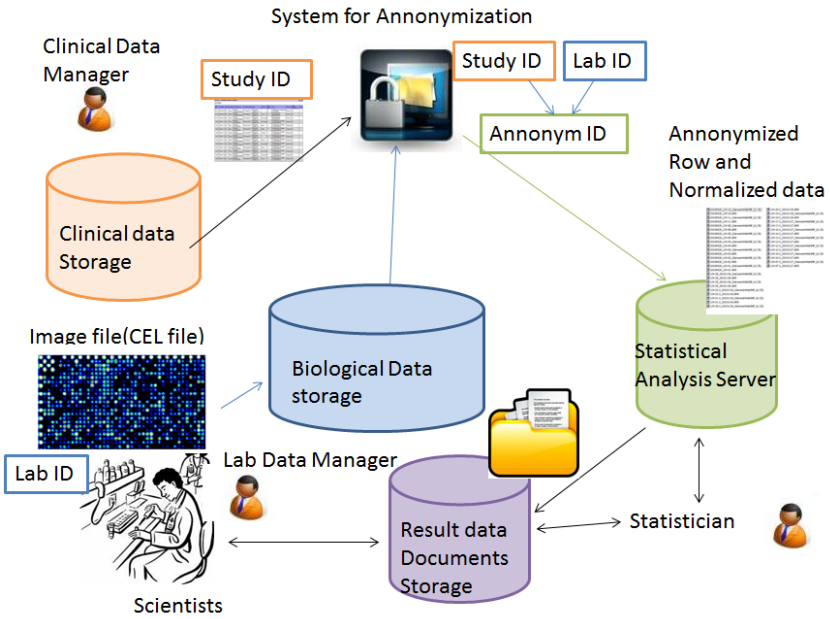
課題番号　　：25指112
研究課題名　：分子生物学的情報を扱う臨床研究の品質管理に関する研究
主任研究者名：田中紀子
分担研究者名：

キーワード　：omics; data management; database; clinical bioinformatics; standardization
研究成果　　：
最終研究成果報告
欧米ではすでに品質管理システムが不十分な施設で行われた臨床研究の結果の承認は行われない。特に米国ではマイクロアレイおよび次世代シークエンス技術などを使った測定についてＦＤＡによりすでに品質管理基準が設けられている。我が国でも施設内での研究の質の管理は今後重要な課題となっていることが、平成 23 年度厚生労働省による臨床研究に関する倫理指針　適合性調査により自己点検という項目がほぼすべての施設において低評価（Ｂ）とされていることからも示唆されていた。また、内部的なアンケートにより、研究所内部からも分子生物学的情報の管理を専門的に行ってほしい要望もあることを把握していた。カルテ情報や検査情報などのデータのみを扱う臨床研究の品質管理についての研究は欧米では製薬企業や臨床試験のセンターを中心として盛んに行われ確立されつつあったが、測定自体が科学の最先端技術を駆使して行われる分子生物学的情報を扱う研究に関する品質管理についてはここ数年で欧米でもその必要性が叫ばれ始め、研究が始まったばかりであった。これは、実験自体が臨床試験のように１目的に１測定のように行われずに、多段階で無計画に行われることが多いことに起因している。そこで、１実験あたりの品質管理は十分に行われていても、いくつもの実験を統合的に管理することが困難であり、ＩＴ技術を駆使した品質管理システムの構築の必要性が高まっている。しかし、分子生物学的情報を扱う臨床研究の品質管理に注目して研究を行っている研究室は皆無に等しい。そこで本研究は主に所内の研究の質および効率の改善を目標として分子生物学的情報を扱う臨床研究の品質管理システムの構築を行った。最終的に目標とされたシステムの概要図は以下のようなものであった。



国立国際医療研究センター研究所および病院、さらに外部の共同研究者が共同で利用できるネットワークシステムが存在しなかったので、まずはシステムのインフラ整備を行わなければならなかった。そこで JCRAC データセンターのネットワークを利用し、解析用サーバー、ウェブサーバー、データストレージサーバーの設置および接続を行った。当初ネットワークはパスワード認証よりも安全性の高い秘密鍵/公開鍵方式を用いた SSH プロトコルでの通信とし、解析用サーバーには SSH サーバーを経由

して接続することとした。しかし、病院のネットワーク環境からＳＳＨサーバーへの接続が不可能であったため、外部及び病院のネットワークからの接続をＶＰＮ接続に切り替えた。

解析用サーバーは、解析データの匿名化、匿名化後データの統計解析処理および高い再現性を目的とした実験履歴を保存するシステムに利用される。このシステムは運用開始後も順次機能追加をする予定である為、新規機能やトラブルシュートの為にも試験環境が必要である。この試験環境の為のサーバー機を購入し、試験環境を構築した。この２つの解析用サーバーに対し、研究履歴保存および定型処理を実行するソフトウェアであるGalaxyをインストールし､その他に統計解析に用いるR及びDNAメチル化解析及びSNPデータの解析に必要なライブラリのインストールを行い、統計解析の環境構築を行った。本研究で構築されたGalaxyをNCGMGalaxyとする。

