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：志村まり　田中康博

キーワード　：omics; data management; database; clinical bioinformatics; standardization
研究成果　　：
平成 25 年度の研究成果報告
（田中（紀）班）
初年度の計画は品質管理システムの基本的なシステム設計、および、複数の研究者間での操作を可能とするために、ネットワークの利用が必要になってくるので、ネットワークの設計も行うことであった。また、欧米においてはこのようなシステムの構築がすすんでおり、すでに運用しているところも多いため、受け入れを承諾してもらえる機関を探し視察も行う予定であった。成果は以下の通りである。

１．ネットワーク設計

外部からもアクセス可能なネットワークをＪＣＲＡＣデータセンター協力のもとに構築し、ウェブサーバーの設置を行った。センター内部および外部からのアクセスが可能なことを確認した。院内からのアクセスも可能かどうかの確認したところ、Softeaser 経由の VPN によるアクセスが可能であることが判明した。このため、各端末に Softeaser のセットアップを行った。

２．生物情報学的データの品質管理システム運用に関する視察

主任研究者は前所属であるハーバード大学公衆衛生大学院 Center for Health Bioinformatics の責任者である Hide 先生に 10 月 21 日にお会いし、ハーバード大学でのデータ管理状況について伺った。そこで、米国では規制の行き過ぎにより臨床生命情報学的研究の遅れが危惧されている状況を伺うことができた。先進的取り組みを行っているカナダのがんセンターへの視察を勧められた。

３．システム設計

既存のシステムを調査したところ、施設内クラスタサーバー用に比較的容易にカスタマイズできるウェブベーストなプラットフォームとして Galaxy（http://galaxyproject.org/）が利用可能なことが判明した。そこで、我々はこの Galaxy を基盤とする新たな内部的プラットフォーム NCGM-Galaxy（仮称）を開発することに決定した。まずはこのプラットフォームの Python で書かれている内部構造を把握し、かなりの改造を加えるために、Galaxy の内部構造に詳しく、いくつかの施設にて実装をお手伝いした実績のあるアメリエフ（株）にコンサルタントを依頼し、11 月から全 4 回のコンサルテーションを受け、東京都立産業技術高等専門学校の望月研究生および東京大学人類遺伝学教室豊岡研究生により基本的なセットアップは完了した。さらに、豊岡・望月研究生は分担研究であるメチル化発現データの標準化方法の見直し項目としてあがったメチル化測定プローブ内に存在するＳＮＰの検出アルゴリズムを開発し、ギャラクシー上で実行可能になるための作業を行った。

４．その他

また、川崎研究員と大津客員研究員は志村分担研究者らからの聞き取り調査により判明した、検体管理段階での品質管理に必要な項目のデータベース化に関する検討を行った。

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 :

Our goals in the first year were:

1. Set up a computer network to control the quality of clinical studies with huge biological data.
2. Understand the movement of QC systems in Clinical Bioinformatics in the world
3. Complete basic system design

