博士論文

ゲノム情報解析の再現性と再利用性を向上させる情報基盤の設計

山中　遼太

1. 序論 4
1.1. 背景 4
1.1.1. 次世代シークエンシングとゲノム情報 4
1.1.2. ゲノム情報を解析するための情報基盤 7
1.1.3. ゲノム情報解析とオープン・サイエンス 8
1.2. 目的 13
1.2.1. ゲノム情報解析の再現性の向上 13
1.2.2. ゲノム情報解析の再利用性の向上 15
1.3. 構成 16
2. 課題の分析 17
2.1. 前提知識 17
2.1.1. ゲノム情報解析の種類 17
2.1.2. ゲノム情報解析におけるデータ処理 19
2.1.3. ゲノム情報解析におけるデータ統合 20
2.2. 問題と先行研究 22
2.2.1. データ処理における問題 22
2.2.2. データ処理における先行研究 24
2.2.3. データ統合における問題 26
2.2.4. データ統合における先行研究 28
2.3. 課題設定 30
2.3.1. 問題を解決するための情報基盤 30
2.3.2. データ処理における課題設定 32
2.3.3. データ統合における課題設定 33
3. 研究成果１（データ処理） 35
3.1. 課題 35
3.1.1. ツールとバージョンの管理 35
3.1.2. ワークフロー管理システムの利用 37
3.1.3. Galaxy の概要と課題 39
3.1.4. 連携されたシステムの提供 42
3.2. 手法 43
3.2.1. 同一の実行環境を共有する方法 43
3.2.2. コミュニティ仮想マシンの開発 46

3.2.3. 開発者会議の定期的な開催 47
3.2.4. 利用者への成果物の提供方法 48
3.2.5. パブリック・クラウドの利用 50
3.2.6. サーバー・クラスタの利用 51
3.2.7. 構築手順のコード化 52
3.3. 結果 54
3.3.1. コミュニティ仮想マシンの公開 54
3.3.2. 利用者への成果物の提供 56
3.3.3. 再現性と再利用性の評価 59
4. 研究成果 2（データ統合） 61
4.1. 課題 61
4.1.1. がんゲノム・データの利用 61
4.1.2. がんゲノム・データのデータ統合 63
4.1.3. RDF データの活用 64
4.2. 手法 66
4.2.1. RDF スキーマの定義 66
4.2.2. RDF データの生成 69
4.2.3. 外部データとの統合 72
4.2.4. データ・ポータルの開発 76
4.3. 結果 78
4.3.1. RDF データの公開 78
4.3.2. データ・ポータルの公開 78
5. 結論 80
5.1. 結果総括 80
5.2. 今後の課題 81
5.3. 将来展望 82
6. 謝辞 84
7. 引用文献 85

1. 序論

1. 1. 背景

1.1.1. 次世代シークエンシングとゲノム情報

1990 年に始まったヒトゲノム計画は国際ヒトゲノム・シークエンシング（配列解析）・コンソーシアムによる国際的協力のもと 10 年以上の歳月をかけてヒトのゲノムの全塩基配列を決定した[1]．その後，2000 年代後半には次世代シークエンシングと呼ばれる様々な技術革新により，ゲノム配列を解読するコストは下がり続け，2015 年のはじめにはヒト全ゲノム配列を 1,000 ドル以下で解析できる「1,000 ドルゲノム」が達成されたと発表された[2]（図 1）．シークエンシングのための解析機器である次世代シークエンサーが研究所や研究室の単位で運用できるようになることで，多様な生物種のゲノム配列の決定に留まらず，配列同士を比較して検出される多型や変異，mRNA や miRNA の定量による転写産物の発現量, ヒストン修飾や DNA のメチル化状態といったエピゲノム情報, など様々なオミックス・データ（ゲノミクス，トランスクリプトミクス，プロテオミクス，などの生体内の分子データの総称）の生成に活用されるようになった．本論文では，これらシークエンシングによって取得できるオミックス・データをゲノム情報と呼ぶこととする．

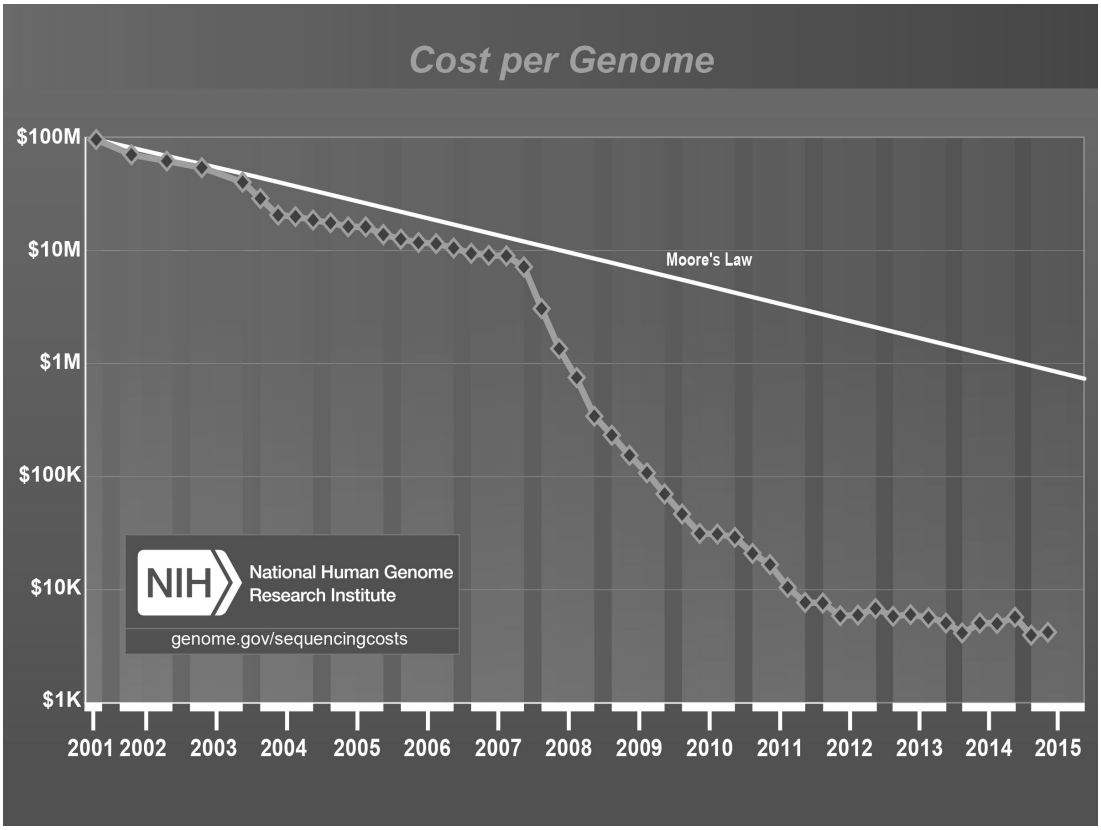

図１：　ヒトゲノムの解読にかかるコスト

ヒトゲノムの解読にかかるコストは 2000 年代初頭には 1 億ドルであったが，その後，ムーアの法則（18 ヶ月で 2 分の 1）と同様の速度で低下し続け，2007 年頃からは次世代シークエンシング技術の登場によりさらに急激な速度でコストが下がり，2015 年には 1,000 ドルを達成したと発表された．(http://www.genome.gov/sequencingcosts/)

今後も以下のような情報が生命科学研究において必要とされているため，さらに広くシークエンシングが使用されるだろう．低コストでゲノム情報を解析できるようになることで，世界中の研究所や医療機関などでゲノム情報解析によるデータが大量に生成されることが予想される．

- 多様な生物種： 既に多くの生物種のゲノムが解読されているが，今後もさらに多くの生物種のゲノム情報が研究に利用されるだろう．特定の生物種ではなく，環境中の微生物のゲノムを同定するメタゲノム解析のような手法も広く活用されてきている．

- 個体間の比較： 人種や個体によってゲノム配列が異なるため，人種差や個体差を知ることによって疾患の研究を精緻化することができる．特に，個人が持つ一般的な SNPs（一塩基多型）や希少な多型の情報を用いた精緻化医療の実現が期待されており，生命医科学におけるトランスレーショナル研究においてゲノム情報解析は重要である．

- 組織間の比較： 個体内のそれぞれの組織の細胞は，ゲノム情報は同一である一方，異なるエピゲノム情報（ヒストン修飾や DNA メチル化）を持っており，組織ごとの遺伝子の機能の違いや発生過程を研究する上で必要とされている．さらに，がん組織のような異なる変異を持つ不均一な細胞集団のゲノム情報を比較するためにさらに多くのサンプルが解析されるだろう．

- 時系列の比較： 細胞に特定の刺激を与えた際の転写産物の発現量の時間推移を観察するためには，時間を追って同じサンプルを何度も解析する必要がある．このような時系列データを使用して細胞内の現象のシステム生物学的な理解を進めるためには，転写因子の結合や転写産物の発現，タンパク質の発現といった多種のオミックス・データ（マルチオミックス・データと呼ばれる）の統合解析が必要となるため，さらに多くのゲノム情報が必要である．

このように，次世代シークエンシングは生命科学研究において様々な用途で利用される一方，今後は健康サービスや医療サービスにも広く活用されることが予想される．例えば，個人が自身のゲノム情報を知ることで罹りやすい疾患を把握して予防に役立てるため，ゲノム情報の取得と共に疾患のリスクに関する情報を提供する，といった健康サービスが既に始まっている．さらに，臨床シークエンシングと呼ばれる医療応用においては，医療機関が患者のゲノム情報を取得し医師がこの情報を用いることで，個々の患者に対してより精緻な診断ができるようになる，といった精緻化医療の発展が期待されている．

1.1.2. ゲノム情報を解析するための情報基盤

次世代シークエンサーによって取得されたゲノム情報はその解釈のために計算機を用いた解析が不可欠であり，そのための情報基盤が必要となる．この解析は「データ処理」と「データ統合」の二つの段階に分けられる．

まず，次世代シークエンサーから得られる情報は配列断片を解析したデータ（リードと呼ばれる）であり，このデータそのものから有用な情報を読み取ることはできないため，配列変異，転写産物の発現量，DNA メチル化状態など，仮説検証に必要な情報を得るために，それぞれの実験系に合わせてリードを処理する．

その後，ここで得られた情報を既存の知識と統合してはじめてこの情報を活用することができる．例えば，発現の高い遺伝子リストに対してこれらの機能に関する情報を付与したり，DNA メチル化領域と遺伝子の既知の制御領域の染色体上の位置関係を確認したりするだろう．また，がん組織の配列変異や遺伝子発現プロファイルといった情報を取得した場合には，これをデータベース中の他の症例と比較して同様の症例の情報を収集することができるだろう．

このように，次世代シークエンサーから得られたゲノム情報は，アラインメントや変異の検出といった「データ処理」とアノテーションやデータベース化といった「データ統合」の二つの段階を経て，最終的に生命科学または医療において分析または解釈可能になる（図 2）．本論文では，計算機上で実施されるこの二つの段階をゲノム情報解析と呼ぶこととする．ゲノム解析全体の質を担保するためには，シークエンシングに使用するサンプルの調製から取得されたデータの解析と解釈までの過程を一貫して管理する仕組みが必要である．

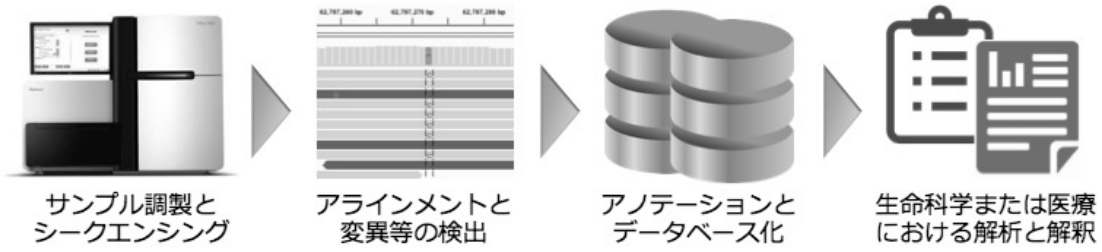

図 2： ゲノム情報解析におけるデータ処理とデータ統合

サンプル調製から解析と解釈までを含めたゲノム解析の過程うち，データ処理とデータ統合の段階は計算機を必要とするゲノム情報解析の過程といえる．

1.1.3. ゲノム情報解析とオープン・サイエンス

大量のゲノム情報が取得されることにより計算機によるデータ解析が必要となり，ゲノム科学における情報解析と情報基盤の重要性が広く認識されるようになった．近年ではゲノム科学のデータ駆動型アプローチの可能性が注目されているが，そこには，従来より多くのゲノム情報が公共データ・レポジトリに収集されていて研究者が自由にアクセスできる環境が整っているという背景がある．ヒトゲノム計画においては各国の公的研究資金を用いた研究プロジェクトが国際的協力に拡大し，1996 年にバミューダで開催された会議において，解読されたゲノム配列は直ちに公開するという原則が合意された（バミューダ原則）．以来，1000 人ゲノム・プロジェクト，ENCODE プロジェクト，国際がんゲノム・コンソーシアムなどの大規模プロジェクトが大量のゲノム情報を公開している．

ゲノム科学に留まらず科学研究全体において，このように公的研究資金を用いた研究の成果を広く一般に公開，利用可能にすることで，より迅速に科学研究とその応用を推進するという，オープン・サイエンスの概念が急速に広まっている．2013 年の G8 科学大臣会合における共同声明で論文のオープン・アクセス化および研究のオープン化に言及されたことを受けて，国内でもオープン・サイエンス推進の必要性が議論されており[3]，今後，公的機関の基本方針が策定されると共にデータ・リポジトリ等の情報基盤整備が検討されるだろう．オープン・サイエンスは多くの既存の概念を包含しており，定義は必ずしも一意ではないが，情報解析の視点では以下のような要素が指摘されている[4].

- オープン・アクセス： 研究結果が公開されていること．特に，学術誌を購入することなく論文を無償で入手可能であることを指すことが多い.
- オープン・ソース： 情報解析に使用されたソフトウェアが公開されており改変や再配布が許可されていること.
- オープン・データ： データが公開されていること．データ処理を伴う情報解析の場合は処理前のデータが公開されていることが望ましい.
- オープン・メソドロジー： 手法の詳細（使用されたツール，データの収集，解析フロー）が公開されていること.

G8 科学大臣会合のオープン・サイエンス推進の議論では，学術誌のオープン・アクセスや科学研究で得られた解析前データを公開するオープン・データに焦点が当てられている．一方で，これらのデータを解析するためのソフトウェアが公開されていることや，学術誌に記載された特定の研究でどのデータにどのような解析手法が用いられたかといった詳細が公開されていることも，解析を再現および共有し，改善するために必須であるといえる．以下の通り，ゲノム科学においては，それぞれの概念は一般的になりつつある.

まず，オープン・アクセスとは，ここでは学術誌に掲載される論文を誰でも経済的または法的な障壁なしに入手できることを意味している．元来，学会や学術誌は本来研究成果の共有を目指したものでありオープン・アクセスの考え方に合致しているが，特に 2000 年代後半にはインターネットから無償で論文を入手できるオープン・アクセスの学術誌が多数創刊され，現在ではオープン・アクセスの概念はこれらの学術誌を指すことが多い．ただし，インターネット上にオープン・アクセスの論文が増加している一方で，現在はまだこれらの学術誌の影響は限定的であるといった議論もある[5]．ゲノム情報解析においては，オープン・アクセスの学術誌である GigaScience 誌が投稿された論文のデータ共有と解析の再現性向上を促進するためにデータ・レポジトリ GigaDB を提供するなど，オープン・サイエンスに対する新しい試みが注目されている[6]．

次に，オープン・ソースとは，多くの場合，誰もが使用，改変，派生物の作成，配布などを許可されたライセンスのもとでソフトウェアが入手可能であることである．このようなソフトウェアは SourceForge や GitHub といった公共のレポジトリを使用して多数の開発者の協業によって開発されることも多く，ゲノム情報解析のツールについてもこれらのレポジトリは利用可能であり，実際に多くのツールがこれらの上で公開されている．例えば，人気の高いマッピング・ツールである BWA や Bowtie2，さらに RNA-seq 解析ツール cufflinks は SourceForge で公開されており，ChIP-seq 解析ツールの MACS や解析プラットフォームの Galaxy は GitHub で公開されている．

オープン・データとは，ここで扱われるデータが公共に入手可能であることである．多くの場合，これはインターネット上で参照できることを意味している．前述の通り，ゲノム科学においてヒトゲノムプロジェクト以降，公共データ・レポジトリを用いてデータが共有されてきた歴史がある．現在も SRA (Sequence Read Archive)というレポジトリに膨大な次世代シークエンシング・データが収集され，これらは公共に入手可能なオープン・データである．

- **Sequence Read Archive**： 次世代シークエンシングによって生成された塩基配列データのデータベースであり，米国の NCBI (National Center for Biotechnology Information)，欧州 EBI (European Bioinformatics Institute)，日本の DDBJ (DNA Data Bank of Japan)によって運営されている[7]．ここには，各研究所の生成した RNA-seq や ChIP-seq のデータの他，1000 人ゲノム・プロジェクトなどの大規模なデータが登録されている．多くのプロジェクトやオープン・アクセス・ジャーナル，また Nature Publishing Group のジャーナルなどが，SRA へのデータの登録を義務付けている[8].