インフラ整備と並行してイルミナ社の４５０Kビードチップで測定された全ゲノムDNAメチル化データの画像ファイルから統計解析用データ作成までのパイプラインに必要なプログラムを研究・開発し、構築されたNCGMGalaxy上へ実装した。研究期間内で検証され、新たに開発されたアルゴリズムは以下の二つである。

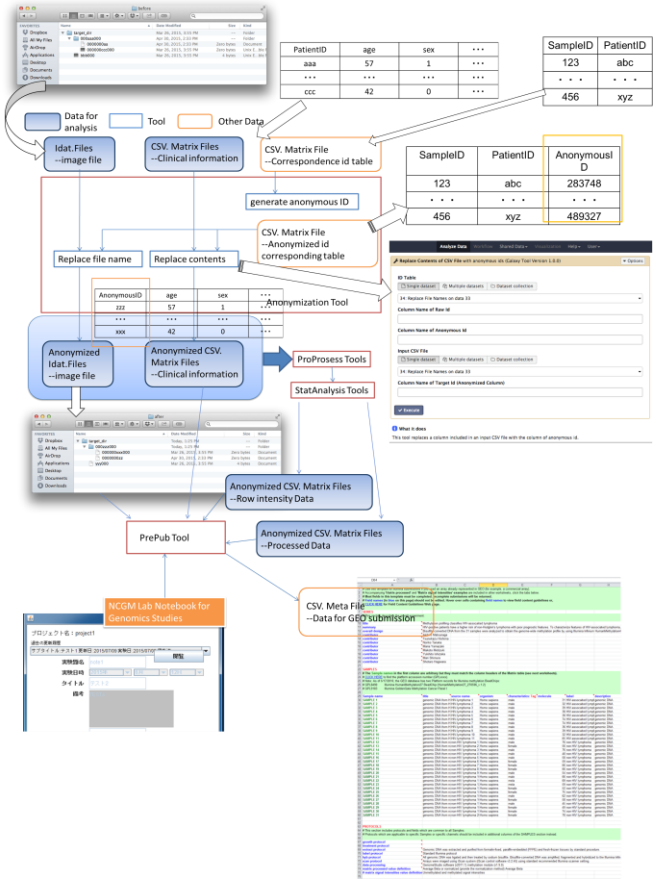
・集団のSNP多型頻度分布を考慮したプローブフィルタリングアルゴリズムの開発

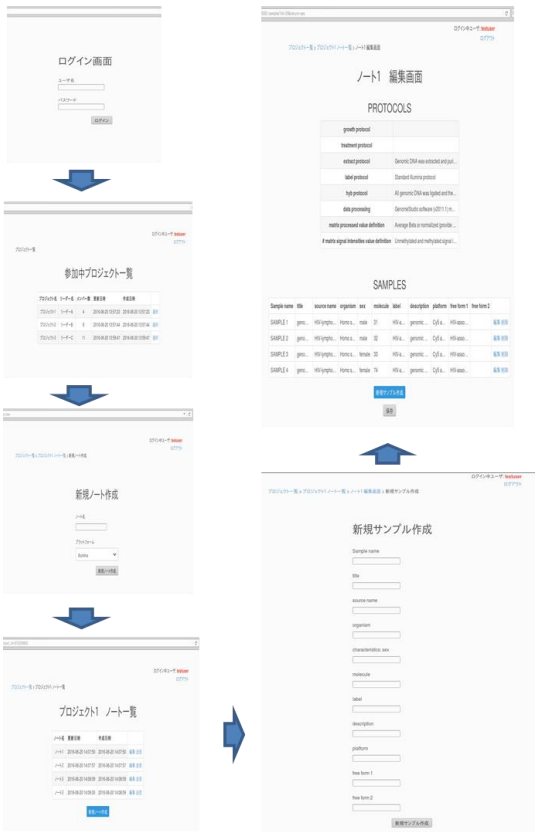
メチル化の測定に関しては、測定プローブ内にSNPが存在すると正しく測定できないことが知られている。そこで、既存のパイプラインにはプローブ内にSNPが測定されたプローブのリストを用いてプローブのフィルタリングを行うプロセスが含まれていたが、この除外プローブリストは主に白人種のSNP分布データに基づいて作成されたものであり、人種によってはリストが正しくない可能性が考えられた。そこで、1000 Genomes Projectデータをリファレンスとし、各SNPのphysical positionの情報と、イルミナ社から提供されている450Kチップに適用されている各プローブのCpGサイトのphysical positionの情報とをマッチさせることにより、各プローブの50bpの範囲内に存在するSNPの数を数えるアルゴリズムをPearl（ver.5.16.3）で開発した。このアルゴリズムを実データに適用した結果、アジア人サンプルでSNPを含むプローブをすべて排除した場合とマイナー対立遺伝子頻度（MAF）が10%以上のSNPを排除した場合ではcoverageに大きな変化は見られなかったことから、SNPを含むプローブをすべて排除する必要はなく､MAFの大きいSNPを含むプローブを排除することで研究の主解析に使用可能なプローブ数が増加することが示された。試験運用は日本人集団データのみならず、アフリカ人集団やヨーロッパ集団サンプルあるいは別の日本人集団サンプルでも問題なく動くかどうかの確認を行い、NCGMGalaxyへの実装が完了した。

