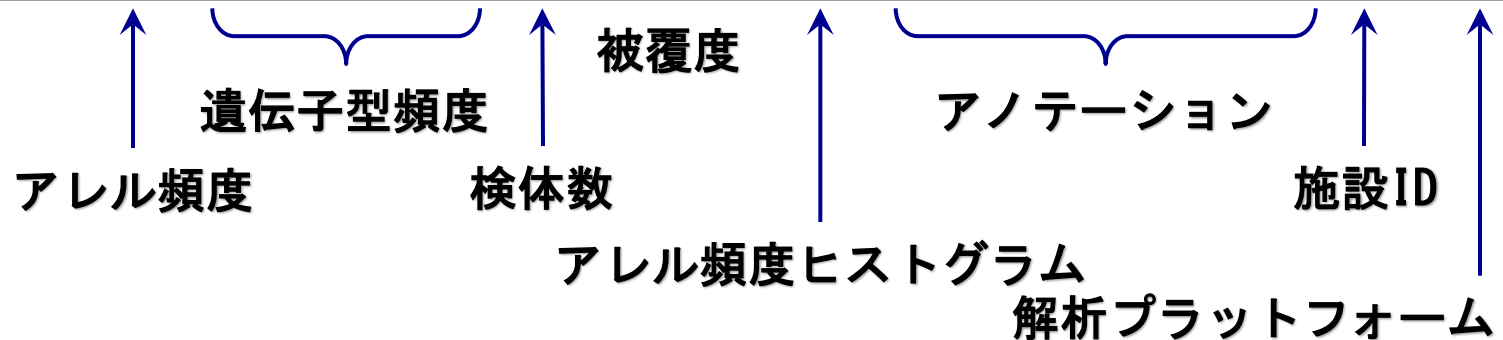
アレル頻度詳細情報

Allele Frequency

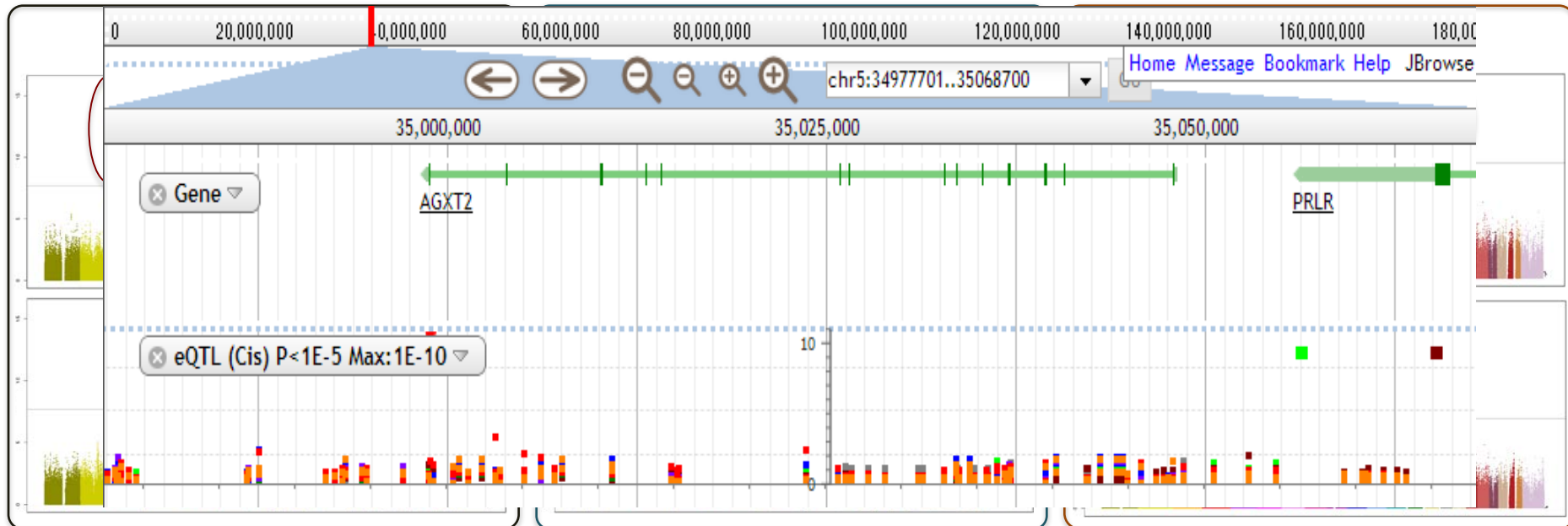
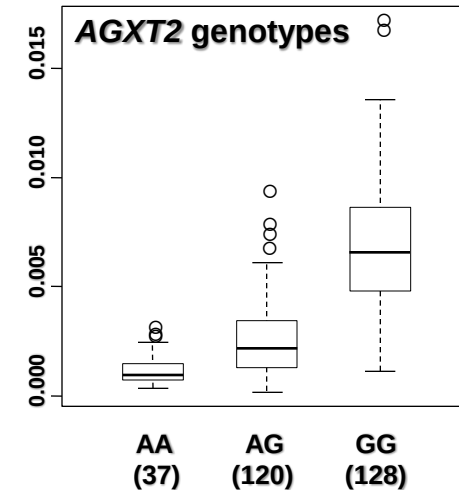