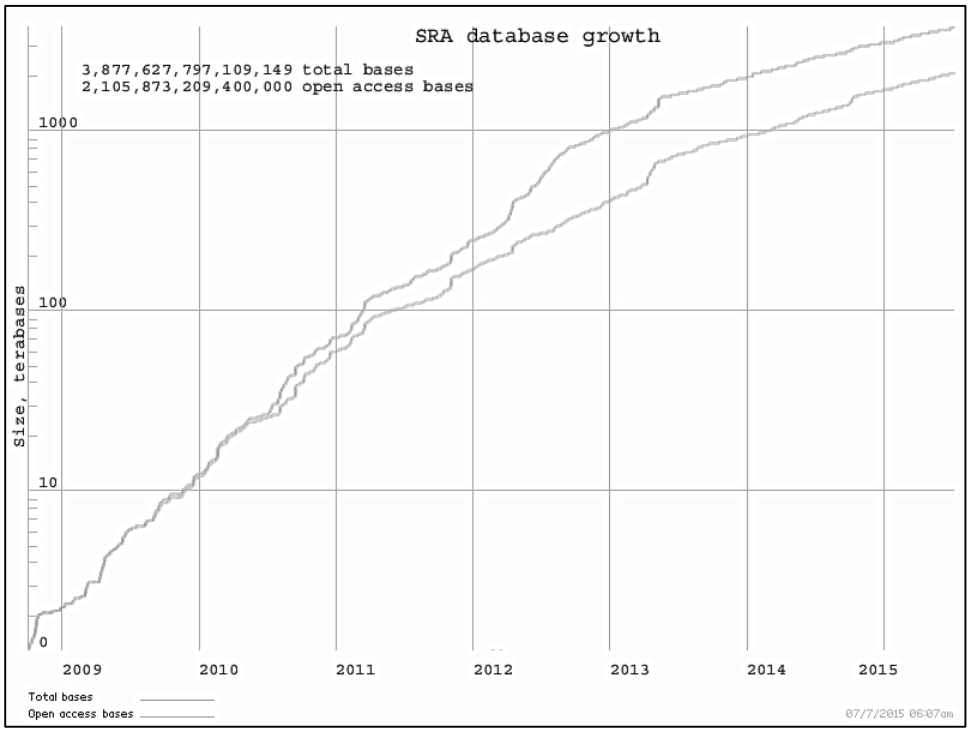

図 3： 次世代シークエンサー・データの増加

2015 年 7 月 7 日現在の，米国 NCBI (National Center for Biotechnology Information) の SRA (Sequence Read Archive) に登録されている次世代シークエンサー・データの合計の塩基数が示されている (http://www.ncbi.nlm.nih.gov/Traces/sra/)．数値の大きい方の線は全てのデータの塩基数，小さい方の線は公共にアクセス可能なデータの塩基数であり，既に数ペタ (= 数千兆) 塩基分のデータが登録されていることが分かる．

主要な生命科学の学術誌における，データ共有のポリシーと実際のデータ共有の状況に関する 2009 年の調査では，多くの学術誌がデータ共有を要件または推奨としているものの，ポリシーが守られていない論文も多く，全てのデータをインターネット上に公開している論文は 9%程度であったという報告もある[9]．ここでは，研究資金の提供機関がデータ公開を研究者に義務付けることなどが提案されているが，近年ではこのような取り組みが広く普及し，公共データ・レポジトリも整備されてきたため，データの共有状況が改善されてきたと考えられる．

以上のように，ゲノム科学におけるソフトウェアの多くがオープン・ソースでありデータの多くが公共データ・レポジトリに管理されている．しかしながら，上記の調査では，データが共有されてもデータ解析方法の記述が完全ではないため結果の再現は難しい場合が多いといった問題も指摘されている．実際，データと解析ツールが入手可能な状況でも，論文に解析手法に関する十分な記載がないなど理由で解析結果を再現できないことがあるが，それでは，この解析手法を他のデータに適用し，新たな仮説を立て，さらに解析手法を改善していくといったオープン・サイエンスの枠組みが実現するデータ駆動型研究のサイクルが断絶されてしまうだろう．

そこで，オープン・メソドロジーといった考え方を導入する必要があると考えられる．オープン・メソドロジーとは，入力データと解析ツールを用いて論文で示した通りの結果を得るために必要な情報を共有することである．これは一見すると，論文に手法をより詳細に記載すれば十分であるようにも感じられるが，本来，論文に記載する手法は解析の完全な再現や共有を目的としたものではない．ゲノム情報解析のような研究領域でデータ駆動型研究が始まることで，データ解析の完全な再現という需要が生まれたといえる．

データ解析の再現に必要な情報とは，解析ワークフローの記述に加え，ソフトウェアのバージョンや実行パラメーター，さらには実行環境の詳細といったものを含んでおり，文書の記載だけでなくパラメーターの値や特定の計算環境の指定といったことも含まれる．このような新たな情報共有を可能にするため，ゲノム情報解析においてオープン・メソドロジーという考え方を明示的に検討することは有意義であると考えられる．

以上の通り，ゲノム情報解析ではオープン・ソースの解析ツールやオープン・データとしてのデータの共有が以前から必要不可欠となっており，オープン・サイエンスを先駆けてきた領域であるといえる．今後はさらに，メソドロジーが共有されてデータ解析が再現可能になり，解析ツールや公共データが再利用されることでコミュニティ全体として効率的に研究が進められ，再利用されたツールやデータさらにその論文がコミュニティによって評価されるといった，オープン・サイエンスの枠組みの成熟が期待できるだろう．

1.2. 目的

1.2.1. ゲノム情報解析の再現性の向上

前節の通り，ゲノム情報の急激な増加に伴い，ゲノム科学研究におけるデータ駆動型アプローチの重要性が高まっている．また，これらの解析にオープン・ソースのツールが使われるとともに公共に入手可能なデータが増加し，ゲノム情報解析は共有されたツールとデータから新たな知見を生み出すオープン・サイエンスの先駆けとなる研究領域と考えられている．このように，解析手法の共有を前提としたアプローチにおいて，解析の再現性を維持することは不可欠であり，再現性のある研究（reproducible research）の枠組みを構築する必要性が示唆されている．

生命科学の学術誌においては，実験の再現性の問題が長く議論されてきているが，解析の再現性についても見直しが求められている．例えば，使用された情報解析ツールが公開されていたり，解析が再実行可能であったりすることを論文掲載の条件とするといった動きがあり，この傾向は今後も強まると言われている[10]．また，以下のような問題が学術誌で検討され始めている．

- 著者は開発した手法を実行するソース・コードを常に提供するべきか.
- 公開されているソフトウェアを用いる場合，最低限提供されるべき情報は何か.
- 学術誌の審査員はソフトウェアを実行したり，ソース・コードを確認したりする責任があるか.
- 審査員は補足として提供されたデータ・セットをダウンロードして計算手法を再実行するべきか.

上記のような問題を考えるにあたって，まずは，学術誌が再現性のある論文の投稿者に対して十分なインセンティブを与えられるようにした上で，再現性が必要条件となるように移行していく必要があるだろう[11]．ひとつの例として，学術誌「Biostatistics」では，掲載の決まった論文に対して「再現性確認委員（AER: Associate Editor for Reproducibility）」がその論文のデータの公開，コードの提供，再現性の 3 点を確認し掲載時にマークをつけるといった取り組みを始めている[12]．さらに，データとコードが入手可能であっても再現可能とは限らないため，再現性確認委員が実際にデータとコードを用いて同じ結果が再現できるかどうかを確認している．

しかしながら，学術誌により実施されている現状の対策だけでは，論文の投稿者および審査員や編集者に対して大きな負担を強いる可能性がある．実際には，情報解析を実施した際の，ツール，データ，実行環境を管理するための手法や情報基盤が整備されることで，研究者コミュニティが積極的に再現性のある研究に取り組み利益を受けられることが望ましい．データが増加し続けているゲノム情報解析において，オープン・メソドロジーの必要要件の定義，およびこれを満たす情報基盤の設計は急務であるといえる．

1.2.2. ゲノム情報解析の再利用性の向上

データ駆動型研究として注目されている研究分野の中には，宇宙望遠鏡や大型加速器など特定の実験機器でのみ取得できるデータを用いる研究や，分子シミュレーションなど最大規模の計算機でのみ計算可能なデータを用いる研究がある一方，次世代シークエンサーは研究所や病院それぞれで保有することができる機器であり，生成されるゲノム情報は多くの場合それぞれの施設の計算機で解析することができる．つまり，データや解析手法が世界中の研究所に分散されていることがゲノム情報解析の特徴の一つであるといえる．

その結果，それぞれの施設の研究者が自分たちのデータを解析しているため，施設ごとに情報系研究者（バイオインフォマティシャン）が開発した異なる解析ツールが使用されることも一般的である．多くのツールが公共に配布されてはいるものの，実際の解析においては既存のツールを使うだけではなく，ツールのアルゴリズムを理解して新しいツールを開発したり，データ資源を設計および管理してデータ同士を統合したりする情報系研究者が必要とされている[13]．その一方，情報解析が各研究の必須要素となることで，生命科学者にも使いやすく共通して使用できるツールが求められるようになり，生命科学者がコマンドラインなしに使用できる解析ツールが多数開発されている[14]．

このように，データ駆動型生命科学の進展のため多くの情報系研究者が必要とされ，情報系研究者以外の研究者も情報解析を理解することが求められる中[15]，限られた数のゲノム情報解析の研究者や開発者で研究成果を得るためには情報解析の中でも特にデータ処理のような定型作業の効率化が不可欠となる．研究所間で解析手法の違いを正確に評価し，解析手法を共有できるようにすることにより，冗長な解析ツールの開発を減らすことも限られた人資源の有効利用になるだろう．

前項の通り，ゲノム情報解析において再現性の向上させるための情報基盤の設計が求められているが，これと同時に，再現性のある解析を複数の研究機関で共有できるような情報基盤の設計を考慮すべきである．再現性の高い環境を解析手法の再利用にも活用するためには，同じデータを用いて解析が再実行できるだけでなく，その解析が開発者以外にも理解でき，異なるデータに対して利用できることが必要となる．

以上の通り，ゲノム情報解析の再現性と再利用性を向上を目指した情報基盤の整備が求められている．再現性のある科学研究の重要性が唱えられる中，再現性を維持することが研究者の負担になるのではなく，再現性と再利用性の向上によって情報解析が効率化される，そのような仕組みを提案することが，本研究の目的である．

1.3. 構成

本章では，本研究の目的を提示した．2 章では，ゲノム情報解析の内容を分析した上で再現性と再利用性に関する問題を指摘し，上記の目的を達成するために解決すべき課題を抽出する．3 章と 4 章では，2 章で設定された 2 つの課題に対する研究成果をそれぞれ記載している．5 章では，結果を総括し，本章で目的と掲げたゲノム情報解析の再現性と再利用性の向上をどれだけ達成できたかを議論する．

2. 課題の分析

2. 1. 前提知識

2.1.1. ゲノム情報解析の種類

次世代シークエンサーの活用方法は多岐に渡るが，ここでは，そのうち利用頻度の高い活用方法[16]を取り上げ，次項以降では，これらの活用方法における情報解析には「データ処理」と「データ統合」がそれぞれ必要であることを示す．主なゲノム解析は以下の通りである（表 1）．

- 全ゲノム配列解析： シークエンシングのコストが急激に下がり比較的低コストでヒトの 32 億塩基の全ゲノム配列解析が可能になった他，全ゲノム配列の 2%以下であるエクソン領域のみを解析するエクソーム・シークエンシングが広く利用されている．シークエンシング対象の領域を少なくすることで，同じコストでも読まれるリード数を増やして実際に解析される領域とその精度を高めることができる．全ゲノム配列解析は多型を検出するための効率的な方法であり，遺伝病の原因遺伝子の探索やがんの変異解析に用いられる．

- De Novo 配列解析： まだアラインメントのためのリファレンスがない生物種など，新しいゲノムの検出のための配列解析．リードのアセンブリが必要になるため，既知のゲノムのシークエンシングよりも計算量の多いゲノム情報解析が必要となる．

- メタゲノム解析： 16S rRNA 遺伝子などのターゲット・シークエンシングによる，環境サンプル中の微生物集団の分類や系統発生の解析．培養が困難なことから従来の方法では取得できなかった微生物のゲノム情報も入手可能であり，腸内細菌や土壌や海中の微生物が解析されている．

- RNA シークエンシング（RNA-seq）： 全 RNA 配列解析により RNA 発現の定量やアイソフォームや融合遺伝子の探索が可能な他，注目している転写産物の発現量の違いやアレル特異的な発現を測定するためのターゲット RNA シークエンシングが使用されている．また，転写産物の中でも短い塩基配列であるノンコーディング RNA やマイクロ RNA の発現量の測定のためには異なるサンプル調製が必要となる．

- ChIP シークエンシング（ChIP-seq）： タンパク質と DNA または RNA の相互作用を検出するため，クロマチン免疫沈降（Chromatin immunoprecipitation，略称 ChIP）とシークエンシングを組み合わせた解析．転写因子の結合領域や修飾ヒストンの検出に広く使用されている．

- DNA メチル化解析：DNA のシトシンのメチル化状態を取得するため，バイサルファイト処理によりメチル化されていないシトシンをウラシルに変換した後，DNA をシークエンシングしメチル化されたシトシンを検出する方法．エピゲノム解析において DNA 制御領域などのメチル化の影響を解析するために使用される．

表 1： 主なゲノム解析

解析	主な目的
全ゲノム配列解析	多型の検出（遺伝病の原因遺伝子の探索，がんの変異解析）
De Novo 配列解析	まだゲノムが解読されていない生物種のゲノムの決定
メタゲノム解析	環境サンプルに含まれる微生物の解析，新規 DNA 配列の検出
RNA-seq	転写産物の発現量の定量，アイソフォームや融合遺伝子の探索
ChIP-seq	転写因子の結合や修飾ヒストンの領域の検出
DNA メチル化解析	DNA のメチル化領域の検出（エピゲノム解析）

2.1.2. ゲノム情報解析におけるデータ処理

次世代シークエンサーから出力されるデータは「ショート・リード」と呼ばれる多数の短い DNA 配列または RNA 配列の断片（リード）であり，いずれの解析の場合にも，実験結果として解釈できる情報を得るためには計算機を用いてこれらリード･データを処理する必要がある. 現在広く使用されている Illumina Hi-seq 2500[17]のような次世代シークエンサーでは一回の配列解析の実行で数億ものリードが取得されるため，これらのデータの処理は計算コストが高く，高速化が求められる傾向がある．前出のゲノム情報解析におけるデータ処理はそれぞれ以下の通りである（表 2）.

- 全ゲノム配列解析： 出力された配列断片をリファレンス配列に対してマッピングすることにより，解析対象のゲノムを再構築する．さらに，リファレンスとの有意な違いを検出することで，SNPs やコピー数の多型や変異を検出する.

- De Novo 配列解析： リファレンスがないゲノムの場合，De Novo アセンブリングと呼ばれるプログラムによって DNA 配列を復元する．一般的に，De Novo アセンブリングはリファレンスがある際のマッピングと比較して，計算量やメモリ使用量が大きくなる.

- メタゲノム解析： アダプタ配列の除去などの処理の後，De Novo 配列解析と同様にアセンブルによってメタゲノム・サンプル中に含まれていた DNA 配列を復元する.

- RNA-seq： リファレンスのゲノムおよびトランスクリプトームに対して配列断片をマッピングし，アイソフォームを考慮して各転写産物の発現量を推定する．また，比較対象のサンプル間で有意な発現量の差がある遺伝子（DEG: differentially expressed gene）を検出する.

- ChIP-seq： 免疫沈降で得られた部分の DNA 配列が取得できるため，これらのリードをリファレンス・ゲノムにマッピングし，マッピングされたリードが有意に多い領域（ピークと呼ばれる）をタンパク質の結合位置として検出する.

- DNA メチル化解析： 全ゲノム配列解析と同様にリードをリファレンス・ゲノムにマッピングし，バイサルファイト処理によってウラシルに変換されていないシトシンを検出する．

このように次世代シークエンサーの出力データを解釈するためにはデータ処理が不可欠である．このデータ処理の方法によって検出の精度に大きな影響が生じることから，データ処理アルゴリズムの開発や適切な手法やツールの選択はゲノム情報解析において重要である．

表 2： ゲノム情報解析におけるデータ処理

解析	データ処理の例
全ゲノム配列解析	マッピング，多型や変異の検出
De Novo 配列解析	De Novo アセンブリング
メタゲノム解析	De Novo アセンブリング
RNA-seq	マッピング，転写産物の発現量の定量，DEG の検出
ChIP-seq	マッピング，ピークの検出
DNA メチル化解析	マッピング，メチル化されたシトシンの検出

2.1.3. ゲノム情報解析におけるデータ統合

現在，生命科学分野では多くの研究機関がデータベースを公開しており，主要なデータベースのデータは研究に不可欠な公共データとして維持されている．毎年データベースのリストを更新している Nucleic Acids Research Database Issue によると 2004 年に 227 だったデータベースは 2014 年には 1,557 まで増加しており[18]，研究者はこれらのデータベースから必要な情報を取得し，統合して利用している．そのため，各データベースの提供者が利用者の使いやすいインターフェイス（GUI や API）を開発する他，複数のデータベースの情報を手元のデータと統合するためのアノテーション・ツールや，ウェブ上で詳細な検索や簡易的な解析を可能にするデータ・ポータルが第三者によっても開発されている．

前項のデータ処理をシークエンシングによって得られたリード・データそれぞれに対して実行することで，遺伝子のリストやゲノム位置情報のリストといった解釈可能な処理結果（縮約データとも呼ばれる）が得られる．一方，こうして得られたデータの意味を理解し知見を見出すためには，多くの場合，データを既存のデータベースの情報と組み合わせて比較する必要がある．前出のゲノム情報解析それぞれで，次のようなデータの結合が考えられる（表 3）．一般的に，表などに格納されたデータセット同士をキーによって組み合わせる処理は結合（join）と呼ばれるが，ここでは実験で得られたデータのアノテーションや統計解析を目的として複数のデータセットを結合しているため，本論文ではこれらの処理をデータの統合（integration）と呼ぶこととする．

- 全ゲノム配列解析： データ処理で検出された多型がデータベースに登録されている既知の多型であるかどうかを検索する．また，検出された変異の中で遺伝子領域や制御領域にある変異を確認する他，影響を受ける遺伝子に関連性の強い機能やパスウェイを検索する．

- De Novo 配列解析： データ処理でアセンブルされた DNA 配列から遺伝子領域を推測し，データベースを用いて相同遺伝子を検索し，この遺伝子の機能をアノテーションする．

- メタゲノム解析： 既知生物種の遺伝子配列を用いて生物種を同定する他，De Novo 配列解析と同様，遺伝子領域を推測し，新規遺伝子の相同遺伝子を用いて検索された遺伝子の機能をアノテーションする．

- RNA-seq： 検出された RNA 配列を既知の転写産物のデータベースと比較し新規アイソフォームなどを探索する．また，有意な発現量の差がある遺伝子のリストを用いて，データベースからこれらの遺伝子に関連性の強い機能やパスウェイを検索する．

- ChIP-seq：転写因子の結合や修飾ヒストンの存在が推定された領域と既知の遺伝子領域や制御領域とを比較する．また，この領域の塩基配列モチーフを抽出し，データベースの既知の塩基配列モチーフを検索する．

- DNA メチル化解析：検出された DNA のメチル化領域と既知の遺伝子領域や制御領域，CpG アイランド領域とを比較する．

表 3： ゲノム情報解析におけるデータ統合

解析	データ統合するデータセットの例
全ゲノム配列解析	既知の多型，遺伝子に関連する機能やパスウェイ
De Novo 配列解析	相同遺伝子とその機能
メタゲノム解析	相同遺伝子とその機能
RNA-seq	既知の転写産物，遺伝子の機能やパスウェイ
ChIP-seq	既知の遺伝子領域や制御領域，塩基配列モチーフ
DNA メチル化解析	既知の遺伝子領域や制御領域，CpG アイランド領域

2.2. 問題と先行研究

2.2.1. データ処理における問題

前節でまとめたゲノム情報のデータ処理のため多くのソフトウェアが開発されている．これらのソフトウェアは，無償で使用できるものやオープン・ソースのものが多い一方で，バージョン更新が不定期なことや大きな動作変更を伴うことも多い．また，生命科学研究における仮説検証のために作成されたソフトウェアであるという性質上，同様の機能のソフトウェアがそれぞれの実験に最適化されて多数開発されることも珍しくない．さらに，ソフトウェア開発者が情報解析の研究者であるために新規性のあるアルゴリズムを積極的に取り入れるといった傾向もあるだろう．

ゲノム情報のデータ処理の再現性においては２つの問題がある．ひとつはソフトウェアやそのバージョンによって同じ解析を目的とした処理でも大きく結果が異なることがあることであり，もうひとつは同じバージョンの同じソフトウェアを使用するつもりでも設定方法や計算環境の違いにより意図せず異なる結果が得られてしまうことがあることである．

同じ解析を目的とした異なるソフトウェアで結果が異なり，標準となるソフトウェアやそのバージョン，入力パラメータを決定することが難しいという問題は多くのデータ処理で指摘されている．例えば，マッピングのためのソフトウェアの比較においても，ソフトウェアにより結果が異なるだけでなく，入力データの些細な違いによっても想定しない結果の違いが生じ，これらのソフトウェアを使用した多型検出ワークフローの最終結果にまで影響が及ぼされるといった報告がある[19]．そのため，ソフトウェアによる結果の違いがあることを考慮して，例えば，がん組織の変異を検出する場合には，同じ入力に対していくつかの異なる検出用ソフトウェアを使用した上でその積集合を正とする，といった手法が採用されることもある[20]．このような複雑なデータ処理方法を利用する場合には，このデータ処理を再現したり他の計算環境で再利用することが更に難しくなるため，これらのソフトウェアをパッケージ化する方法も提案されている[21]．

多くのパイプラインがオープン・ソースで公開されているものの，ソフトウェアの依存関係や環境依存の設定があるため，パイプラインを配布することが難しいことが指摘されている[22]．その一方，技術的には，同じソフトウェアを全く同様に設定することで異なる計算環境に同一のパイプラインを設置することは可能と考えられるため，データ解析が再現できない場合，ソフトウェアの問題かそのソフトウェアを設定した作業者のミスとされることが多い．しかし，実際には，解析の再現性や再利用性を向上させるインセンティブ以上に，計算環境による結果の違いを検証したりこれを修正したりすることに労力がかかってしまうことから，ここには情報基盤の改善余地があるとも考えられる．

これを技術的に解決するために，仮想マシンを利用すると共にこれを実行するクラウド環境の整備することで，システムそのもののスナップショットを共有できるようにする（whole system snapshot exchange）といった情報基盤が提案されている[23]．このような情報基盤があれば，データ解析の再現を簡単に実現できるようになるだろう．さらに，データ解析を再現するだけでなく再利用して継続的に改善していくには，入力ファイルやパラメータさらにソフトウェア自体の変更が可能な環境を配布する，ソフトウェアの使用例や仕様を共有する，といったことも重要であり，メソドロジーの共有方法を包括的に設計する必要がある[24].