・混合分布を用いたサンプルフィルタリングアルゴリズム

測定原理からDNAメチル化の測定値の分布は３混合β分布であることが知られていたが、本研究で実例としたリンパ腫組織から測定されたデータには４以上の混合数が考えられるメチル化状態にあるサンプルが観察された。これらのサンプルは測定誤差によるものなのか、分子生物学的意義があるのかに関わらず、フィルタリング対象としたほうがよいことが推察されたため、３混合β分布をあてはめることにより、サンプルごとのメチル化状態をあらわす分布のピーク数推定に基づくサンプルフィルタリングプロセスの提案を行った。そこで、まずこの混合数が多い検体の特徴抽出について発表し、混合数推定の方法を混合β分布より求める以外に、変数変換からの混合正規分布あては、および、カーネル関数による平滑化によるピーク数推定方法の３種類の方法をあてはめた結果、カーネルによるピーク数推定は当てはまりすぎをうまく抑えることにより、もっとも混合数の多い検体の特徴を抽出できることが示された。全てのアルゴリズムはRで開発され、NCGMGalaxyへ実装された。

上記二つのアルゴリズムを含めた新たなパイプラインを提案し、すべて NCGMGalaxy へ 450Ktool として実装を行った。ツールの全体としては、匿名化→前処理→（統計解析：本研究では開発していない）→出版用データ準備という流れで以下のように開発された。匿名化ツールは JAVA の並べ替え関数を用いて重複を許さない乱数列を生成し、匿名化 ID に利用している。またデータ出版用ツールは 450K データを GEO に公開するためのツールである。研究プロトコル情報や、匿名化済みの解析に利用した画像データおよび表形式のデータからサンプルの情報も抽出可能とし、手作業によるアップロードファイルの準備を減らし、作業効率を上げる工夫が施されている（下左図）。
この研究プロトコルや実験準備段階で入力できるサンプル情報に関して、実験前あるいは実験後すぐに研究者によって入力し記憶によるバイアスを減らし、実験情報の管理に役立てるようなウェブブラウザベースの電子実験ノートを開発した。実験ノートの画面遷移を下右図に示す。





開発した NCGMGalaxy のトップ画面は下に示した。

これらのツールを開発した結果、画像データから統計解析用データへの加工、および加工したデータの出版用データ作成までのプロセスを履歴を残しながら一元管理できるようになった。今後はこのツールをセンター内公開し、研究者に使っていただきながらバリデーションと改定を行うことが必要である。

Subject No. : 25 指 112

Title : A study for designing quality control system for studies in clinical bioinformatics

Researchers : Noriko Tanaka, Mari Shimura, Yasuhiro Tanaka

Key word : omics; data management; database; clinical bioinformatics; standardization

Abstract : States-of-the-art omics technologies, such as microarrays, proteomics or next-generation sequencing, can generate vast amounts of raw data in a single experiment, as well as summaries in the form of lists of sequences, genes, proteomics, metabolites, or SNPs. Controlling the quality of clinical studies with the large omics data present challenges. Here we developed a study management system which enables to track data management activities for research in Clinical Bioinformatics, from anonymization to submission to public database for publication, via a web-enabled Galaxy interface developed by a team in Univ. Pennsylvania. We tested our tools with data from an epigenomewide study for lymphoma patients.

Every researcher who conducts biological experiments knows there are so many potential sources of variability and errors in biological experiments and they should be controlled to be minimized as much as possible before and after experiment. So thousands of papers are published which described how to control the quality of the data from biological experiments, however, most of papers and books just focus on how to control the lab data or how to control the samples in the lab and little attention has been paid for controlling quality of whole processes of the study. General instruction for quality control has been developed in industrial fields, and the concept was accepted and applied to clinical researches. The fundamental premise is that quality considerations should be integrated into every phase of the study from initial hypothesis formulation to the final publication of findings and archiving of data (Rajaraman and Samet, 2014).The term 'quality control' is defined as the activities that occur during and after data collection to correct data errors, while the term 'quality assurance' means those to ensure quality of the data before data collection (Moyses Szklo 2007). According to their description, here we consider not only data quality and also guidelines for good practice describing both quality control and quality assurance activities in clinical bioinformatics or translational research, thus for the clinical researches with human biological data(Fig1) .

Our goals in this study:

1. Set up a computer network to control the quality of clinical studies with huge biological data.
2. Construct basic web-based system to manage the various information from the studies.
3. Develop a web-based experiment notes to control the quality of experiments for epigenomic studies.