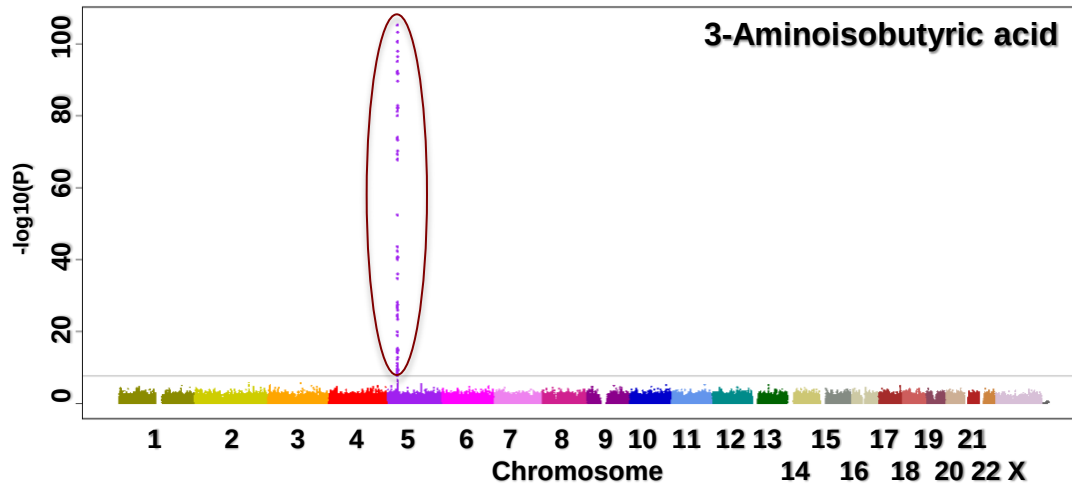
Allele Color Code: ■:A ■:G ■:C ■:T ■:Others

Variation Color Code: ■:nonsense, splice ■:missense ■:synonymous ■:indel ■:non-coding, intron, other