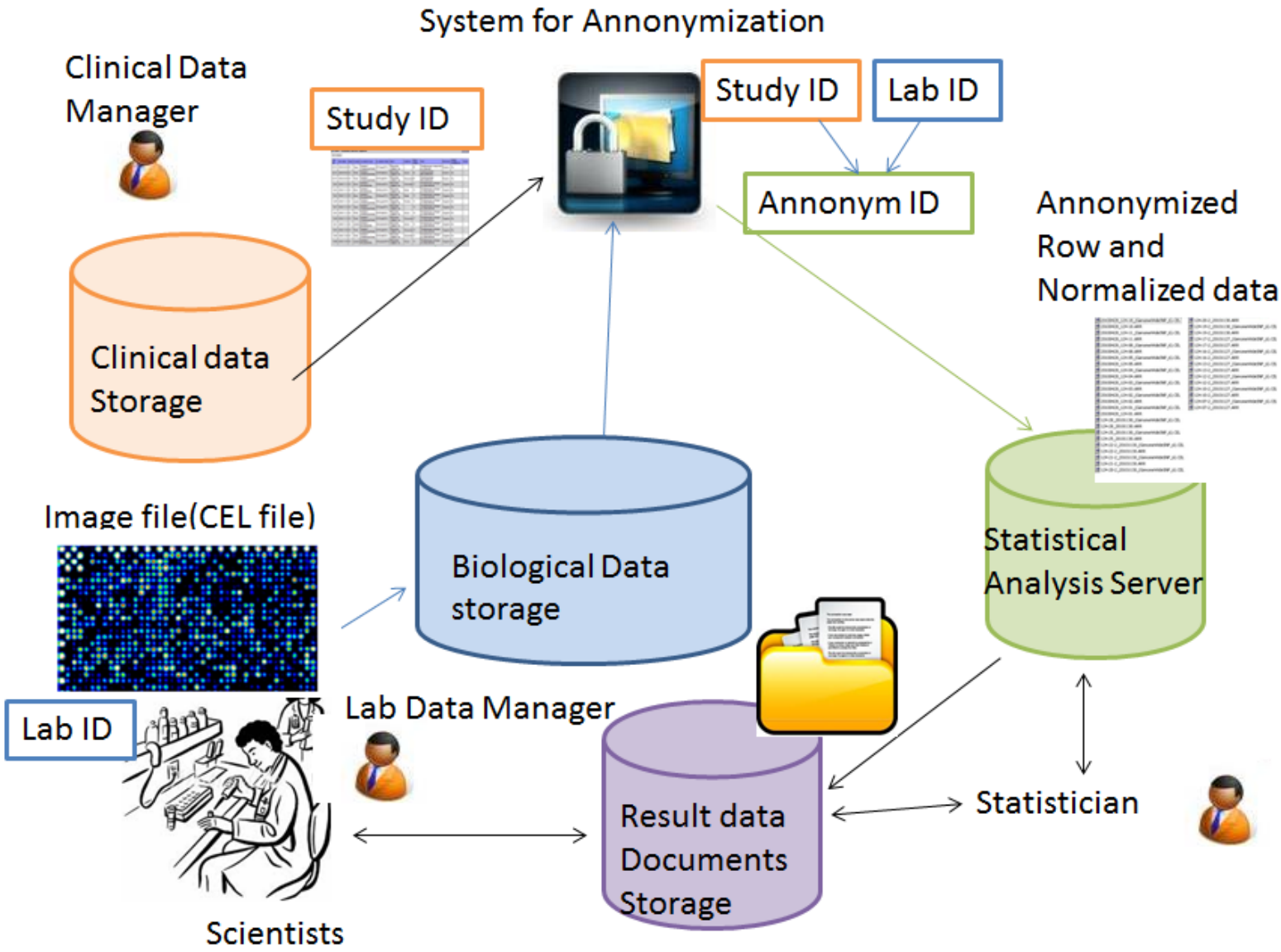
For no.1, we set the web and computing server in JCRAC data center, and set up a computer network for some terminal users in the hospital and the research institute to have access to those central servers. We have established a VPN connection for the terminal users in the hospital.

For no.2, PI (TanakaN) visited Center for Health Bioinformatics in Harvard University in Boston, USA to meet Dr.Winston Hide on 21th October, 2013. Tanaka asked Dr.Hide about QC system for biological and genetic data from clinical studies at Harvard, and knew that tight regulations for clinical bioinformatics research in the US discourage researchers from conducting analysis of human genetic data with clinical outputs. Then many bioinformaticians at Harvard currently did not analyze genetic data with clinical information, which means, they were analyzing only biological data. So Dr.Hide recommended visiting one of Canadian cancer institute to see how they control clinical data with genetic information in the hospital.

For no.3, we choose Galaxy, which is an open, web-based platform for data intensive biomedical research. We set up the test environment for testing NCGM Galaxy customized for our purpose, QC for DNA methylation data. Testing NCGM Galaxy and related algorithms in the test environment, we completed an advisory contract with Amerief.co.ltd on the setting up Galaxy on the main computing server and Mr.Toyooka and Mr.Mochizuki, who were both student fellows at Biostatistics Section, constructed NCGM Galaxy with help of Amerief.co.ltd. Mr.Toyooka and Mr.Mochizuki also developed an algorithm to detect SNPs on the probes for Genome-Wide DNA methylation studies, and tried to implement it in NCGM Galaxy.

Additionally, Dr.Kawasaki and Dr.Ohtsu assessed items which need to include in the management system for experimental samples.