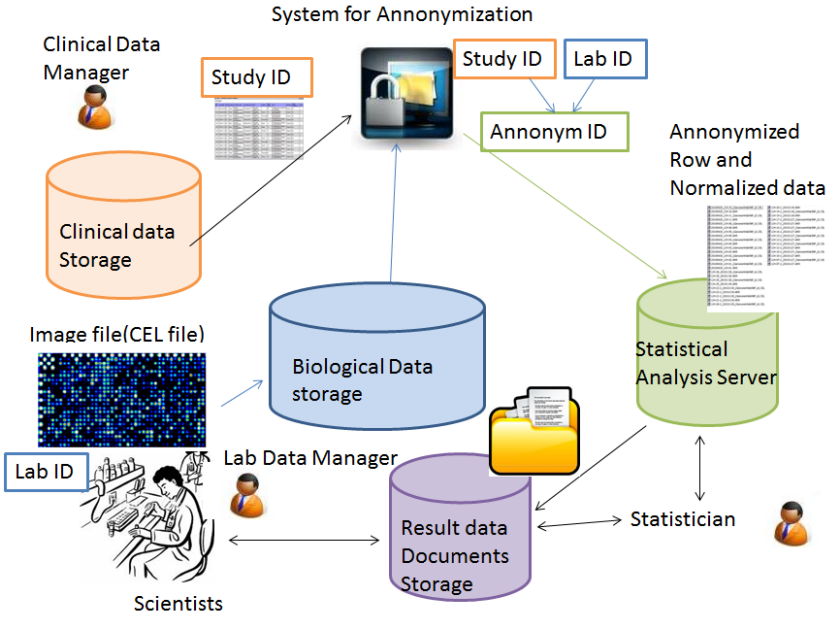


Fig 1: Study and Data flow of studies in Clinical bioinformatics

For no.1, we set the web and computing server in JCRAC data center, and set up a computer network for some terminal users in the hospital and the research institute to have access to those central servers. We chose Galaxy, which is an open, web-based platform originally developing by the Galaxy team at Penn State University and Johns Hopkins University (https://galaxyproject.org/) to manage for data intensive biomedical research to manage study information and omics measurements, and set it up on the central server. VPN connection has established to access Galaxy server for the terminal users in the hospital.

For no.2, we set up the test environment for testing NCGM Galaxy (Fig 2) customized for our purpose, QC for DNA methylation data. We tested NCGM Galaxy and related algorithms in the test environment. We developed four new algorithms to be implemented on the NCGM Galaxy. Dr.Toyooka and Mr.Mochizuki developed an algorithm to detect SNPs on the probes for Genome-Wide DNA methylation studies. Mr.Uesato developed tools to annonymize meta-files generated from omics experiments. Mr. Kurosawa developed algorithms to count peaks of multimodal distributions for filtering unexpected samples. Mr.Kurosawa, Dr.Yamazaki and Mr.Uesato developed tools for normalizing the data from Illumina Infinium assay combining with the probe and sample filtering algorithms we developed in our project. The concept of the system and data flow are shown in Fig.3. The person in charge of the anonymization can create a new correspondence table with anonymized new id, study id and lab id using Anonymization tool. Subsequently, anonymized experimental metadata and anonymized clinical data can be created for the person who will analyze the data for publications. Prior to statistical analysis, image (raw) data from Illumina Infinium assay are converted into text data and normalized and filtered using 450KPreProcess tools. After statistical analysis, experimental metadata and the format files are provide using PrepPub tool to submit the data for publication into public repositories, GEO.