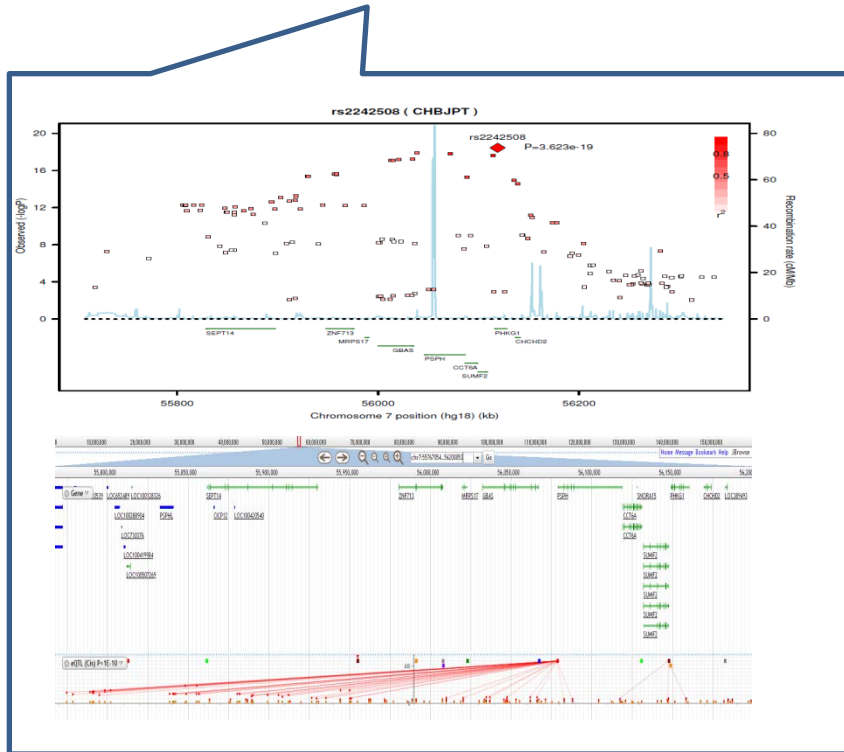
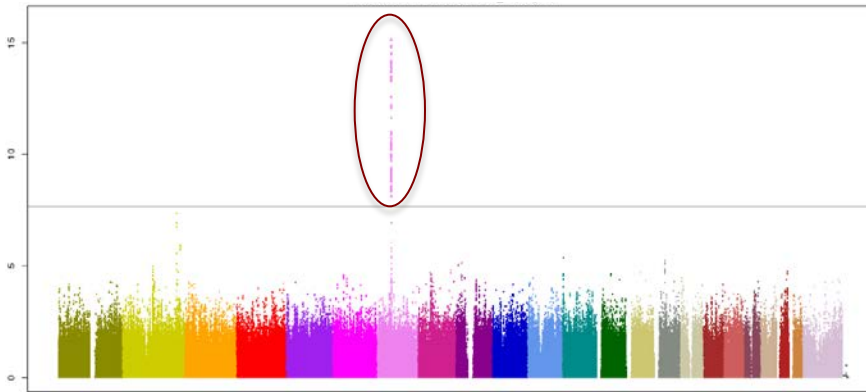
Variant position▲▼	rsID▲▼	Ref/Alt▲▼	Frequency of alternative allele▲▼	Genotype count			Number of Samples Covered▲▼	Average sample read depth▲▼	Alleles	Gene▲▼	mRNA Accession#	Function				Org▲▼	Platform▲▼
				Ref/Ref	Ref/Alt	Alt/Alt						Codon		AA			
												Ref	Alt	Ref	Alt		
chr5:34840841 U	rs6451173	A/G	0.153	219	70	11	300	59.47±23.93									