今後は，大規模なコホートを対象にしたゲノム情報解析など，多数のサンプルに対して同様のデータ処理が必要になる他，より多くの研究機関や病院で次世代シークエンサーが使用されるため，データ処理を再利用することで低コストで効率的に解析できる情報基盤が必要になるだろう．また，大型プロジェクトでは，複数の研究機関で同じデータ処理方法を用いるべきか，データ処理の再現性の品質をどのように管理するか，といった問題が今まで以上に議論されるだろう．

2.2.2. データ処理における先行研究

ゲノム情報解析におけるデータ処理のワークフローを記述し，これを管理するためのシステムとしていくつかのワークフロー管理システムが開発されており，その中には既にクラウド計算環境上で使用されているものもある．既存のワークフロー管理システムについては「3.1.2 ワークフロー管理システムの利用」で紹介することとし，ここではワークフロー管理システムの一例としてmyExperiment における問題点を取り上げる．

myExperiment[25]はゲノム情報解析に限らずバインインフォマティクスに関するソフトウェアのワークフローを管理するための代表的なワークフロー管理システムの一つである．このシステムでは一定期間運用された後の調査で，レポジトリにある 92 のワークフローのうち 80%は，内容が理解できないものであるかデータの欠如や参照先システムの変更によって実行できないものであり，事実上「使えなくなった」ワークフローであった．これには以下のような原因があると分析されている[26].

- **不十分なドキュメント：** ワークフローの入出力や中間ステップについて十分な記載がないため，利用者はワークフローによって実装された解析を理解することができない.
- **例となるデータの欠落：** ワークフローによって実装された解析が理解できても，どのようなデータやパラメーターを入力することで正しくワークフローを実行できるかを把握することが難しい.
- **外部リソースの変化：** ワークフローが外部リソースに依存している場合（API 経由で外部のサービスを利用するなど），これらの外部リソースが利用できなくなった場合にワークフローも実行できなくなる.
- **実行環境：** ワークフローの実行には依存関係のあるソフトウェアが必要になる場合があり，このための環境構築が難しい.

これらは他のワークフロー管理システムを使用した場合にも同様に生じ得る問題である．本研究で取り上げるワークフロー管理システム Galaxy[27]においても，これらの問題は当てはまっており，特に Galaxy の公共サーバー上では匿名の登録ユーザーがワークフローを公開できることから，より多くの「使えなくなった」ワークフローが存在していると推測できる.

2.2.3. データ統合における問題

前節の通り，多くの公共データベースの情報を活用するため，シークエンシングとデータ処理によって得られた縮約データと公共データベースのデータとのデータ統合が必要であることがわかる．ここで，ゲノム情報解析におけるデータ統合として挙げた例を見てみると，一般的な統合方法として次の 2 つがあることがわかる．

- ゲノム位置情報による統合： ゲノム配列上のデータは同じリファレンス・ゲノム上の位置にマップすることで統合できる．例えば，ChIP-seq で検出されたピーク（タンパク質の結合位置）と遺伝子制御領域に関する既知情報はゲノム位置情報によりデータ統合されるといえる．

- 遺伝子や分子の ID による統合： 遺伝子やタンパク質分子にはデータベースによって管理された ID があるため，これらを用いて統合することができる．同じ遺伝子でも複数のデータベースでそれぞれ独立した ID が割り当てられているため，クロス・リファレンスによる「名寄せ」が必要な場合もある．

ゲノム位置情報を用いた統合を用いた代表的なアプリケーションにはゲノム・ブラウザや多型のアノテーション・ツールがある．例えば，UCSC Genome Browser[28]は位置情報を軸に複数のデータ（例えば，遺伝子と結合位置）を並べて表示するため，これらのデータを直接的に統合せずに，視覚的に近傍の情報を把握することができる．一方，Annovar[29]のような多型のアノテーション・ツールは検出された多型とデータベースに登録されている多型を比較して，アノテーションを付加するため，より直接的なデータである．ゲノム位置情報によるデータ統合は，特定のリファレンス・ゲノムを用いることで位置を一意に特定できるため，データ処理を含めた解析過程で同じバージョンのリファレンス・ゲノムを使うことが必須である．このような多型のアノテーション・ツールの場合，参照するトランスクリプト・データベースの違いやツールの用いるアルゴリズムの違いにより結果が大きく異なることが報告されている[30].

パスウェイ解析や機能解析の際には，遺伝子やタンパク質の分子の ID を用いてデータベースを検索する．例として，機能解析ツールの DAVID [31]の場合には，与えられた遺伝子名のリストに対して，これらの遺伝子と関連の強い機能やパスウェイを調べることができる．この際，機能は Gene Ontology データベース，パスウェイは KEGG データベースなどを参照している．遺伝子の ID は refSeq や Emsembl ID，など複数の異なるデータベースで定義されているため，DAVID は与えられた遺伝子名に対して ID を検索し，これと同じ遺伝子を複数のデータベースで検索して結果を生成している．このような ID のマッチングは各研究者が個別に作業するよりも，このようなウェブ・アプリケーションで実行できる方が効率的である．

さらに，参照先のデータセットがリファレンス・データだけでなく，今までに蓄積された多数のデータそのものである場合には，これらのデータの集積と比較が必要となる．例えば，ICGC Data Portal[32]は，今なお登録され続けている多数のがん症例のデータを蓄積しており，その中からある遺伝子に変異のある症例を検索したり集計したりすることができる．ゲノム情報のデータ生成量の増大によって未解析のデータが共有され，データ駆動型研究の可能性が期待される一方，これらの大規模データをダウンロードしたり手元で解析したりすることがデータ・サイズやリアルタイム性またはアクセス・コントロールの観点から難しくなるため，ウェブ上で利用可能なデータ・ポータルの開発が求められている．

ここで，アノテーション・ツールやデータ・ポータルのようなアプリケーションの開発者は，各データベースのスキーマを理解し，双方のデータベースに含まれる ID やゲノム位置情報をキーとしてデータベースを結合する．しかしながら，このようなアプリケーション開発者によるデータ統合には以下の二つの問題があると考えられる．

第一に，データ統合のプロセスが必ずしも再現できないことが挙げられる．例えば，異なるデータベース由来のデータセットを遺伝子 ID で結合するとき，結合に使用できる遺伝子 ID が数種類ある場合や，文字列値を変換したもの同士を結合するといった手続きが含まれる場合には，結合後のデータセットからそのプロセスを追うことができない．どの値がどのデータベース（例えば，Ensembl や RefSeq）で定義されたキー値であるかを見分けることにも分野の知識が必要であり，機械的なデータ処理の弊害になっている．

第二に，統合後のデータセットを容易に共有できないことが挙げられる．あるデータ・ポータルで使用している統合済みのデータを他のアプリケーションで使用したい場合には，その統合済みのデータをスキーマとともに公開している必要がある．しかし，この方法では，個々のアプリケーションに最適化したデータ・スキーマを他のアプリケーションの開発者が解釈することに労力がかかるだけでなく，統合するデータセットが増えていくほどスキーマが拡張して汎用性に乏しくなる．

2.2.4. データ統合における先行研究

複数のデータセットを統合するためのデータベースは一般的にデータ・ウェアハウスと呼ばれ，近年ではデータ分析の要求に応えるため多くの ICT システムの基幹システムに実装されている．ゲノム情報解析とそれを取り巻く生命科学情報のデータセットはより多様性が高く，大量で，情報の提供者と利用者が世界中に分散しているため，データ・ウェアハウスの構築が難しく，現在まで多くの先行研究が行われてきた．

その中でも，頻繁に利用されているものとして，UCSC Genome Browser Database，BioMart Central Portal[33]，BioGRID[34]，InterMine[35]などが挙げられる．UCSC Genome Browser Database は複数のデータのゲノム位置情報により可視化する他，リファレンス・データのメタ・データを表形式で保持しており，リファレンス･データの入手先として活用されている．BioMart Central Portal や BioGRID は既存の複数のデータベースの情報を統合して保持することで，一括したデータ検索や統合されたデータセットの取得を可能にする．さらに，InterMine はデータ・スキーマの柔軟な変更を可能にするデータベース構造と複雑な検索ができるウェブ・インターフェイスを持ったフレームワークであり，研究分野ごとに特化したデータベースの作成（YeastMine，FlyMine，TargetMine 等）に利用されてきた．

これらのアプリケーションはデータ統合を可能にするものの，ここで統合されたデータセットのデータ・スキーマはアプリケーションに依存しているため，アプリケーションが保守されなくなった場合には利用できず，アプリケーション横断的にデータ統合することも難しい．そこで，アプリケーションに依存しないデータ統合の方法としては，セマンティック・ウェブを用いたデータ統合が注目されてきた[36]．セマンティック・ウェブの場合には，あらゆるデータを RDF（Resource Description Framework）というフレームワークで表現し，そのデータ・スキーマはオントロジーを用いてアプリケーションと分離して記述することができる．

このため，アプリケーションとは独立に多くのオントロジーが設計されて BioPortal[37]や Ontology Lookup Service[38]に登録されている他，これらのオントロジーを参照する RDF データが公開されている[39]．さらに，セマンティック・ウェブを利用したアプリケーションとして，タンパク質データベースの UniProt[40]では差分データの追加時に RDF データで柔軟な入力を可能にしている他，TogoTable[41]はユーザーがアップロードした表データに対して RDF で統合された複数データベースの情報を用いてアノテーションを付与することができる．

2. 3.　課題設定

2.3.1.　問題を解決するための情報基盤

前節の通り，ゲノム情報解析のデータ解析とデータ統合にはそれぞれ再現性と再利用性の維持に問題点があることがわかる．これらの再現性と再利用性を向上させることにより，オープン・ソースの解析ツールと公共レポジトリのデータを用いて，論文などによって共有された知識を再現できるようにするという試みは，前章で取り上げたオープン・メソドロジーの考え方に合致するものである．この目的を達成するため以下の 3 点を情報基盤の課題として設定する．

- 再現性と再利用性の向上によるデータ解析手法の継続的な統合
- グローバルなデータ空間で永続的にデータを共有できる仕組み
- ツールやデータの提供者と利用者のコミュニティの構築

まず, データ解析手法について, 公共に入手可能なツールを活用するためには, データ解析手法とそのための解析ツールを継続的に改善できるサイクルが必要である．そのようなサイクルがない場合，同じ機能を持った多くの解析ツールがインターネット上に氾濫し，再利用されることのないまま古くなったツールが山積し, 実際に利用できるツールを探すことも難しくなるであろう. 例えば, 現時点では Wikipedia に掲載されている RNA-seq の解析ツールは 200 以上[42], Galaxy のツール・レポジトリである Galaxy Tool Shed[43]に登録されているツールは 3,500 以上もあり，ここから保守されているツールや自分の研究に最適なツールを見つけることは難しい.

ソフトウェア開発においては，ソフトウェアに変更を加える場合，これを変更前と同じデータを用いてテストすることでその変更が正しく実装され，システム全体に与える影響範囲が適切であることを確認できる．このようなテストを継続的に実施することで設計を改善しながら開発を進めていく手法を「継続的な統合（continuous integration）」と呼ぶ[44]．この考え方を広義に捉えることでソフトウェア開発だけでなくデータ駆動型研究にも応用できるだろう．つまり，仮説検証のために特定のデータ解析手法を開発した際には，この解析手法の再現性と再利用性を維持し，新しいデータに対して利用することで，継続的に解析手法そのものの改善を検討できると同時に，既存のテストで従来のデータ解析との結果の差異が確認できる．この場合，解析手法の変更（ツールやパラメーター，またはワークフローの変更）の影響範囲や以前の解析との整合性は前回と同じデータを用いてテストすることによって確認できるため，場当たり的な解析手法の変更を防ぐことができ，この変更の繰り返しによって標準となるデータ解析手法を構築することができる．

次に，データ共有の仕組みについて，前述の通り，次世代シークエンサーを使ったゲノム情報は多数の研究機関で並行して取得され，解析されている．そのため，データ解析手法の継続的な統合のためには，使用されたツールやデータが共有されるだけでなく，グローバルなデータ空間で一意に特定でき，テストに使用できる必要がある．つまり，ツールやデータは永続的で一意の ID を持ち，その ID を用いて特定のバージョンのツールや以前に解析したデータの解析前データを取得できなくてはならない．このようなウェブ上のデータの識別子としてはリンクト・データという概念が提唱されており[45]，生命科学データベースで採用が始まっている．

さらに，情報解析の再現性と再利用性の問題の解決のために，技術的な課題と同時に社会的な課題に取り組まなければならない[46]．つまり，再現性と再利用性の向上を支援するための情報基盤の完成度が一定レベル以上の水準となり，かつ，実際に研究者がこの情報基盤を使用して意識的に再現性のある情報解析を採用してはじめて，この問題は解決される．

そのため，開発者と利用者を含むコミュニティの構築も必要である．オープン・ソース・ソフトウェアの開発におけるコミュニティの役割のひとつは，各開発者に対してオープン・ソース・プロジェクトに対する貢献度に応じた評価を与えることである．このような業績主義的な評価システムが動機付けとなることによりオープン・ソース・プロジェクトが継続されることから，オープン・ソースの枠組みにとってコミュニティが必須であると考えられている[46]．これは，オープン・データのデータ提供者など，オープン・サイエンス全体についても同様のことがいえるだろう．

以上より，前節であげた再現性と再利用性の問題点を解決するためには，ゲノム情報解析の継続的な統合を可能とするプラットフォームおよび永続的に再利用可能なデータを保持するための情報基盤の構築が必要であると考えられる．さらに，このような基盤上で研究を遂行することが推奨され，開発者と利用者の多くが評価とフィードバックを得られるコミュニティが構築されることによって，再現性と再利用性が向上し，解析手法の継続的な統合とリンクト・データの形成が実現されるだろう．本研究では，その技術要素として，データ処理のためのワークフローの共有とデータ統合のためのリンクト・データの利用を検証している．

2.3.2. データ処理における課題設定

問題点として，データ処理のためのワークフローの共有が不十分で，ワークフローの実行結果も再現性に乏しい場合があることを指摘した．次世代シークエンス・データのデータ処理についてはいくつかのワークフロー管理システムが利用されているが，前節で紹介した先行研究の例でもわかる通り，このようなシステムの運用を通して再現性と再利用性を維持することも容易ではない．

そのため，本研究ではワークフロー管理システムのひとつを選択し，これを利用してデータ処理の再現性と再利用性を向上させる枠組みを提案し，これに必要とされる情報基盤を開発することを課題とする．研究対象のワークフロー管理システムとして，オープン・ソースでゲノム情報解析に適しており利用者と開発者の数が多い「Galaxy」を選択しているが，このシステムの詳細や，その他の主なワークフロー管理システムについては「3.1 課題」に記載している．

本研究で設定する課題は次の通りである．現時点では Galaxy のようなワークフロー管理システムの利用者は限られているため，まず賛同する開発者を集め，少数のワークフローを用いて検証を実施する．

- Galaxy の運用が継続できない場合にも，Galaxy 上のワークフローとその実行結果が再現できるような情報基盤を設計する
- Galaxy 上のワークフローを異なる研究機関で共有できるようにし，異なる計算環境上でも同じワークフローと同じ入力データを用いて同じ結果が得られるような情報基盤を設計する
- 上記のような情報基盤上で継続的にワークフローを開発し，ユーザーのフィードバックを得るためのコミュニティを構築する

2.3.3. データ統合における課題設定

問題点として，複数のデータベースのデータの統合はそれぞれのアプリケーション開発者によって実施されているため，データ統合のプロセスが再利用できず冗長な作業が発生するだけでなく，統合後のデータセットの再現性が乏しいことを指摘した．「4.1.3 RDF データの活用」で詳しく記述する通り，セマンティック・ウェブの技術要素である RDF を使用することでデータ統合のプロセスを再利用することが可能にできることが注目され，多くの生命科学データベースのデータが RDF データとして提供されている．その一方で，がんゲノム・データのポータルといった実用的なウェブ・アプリケーションはまだ RDF データのデータベース上に構築されていない．

そこで本研究では，がんゲノム・データベースのデータを用いて，表形式のデータを RDF データに変換する方法や，このデータを他の公共 RDF データと統合する方法，さらにこの RDF データのデータベース上にウェブ・アプリケーションを実装する方法を調査する．これによって，公共 RDF データを使用して，将来的に実用的なデータ統合とシステム開発が可能であることを示す.