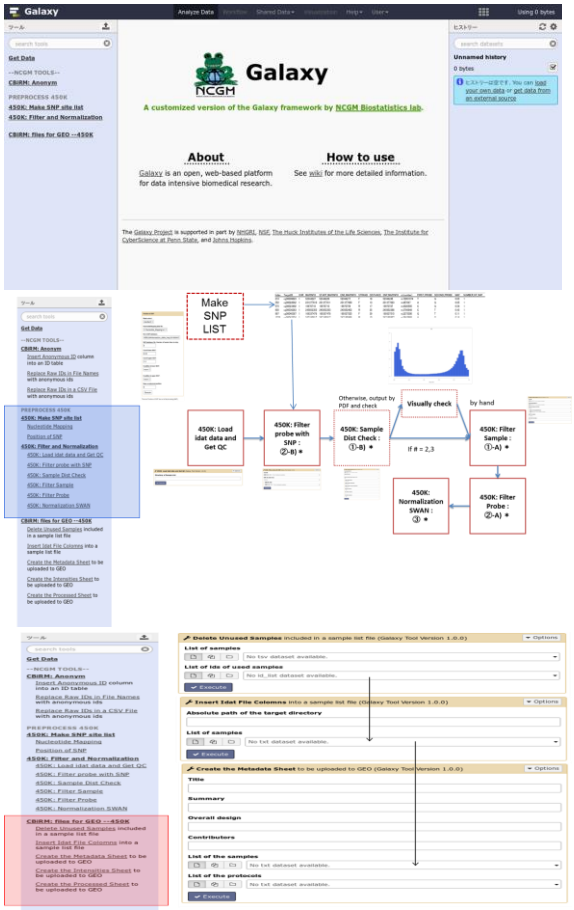


Fig 2. NCGMGalaxy

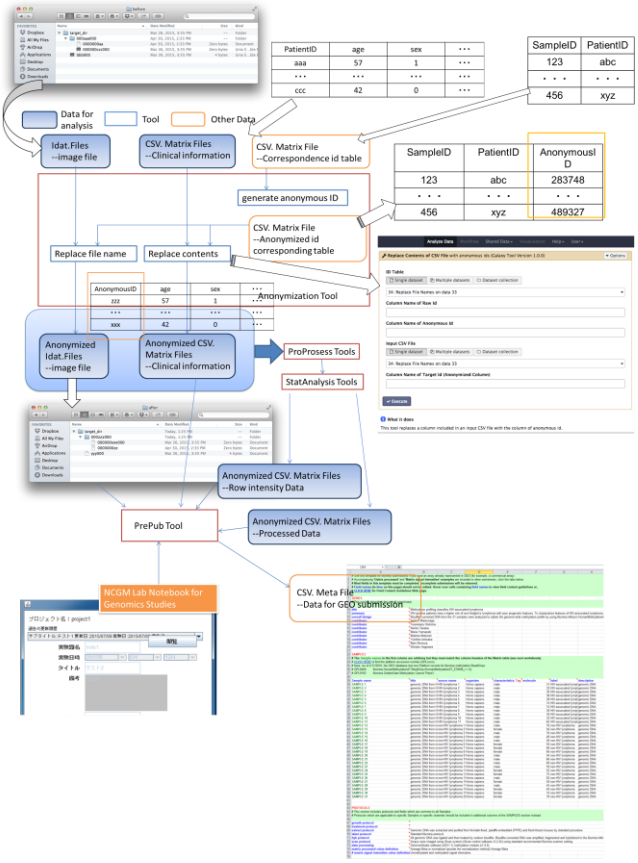


Fig 3. Concept of the system

For no.3, Mr.Mochiduki and Shitaoka designed and developed a web-based experiment notes to control the quality of experiments for measuring