血漿中の代謝物量はゲノム多型の影響を受ける ～QTL (Quantitative Trait Locus) 解析の一例～

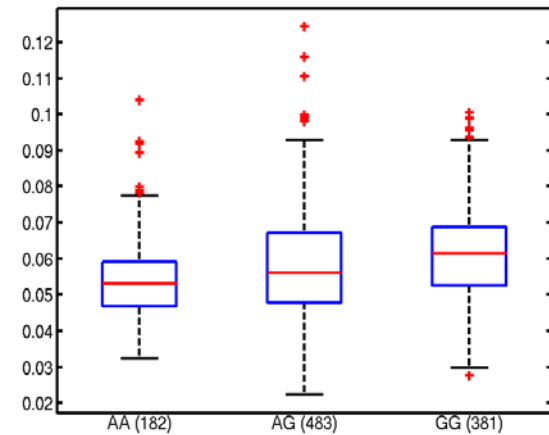


ゲノム多型～遺伝子発現～代謝物の血中濃度



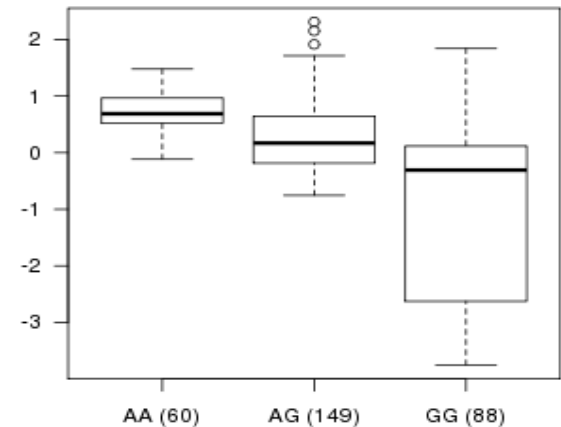
QTL analysis

rsXXXX404 – amino acid A
 $p=1.18 \times 10^{-10}$



eQTL analysis

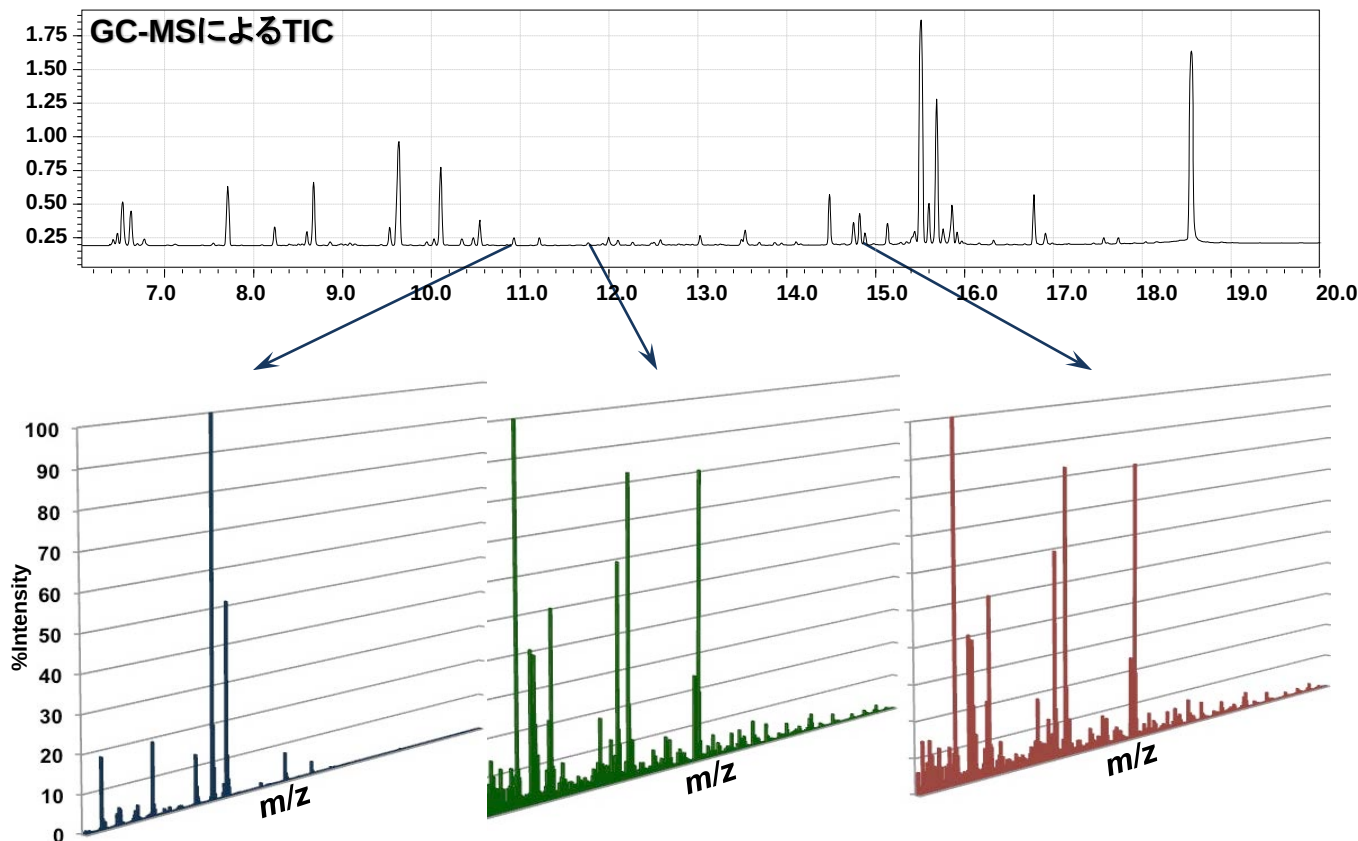
rsXXXX404 – gene A
 $p=2.0 \times 10^{-24}$