- 既存データから RDF データを生成するための手法を確立するため，大量のゲノム情報を蓄積しているがんゲノム・データベースのデータから RDF データを生成する
- 再現性のあるデータ統合を検証するため，生成した RDF データと公共 RDF データを統合する
- 統合されたデータの再利用性を検証するため，統合された RDF データのデータベース上にウェブ・アプリケーションを実装する

3.　研究成果１（データ処理）

3.1.　課題

3.1.1.　ツールとバージョンの管理

2 章で分析した通り，ゲノム解析におけるデータ処理の再現性と再利用性を向上させるための解析環境が必要である．ここで再現性とは，次世代シークエンサーから出力された配列データから情報として有用な縮約データ（例えば，ある遺伝子領域の変異の有無）を取り出す処理を，時間を経ても繰り返し実行することができ，同じ結果が得られることを指している．また，再利用性とは，同一のデータ処理が他の研究所や医療施設においても実行可能であることを指している．これら再現性と再利用性を向上させるための方法が必要である．

例えば，データ処理において，意図していないソフトウェアのバージョンの違いが出力結果に大きく影響することがある．ある研究で実施されたデータ処理を異なる計算環境で再現する場合，当然ながら以前の環境で使用されたソフトウェアをバージョンまで考慮して導入する必要があるが，これは論文等に記載されていない場合も多い．さらに，それらが十分に記述されている場合にも，ダウンロード先のするソフトウェア自体が変更されている，環境構築を手作業で実施することによって手順の違いからソフトウェアの動作に何らかの違いが生じるといった可能性が考えられる（図 4）

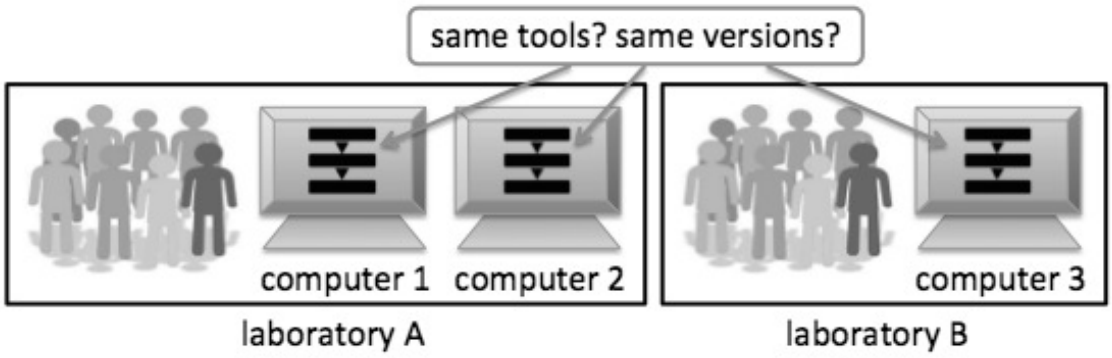

図 4： ツールとそのバージョンの差異

ツールとそのバーションの差異によってデータ処理の実行結果は異なるが，これらを異なるコンピュータ間または異なる研究施設間で揃える方法は提供されていない．

通常，ソフトウェアがそのハードウェアをサポートしている限り，このソフトウェアの出力結果はハードウェアの違いに依らず同じである．しかし，ハードウェアの運用上の問題で意図せず異なるバージョンのソフトウェアが使用されるといったことに注意が必要である．例えば，サーバー・クラスタにおいて，クラスタ・ノードの増設の際に新しいバージョンのソフトウェアがインストールされたため，このサーバー・クラスタに同じジョブを投入していても実行されるソフトウェアのバージョンがクラスタ・ノードによって異なるといった問題が起こり得る．再現性を向上させるためには，このような問題が発生しにくい設計が求められる（図 5）．

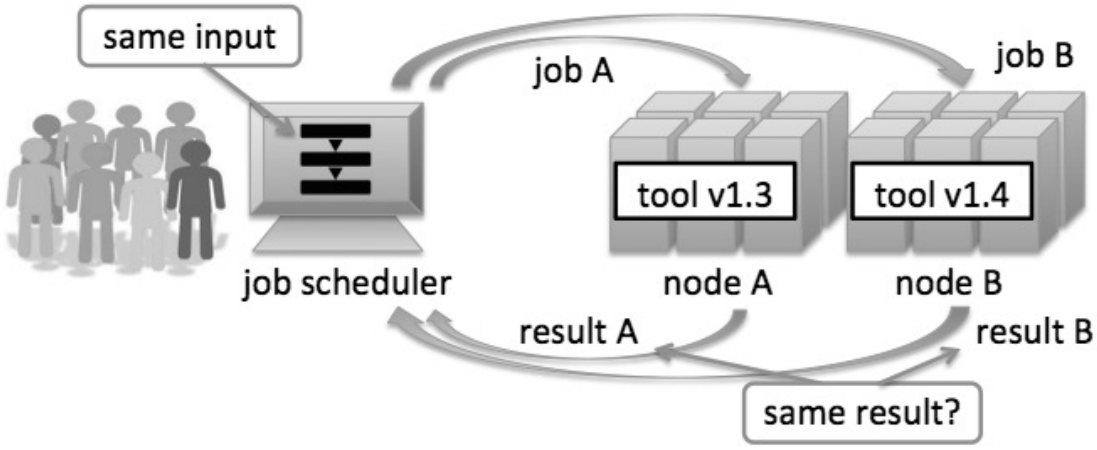

図 5： クラスタ環境におけるツールのバージョンの差異

クラスタ環境においては，同じコマンドで意図せず異なるバージョンのツールが実行されるといったことが生じ得る．このように，コマンドを保存しておくだけでは再現性が十分に担保できない場合もあり，実行環境を管理する方法が必要になる．

3.1.2. ワークフロー管理システムの利用

同じデータ処理を再現するためには，実行されたツールが同じものであるかどうかだけでなく，これらのツールが同じパラメーターを入力として同じ順序で実行されている必要がある．各ジョブがツールと入力ファイルさらにその実行時パラメーターで定義されるとき，ワークフローは複数のジョブの一連の流れ，つまり実行順序と入力データと出力データの関連を記述している（図 6）．ワークフローの出力結果を再現するためには，ワークフローを保存して再実行する仕組みが必要であり，さらに，ワークフローを再利用するためには，複数の利用者が同じワークフローを異なる環境でも実行できる仕組みが必要である．

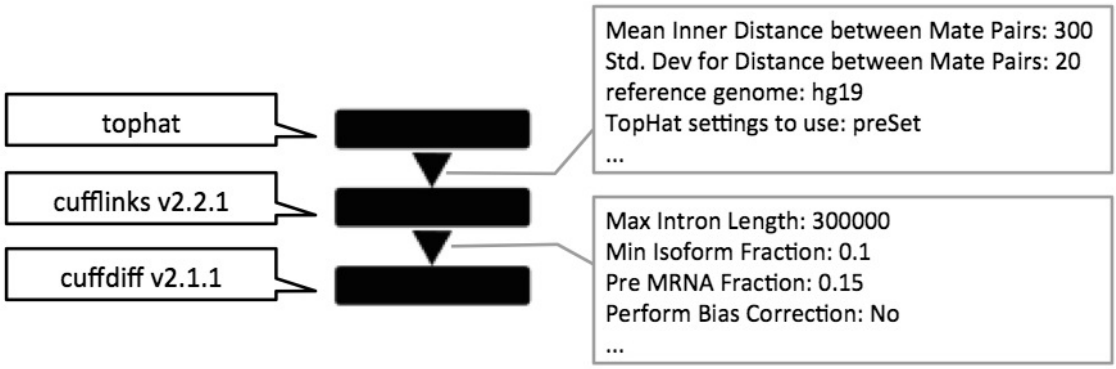

図 6：　ワークフローの構成

この図は RNA-seq のデータ処理で使用されるワークフローの例である．各ジョブはツールと入力ファイルさらにその実行時パラメータで定義され，ワークフローは複数のジョブの一連の流れを記述している．

ワークフロー管理システムは，ワークフローを保存することで再実行することを可能にし，また，ワークフローを異なる計算環境で共有することを可能にするもので（図 7），通常，多くのツールで構成された様々なワークフローに対して汎用的に利用できるように設計されている．

現在，いくつかのワークフロー管理システムが使用されているが，非商用かつオープン・ソースのシステムの代表的な例として，Knime[47]，Taverna[48]，Galaxy[49][50]といったウェブ・システムが挙げられる．それぞれに得意とする研究分野があるが，特に次世代シーケンシング・データの解析には Galaxy が用いられてきた[51]．また，商用のシステムとしては KDE (Inforsense)，Pipeline Plot (Accelrys)，VIBE (Incogen) などが販売されている[52]．また，2011 年に次世代シークエンサーの開発元である Illumina 社が Amazon Web Service のクラウドを利用した従量課金の解析環境 BaseSpace の提供を始めたが，このプラットフォームにもワークフロー管理機能が追加され，商業的にも注目されていることがわかる．

上記のワークフロー管理システムは GUI を備えることで，通常はコマンド（CUI）で実行されるツールをコマンド入力なしで操作できるように設計されている．システムの潜在的な利用者の多くがコマンド入力に慣れない実験研究者であることから，システムの普及の促進という観点からも GUI を備えたシステムの開発は必要であると考えられる．ただし，再現性と再利用性の向上のためにデータ処理内容を記録するという目的のためには，必ずしも GUI を使用する必要はなく，IPython Notebook, RStudio, knitr のようなソフトウェアを用いてスクリプトで書かれた処理を記録することも可能であり，これらのソフトウェアはより柔軟なカスタマイズが頻繁に必要な処理の管理に適している．

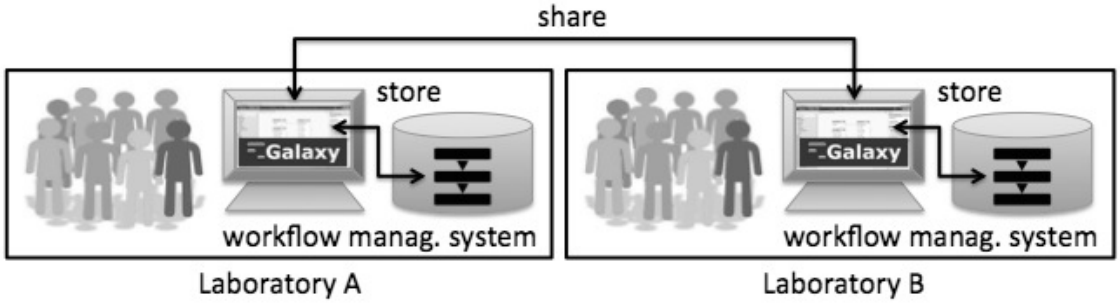

図 7： ワークフロー管理システム

ワークフロー管理システムは，ワークフローを保存することで再実行することを可能にし，また，ワークフローを異なる計算環境で共有することを可能にする．

3.1.3. Galaxy の概要と課題

Galaxy はゲノム情報のデータ解析ために無償で利用できるワークフロー管理システムである．公開されているツールを組み合わせてワークフローを構築し，それらのワークフローの再実行や結果の簡単に共有できるため，データ処理の再現性と再利用性を向上させるためのウェブ・システムとして広く利用されてきた実績がある．本研究で Galaxy を採用した理由として以下の 3 点がある．

- **オープン・ソース：** Galaxy はオープン・ソース・ライセンスと呼ばれるライセンスの一つである Academic Free License (version 3.0)で提供されており，このライセンスの下ではソース・コードが共有され使用者がこの派生物を作成することが許可されている．再現性と再利用性の観点から, オープン・ソースであることは必須条件といえる. 再現性の観点から，このプラットフォームが無期限で使用できる必要があり，そのためにはプラットフォームの提供が終了しても改変して使用し続けられるライセンスである必要がある．また，データ処理に問題が生じた際に問題を特定するために，ソース・コードが公開されていることが望ましい．さらに，再利用性の観点から，このプラットフォームが営利・非営利を問わず使用できる必要があり，さらに無償で提供されることで利用を促進することができる．

- **ゲノム情報解析との適性：** 現在，生命科学研究においていくつかのワークフロー管理システムが使用されているが，Galaxy は次世代シークエンサー・データを中心としたゲノム情報の解析に焦点を置いて開発されており，またそのように利用されている．ワークフロー管理システムは必ずしも大容量のデータを扱うように設計されているわけではない．例えば，Taverna の場合には複数のツールがウェブ・サービスを介して異なるシステム間でデータを転送するが，Galaxy の場合は各ジョブの入出力は基本的にはシステム内に閉じている．ゲノム情報は大容量でありデータ処理の計算コストも大きいため，これらのデータ処理用のツールはウェブ・アプリケーションではなくソフトウェアとして配布されている事が多い．そのことから，現時点では Galaxy のような設計がゲノム情報解析に適していると考えられ，利用されている．

- **利用者数と開発者数：** 現在，Galaxy は多数の利用者に利用され，多数の開発者により Galaxy 上で動作するツールが提供されている．Galaxy プロジェクトの Wiki によれば [53]，Galaxy プロジェクトの運用している公共サーバーでは月間およそ 20 万件近くのジョブが処理されており，2015 年の 7 月の時点でこれ以外に 73 の公共サーバーが異なる研究機関で公開されている．ツールのレポジトリである「Galaxy Tool Shed」に 3,000 以上のツールが登録されており，この中には使用されないツールも多いと考えられるが，一方で Galaxy プロジェクト自体は 2014 年の一年間には 700 以上の論文に引用されており，一部のツールは確実に活用されているといえる．また，利用者および開発者のための国際会議「Galaxy Community Conference」には 2012 年から 2015 年の 4 年間は毎年 200 名以上が参加しており，ワークフロー管理システムおよびそのツール開発のコミュニティとして現時点で非常に活動的である．

再現性と再利用性の向上という観点から，情報解析研究者に留まらず多くの利用者が共通のインターフェイスを使えることは重要であり，その点で Galaxy は既に海外を中心に開発者や利用者が多いという優位点がある．一方で，Galaxy の利用には次の 2 つの課題があると考えられる．

- Galaxy 環境の再構築が必要となるとき，全てのツールのバージョンを以前と揃えて構築することが難しいため，事実上，過去に実行されたワークフローの再現性が失われてしまう．
- 各研究機関が Galaxy 上で利用しているツールやワークフローの情報が十分に共有されていない，または，公開はされていても情報が不十分のため結果的にあまり共有されていない．

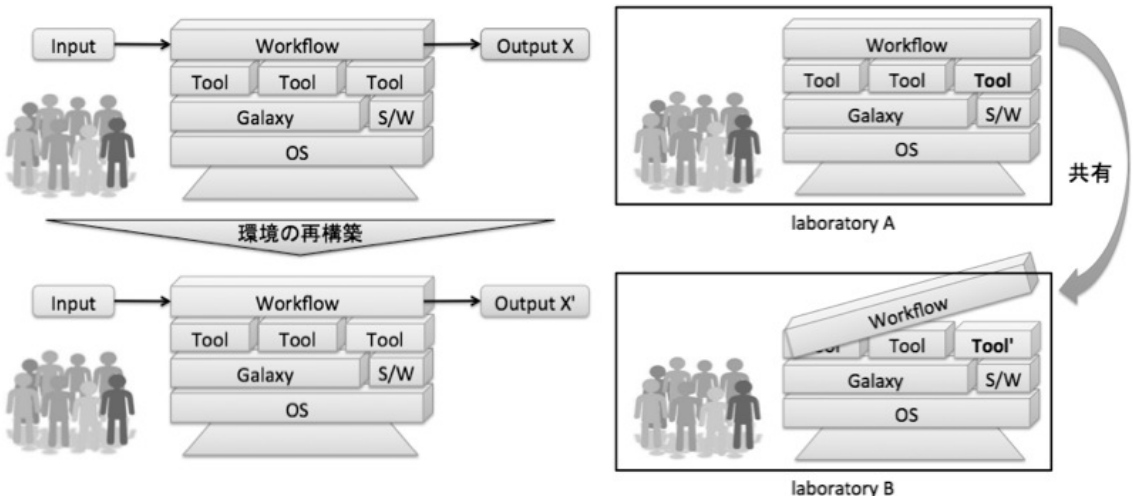

図 8： Galaxy システムの課題

Galaxy と同じ OS 上に異なるバージョンのツールをインストールした際に，同じワークフローを実行していても異なる結果が出力され，結果的にワークフローの再現性が失われる場合がある（左図）．また，異なる計算環境でインストールされていたツールが異なるために，共有されたワークフローが実行できず，ワークフローの再利用ができない場合がある（右図）．

3.1.4. 連携されたシステムの提供

Galaxy はゲノム情報解析のプラットフォームとして広く用いられてきた経緯から，ウェブ・システムとしての解析環境の作成において必須のコンポーネントといえるが，Galaxy のみで全てのゲノム情報解析を網羅できるわけでなく，解析環境に Galaxy 以外のシステムや Galaxy の不得意な機能を補完するソフトウェアも加えることで，用途を広げることができるだろう．例えば，実験研究室向けの情報管理システム（LIMS: Laboratory Information Management System）を組み合わせることで，シークエンサーから出力されたデータの管理を一元化でき，解析しているデータからサンプルの情報を遡って取得することができるだろう．また，Python や R といったプログラミング言語でインタラクティブにデータ解析ができるウェブ・アプリケーションである iPython Notebook や Rstudio と連携させることで，一度限りプログラマブルに解析したログについても保存しておくことができる．

3. 2.　手法

3.2.1.　同一の実行環境を共有する方法

前節の通り，ワークフロー管理システムによりワークフローを管理しても，これらのワークフローの実行環境が異なることで再現や再利用が実現していないといった課題がある．そこで，複数の利用者が同一の実行環境を使用する方法を検討する．まず，同一の実行環境を共有する方法として，プラットフォームをサービスとして提供する方法と配布する方法がある．

サービスとして提供する場合には，利用者管理機能を持った単一の環境を公共サーバーとして公開する方法と，利用者ごとに異なる環境をクラウド環境上で起動できるようにする方法がある（表 4）．どちらの場合も，基盤の構築と運用に多くの開発コストと計算リソースが必要となる．

Galaxy の場合，単一の環境を公共サーバーとして公開している例として，Galaxy プロジェクトの公開しているサーバーが挙げられる．このサーバーは現在 5 万人の利用者が登録されており，月に平均 20 万件程度のジョブを処理している[53]．これらのジョブを複数のサーバー・センターに分散することにより，最大 40 万コア以上の CPU で処理されている[54]．このような環境の場合，利用者のメリットとして利用者間でデータやワークフローを共有することが可能である一方で，不要なデータの整理や可用性の担保などの運用コストが高い．

また，利用者ごとに異なる環境をクラウド環境上に起動する例として，Genomics Virtual Laboratory プロジェクトが挙げられる．利用者は大中小の大きさのクラウド環境を選択して，利用者ごとに別々の環境を起動することができる．このような環境の場合，利用者のメリットとしてそれぞれの利用者がツールの追加など環境をカスタマイズすることができ，運用者のメリットとして不要なデータを環境ごと削除するといった運用が可能になる．クラウド上の仮想化やアプリケーション・コンテナの技術の進歩とクラウド環境の低価格化により，今後はクラウド環境の利用が進むと考えられる．

表 4： 同一の実行環境を共有する方法（サービスの提供）

方法	利用者間のデータの共有	環境のカスタマイズ	その他
公共サーバーの提供	可	不可	仮想化によるオーバーヘッドがない
クラウド環境の提供	不可	可	使用されていない環境の削除が容易

各々の利用者が計算環境を用意してここに実行環境を配布する場合には，単純に構築手順を共有する方法，この手順をコード化することで共有を容易にする方法，環境構築済みの仮想マシンやアプリケーション・コンテナを配布する方法がある（表 5）.

まず，インストール手順の共有やコード化は，仮想化による性能劣化がなく，多くの計算環境で利用可能な方法であり，また仮想化ソフトウェアに対するサポートを必要としないといったメリットがある．その一方，インストール手順が同じでも同一の実行環境にならない可能性があるため，同一の実行環境が構築されているかどうかを，ワークフローの実行結果を基にテストするような方法が必要となるだろう.

仮想マシンを使って実行環境を配布する方法は，構築済みの環境をそのまま配布することができるため，利用者の構築作業を必要としない．利用者は仮想化ソフトウェアを用いて任意の OS 上で同一の実行環境を手軽に起動することができる．デメリットとして，仮想マシン上の計算はエミュレートによるオーバーヘッドにより性能劣化を伴うことが挙げられる．特に性能を重視するゲノム情報解析においては，この性能劣化が許容できない場合が考えられる.