DNA methylation levels using Illumina 450K bead chips. We used Ruby on Rails to develop the web application. We show the schreen transition of the web application experimental note in Fig 4.

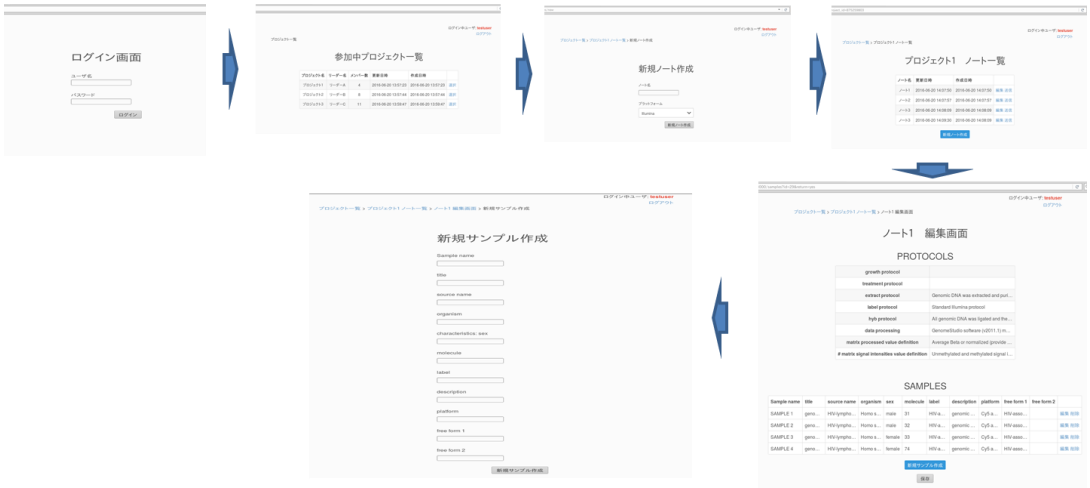
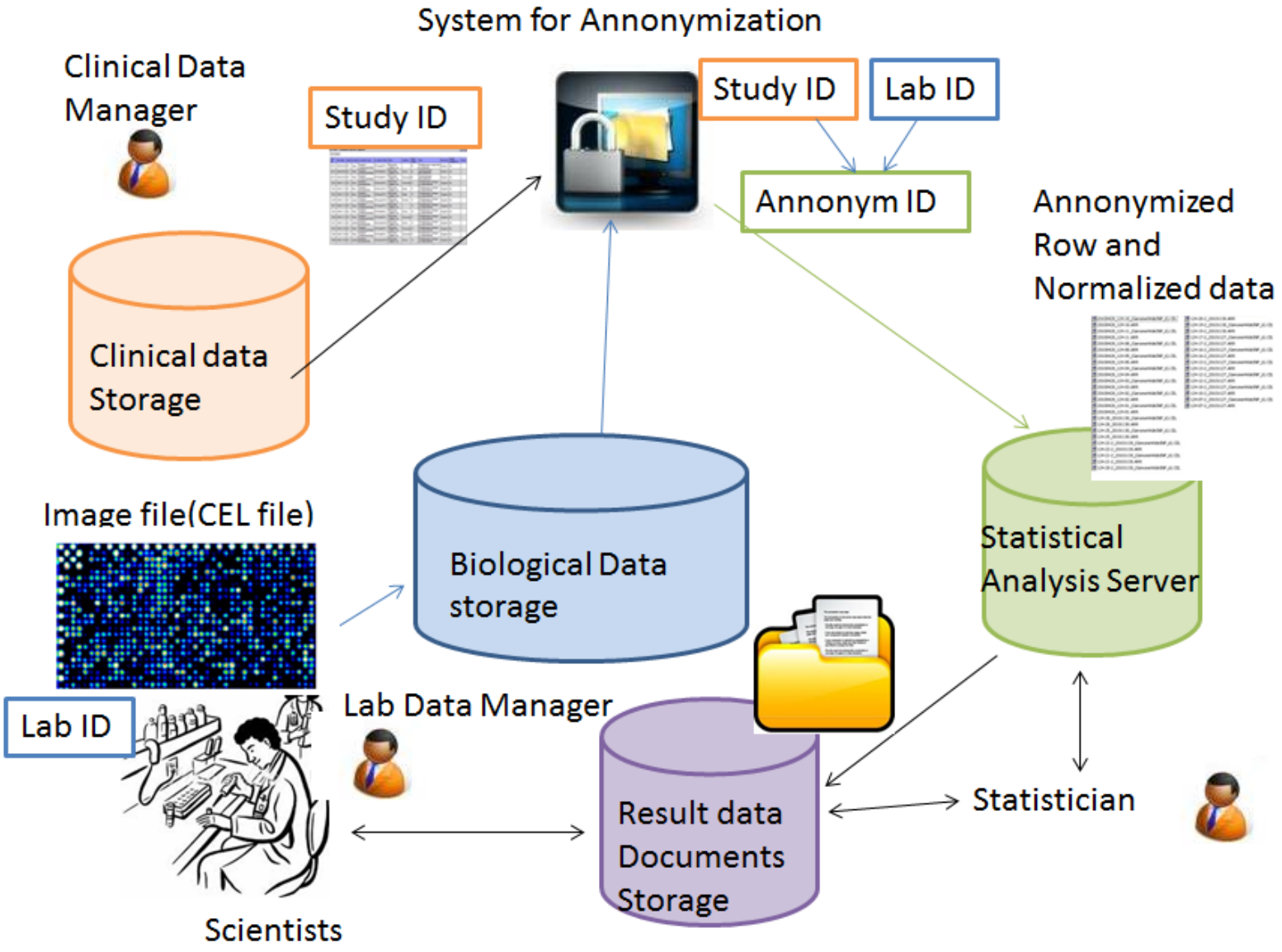