ゲノム多型～遺伝子発現～代謝物の血中濃度

GC-MS	Ch r	Position	rs ID	Ref.	Alt.	P (QTL)	P (eQTL)	Annotation for SNP	Annotation for probe
Amino acid A	7	55XXXXX	rs12345	G	T	1.33E-10	0.0002086	1234,ABCDE....	Homo sapiens gene....
Amino acid A	7	55XXXXY	rs12346	G	A	1.18E-10	0.0002075		
Amino acid A	7	55XXXXZ	rs12347	A	G	8.12E-10	0.0005998		
Amino acid A	7	55XXXYX	rs12348	G	A	1.45E-09	0.0005998		
Amino acid A	7	55XXXY Y	rs12349	C	T	1.04E-09	0.0004905		
Amino acid A	7	55XXXYZ	rs12350	C	T	3.79E-13	0.0007183		
Amino acid A	7	55XXXZX	rs12351	G	T	1.33E-10	1.69E-24		
Amino acid A	7	55XXXZY	rs12352	G	A	1.18E-10	2.00E-24		
Amino acid A	7	55XXXZZ	rs12353	A	G	8.12E-10	1.57E-19		
Amino acid A	7	55XXYXX	rs12354	G	A	1.45E-09	1.57E-19		
Amino acid A	7	55XXYXY	rs12355	C	T	1.04E-09	1.54E-21		
Amino acid A	7	55XXYXZ	rs12356	G	A	7.38E-11	1.25E-09		
Amino acid A	7	55XXYYX	rs12357	C	A	4.90E-11	1.43E-13		
Amino acid A	7	55XXYYY	rs12358	A	C	2.82E-12	5.10E-10		
Amino acid A	7	56XXXXX	rs12359	T	C	1.11E-13	4.56E-10		
Amino acid A	7	56XXXXY	rs12360	T	C	6.11E-14	1.65E-08		
Amino acid A	7	56XXXXZ	rs12361	C	T	3.79E-13	8.12E-06		
Amino acid A	7	56XXXYX	rs12362	A	C	1.40E-09	0.0005151		