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



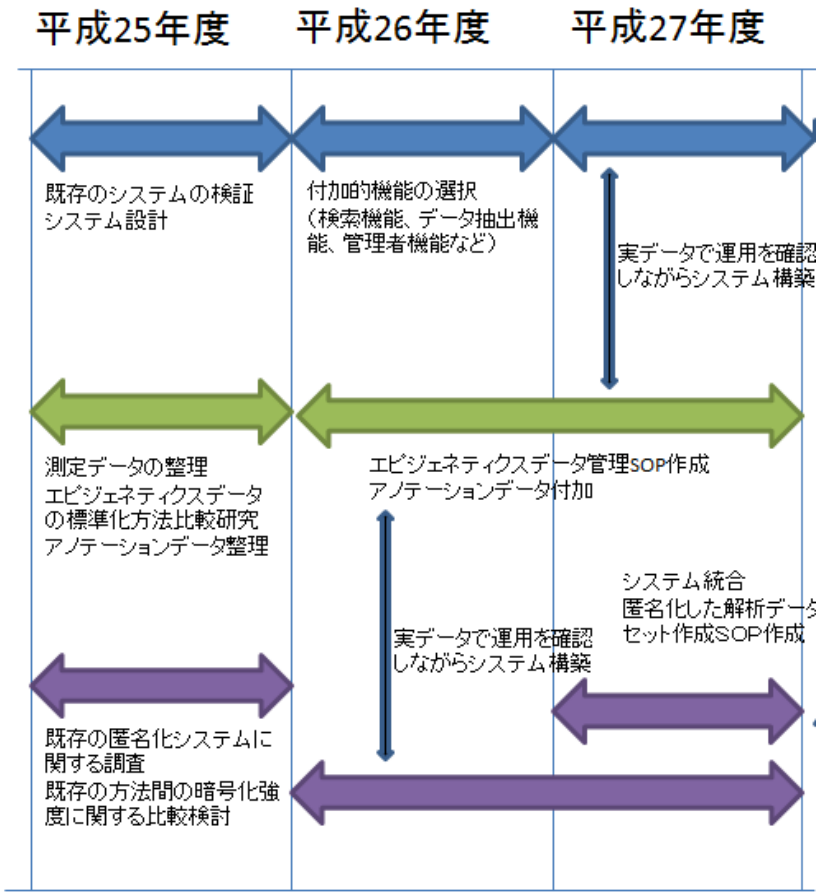
研究計画と初年度達成内容

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



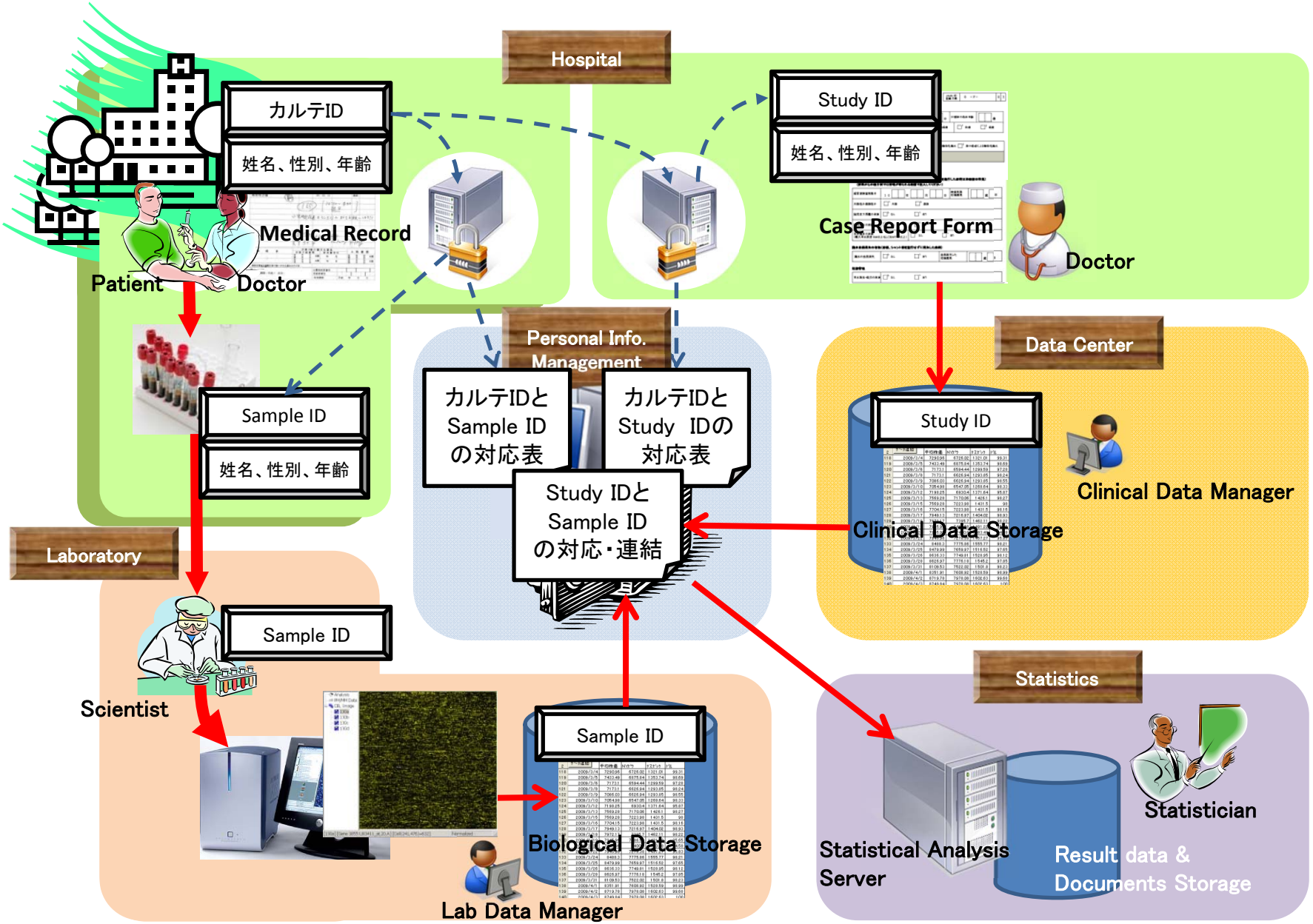
初年度達成事項

・システムデザインの核となる既存プラットフォーム（Galaxy）の構築
・Galaxy上でいくつかのRコードが動くことの確認

・既存の標準化プロセスの問題点（SNPの存在、4ピークの検体の扱い）を発見
・SNPをdetectするアルゴリズムの提案
・PCRによりいくつかのプローブでアレイデータの信頼性確認
・B細胞性リンパ腫特有のメチル化発現分布の発見

個人情報の匿名化の方式、アルゴリズムの現状についての文献調査

【25指112分担研究課題名】
匿名化した臨床データを結合した解析セット作成システムに関する研究



匿名化システム（市販）

製品名	販売／開発元	導入実績等特記情報
BIOPRISM	NEC	国立循環器病研究センター、国立国際医療研究センター、国立成育医療研究センター等
匿名化システムAnonymity	メディビック	内資系製薬企業、国立大学附属病院、クリニック等多数。「検体管理システム SATS」との連携カスタマイズ
匿名化番号作成システムANCS	日本システム開発	国立がん研究センター中央病院/東病院等 ソフトウェア：¥600,000 保守費：¥120,000/年
遺伝子情報匿名化システムDAHLIA	有限会社エッチ・アンド・ティー	虎ノ門病院（本院・分院） 250万円（最小構成端末2台含む）～
匿名化情報管理サービス 匿名バンク	日立ソリューションズ	（株）ファンケル等
医用画像の個人情報匿名化ソフトウェアWhiteBerry	三菱スペース・ソフトウエア	詳細不明 希望小売価格：「WhiteBerry」150万円(税別)、 「WhiteBerry/Lite」14万円（税別）
遺伝子情報提供者匿名化システムSCTS21	三井情報	現在販売中止（自治医科大学、国立循環器病研究センター、三重大学等）
臨床検査匿名化システム	NTT データ	現在販売中止（大阪大学等）