Fig 4. Screen transition of the web-based experiment notes

Our system developed in this study provides 1:secondary anonymized meta-datasets, so researchers do not need to re-anonymize meta-data for publication, 2:history of data anonymization, or normalization process, statistical analysis and submission of meta-data for publication so that investigators can manage and control research quality.

Researchers には、分担研究者を記載する。

研究目標-
分子生物学的情報を扱う臨床研究の品質管理システムの構築



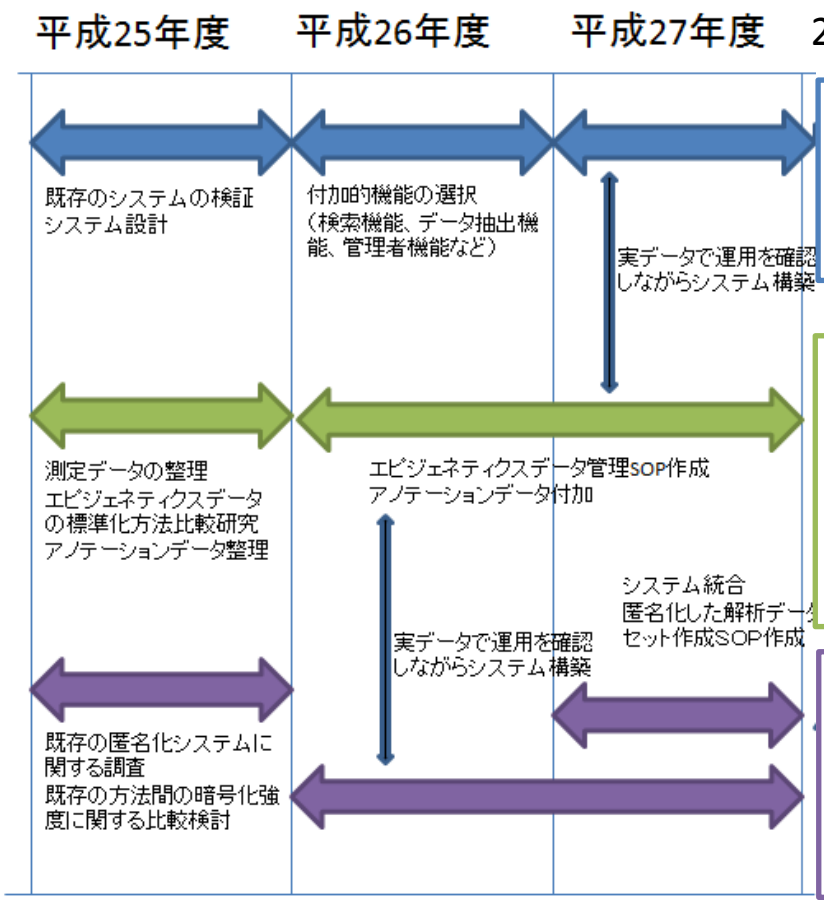
研究計画と達成内容

研究目的：分子生物学的情報を扱う臨床研究の品質管理システムの構築

Specific Aim 1:
ネットワークを利用した多層かつ異質的分子生物学的情報を格納するストレージシステムの構築

Specific Aim 2:
エピジェネティクスデータ（実データ）の品質管理工程の確立

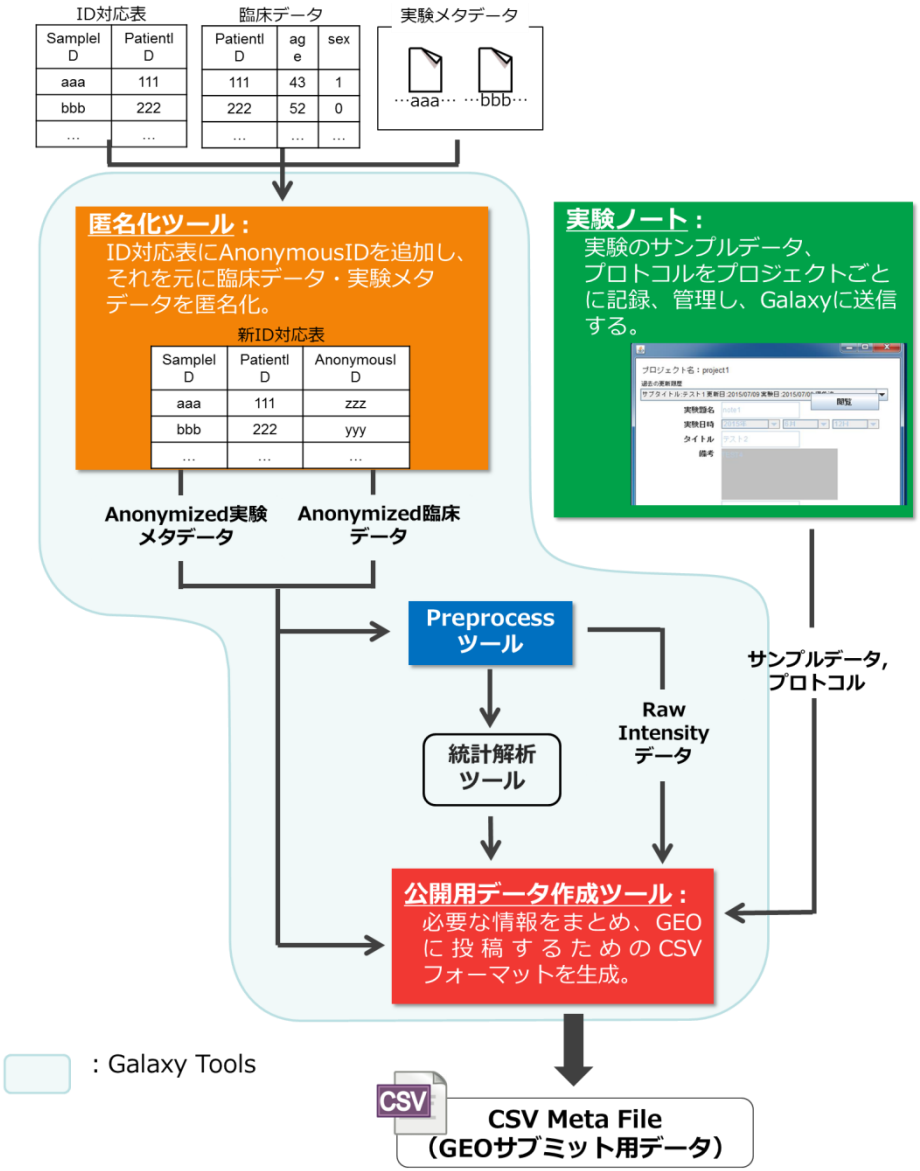
Specific Aim 3:
連結可能匿名化により臨床データとの連結を行い、解析セットを作成するシステムの構築