GC-MS による代謝物の網羅的解析

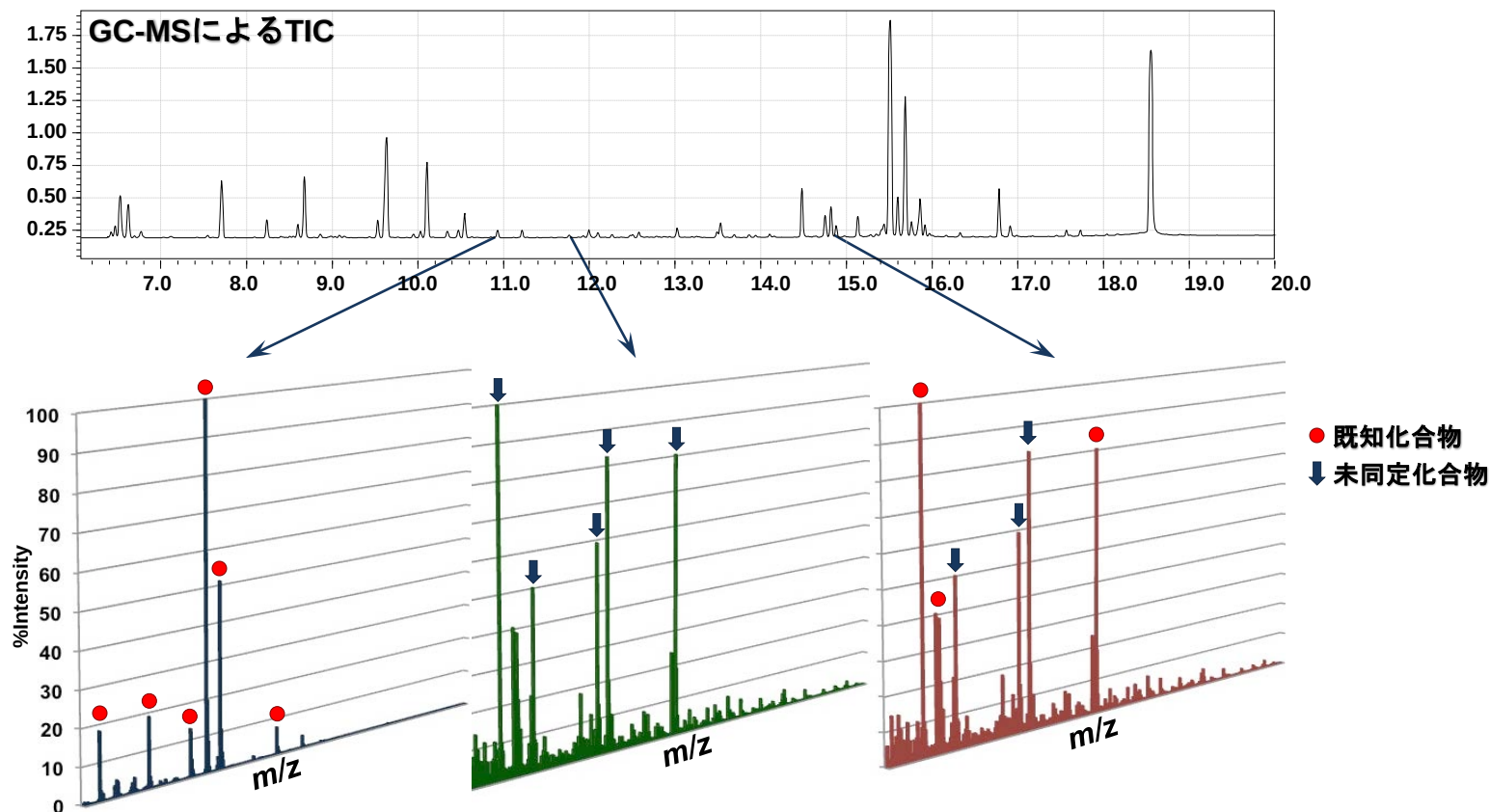


使用機器

島津製作所製 GCMS-QP2010 Ultra

- 保持時間(RT): 290秒～1,470秒を0.3秒刻みで分離
- m/z値: 85～501

網羅的メタボローム解析における情報抽出の問題点



限定された数（高々100種前後）の既知化合物の情報のみを収集し、低いピークを含む多数の化合物の情報は利用されず

見過ごされているデータをいかに抽出し、解析に利用するか

測定値の生データを抽出しマトリックスを構築

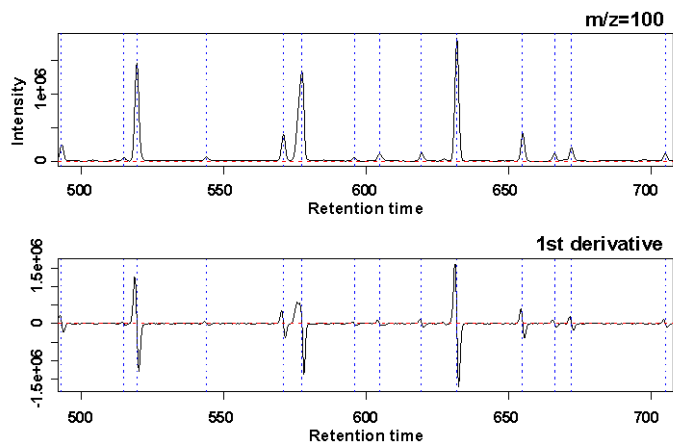
使用機器のパフォーマンス

GCMS-QP2010 Ultra(島津製作所)