主なアルゴリズム開発の取り組み事例

研究/技術開発概要	開発者
データを暗号化したまま高速で内積計算を可能にする技術開発　準同型暗号、複数ビットの一括暗号化方式	富士通研究所　小暮淳
プライバシーを含むLinked Data（RDF）の秘匿分析技術開発　準同型暗号	富士通研究所　津田宏、伊藤孝一
JALSG（日本成人白血病治療共同研究グループ）　医療統計処理における秘密計算技術を実証　秘密分散、マルチパーティ計算	NTT
パーソナルデータをより高速で匿名化するプライバシー保護技術を開発　k-匿名化	NEC
データベースの情報を暗号化したまま処理できる秘匿計算技術を開発 Request-Base Comparable Encryption	NEC　古川潤
秘密計算による化合物データベースの検索技術 筑波大との共同研究 加法準同型暗号	産総研CBRC
Sharemind 秘密分散、マルチパーティ計算	University of Tartu エストニア

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
　　エピジェネティクスデータの品質管理工程の構築に関する研究
主任研究者名：田中紀子
分担研究者名：志村まり　田中康博

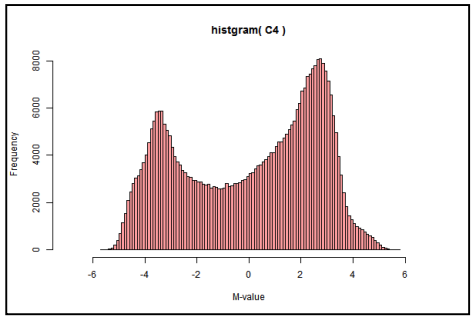
キーワード　：omics; data management; database; clinical bioinformatics; standardization
研究成果　　：
平成 25 年度の研究成果報告
（志村班）
初年度の計画は実データとして実装するエピジェネティクスデータに関して、データ標準化方法の標準化のために、ＤＭＰおよびＳＯＰ作成を目標として標準化方法についての研究を行い、そのために必要となる DNA メチル化アレイ解析の信頼性評価に際し、個々の評価遺伝子のいくつかについて、PCR、DNA sequence を行うことであった。平成 25 年度においては以下の 3 点について実施した。
1.　　DNA メチル化データの品質管理工程の標準化に関する研究
イルミナ社の４５０Ｋパネル測定データのノーマライゼーションの必要性については、Dedeurwaerder ら(Epigenomics2011)が４５０Ｋパネルで用いられている二つのアッセイ間に測定として本質的な差があることが指摘されてから、アッセイ間の分布を揃えるピークシフト法と呼ばれる方法がいくつか提案され、用いられてきた（Dedeurwaerder et al.,2011; Maksimovic et al., 2012; Touleimat et al., 2012）。しかし、我々の実際のデータでは既存の方法だけでフィルタリングしきれない質の悪いサンプルやプローブの存在の可能性が示されていた。その問題は
１）　BMIQ では３ベータ混合分布（3 峰性）を前提にノーマライゼーションが実行されるが、単峰性や多峰性も松永ら（AIDS 投稿中）のデータでは観測された。
２）　ＳＮＰフィルタリングのリストが、どの人種のデータを基にして作成されたのか不明だったので、本当にフィルタリングアウトしたプローブ中にＳＮＰがあるのかないのかを確認するのが難しい（無駄なフィルタリングをしている可能性）
１）については、まず現在の品質管理プロセスで、３峰性以外のもの（測定がうまくいっていない可能性が高いサンプル）が除外されているかどうかについてＮＣＧＭおよび GEO に登録済みの二つのデータセットで確認したところ、単峰性については品質管理基準をデータセットごとに変えることによって除外可能であったが、多峰性のサンプルに関しては不可能であった。そこで、BMIQ アルゴリズムのソースコードから混合分布パラメタを取り出すことが可能かどうかの検討を行ったところ、そもそも BMIQ アルゴリズムでは混合分布の推定方法に問題がある可能性を発見した。そこで、田中紀子医学統計研究室長および川崎洋平上級研究員によって BMIQ アルゴリズムの混合分布推定部分について改善した新しいアルゴリズムの開発を行った。さらに、多峰性を示したサンプルが、B 細胞性リンパ腫の臨床検体のみであることを GEO のサンプルの解析で明らかとした。
２）については lumi package より配布されている SNP リストの配布元研究者に確認したところ、情報が古く、人種も考慮していないという回答を得たため、まずはターゲットポピュレーションに適当でないＳＮＰリストを用いた場合のフィルタリングおよびノーマライゼーションを行ったことによる研究結果への影響を調べることとした。この結果については 11 月の日本人類遺伝学会にて発表した。
2. DNA メチル化アレイ解析の信頼性評価
PCR、DNA sequence の方法を確立し、既に各解析について２～５target についての DNA メチル化の程度について、PCR、DNA sequence での確認を行った。結果、これまで解析した target について、いずれも DNA メチル化解析および統計解析結果を反映している結果となったことが示唆された。