2016年時点達成事項

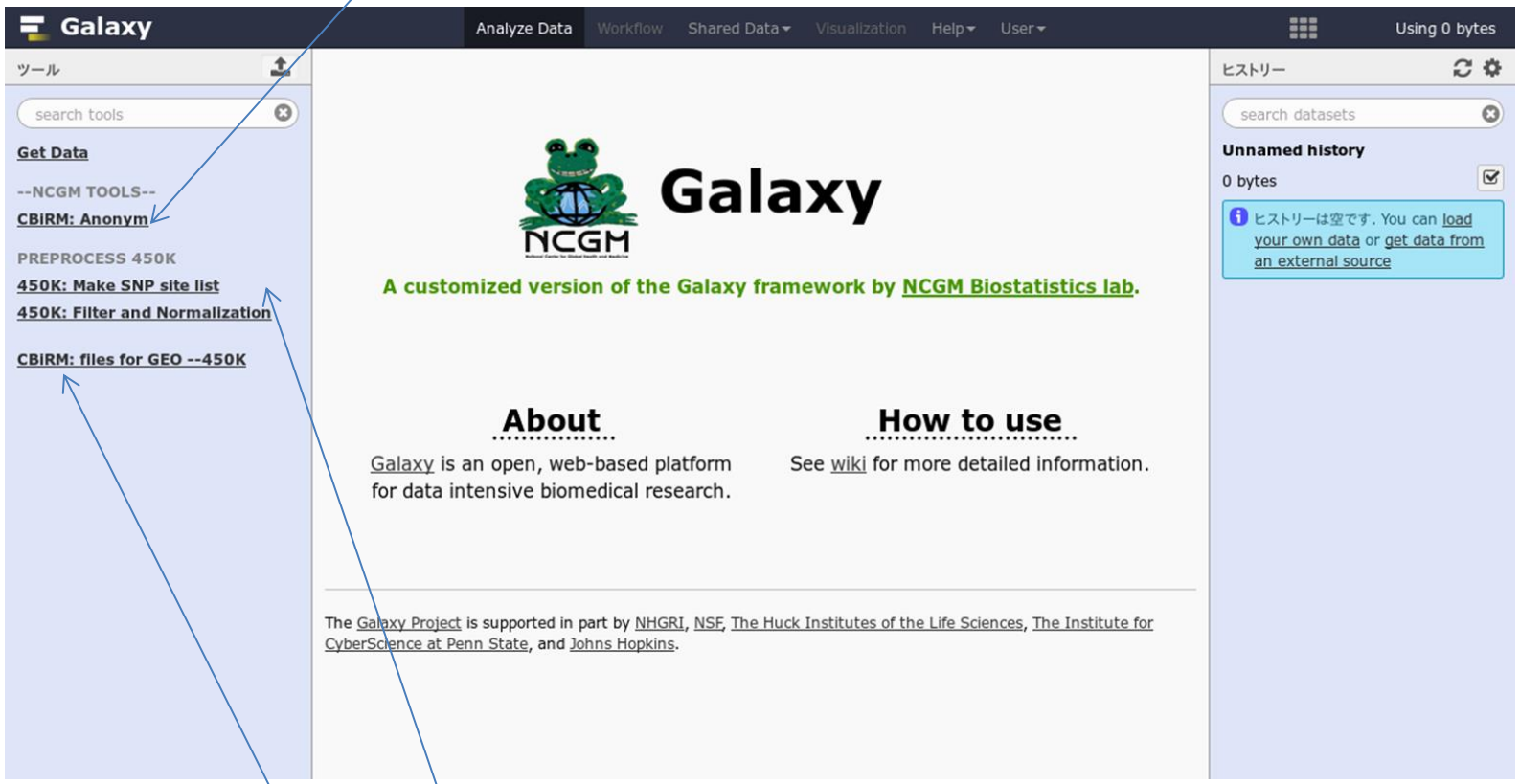
- 採用したGalaxy(プラットフォーム)上へ実装予定の付加機能の実装とテスト
- ・新規フィルタリング法の提案
  ・データ管理工程設計およびRによるツールの開発、テスト版作成
  ・電子実験ノート作成
- 匿名化アルゴリズムをJAVAで作成、Galaxyへの実装
  ファイル管理システム設計

Galaxyを利用した研究管理の流れ



構築されたNCGM-Galaxy

データの2次匿名化ツール



イルミナ450K測定キット用ツール

出版用データ作成ツール

研究発表及び特許取得報告について

課題番号： 25指112

研究課題名： 分子生物学的情報を扱う臨床研究の品質管理に関する研究

主任研究者名： 田中紀子

論文発表

論文タイトル	著者	掲載誌	掲載号	年
該当なし				

学会発表

タイトル	発表者	学会名	場所	年月
ヒト検体からの遺伝情報を伴う臨床研究の研究管理ツールの開発.	田中紀子,上里和也, 黒澤匠雅,下岡純也, 望月尊仁, 豊岡理人, 田中康博, 大津洋, 志村まり.	日本臨床試験学会第７回学術集会総会	名古屋	2016年3月
CBiRMTools: a Galaxy toolbox for research management tools for research in Clinical Bioinformatics.	Tanaka N, Uesato K, Mochizuki T, Kurosawa T, Tanaka Y, Shimura M, Ohtsu H.	Japanese Society of Medical Informatics 2015.	沖縄	2015年11月
Prognostic prediction based on genome-wide DNA methylation distribution in non-Hodgkin’s B-cell lymphomas.	松永章弘、豊岡理人、吉田塁、石坂幸人、田中紀子、志村まり	第37回日本分子生物学会年会	横浜	2014年11月
FILTERING SAMPLES BASED ON BETA-MIXTURE MODEL FOR DNA METHYLATION DATA QUANTIFIED BY BISULPHITE MICROARRAYS .	Tanaka N, Kurosawa T, Inaba Y, Toyo-oka L, Yoshida L, Kawasaki Y.	International Biometric Conference	Florence. Italy.	2014年7月
ゲノムワイドDNAメチル化解析において、中間での高い分布の解析.	松永章弘, 豊岡理人, 吉田塁, 石坂幸人, 田中紀子, 志村まり.	第66回日本細胞生物学会.	奈良	2014年6月
メチル化アレイ測定データの分布に基づくサンプル品質管理及び人種差を考慮したプローブフィルタリングの妥当性の検討	豊岡理人 松永章弘 山崎茉莉亜 志村まり 田中紀子	日本人類遺伝学会	仙台	2013年11月
The choice of smoothing parameter and the number of permutation in estimation of multidimensional local false discovery rate based on the subset of high-dimensional genomic data.	山崎茉莉亜 豊岡理人 田中紀子	日本計量生物学会	福島	2013年5月

研究発表及び特許取得報告について

その他発表(雑誌、テレビ、ラジオ等)

タイトル	発表者	発表先	場所	年月日
該当なし				

特許取得状況について　※出願申請中のものは(　　)記載のこと。

発明名称	登録番号	特許権者(申請者) (共願は全記載)	登録日(申請日)	出願国
該当なし				

※該当がない項目の欄には「該当なし」と記載のこと。
※主任研究者が班全員分の内容を記載のこと。