コンテナの技術を使って実行環境を配布する方法は，仮想マシンの配布と同様インストール作業が不要のため同一の実行環境が構築でき，仮想マシンと異なり仮想化による性能劣化もない．現時点では，仮想マシンの利用の方が一般的であり多くの利用者に受け入れられやすいと考えられるが，今後の普及が期待される.

- **仮想マシンとコンテナ：** どちらも構築済みの環境のイメージを展開することができるが，仮想マシンではハイパーバイザまたはホスト OS のリソースの仮想化のオーバーヘッドがあるが，コンテナでは OS 上のリソース管理をするのみなのでオーバーヘッドがほとんどなくディスク使用量も少ないという利点がある．仮想化ソフトウェアの VMWare や VirtualBox は既に普及しているが，コンテナ管理ソフトウェアでは 2013 年に登場した Docker が急速に広まっている．

表 5： 同一の実行環境を共有する方法（実行環境の配布）

方法	仮想化による性能劣化	利用者の構築作業	その他の問題点
構築手順の共有	なし	要	手順を再現することが難しい
構築手順のコード化	なし	要	手順をコード化することが難しい
仮想マシンの配布	あり	不要	特になし
コンテナの配布	なし	不要	多少のコマンド操作が必要

本研究では，多くの計算環境で利用でき，最も簡単に同一の実行環境を構築できる仮想マシンを利用することとする（図 9）．その一方，他の方法の利用についてもそれぞれメリットがあるため，今後検討していくことが望ましい．

仮想マシンを定期的に更新することで新たなワークフローの追加が可能だが，その一方で，全てのバージョンの仮想マシンは一定期間保存されている必要がある．仮想マシンを更新した際には，更新前のツールやワークフローが動作しなくなる可能性もあるが，その場合も以前のバージョンの仮想マシンを使用することで，以前のデータ処理が再現可能である．今回は，多くの計算環境で導入できるように，無償で入手可能で，Windows，Macintosh，Linux 上で動作する仮想化ソフトウェアとして，VirtualBox（無償）および VMWare Player（非営利の場合のみ無償）の双方に対応する構成とした．

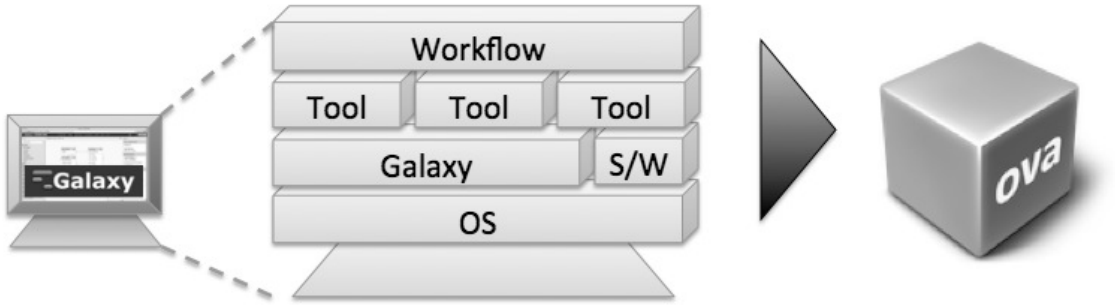

図 9： 仮想マシンの開発

仮想マシンを使用することで，OS，OS パッケージ，Galaxy，Galaxy 上のツールの全てをパッケージ化し，これらが同一の実行環境を異なる計算機上に複製できる．

3.2.2. コミュニティ仮想マシンの開発

このような解析環境のパッケージ化は以前から取り組まれており， Galaxy が含まれた仮想マシンとしては，Galaxy プロジェクトが公開している仮想マシンのリストがある他[55]，国内でもライフサイエンス統合データベースセンターの「DBCLS Galaxy パッケージ」のように入手可能な仮想マシンおよび Amazon Machine Image があった[56]．Galaxy 以外のシステムでも，医薬品化合物データベースの ChEMBL の myChEMBL[57]のようにデータベースとそれを検索するウェブ・アプリケーションを設定済みの仮想マシンを配布している例がある．

ゲノム情報解析におけるオープン・ソースのツールの配布の方法としては既に Bioconductor[58]や Galaxy Tool Shed といったレポジトリが広く使用されているが，仮想マシンは解析ツールに加えてデータやワークフローを含めることができることから，仮想マシンの配布は「1.1.3 ゲノム情報解析とオープン・サイエンス」におけるオープン・メソドロジーを目的としたものであるといえる．

しかしながら，これらの仮想マシンは，ある開発元が単独で提供しているものであり，含まれているツールやワークフローは限られており，また，また開発元が提供を止めると，利用者は以前の仮想マシンを使用して解析結果を再現することができなくなるといった問題がある．そのため，ツールやワークフロー

を提供する開発者とそれを活用する利用者のコミュニティによってこの仮想マシンを運用することを提案している．これは，Linux OS や Apache プロジェクトをはじめ多くのオープン・ソース・ソフトウェアがディストリビューターと呼ばれる団体によってパッケージ化され，有償または無償によって配布されたり，サポート・サービスが提供されたりすることによって信頼性を向上させていることと同様の仕組みである[59][60].

この仮想マシン上にワークフローと必要なツールをインストールして配布することで，ワークフローの共有が可能になる．このため，ワークフローの共有のための環境を必要とする複数の開発者からワークフローを収集し，ひとつの仮想マシンにインストールしたものを「コミュニティ仮想マシン」として共同で開発した.

今までひとつの研究機関の解析環境にのみインストールされていたワークフローも，このコミュニティ仮想マシンにインストールすることにより，他の研究機関に配布されてすぐに使える状態になる．このワークフローを編集して利用するなどの再利用も容易である.

3.2.3. 開発者会議の定期的な開催

コミュニティ仮想マシンの開発のため，月次の開発者会議を開催する．この会議では主にツール開発者とデータ処理基盤の開発者が情報交換し，コミュニティ仮想マシンの仕様と搭載すべきツールやワークフローを検討する（図 10）．この取り組みの継続的な実施により，次のような効果が得られると期待できる.

第一に，コミュニティ仮想マシンの継続運用が可能になる．単独の研究者やそのプロジェクトによって運用されている場合，その研究者の都合によって運用が中止される可能性が高くなる．複数の研究者によって運用されることにより，利用者はコミュニティ仮想マシンが継続的に運用されることを期待できる.

第二に，情報共有による生産性の向上が可能になる．今までそれぞれの研究機関で別々に作成していた同様のツールについて情報共有することにより，冗長な作業を削減し，次に必要なツールを共同開発することができる．ゲノム情報解析の研究者や開発者が不足している現状において，このような情報共有の場の提供が必要と考えられる．

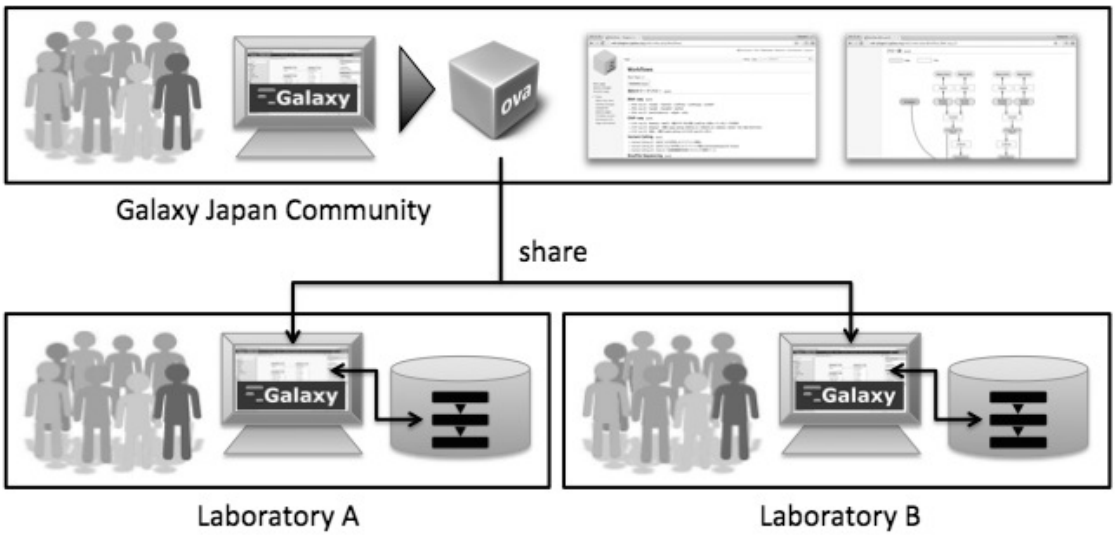

図 10： 開発者会議の開催

開発者会議を定期的に開催することで，複数の研究機関で開発されたツールやワークフローを継続的に追加しながらコミュニティ仮想マシンを更新することができる．

3.2.4. 利用者への成果物の提供方法

ここでは，開発したコミュニティ仮想マシンを利用者に提供する有効な方法を検討する．「3.2.1. 同一の実行環境を共有する方法」で議論した通り，実行環境を利用者に提供する際には，サービスとして提供するか，実行環境そのものを配布するかといった二通りが考えられる．また，利用者がこの実行環境を使用するためには，実行環境の使い方だけでなく，ワークフローそのものの解説，つまり，ワークフローがどのようなデータに適用でき，それぞれのジョブがどのような処理をしているかといった説明が必要不可欠である．

先に紹介した，Genomics Virtual Laboratory [61] プロジェクトでは利用者へのリソースと知識の提供を「USE, GET, LEARN」の 3 つに分類している．USE は Galaxy 公共サーバーのサービスとしての提供であり，GET は利用者ごとにクラウド環境を用意することで実行環境を配布しており, LEARN は各ワークフローの開発者が作成したチュートリアルをウェブサイト上にまとめており，これらのチュートリアルは大学の講義でも利用されている．（表 6）．本研究の取り組みでも, Genomics Virtual Laboratory プロジェクトの「USE, GET, LEARN」のモデルに則って，リソースと知識を利用者に提供することとする．

まず，USE として，ワークフローをテストすることのみを目的とした小規模な公共サーバーを運用することとする．これは，開発者会議をベースとした有志のコミュニティでは大規模な公共サーバーを運用する計算リソースおよび人的リソースを維持することができないためである．一方で，より大規模な計算が可能な公共サーバーを公開するため，計算リソースを持つ研究機関との共同研究を推進する．

次に，GET として，上述のコミュニティ仮想マシンを配布する．Genomeics Virtual Laboratory プロジェクトでは利用者にクラウド環境を提供しているが, 本研究の取り組みでは, 計算リソースは提供しないため, 利用者が手元の PC やサーバーを使用することとなる. また, 後述の通り, クラウド環境である Amazon Web Service を使用して環境を起動することもできる．

最後に，LEARN として，各ワークフローの解説，使い方，テスト・データといった文書をワークフローの開発者が記載できるウェブサイトを作成する．利用者はこのサイトを参照することで，コミュニティ仮想マシンで利用できるワークフローを確認するだけでなく，各ツールの動作など詳細な情報を得ることができる．

表 6： 利用者への提供

	Genomic Virtual Lab	今回の取り組み
USE	プロジェクトに参加している研究機関が運用している公共サーバー	試験用途にのみ使用できる小規模な公共サーバー．より大規模な計算が可能な公共サーバーを公開するために共同研究を計画中
GET	プロジェクトの管理するクラウド上に利用者ごとに設定済みの環境を起動できる	仮想マシンおよび Amazon Machine Image．利用者は手元の計算環境かまたは Amazon Web Service を使ってこの環境を起動する
LEARN	各ワークフローについて解説された文書を収集したウェブサイト	各ワークフローについて解説された文書を収集したウェブサイト（Wiki を利用）

3.2.5. パブリック・クラウドの利用

大量のゲノム情報のデータ処理とデータ共有のため，伸縮性のある計算環境であるクラウド計算環境の利用が注目されている．しかしながら，サイズの大きいデータをクラウド計算環境に転送する時間的コストや，コンプライアンス上の制約から研究機関の外の計算機で扱うことが許されていないデータが多いといった問題があることから，ゲノム情報のデータ処理におけるクラウド計算環境の利用は現時点では限定的である．

このため，コミュニティ仮想マシンは基本的にはデータが保管されている研究機関内の計算機で使用されることを想定し，仮想マシンとして配布している．その一方で，利便性の向上のため，パブリック・クラウドである Amazon Web Service（AWS）でも同じ解析環境を使用可能にした．仮想マシンのイメージを AWS 用に変換して設定を加えた上で一般公開しており，AWS にアカウントを持つ利用者はこのイメージを数分で起動して，コミュニティ仮想マシンと同一の解析環境を利用することができる．

パブリック・クラウドを使うことにより，十分な性能の計算環境を持たない利用者がワークフローの動作を試す，または複数の利用者がワークフローやデータを共有する，といったことが容易に実現できる．

3.2.6. サーバー・クラスタの利用

サーバー・クラスタを使用した処理の分散には二つの方法がある．一方は，多数のリクエストを処理するため，多数のデータ処理を複数のノードで並列に実行する場合．もう一方は，一つの処理を高速化するため，一つのデータ処理を複数ノードで分散処理する場合である．

現在，Galaxy には後者のような一つのデータ処理を分散処理するといったツールがないため，前者の使い方が一般的である．一方，Galaxy にはワークフロー上で複数データ（データセット）を入力や出力とする機能が既に実装されており，大容量の配列データを分割して複数ノードによる分散処理を可能にするといったツールの開発が取り組まれている．この方法は大容量の配列データを扱う際には不可欠である．

そこで，コミュニティ仮想マシンをサーバー・クラスタ上で使用する構成についても検討し，設定方法を作成した（図 11）．この構成では，クラスタ管理ノードとして仮想マシンを用いているためデータ処理の再現性は保たれるが，ジョブはクラスタ・ノード上で実行される．仮想マシン上の処理はオーバーヘッドがあるのに対して，クラスタ・ノードに投入されたジョブは実マシン上で実行されるためオーバーヘッドがないといった特長がある．

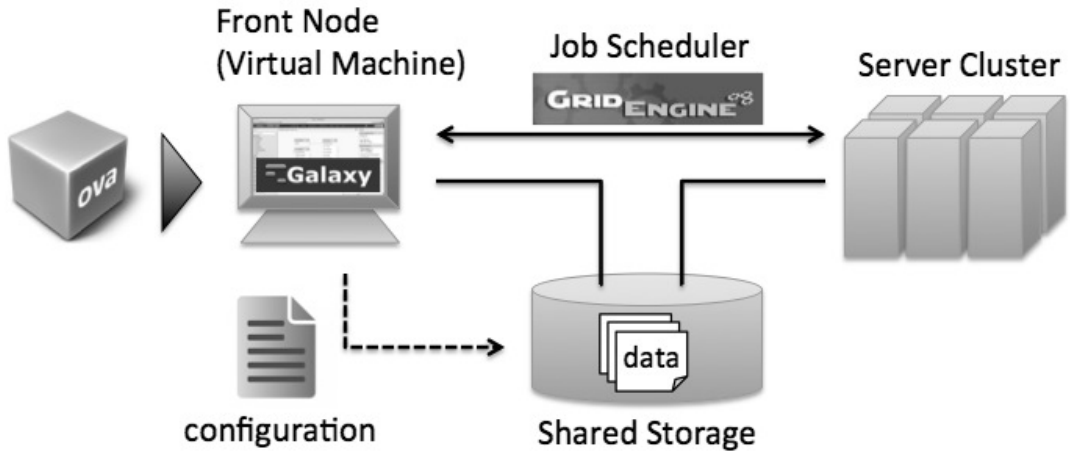

図 11： サーバー・クラスタの利用

サーバー・クラスタ構成で仮想マシンを利用する方法として，共有ストレージ上にデータを格納し，ジョブ・スケジューラーでクラスタ・ノードにジョブを投入する構成を検証した．この設定はウェブサイトで公開している．

3.2.7. 構築手順のコード化

計算環境の構築を人手で実施すると，手間がかかる上に誤った設定をしてしまう場合が多く，さらに近年は多数の計算機を利用するシステムが増えたため，計算環境の構築を自動化する手法が多く開発されている．コミュニティ仮想マシンの構築においても，以下の理由から，構築手順をコード化している．

- **構築手順の再現性**：コミュニティ仮想マシンはその上で実行されるデータ処理の再現性を向上させることが目的であり，仮想マシン自体の構築も再現性が高いことが望ましい．仮想マシンのイメージが保管されることで既に解析環境の再現は可能だが，それに加えて仮想マシンの構築手順がコード化されることで，作業者に依らず同じコミュニティ仮想マシンを再構築できる．

- **構築手順の公開**：コミュニティ仮想マシンの配布を利用者が使用する際，仮想マシン上の OS やソフトウェア，それらの設定が公開されている方が望ましい．例えば，利用者が計算環境の設定変更が必要と考えたときに，仮想マシン構築時の設定がわかる必要がある．構築手順を文書として公開することも可能だが，コード化されることでより明確になる．

- **構築作業の自動化**：コミュニティ仮想マシンは定期的に更新されるものであり，OS，必須パッケージ，Galaxy 等のバージョンを更新する際には，その度に仮想マシンの再構築が必要になる．しかしながら，これは手間のかかる作業であるため，仮想マシンの構築を自動化することで，より早いサイクルで仮想マシンを作成することが可能になる．

- そこで，Galaxy 上のツールのインストールのため，Galaxy Project が提供しているツール・レポジトリーである Galaxy Tool Shed と，Galaxy の API を使用し，手順をシェル・スクリプトのコードに記述した．開発したツールを Galaxy Tool Shed に登録することで，ツールの本体のインストールと Galaxy への組み込みが併せて実行できる．これらのコードは全て GitHub（コード・レポジトリー・サービス）にて公開しており，利用者はこのコードを用いてコミュニティ仮想マシンを自分自身で構築することもできる（図 12）．

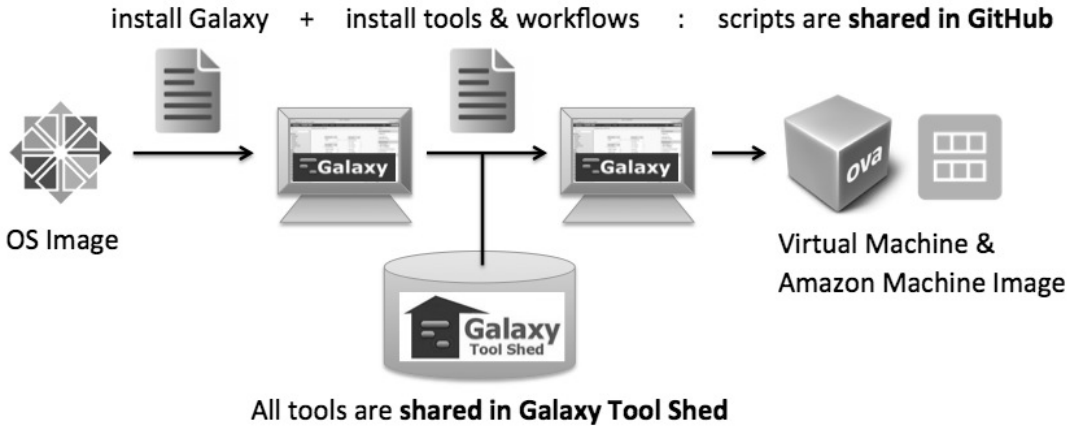

図 12： 構築手順のコード化

Galaxy の提供する Galaxy Tool Shed に必要なツールを格納し，API 経由でツールの本体と Galaxy への組み込みを実行する．ツールのインストールを含めた構築手順をコード化することで，仮想マシン上のツールやそのバージョンを継続的に管理できる．

3. 3.　結果

3.3.1.　コミュニティ仮想マシンの公開

コミュニティ仮想マシンを作成するため，賛同する開発者を募り，2014 年 11 月から月次の開発者会議を開催した．異なる研究機関から毎月 7〜10 人程度の参加者があり，この会議は現在まで継続している（図 13）（図 14）.