エピジェネティクスデータの品質管理工程の構築に関する研究

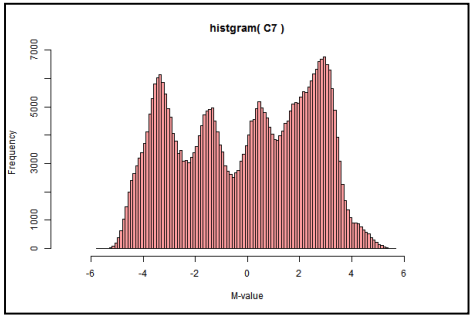
25指112分担研究
研究所難治性疾患研究室
志村まり

異常なピークを示したDNAメチル化分布の解析
―B細胞性リンパ腫瘍臨床検体―

2ピーク



>2ピーク

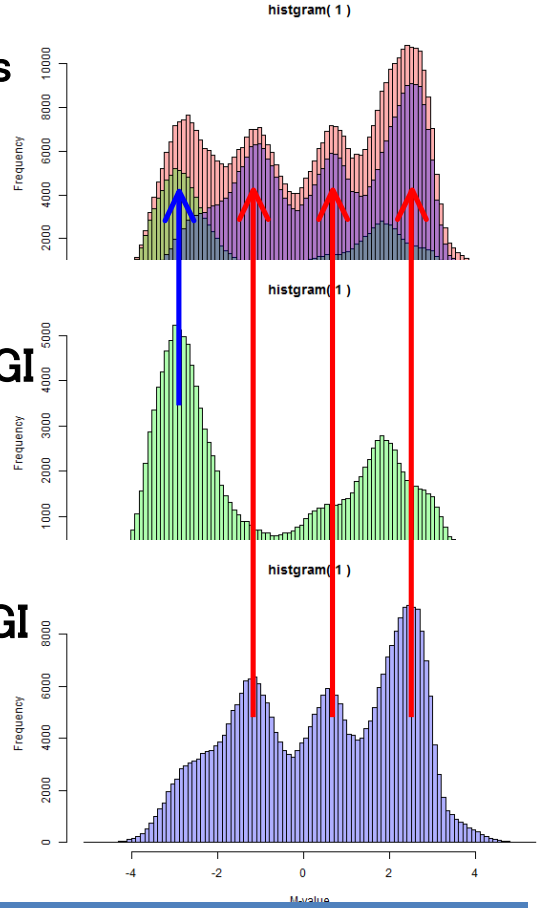


DNAメチル化分布	検体数(%)
2ピーク	6 (21.4%)
3ピーク	10 (35.7%)
>3ピーク	12 (42.9%)
合計	28

All Probes

CGI

Non-CGI



1. 40%の検体で異常な複数ピークが見られた。
2. Non-CGI領域のメチル化状態が複数ピーク形成に寄与していることが判明した。

異常なピークはB細胞性リンパ腫瘍特有？

米国NCBI データベース検索

GSE37362 (びまん性大細胞型B細胞リンパ腫) → 2,3または複数ピーク

2&3ピーク: 8検体

複数ピーク:23検体

GSE46306 (頸部上皮内癌) → 2ピーク

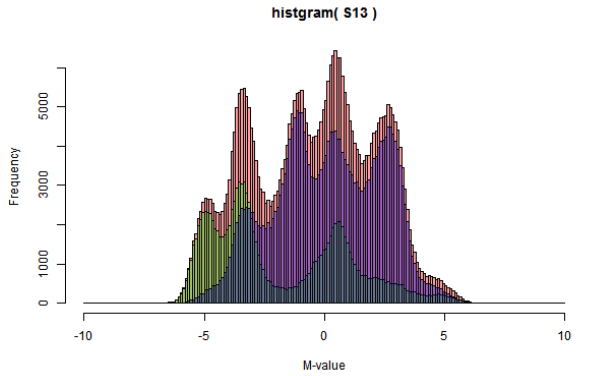
GSE38268 (頭頸部扁平上皮癌) → 3ピーク

GSE4005 (頭頸部扁平上皮癌) → 2ピーク

GSE44837 (乳癌) → 2ピーク

GSE49656 (胆管癌) → 2または3ピーク

複数ピーク例 (GSE37362)