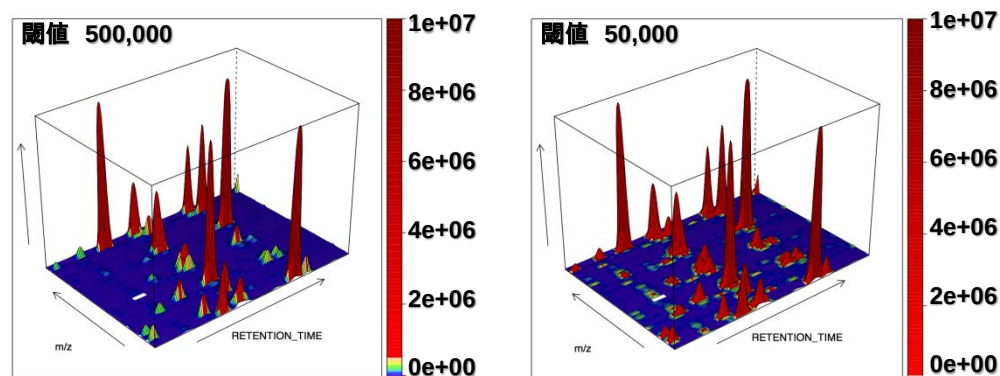
- 保持時間(RT): 290秒~1,470秒を0.3秒刻みで分離 (3,937フラクション)
- m/z値: 85~501 (417フラクション)

マトリックスの構築

- 測定値の生データ(無加工)を使用
- RTを10秒刻み、m/zを10刻みにグリッド化し、4,998グリッドに対応するピークを取得
- 4,998グリッドの二次元マトリックスを1,173検体につき調査



計測値から再構成した代謝物ピーク
(RT: 300~350sec, m/z: 85~150)



閾値	Day 0	Day 1	Day 7	Day 90
500,000	217.3 ± 24.4	227.3 ± 42.7	213.5 ± 41.8	180.0 ± 28.5
100,000	1461.1 ± 205.1	1552.8 ± 240.6	1443.7 ± 240.6	1153.4 ± 235.0
50,000	2582.4 ± 464.9	2731.9 ± 443.3	2522.9 ± 489.3	2007.7 ± 409.6
10,000	9480.5 ± 2830.1	10006.8 ± 2493.2	9069.0 ± 2497.8	6799.2 ± 1814.2

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

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

Day 1 Exome and whole genome sequencing using massively-parallel sequencers

**Day 2 Application to RNA analysis
Data submission**

Day 3 Practicals for exome and RNA seq

Day 4 Human genetics

Mark Lathrop (CEPH, France)

Daniel E. Weeks (Pittsburgh, U.S.A.)

Rajesh Radhakrishnan (EBI, U.K.)

Marcella Devoto (Pennsylvania, U.S.A.)

Makoto Mark Taketo (Kyoto, Japan)