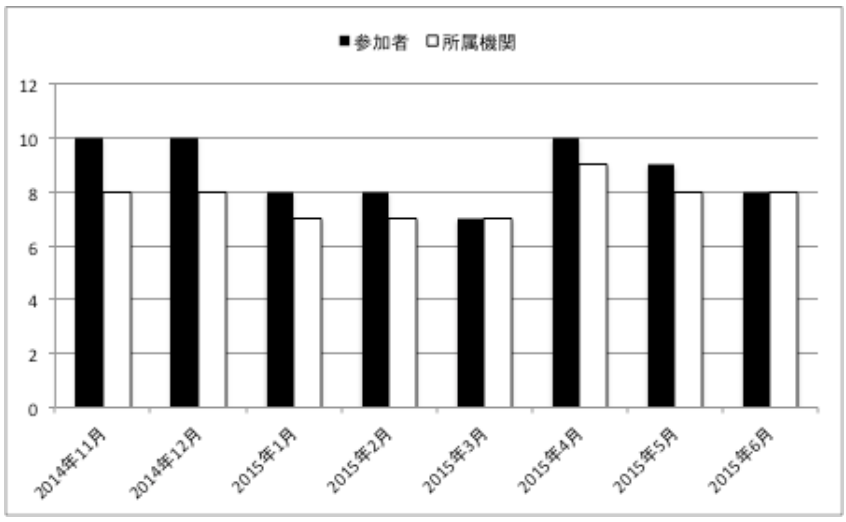

図 13：　開発者会議の参加人数

開発者会議の参加者の参加人数と所属機関の数．それぞれ異なる研究機関から 7〜10 人の開発者が参加している．

図 14： 開発者会議の様子

開発者会議「Galaxy Meetup」は 2014 年 11 月より月に一度，終日開催している．

このコミュニティ仮想マシンは，開発者会議で新しいワークフローが追加された際に更新している．現在，この仮想マシンには 8 つの異なる研究機関で作成された 10 のワークフロー, それらに必要なツールがインストールされている（表 7）．ツールのバージョンやリビジョンは全て管理シートに記載されウェブサイト上に公開されている．さらに利用者は以下の 3 つの方法でこれを使用することができる．

- 仮想マシンをダウンロードして自分の PC やサーバーで使用する
- Amazon Web Service でクラウド・インスタンスを起動する
- 公開されているテスト用サイトにアクセスする

表 7： ワークフローとツールの数

項目	値
ワークフローの数	10
ツールの数（レポジトリ単位）	22
うち独自に開発したツール	14
ワークフローを提供した開発者	8

コミュニティ仮想マシンにインストールされ，共有されているワークフローは以下の表のとおりである（表 8）．ChIP-seq, RNA-seq, Bisulfite-seq, Variant Calling と異なる実験系に対するワークフローが収集されており，これらのワークフローについてウェブサイト上で詳しく解説されている．

表 8： ワークフローの一覧

項番	実験系	詳細
1	ChIP-seq	結合ピーク領域と遺伝子のコーディング領域や転写開始点領域との重なりの抽出
2	ChIP-seq	複数のピーク検出アルゴリズムを使用したピーク検出（bowtie2 によるマッピング）
3	ChIP-seq	複数のピーク検出アルゴリズムを使用したピーク検出（bwa によるマッピング）
4	RNA-seq	発現量の異なる遺伝子の検出
5	RNA-seq	Sailfish を使用したアイソフォームの定量
6	RNA-seq	エンリッチメント解析，ヒートマップやタイムコース・プロットの作図
7	BS-seq	Bisulfighter を使用したメチル化領域の検出
8	Variant calling	GATK を使用したバリアントの検出
9	Variant calling	異なるマッピング・ツールを使用した際の，がんの変異検出結果の比較
10	Variant calling	HLA（ヒト主要組織適合抗原）のタイピング

3.3.2. 利用者への成果物の提供

ここでは，「3.2.4. 利用者への成果物の提供方法」の「USE, GET, LEARN」それぞれのカテゴリで成果物を利用者に提供している．まず，「3.3.1. コミュニティ仮想マシンの公開」に記載した通り，コミュニティ仮想マシンの配布とその解説の公開は「GET」および「LEARN」に該当している．

さらに積極的な成果物の提供として，2015 年 4 月にハンズオンを含むワークショップ「Galaxy Workshop Tokyo 2015」を開催し，およそ 90 人の参加者を対象にハンズオンやワークフローの紹介を実施した（図 15）．海外では，Galaxy プロジェクトの主催のカンファレンスの他，Galaxy をテーマとしたワークショップが数多く開催されているが，国内では今回が初めてのワークショップである．

このワークショップで，まず，ハンズオン・セッションではコミュニティ仮想マシンのインストール方法や Galaxy の基本的な使い方を紹介しており，これは「GET」の提供を意図している．また，ワークフローの紹介のセッションでは，それぞれのワークフローの内容を紹介しており，これは「LEARN」の提供を意図している．

実施したアンケートでは，参加者のおよそ半数が実験をしてデータを出す研究者，それ以外がデータ解析の研究者や計算機の販売に携わる企業の従業員であった．現在は Galaxy を使用していない参加者が多かった一方，仮想マシンの使用を検討したい，ワークショップが役に立つ，といった回答が多く，このような解析環境のニーズがあると考えられる（図 16）．

図 15：　Galaxy Workshop Tokyo 2015 の様子

ワークショップにはおよそ 90 人以上が参加し，午前中はコミュニティ仮想マシンのインストールのためのハンズオン，午後は各ワークフローの開発者による解析内容の説明のセッションが開催された．

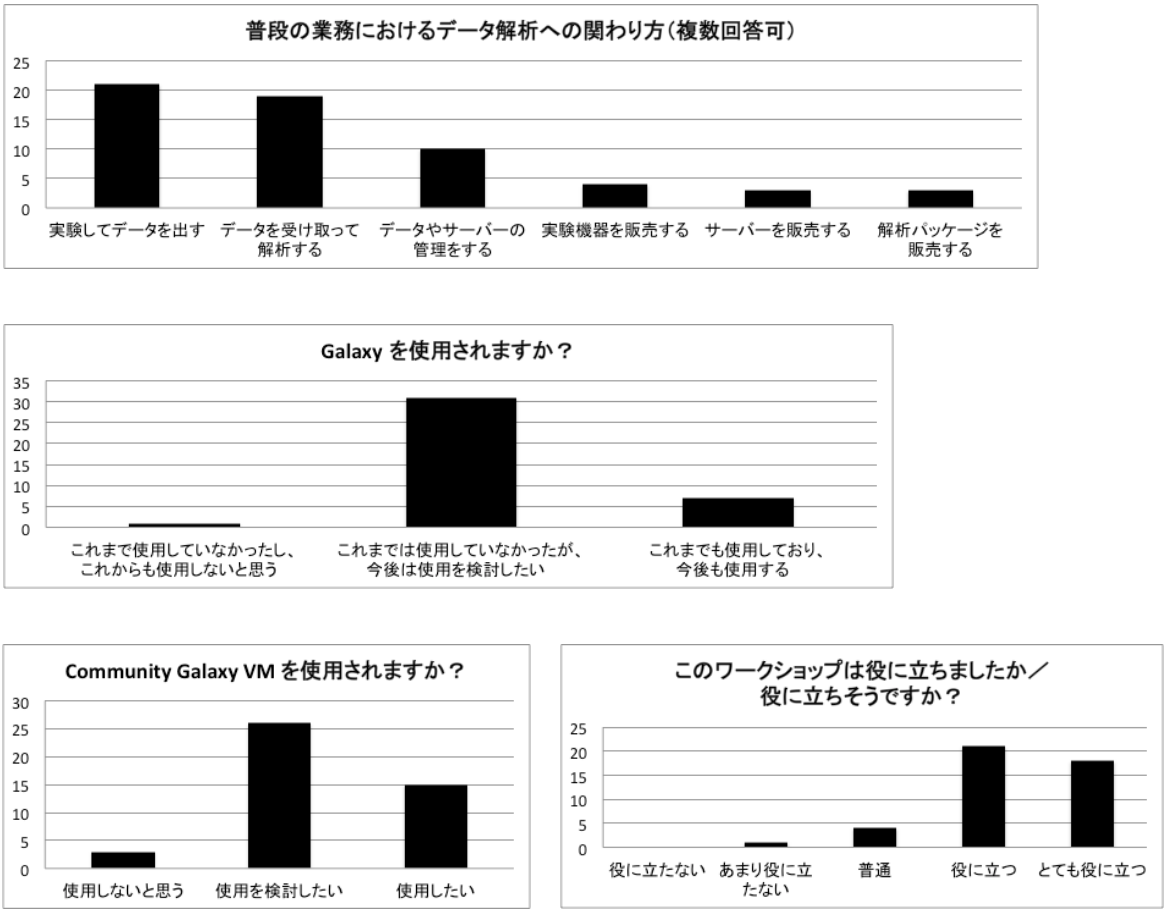

図 16：　ワークショップのアンケート結果

ワークショップ終了後にウェブ・フォームによりアンケートを実施した．有効回答数 44 のうちの回答数を縦軸としている．今まで Galaxy を使用していなかったが，今後，Galaxy およびコミュニティ仮想マシンの使用を検討したいといった参加者が多かったことが分かる．

さらに，「USE」として，公共サーバーを用意した．ワークフローの動作確認のみを目的とした小規模な公共サーバーを公開している他，同じワークフローが使用できる大規模な計算環境を公開できるように，計算リソースを持つ研究機関と共同研究に取り組んでいる．

まず, Galaxy プロジェクトがウェブサイトにリストしている 60 の公開サーバーのうちの日本の 2 つとして，ライフサイエンス統合データベースセンターおよび国立遺伝学研究所が以前より Galaxy サーバーを公開しているが[56][62]，うち国立遺伝学研究所の Galaxy サーバーでコミュニティ仮想マシンに置き換えるための作業が始まっている．さらに，アプリケーション・コンテナを利用したクラスタ上で同じワークフローが実行できる環境を国立情報学研究所と共同で試験している．

3.3.3. 再現性と再利用性の評価

再現性と再利用性の高いデータ処理環境を構築するという本研究の課題に対して，この評価方法を検討した．第一の評価として，上記の設計に基づく環境上で共有されているワークフローを実行した際に，異なる環境で同じ結果が得られることを確認する．この際，利用者が容易に誤った操作をするといった場合にも，実際上の再現性は低下するといえる．第二の評価として，利用者の数と利用頻度を確認する．この指標は必ずしも再現性と再利用性の向上を示すものではないが，再現性と再利用性が必要とされる結果として間接的に利用者の数と利用頻度が増加すると考えられる．さらに，第三の評価として，論文へのデータ解析方法の記載のために仮想マシンとそのワークフローが引用されている頻度を確認する（表 9）．

将来的には，この環境上のワークフローを論文に引用することでデータ処理方法を査読者や他の研究者と共有するといった利用方法が想定される．それによって，今までの論文の記述では十分にデータ処理を再現または再利用するための情報を得られなかったワークフローが，簡単に再実行可能になるだろう．既に学術誌「Giga Science」では投稿された論文で引用された解析のヒストリーを公共 Galaxy サーバー上でホストするといった試みが始められている[63]．今後は，学術誌それぞれの取り組みと，本研究のような開発者コミュニティの取り組みが協力する必要があるだろう．このように，今後はワークフローが論文に引用され，その引用数が情報基盤の再現性の向上に対する貢献の直接的な指標となると考えられる．

表 9： 再現性と再利用性の評価

評価指標	評価方法	着手状況
異なる環境で同じ結果が得られること	出力データを比較するツールの追加	済（今後，クラスタ環境やコンテナ環境で着手）
利用者の数と利用頻度	仮想マシンとワークフローの使用状況を監視するツールの追加	未着手
論文への引用	論文への引用方法の検討 引用数の調査	未着手

上記で挙げた第一の評価として，この解析環境を使用することで再現性のある解析，つまり，異なる環境で同じワークフローを実行した際に同じ結果が得られることを確認する．本研究では仮想マシンを使用しているため，設計上は同じ結果が得られることが期待できるが，ベアメタル・サーバーへのインストールやコンテナ用のイメージを作成した場合に，同じ結果が出力されるかどうか検証する方法が必要になる．

そこで，結果の検証のため，Galaxy ツールとしてデータのチェックサム（MD5 ハッシュ値）を計算するツールを作成し，これを用いてワークフロー単位で実行結果が一致していることを確認できるようにした．その結果，ラップトップ，ワークステーション，AWS クラウドの全ての環境で期待通り結果が一致していることが確かめられた（表 10）．このツールを用いた実行結果の検証方法は仮想マシン以外の環境で使用できると考えられる．

表 10： 異なる環境における再現性の確認

ワークフロー	実行環境	結果
ChIP-seq 02 v002	ラップトップ	一致
	ワークステーション	一致
	AWS クラウド	一致
RNA-seq 01 v002	ラップトップ	一致
	ワークステーション	一致
	AWS クラウド	一致

4.　研究成果２（データ統合）

4. 1.　課題

4.1.1.　がんゲノム・データの利用

現在入手可能な公共のゲノム情報として，最も大規模なデータのひとつががんゲノム・データである．次世代シークエンサー技術を用いた全ゲノムまたは全エクソンのシークエンシングによりハイスループットに塩基配列変異を検出できるようになり，がんゲノムアトラス（TCGA; The Cancer Genome Atlas）や国際がんゲノムコンソーシアム（ICGC; The International Cancer Genome Consortium）などの大規模ながんゲノム解析プロジェクトによって数十のがん種それぞれの変異情報がカタログ化されつつある[64][65]．シークエンシングコストが低下することで出力されるデータ量は急増し，これらのデータベースに大量のデータが登録されている．

さらに，がんゲノム・データベースの特徴として，塩基配列変異データのみならず，同一症例から RNA，miRNA，DNA メチル化，コピー数，タンパク質リン酸化など種々のオミックス・データが収集されていることが挙げられる．サンプルに対する複数オミックス・データとしては，現在，最も多くの症例が登録されているデータベースであり，現在も登録データが増え続けている（

表 11)．

本研究では，データ統合の一例として，ICGC のデータとデータ・ポータルについて着目している．この理由として，第一に，ICGC が大量のデータを保管していてかつこれらがダウンロード可能であること，第二に，高機能なデータ・ポータルが提供されていてかつ他のデータ・ポータルを作成および公開することを許可していることが挙げられる．ここで，データ・ポータルからダウンロードできる ICGC のデータはタブ区切りのテキスト・ファイルであり，通常，これらのデータをデータベースに格納する場合には，表データとしてリレーショナル・データベースで扱われることを想定している．

表 11： がんゲノム・データベースのデータ内訳

	ICGC (available)	TCGA (plan)
Projects	55	89
Donors	12,979	20,985 (9,010 complete)
SSM	8,038	14,597
CNSM	9,865	17,830
mRNA seq	8,143	10,858
mRNA array	3,135	5,132
miRNA-S	8,190	1,725
METH-A	9,089	4,316
Protein Exp	3,165	6,551

ICGC データの各値は 2015 年 6 月 16 日更新時（リリース 19），TCGA データの各値は目標としている数値で，2015 年 5 月 8 日で 9,010 の症例が完了している．

4.1.2. がんゲノム・データのデータ統合

2 章で分析した通り，ゲノム情報のデータ統合の再現性と再利用性を向上させるための手法の検証が必要である．ここで再現性とは，異なるデータベースのデータセットを統合した際に，統合後のデータセットが毎回同一であり，このデータセットを対象に同じ検索を実行した際に同一の結果が得られることを指している．また，再利用性とは，アプリケーション開発者やデータの公開者が，あるデータセットと他のデータセットとの統合方法（例えば，ある遺伝子がどの遺伝子に対応付けられるか）を明示的に記述することができ，その統合後のデータセットを共有することができることを指している．

まず，データ統合は一つのデータベースの中のデータにおいても必要である．例えば，ICGC データ・ポータルから複数の表データとしてダウンロードされたデータは，適切なキー列で結合することができる．症例の表と体細胞変異の表にはどちらにも症例の ID の列があるため，これをキー列として二つの表を結合できる．多くの異なる種類のデータを格納しているがんゲノム・データベースでは，多くの表が定義されているが，その表同士がどのように結合されるか（どのキー列を用いるか）はデータの利用者に任されている場合が多く，一つのデータベース由来のデータでもデータ統合の再現ができない場合がある．

さらに，ICGC データ・ポータルの「Advanced Search」で表示される内容には，ICGC データ・ポータルからダウンロードしたデータには含まれていない情報がある．例えば，遺伝子のアノテーションやパスウェイの情報は「Advanced Search」に表示されるが，ICGC データセットには含まれない．これは，遺伝子のアノテーションやパスウェイの情報が外部データベースに由来するものだからである．「Advanced Search」では前処理によってそれらのデータを統合しており，その結果を表示している．

この場合，データ統合はアプリケーション作成者によって実施されたものなので，利用者がこのデータ・ポータル以外で同様のデータ統合結果を扱いたい場合には，再度それぞれのデータベースからデータを取得し，データ統合を実施しなければならない．また，データ統合の処理が異なる場合，例えば異なるキー列で結合している場合，にはそれが同じデータ統合結果にならない可能性もある．

4.1.3. RDF データの活用

RDF（Resource Description Framework）は，セマンティック・ウェブを構築するための要素技術として提案されたデータ記述のフレームワークであり[66]，このフレームワークに沿って生成されたデータを本論文では RDF データと呼ぶこととする．RDF は 1999 年には W3C によって規格化されているが，計算機の性能向上とデータ分析のニーズに伴い，近年になって IT システム全般において RDF の活用が注目されている．生命科学においても，Bio2RDF[67]，EBI RDF Platform[68]，DBCLS BioHackathon[69] といったプロジェクトにおいて，RDF データが生成されている（表 12）．

- **セマンティック・ウェブと RDF**： セマンティック・ウェブは機械により処理可能なデータのウェブを目指した現在のウェブの拡張として 2001 年に提唱され[66]，W3C によって策定されているまたは策定予定の規格の一群である．その中で，共通データ・フォーマットの RDF やクエリ言語の SPARQL などが既に策定されている[70]．

表 12： 生命科学 RDF データを公開している主なプロジェクト

プロジェクト	URL	データセット
EBI RDF Platform	https://www.ebi.ac.uk/rdf/platform	BioModels BioSamples chEMBL Expression Atlas Reactome UniProt
Bio2RDF	http://bio2rdf.org/	DrugBank GO Annotation HUGO Gene KEGG PubMed　など
BioHackathon	http://data.allie.dbcls.jp/	Allie
myExperiment	http://rdf.myexperiment.org/	myExperiment

通常利用される表データではなく RDF データをデータベースに格納して利用する理由として，一般的には以下のような優位点が挙げられる．本研究では，特にはじめのデータ統合における特長に注目している．