3. 公開マイクロアレイデータベース検索より、私たちのマイクロアレイ解析データと同様の傾向は、B細胞性リンパ腫瘍のみで確認された。

4. 以上より、特にB細胞性リンパ腫瘍では、Non-CGIとCGI領域に別けた、データ解析が望ましいと考える。臨床解析にどのような影響を与えるか、今後GO分析を行い検討する。

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
　　匿名化した臨床データを結合した解析セット作成システムに関する研究
主任研究者名：田中紀子
分担研究者名：志村まり　田中康博

キーワード　：omics; data management; database; clinical bioinformatics; standardization
研究成果　　：
平成 25 年度の研究成果報告
（田中（康）班）
初年度の計画は現在国内外において、どのような匿名化システムが使用され、あるいは構築され実際の研究に利用されているのか調査を行い、また実際に使われているシステムとそこに実装されているアルゴリズムの比較を行うことであった。平成 25 年度においては以下の２点について実施した。
（A）現在国内外において、どのような匿名化システムが使用されているか。あるいは構築され実際の研究に利用されているのかの調査をまとめた。
（B）実際に使われているシステムとそこに実装されているアルゴリズムの比較調査を行った。
臨床データは、個人情報に関係するため、個人情報の匿名化の方式、アルゴリズムについての現状について、文献調査を中心に情報を収集してきた。
特にクラウド・コンピューティングに代表されるインターネット上で大量のデータが誰でも自由に扱える状況になったため、個人情報やプライバシー情報の取り扱いや匿名化のための研究、技術開発が活発に行われていると思われるため、先ず（財）日本情報処理開発協会やシンクタンク（三菱総合研究所等）の公開されている調査レポートにより、技術的な動向を調べた。
現時点で分かる範囲で大きく分けると、着目する視点により以下の分類、アプローチになるように思われる。
（1）個人情報を含む情報ないしデータに対する技術的処理の如何に関わらず、頑強なセキュリティ、ネットワーク監視システムの下での利用に制限をかけた個人情報の利用
（2）セキュリティ対策やネットワーク監視が施されたシステムではあるが、万が一の情報漏洩がなされても被害が発生しない、あるいは発生しても被害が最小限にとどまるよう、個人情報も含む情報ないしデータに対して匿名化等の技術的処理を施す
（3）ネットワークシステムに依存せず、個人情報を含む情報ないしデータが安全かつ頑強にセキュリティが保たれるように、データ変換等の技術的処理を施す。
臨床データを扱う医療分野においても基本的な捉え方は同様と思われる。
特に本研究の（B）に関係する（3）の研究動向を中心に情報を集めた。
現状をまとめると、以下のようなアルゴリズム等の研究がおこなわれているようである。
（1）プライバシー保護データ公開
ランダム化（マイニングに影響を与えないようなノイズを挿入）、削除（マイニング結果に影響を与えない稀少なデータを削除）、曖昧化（k-匿名化、l-多様性、t-closeness 等）といった手法により、データ（データベース）そのものに雑音を加えたり、情報を間引く。公開データを組み合わせることにより、プライバシー情報を推測される（link attacks の脅威の）可能性があり得る。
（2）セキュア計算（プライバシー保護データマイニング）
　　データを流通させずに、分析結果のみを開示する仕組みを狙い、分散したプライバシー情報の暗号化処理で、公開鍵暗号やデータ分析手法等を組み合わせる。
以上を総合すると以下の、
（1）市販、商用システムとしてそのまま組み込めるものは数が限られる。仮に組み込むとしても利用に合わせたカスタマイズが必要になる。
（2）実用レベルのアルゴリズムについては、技術情報が開示されているものは殆ど無く、プライバシー保護データマイニングやセキュア計算等の研究段階のアルゴリズムにとどまっている。

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
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

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。