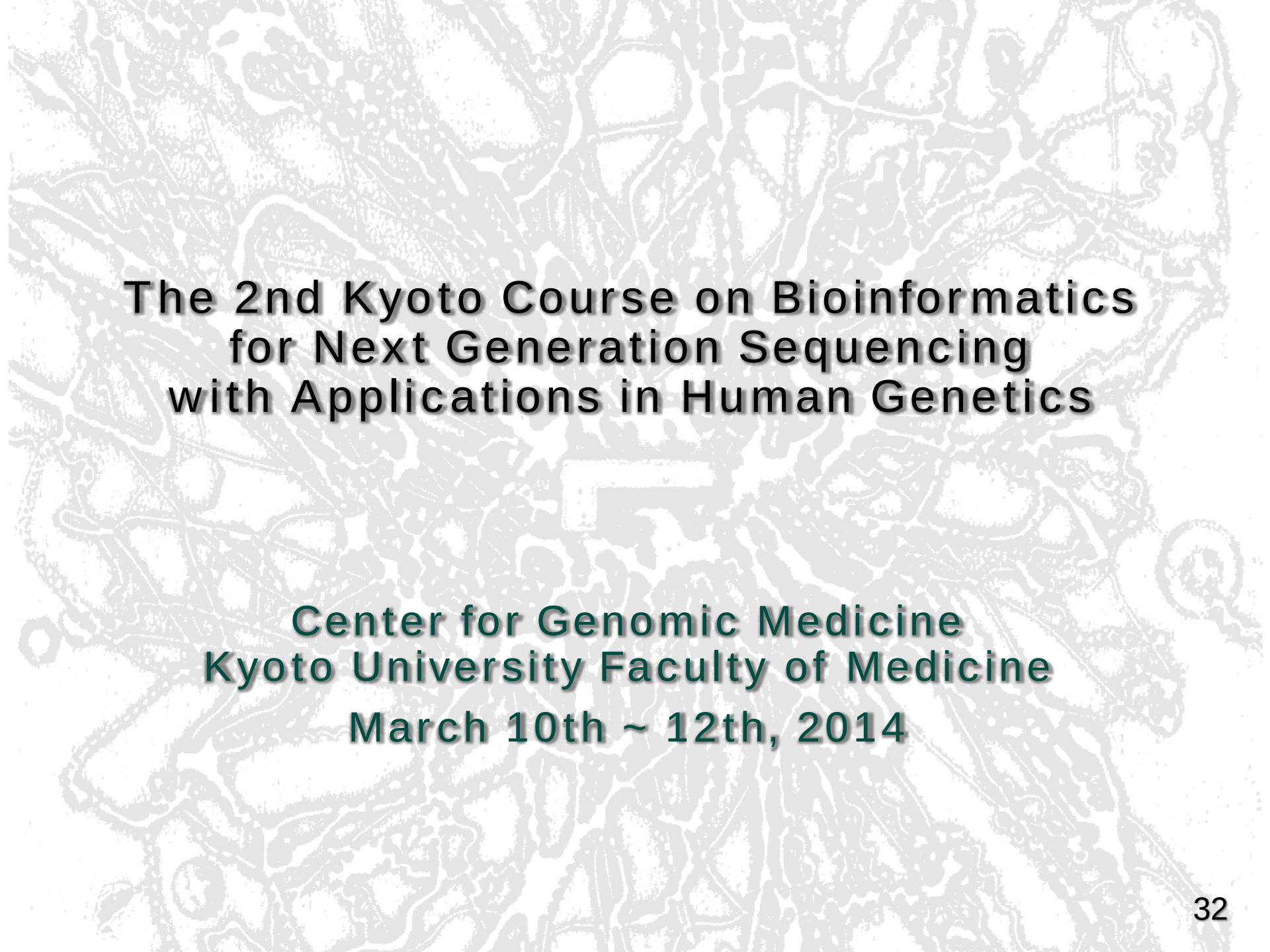
Guillaume Bourque (McGill, Canada)

Mar González-Porta (EBI, U.K.)

Jurg Ott (Rockefeller, U.S.A.)

Derek Gordon (Rutgers, U.S.A.)





The 2nd Kyoto Course on Bioinformatics for Next Generation Sequencing with Applications in Human Genetics

Center for Genomic Medicine
Kyoto University Faculty of Medicine
March 10th ~ 12th, 2014

ながはま0次コホート事業

京都大学医学研究科

健康情報学 医療倫理学
ゲノム医学センター 呼吸器内科学
歯科・口腔外科学 眼科学
糖尿病・栄養内科学 臨床免疫学
泌尿器科学 整形外科学
内分泌内科学 消化器内科学
神経内科学

長浜市民の参加者の方々

長浜市役所 健康づくり0次クラブ
長浜市立病院 長浜赤十字病院
湖北医師会・湖北歯科医師会

協力企業

フクダ電子株式会社 株式会社SRL
滋賀保健研究センターライオン株式会社
味の素株式会社

統合オミックス解析

京都大学医学研究科附属ゲノム医学センター

山田 亮 田原康玄 日笠幸一郎 寺尾知可史 高橋めい子 川口喬久
清水正和 小久保美紀 采女寛行 瀬藤和也 奈良原舞子 **Marina PENOVA**
Gunalini SARAVANAMUTHU

島津製作所 基盤技術研究所 ライフサイエンス研究所

佐藤孝明 九山浩樹 園村和弘 松尾英一

Fondation J. Dausset, Centre d'Etudes du Polymorphisme Humain

鎌谷洋一郎 Mark LATHROP