- 再現性と再利用性の高いデータ統合が可能であること
- スキーマの記述性が高くかつスキーマの変更に対応しやすいこと
- ウェブ上の分散データベースに対するクエリを記述できること
- 複数のアプリケーションがスキーマを共有することが容易であること（インター・オペラビリティ）

再現性と再利用性の高いデータ統合が可能になる理由は，データの生成時に既存リソースのグローバルな ID（URI で表記される）を再利用することができ，生成するデータと既存データとの関係性を記述するデータを作成できるためである．異なるデータベース由来のデータセット同士が新たなデータセットによって関係性を定義されているため，これらのデータセットをひとつのデータベースに収集するだけでデータ統合済みのデータセットを取得できる．

このような特長が注目され，多くの生命科学データベースのデータが RDF データとして提供されている．例えば，蛋白質データベースの UniProt は最も大きな公共 RDF データのひとつであり，2015 年 7 月の時点で 140 億トリプル（トリプルは RDF におけるデータの単位）のデータが公開されている[71]．その一方で，実際の解析のために公共のデータベースから入手できるデータは表形式の場合が多く，また，ウェブ・アプリケーションが RDF データのデータベース上に構築される例は珍しい．

そこで本研究では，がんゲノム・データベースのデータを用いて，表形式のデータを RDF データに変換する方法や，このデータを他の公共 RDF データと統合する方法，さらにこの RDF データのデータベース上にウェブ・アプリケーションを実装する方法を調査する．これにより，現在入手可能な RDF データを使用して，今後は実用的なデータ統合とシステム開発が可能であることを示す．

4.2. 手法

4.2.1. RDF スキーマの定義

ICGC データの表定義を参考に ICGC データが格納できる RDF スキーマを定義した．RDF スキーマも RDF によって記述することが可能であり，オントロジー編集ソフトウェア「Protégé」[72] で作成した RDF スキーマを RDF/XML 形式のファイル「icgc.owl」として公開している．RDF データの生成やクエリの開発時には RDF スキーマを参照するが，（それ自体も RDF で記述された）RDF スキーマは可読性が低いため，通常は替わりに下図のようなネットワーク図を用いてクラスやプロパティを参照する（図 17）．

- RDF スキーマとオントロジー： RDF スキーマはリソースのクラスとプロパティを記述するための語彙であり，通常オントロジーを記述するウェブ・オントロジー言語（OWL）は RDF スキーマに加えてより表現力の高い語彙，例えばクラス間の関係性やカーディナリティ，プロパティの特性などの語彙を含む．本研究では，表データから RDF データを生成しているが，この際に必要となる語彙は RDF スキーマで十分であるため，オントロジーではなく RDF スキーマと呼んでいる．

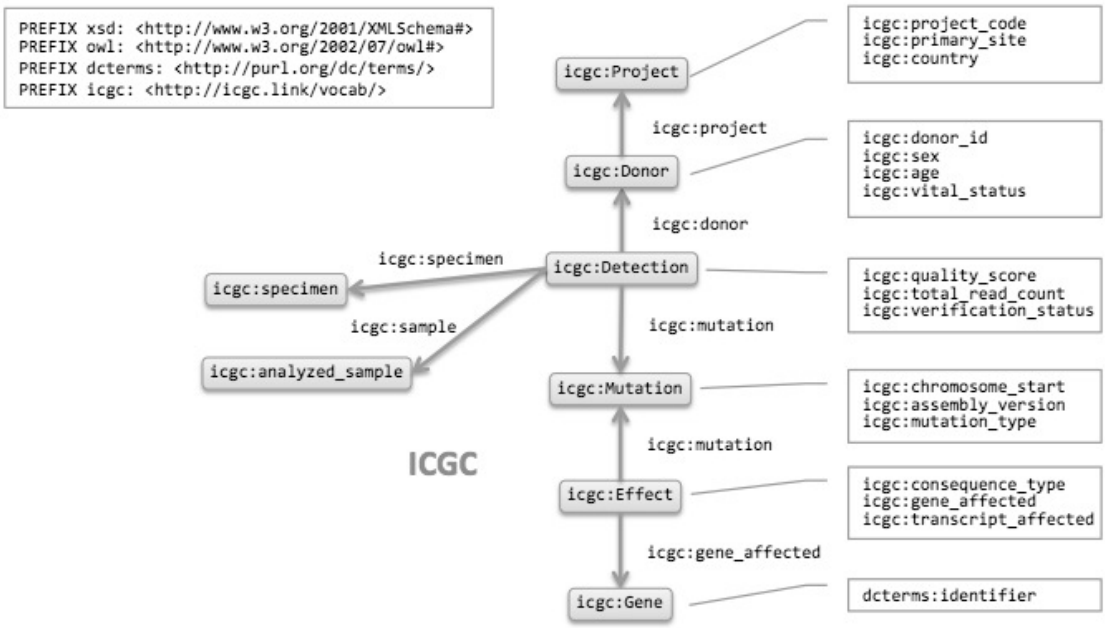

図 17： RDF スキーマ

本研究で定義した ICGC データの RDF スキーマをネットワーク・グラフで示している．ノードはクラスを表しており，エッジとボックスはそれぞれ URI とリテラル値を値域とするプロパティを表している．

設計および公開された RDF スキーマが，データ・プロバイダーやアプリケーションの間で共有されることで，データの統合が用意になる．このため，データの利用者が理解しやすい RDF スキーマや URI の名前規則が用いられることが重要である．多くの利用者が使いやすい RDF スキーマや URI の名前規則を作成するためには，少なくともそのデータの対象利用者の分野においてガイドラインが共有されていることが望ましい．

そのため，W3C においても RDF データの公開におけるベスト・プラクティス[73] が公開されている．さらに，生命科学といった各研究領域においては，今後，より実用的なガイドラインが策定，共有されるだろう．例えば，2014 年の国際会議 BioHackathon 2014 [74] においては，「Ten simple rules for publishing RDF data for the Life Sciences」を策定するといった提案がなされ，現在まで議論が継続している．本研究の RDF スキーマの定義においては，ライフサイエンス統合データベースセンターのガイドライン [75] に則って，以下のようなテンに配慮してデータを生成している．

- リソースを示す URI には識別するための ID を URI の末尾に記述する
- URI リソースには rdf:type でオントロジーのクラスをタイプ指定する
- URI リソースには rdfs:label でラベルをつける
- URI リソースには dcterms:identifier で ID を記述する

4.2.2. RDF データの生成

ICGC のデータ・ポータルから入手可能な表データから RDF データを生成する．ここでは，宣言的で変換ルールが分かりやすい方法を採用している．まず，表データをリレーショナル・データベースにロードしクエリ言語（SQL）を使用して表データを第 3 正規形に正規化する．これにより，その後の処理において冗長なトリプルの生成を省くことができる．次に，正規化された表データと生成する RDF データとのマッピングを定義し，変換エンジン（D2RQ）を使って RDF データを生成した．この手法は表データから RDF を生成する際に汎用的に使用できる．

この変換ルールでは，表データのある列を主語としてその他の列を目的語とする RDF データを生成することができる．そのため，表データをまず第 3 正規形にすることで主キー列となる列の値を主語として従属列の値を目的語とするトリプル（RDF データの最小単位で，主語，述語，目的語の組を意味する）を生成している（図 18）．

```
CREATE TABLE `clinical` (                        <-- original, un-normalised table “clinical”
  `icgc_donor_id` varchar(20) DEFAULT NULL,
  `project_code` varchar(20) DEFAULT NULL,
  *snip*
  `donor_age_at_diagnosis` int(3) DEFAULT NULL,
  *snip*
);

LOAD DATA LOCAL INFILE "clinical.tsv"             <-- tab-delimited text data
  INTO TABLE clinical IGNORE 1 LINES;

CREATE TABLE donor (                              <--  new, normalised table “donor”
  SELECT DISTINCT
    icgc_donor_id AS donor_id
  , project_code
    *snip*
  , donor_age_at_diagnosis
    *snip*
  FROM clinical
);
```

図 18： 表データの正規化

ICGC データ・ポータルからダウンロードされたデータは正規化されていないため，そのまま RDF に変換するとマッピング定義が複雑になる上，重複が多く生成される．そのため，RDF への変換前に，第三正規形の表にロードしておく．

表データを RDF に変換する方法として，変換ツールとマッピング定義を使用する方法と，マッピング定義を使用せずに独自のプログラムを作成する方法がある．今回は，既存の公開データを変換対象としていることから，データの利用者から見たデータ変換の信頼性や再利用性を高めるために，マッピング定義を使用することとする．

マッピング定義には数種類の変換ツールの独自仕様および W3C 勧告となっている仕様があり [76]，特によく利用されるものとして，利用実績が多く表現力の高いオープン・ソース・ソフトウェアの変換ツール D2RQ [77] の独自仕様の他，W3C 勧告となっている Direct Mapping [78] と R2RML [79] がある（表 13）．それぞれのマッピング定義の仕様上，変換先の RDF スキーマや URI の名前規則が既に決まっている場合には D2RQ または R2RML を使用する必要があり，現状では実績のある変換ツールである D2RQ が使いやすいが，今後は W3C 勧告である R2RML とこれに対応した変換ツールの使用が望ましいと考えられる．

表 13： 表データを RDF に変換する方法の概要

変換方法	概要
D2RQ	変換に使われることの多いオープン・ソース・ソフトウェアだが，独自仕様のマッピング定義を使用している．
Direct Mapping	W3C 勧告の仕様で，表データをマッピング定義なしに変換することができるが，変換後の RDF スキーマや URI の名前規則が決まっている場合には対応することができない．
R2RML	W3C 勧告のマッピング定義仕様．現在はこれに対応した変換ツールが少ないが，今後の標準となることが望まれる．
その他の方法	マッピング定義を使用せずに，表データを変換するプログラムを作成することは可能だが，データ変換の信頼性や再利用性を高めるために，マッピング定義を使用することが望ましい．

本研究では，マッピング定義を作成するためにウェブ・ツールである D2RQ Mapper [80] を使用した．D2RQ Mapper は既存のリレーショナル・データベースに接続して表データのスキーマを抽出した後，利用者はこのスキーマを参照しながらウェブ上の UI を使用してマッピング定義を設計することができ，このマッピング定義は D2RQ 独自仕様または R2RML で出力することができる（図 19）．D2RQ の独自仕様および R2RML のマッピング定義は，それ自体も RDF で記述されたファイルであるため，手作業でこれらのファイルを作成することが可能である一方，これらのファイルの可読性は低い（図 20）．そのため，ウェブ・ツールの UI によって可読性が向上し，修正が容易に可能になることで，マッピング定義の再利用性が高まると考えられる．こうして得られた D2RQ 独自仕様のマッピング定義を使用して D2RQ でデータを変換した．

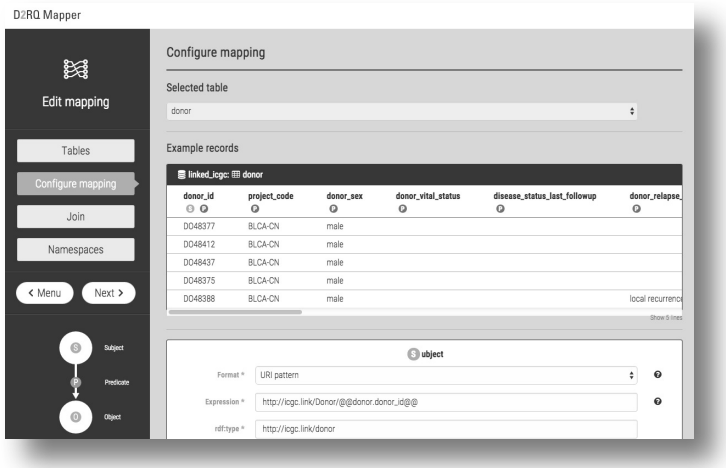

図 19： ウェブ・ツール「D2RQ Mapper」

ライフサイエンス統合データベースセンターで開発中のウェブ・ツール D2RQ Mapper を試用して，マッピング定義を設計した．接続先のリレーショナル・データベースの各表が表示され，これらの各列に対してマッピングを定義する．

```
map:donor a d2rq:ClassMap;                                          <-- “donor” class
          d2rq:dataStorage map:database;
          d2rq:uriPattern "Donor/@@donor.donor_id|encode@@";
          d2rq:class <Donor>;
          .
map:donor_donor_age_at_diagnosis a d2rq:PropertyBridge;             <-- “age” of the donor
          d2rq:belongsToClassMap map:donor;
          d2rq:property <vocab/donor_age_at_diagnosis>;
          d2rq:column "donor.donor_age_at_diagnosis";               <-- dependent column
          d2rq:datatype xsd:integer;
          .
map:donor_project_code__ref a d2rq:PropertyBridge;                  <-- “project” of the donor
          d2rq:belongsToClassMap map:donor;
          d2rq:property <vocab/project>;
          d2rq:refersToClassMap map:project;
          d2rq:join "donor.project_code => project.project_code";   <-- reference key column
          .
```

図 20： D2RQ 独自仕様のマッピング定義（一部抜粋）

D2RQ の独自仕様によるマッピング定義．表の主キー列を各クラスの URI リソースと指定して，その従属列をデータ・プロパティに指定している他，他の表の主キー列をオブジェクト・プロパティに指定している．

4.2.3. 外部データとの統合

ICGC データ・ポータルのようなアプリケーションを作成するためには，生成した ICGC の RDF データと，公開されている他のデータベースの RDF データを統合する必要がある．例として，ICGC データ・ポータルから入手できるデータセットに含まれる遺伝子名は Ensembl ID のため「ENSG00000155657」のような表記であり一般的な遺伝子名は分からない．そこで，外部のデータベースである HUGO Gene Nomenclature Committee (HGNC) [81] のデータを統合することによって，これが Gene Symbol「TTN」であることがわかる．このように，ICGC データセットと外部の RDF データを同じデータベースに格納することで，ICGC データセットに含まれていない遺伝子名の情報を統合して，検索などに用いることが可能になる．

具体的な手法としては，生成したICGCのRDFデータを一度トリプルストアに格納し，これに対しSPARQLクエリを実行してICGCデータと外部データとの関係性を定義する新たなデータセットを出力し，このデータセットをICGCデータと外部データと併せてトリプルストアに格納する．トリプルストア内にルールを作成して実行する等，新たなデータセットを出力しない方法もあるが，今回は統合用のRDFデータを個別のファイルとして公開するため，SPARQLクエリのCONSTRUCT文を用いて出力している（図21，図22）．

- **トリプルストア：** RDFを格納しSPARQLクエリを実行するデータベース・マネジメント・システム．表データを格納する場合のリレーショナル・データベースに相当する．オープン・ソースのトリプルストアの代表的なものにはApache Jena TDBやSesame，Virtuoso Open-Source Editionなどが挙げられる他，多くの商用製品がある．

- **SPARQL：** RDFを格納するトリプルストアに対してデータの参照や更新を実行するためのクエリ言語で，リレーショナル・データベースのSQLに相当する．W3C勧告として2013年にSPARQL1.1の仕様が策定されている．

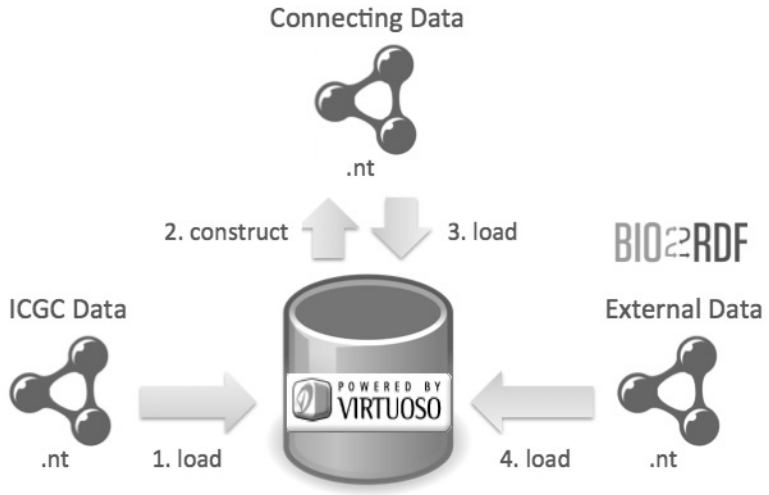

図 21： 統合用 RDF データの作成

ICGC データは他のデータベースのリソースを参照していないので，ICGC データから生成された RDF はそのままでは他のデータベースの RDF とは統合されない．そのため，ICGC データから外部の RDF データを参照する RDF を作成し，統合したいデータセットとともにデータベースにロードする．

```
CONSTRUCT {
  ?uri_icgc dcterms:identifier ?postfix .      <-- New triple for identifier
  ?uri_icgc owl:sameAs ?uri_uniprot .          <-- New triple pointing UniProt gene
  ?uri_icgc owl:sameAs ?uri_bio2rdf .          <-- New triple pointing HGNC gene
}
WHERE {
  ?s icgc:gene_affected ?uri_icgc .
  FILTER(!(?postfix=""))
  BIND (REPLACE(str(?uri_icgc), "^.*Gene/", "") AS ?postfix)
  BIND (IRI(CONCAT("http://purl.uniprot.org/ensembl/",?postfix)) as ?uri_uniprot)
  BIND (IRI(CONCAT("http://bio2rdf.org/ensembl:",?postfix)) as ?uri_bio2rdf)
}
```

図 22： 統合用 RDF データの作成のためのクエリ（一部抜粋）

SPARQL クエリの CONSTRUCT 文を使用して，統合用の RDF データを作成する．ここでは，ICGC データの遺伝子オブジェクトに，ID としてリテラル値を参照させるトリプルの他，外部データベースの遺伝子オブジェクト同士を owl:sameAs で参照させるトリプルを作成している．

ICGC Data Portal では，特定のパスウェイや遺伝子オントロジーに関連する遺伝子上の変異やその変異を持つ症例を検索することができる．このような機能を実装するために，パスウェイ・データベースの「Reactome」[82] および遺伝子オントロジー・データベースの「Gene Ontology Annotation」[83] の RDF データもデータ統合されている必要がある．

例えば，Reactome のデータは EBI RDF Platform プロジェクトにより RDF データが提供されており，これを HUGO Gene Nomenclature Committee のデータと共にデータベースに格納することで, ICGC データと統合されることがスキーマの図からわかる（図 23）．

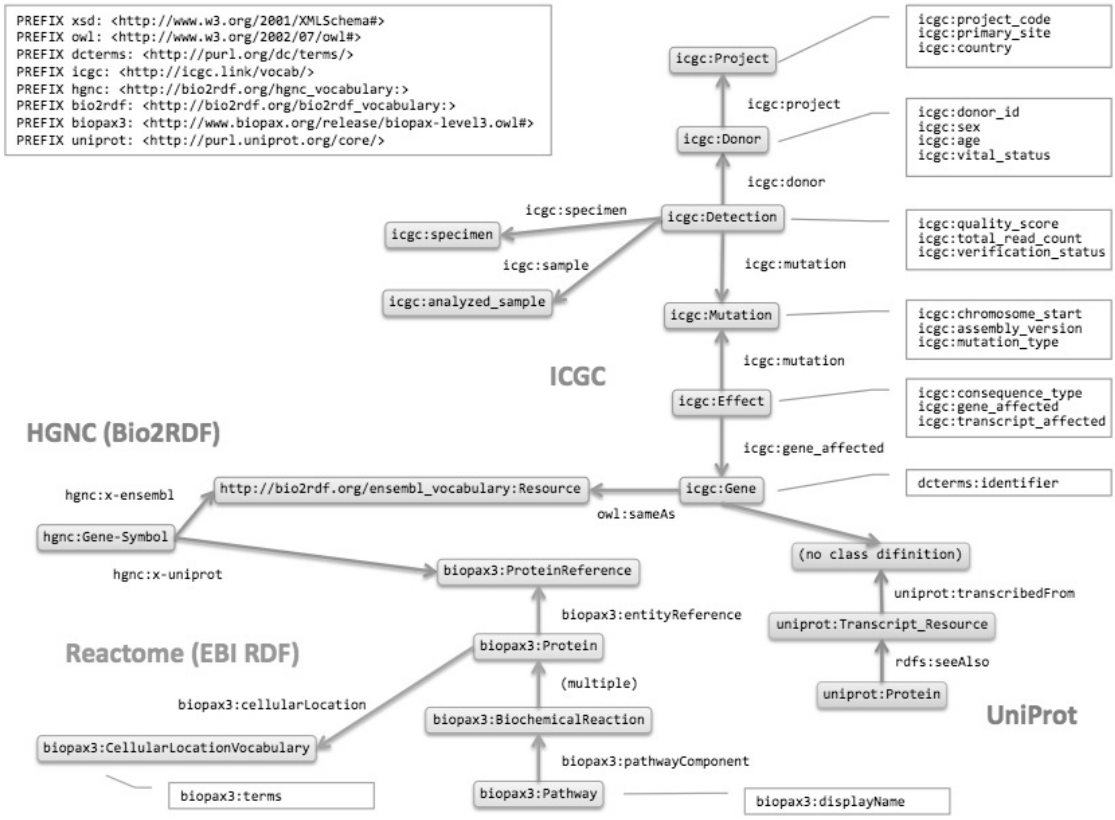

図 23： 外部データとの統合後の RDF スキーマ

前出の ICGC データの RDF スキーマに，遺伝子クラス（icgc:Gene）を介して外部データの RDF スキーマが統合されている．外部データ（HGNC, UniProt, Reactome）の RDF スキーマはここには一部のみを記載している．

4.2.4. データ・ポータルの開発

データ統合済みのデータセットを格納したデータベースに対して SPARQL クエリを実行して結果を可視化するセマンティック・ウェブ・アプリケーションを実装した．セマンティック・ウェブ・アプリケーションの開発において，アプリケーション開発者はアプリケーションに依存したデータ・モデルを開発するのではなく，既存のデータの RDF スキーマを参照してクエリを設計する．

実装においては，RDF を格納するトリプルストアとしてオープン・ソースの Virtuoso Open-Source Edition を用いており，検証という目的から SPARQL クエリの内容が利用者に分かるようにサーバー・サイドではなくクライアントでクエリを生成している．さらに，クエリ実行結果の可視化のために開発された JavaScript ライブラリ「D3SPARQL」[84]を用いて開発コストを抑えるとともに，クエリ結果のキャッシング機能を実装した中継サーバー[85]を使用することによって実用可能なレスポンスを実現している．（図 24）

- **JavaScript ライブラリ**： SPARQL エンドポイントのクエリ結果は JSON 形式にシリアライズする際の形式が標準化されているため[86]，この結果を可視化するためのライブラリが作成されている．D3SPARQL はデータ可視化 JavaScript ライブラリである D3.js を使用している．

- **クエリ結果のキャッシング**： データ・ウェアハウスの多くではデータ更新はバッチ処理で実施され通常時のクエリは参照のみであるため，クエリ結果をキャッシングしておくことができる．本研究ではトリプルストアの性能を評価することは目的ではないため，クエリ結果をキャッシングすることによりウェブ・アプリケーションが実用的な応答時間でクエリ結果を得られるようにしている．

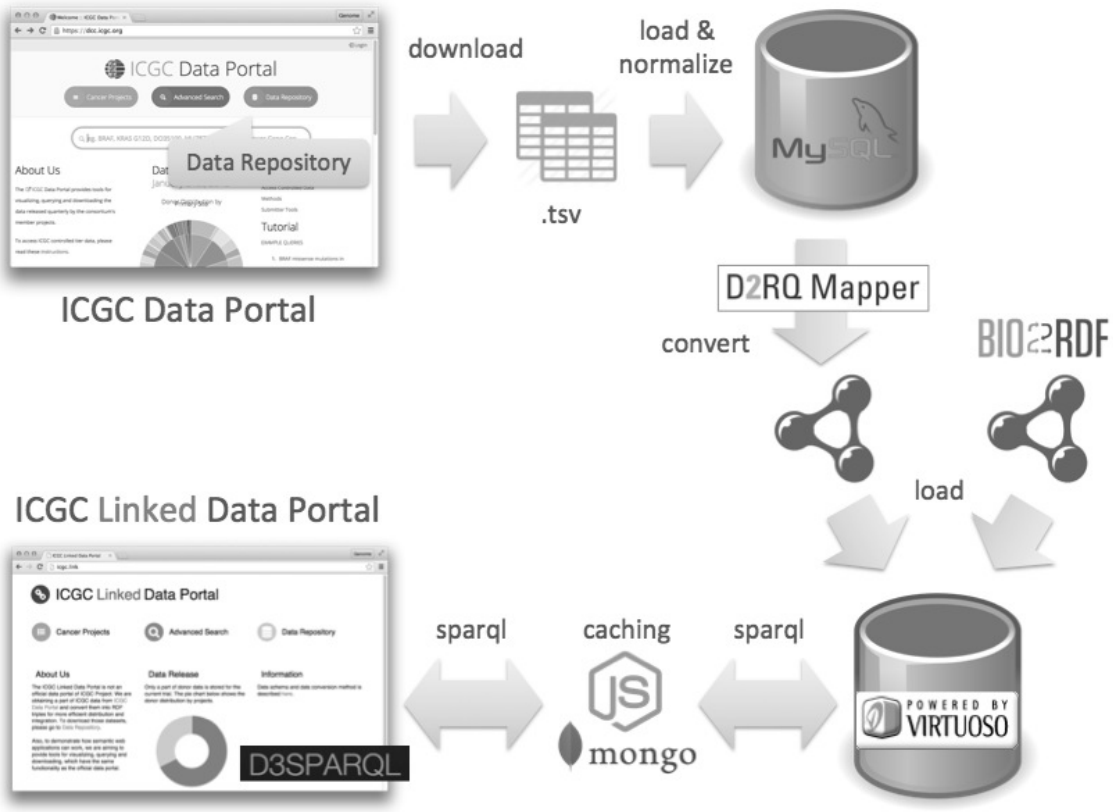

図 24： ICGC Linked Data Portal の開発

RDF データの生成から，外部 RDF データとの統合，そのデータを用いたデータ・ポータルの実装まで，一連の開発のフローを示している．最終的に，データ・ポータルに必要なコンポーネントは下段のポータルおよび RDF ストアとしてのデータベース，さらにその中継サーバー（本研究では，クエリ結果のキャッシングや更新クエリの除外が可能なサーバーを実装している）である．

4.3. 結果

4.3.1. RDF データの公開

ICGC データセットの RDF データと RDF スキーマを ICGC リンクト・データ・ポータル（http://icgc.link/）に公開し，統合用の RDF データを用いて，外部のプロジェクト（HGNC および UniProt）の RDF データと容易に統合できるようになった．これにより，データの利用者は既にデータ統合されたデータセットを入手することができ，これらのデータセットをデータベースにロードして，RDF スキーマを参照することで，データベース横断的な検索クエリを作成および実行することができる（表 14）．

表 14： 生成したデータセットのトリプル数

項目	トリプル数
ICGC データ	606,046,773
統合用データ（HGNC および UniProt との統合）	173,319
合計	606,220,092

今回作成した RDF データを使用することにより，データ統合時のアプリケーション間の差異や誤りが減少し，アプリケーションの開発工数が削減できることが期待できる．これにより，この RDF データが他のアプリケーションでも利用されることで，このデータ統合の再現性と再利用性が向上したといえる．

4.3.2. データ・ポータルの公開

さらに，本研究では作成した RDF データを用いて，実用可能なセマンティック・ウェブ・アプリケーションを実装できることを示した．ICGC データ・ポータルの「Advanced Search」は指定した条件で症例，遺伝子，変異を絞り込むことができる，ポータルの主要な機能であり，この機能をセマンティック・ウェブ・アプリケーションとして実装できたことで，より多くのアプリケーションでセマンティック・ウェブが利用されることが期待できる．

データ統合の再利用によってアプリケーションの開発工数が削減できるかどうかを次の比較により評価する．RDF データが入手可能な場合は，データベース上のデータ・モデルの定義の作業が削減できることが分かる．

- **本研究で作成した RDF データがない場合**：ICGC データ・ポータルのデータを表データとして入手する．このデータ形式に従ったデータ・モデルをデータベース上に定義し，データをロードする．その後，ER 図を参照しながら複数の表を結合するキー列を意識してクエリを実装する．

- **本研究で作成した RDF データがある場合**：RDF データを入手する．データ・モデルを定義することなく，データベースにデータをロードする．その後，RDF スキーマを参照しながらクエリを実装する．

アプリケーション実装においては，この検索結果を実用可能なレスポンス時間で返すことが求められるが，このレスポンス性能は，RDF データを格納するデータベース・マネジメント・システムの仕組みに大きく依存している．データベースの設定やSPARQLクエリのチューニングなどによりレスポンス性能を改善できる可能性があるが，性能の改善は本研究の目的ではないため，技術的課題として指摘するのみに留める．

一方，このアプリケーションには，クエリとその実行結果のキャッシングの機能を追加したため，実際の画面操作におけるレスポンス時間は非常に短くなっている（表 15）．ICGC データ・ポータルのデータは通常は更新されず，定期的に全てのデータセットが一度に更新されるものなので，通常は参照が主である．このようなシステムでは，よく使用されるクエリをキャッシュに保存しておく方法は非常に効果的である．今後，データベースの性能課題を補うための手法として検討が必要である．

表 15：　レスポンス性能の例

項目	キャッシュ無	キャッシュ有
条件の入力前（初期画面）	14 秒	1 秒未満
検索 1（症例のプロジェクトによる絞込み）	5 秒	1 秒未満
検索 2（遺伝子のがん種による絞込み）	3 秒	1 秒未満
サマリ画面 1（特定のプロジェクト）	7 秒	1 秒未満
サマリ画面 2（特定の遺伝子）	1 秒	1 秒未満

5. 結論

5.1. 結果総括

本研究では，ゲノム科学におけるデータ駆動型研究を進めるために，データ解析の再現性と再利用性を大幅に向上させる必要があることを指摘し，そのためのデータ処理とデータ統合の情報基盤の設計をそれぞれ議論した上，実際にゲノム情報解析に携わっている利用者が活用できる仮想マシンや，現行のゲノム情報データベースのデータを用いた技術検証のためのウェブ・アプリケーションを作成した.

第一に，ゲノム情報のデータ処理については，開発者会議を立ち上げて複数の研究機関のデータ処理ワークフローを収集し，これらを実行するためのコミュニティ仮想マシンを作成した. 今まで Galaxy のようなワークフロー管理システムが使用されてきたが，これだけでは再現性と再利用性は十分に得られていなかった．コミュニティ仮想マシンとこれを運用するディストリビューターとしての開発者会議を設けられたことで，今後，再現性と再利用性の高いデータ解析の実行環境を利用者に提供できると考えられる．この仮想マシンは，利用者向けのワークショップを開催して紹介された他，大規模計算環境を利用した公開サーバーとしても利用される予定である.

第二に，ゲノム情報のデータ統合については，大規模な公共ゲノム情報であるがんゲノム・データベースのデータを用いて，公開されている生命科学 RDF データの活用事例を作成した. 数多くの生命科学データベースが既に RDF データとして公開されているが，これらのデータを二次活用したウェブ・アプリケーションは限られていた．本研究では，表データとして公開されているがんゲノム・データをもとに，RDF データを生成し，さらに，このデータを用いてデータ・ポータルを実装した．このデータ・ポータルは，外部データベースの RDF データを統合したデータベース上で動作するセマンティック・ウェブ・アプリケーションであり，性能面などの技術的な課題は残っているものの，データ統合の再利用という手法が可能であることを示した.

5.2. 今後の課題

ここでは「2.3 課題設定」で提示した課題がどの程度達成できているかを評価するため，現在の進捗状況と今後の課題を確認する．

まず，データ処理について，コミュニティ仮想マシンの開発体制ができ仮想マシンの配布の開始したものの，現在のところ限られた利用者からのフィードバックしか得られておらず，今後活動を続ける中で利用者を把握してフィードバックを開発に活かす仕組みを構築する必要がある．また，技術面においては，仮想マシン上で共有するワークフローの数が増えた際に仮想マシンのパッケージングと依存関係の確認のためのコストが増大するという問題があり，これを解消するために Docker コンテナを用いたツール配布の仕組みを開発中である．さらに，海外コミュニティともお互いのワークフローを共有するため，現在，オーストラリアの Genomic Virtual Laboratory の開発者と検討を進めている．このためには，コンテナによるツール配布や仮想マシンの共同開発を可能にするクラウドの整備が必要になると考えられる（表 16）．

表 16： 研究成果 1 の進捗状況と今後の課題

項目	進捗状況
開発者コミュニティの立ち上げ	○：開発者会議を継続的に開催している
コミュニティ仮想マシンの開発（GET）	○：仮想マシンそ配布している
ドキュメントの整備（LEARN）	○：Wiki を継続的に更新している
公共サーバーの設置（USE）	△：高性能のサーバー・クラスタへ移設検討中
利用者への説明	○：ワークショップや勉強会を実施している
利用者からのフィードバック	×：まだフィードバックが得られていない
コンテナによるツール配布	×：Docker コンテナを用いて開発中
海外コミュニティとのワークフロー共有	×：オーストラリア・コミュニティと検討中

また，データ統合について，新しく生成した RDF データと共に既存の生命科学 RDF データが有効活用できることを示したが，生成したがんゲノム・データベースの RDF スキーマや RDF データはまだ外部からは参照されていない．今後のアプリケーション開発にこのデータが活用されることで，データ統合の再現と再利用が実現し，ウェブ上のリンクト・データが拡大することから，公開されているスキーマやデータが外部の開発コミュニティに認知されることは重要である．そのため，BioPortal[37]のような公共ポータル・サイトにスキーマやデータを登録することが有効である．また，現在も公共がんゲノム・データは増え続けているため，RDF データの生成を継続するために手順を自動化するなどの工夫が必要である．（表 17）

表 17： 研究成果２の進捗状況と今後の課題

項目	進捗状況
RDF スキーマの定義	○：RDF スキーマを公開している
RDF データの生成	○：RDF データを公開している
公共 RDF データとの統合	○：複数の公共 RDF データと統合されている
データ・ポータルの開発	△：まだ全機能は実装されていない
公共ポータルにおける公開	×：公共ポータルにはまだ登録されていない

5.3. 将来展望

本研究では，ゲノム情報解析全体の再現性と再利用性を向上させるために，データ処理とデータ統合という二つの段階に分けて，それぞれに課題を設定して取り組んだ．これらの段階はそれぞれ，オープン・ソースの解析ツールと公共データベース上のオープン・データを用いている一方，再現性と再利用性の向上のためにはワークフロー管理システムやリンクト・データを活用する必要があることがわかる．これらは手法の共有を目指すオープン・メソドロジーのための技術要素であり，再現性と再利用性を前提としたデータ駆動型研究に必須であると考えられる．

一方，オープン・メソドロジーを考える上で，データ処理とデータ統合は相互に関係しているため，ゲノム情報解析のフロー全体における再現性と再利用性を再度考慮する必要があるだろう．例えば，データ・ポータル上の特定のサンプルの情報がどの次世代シークエンサーを使用してどのようなデータ処理によって得られた縮約データであるかを遡って知る必要もある．この場合，特定のサンプルという情報とデータ処理ワークフローの情報がデータベース上で統合されている必要がある．データ処理ワークフローのデータ形式については，複数のプラットフォームの開発者によって RDF を用いたワークフロー記述の標準（Common Workflow Language）を策定するといった試みもなされており[87]，将来的にはデータ統合が可能になるだろう．

本研究では，再現性と再利用性を向上させる情報基盤として，仮想マシンを使ったワークフロー管理システムの運用やセマンティック・ウェブを使ったゲノム情報の統合を提案した．今後，実際にこれらの手法を利用して，多くの研究者のための情報基盤を構築するためには技術的な進展も必要だろう．例えば，大量のデータ処理を伴うデータ解析手法を解析ツールの変更の度にテストすることのできるクラウド環境，また，セマンティック・ウェブで統合された多数のデータセットを高速に検索するデータベース，などは今まさに急ピッチで開発が進んでいる技術である．これらの新しい技術をいち早く活用すると共に，再現性のある研究にインセンティブを与える取り組みや，オープン・ソースやオープン・データに貢献するコミュニティの構築を推進することで，大量のゲノム情報を有効活用するオープン・サイエンスの枠組みが構築されるだろう．

6. 謝辞

本研究を進めるにあたり，様々なご指導を賜りました，東京大学 先端科学技術研究センター 油谷浩幸教授に心より御礼申し上げます.

東京大学 先端科学技術研究センター 森川博之教授，井原茂男教授，谷内江望准教授，ライフサイエンス統合データベースセンター 山口敦子准教授には，ご多忙にもかかわらず本学位審査をご快諾頂き，貴重なご意見を戴きましたこと，深く御礼申し上げます.

また，日常の議論を通じて多くの知識や示唆を戴いた東京大学 先端科学技術研究センター ゲノムサイエンス分野の皆様，理化学研究所 統合生命医科学研究センター 疾患システムモデリング研究グループの皆様，ライフサイエンス統合データベースセンターの皆様に感謝申し上げます.

7. 引用文献

[1] Eric S. Lander, et al. "Initial sequencing and analysis of the human genome." Nature 409.6822 (2001): 860-921.

[2] Kris Wetterstrand. "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)." Web. 20 Aug. 2015. <http://www.genome.gov/sequencingcosts/>.

[3] "我が国におけるオープンサイエンス推進のあり方について." Web. 19 Aug. 2015. <http://www8.cao.go.jp/cstp/sonota/openscience/150330_openscience_summary.pdf>.

[4] Peter Kraker, et al. "The case for an open science in technology enhanced learning." International Journal of Technology Enhanced Learning 3.6 (2011): 643-654.

[5] Peter Binfield. "Open Access MegaJournals – Have They Changed Everything?" Creative Commons. Web. 20 Aug. 2015. <http://creativecommons.org.nz/2013/10/open-access-megajournals-have-they-changed-everything/>.

[6] Tam P. Sneddon, et al. “GigaDB: Promoting Data Dissemination and Reproducibility.” Database: The Journal of Biological Databases and Curation 2014 (2014): bau018. PMC. Web. 20 Aug. 2015.

[7] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. "The sequence read archive." Nucleic acids research (2010): gkq1019.

[8] "Availability of Data, Material and Methods." Web. 20 Aug. 2015. <http://www.nature.com/authors/policies/availability.html>.

[9] Alawi A. Alsheikh-Ali, et al. "Public availability of published research data in high-impact journals." PloS one 6.9 (2011): e24357.

[10] "Rebooting review." Nature Biotechnology. 33(4):319. 2015

[11] Cameron Neylon, et al. "Changing computational research. The challenges ahead." Source code for biology and medicine 7.2 (2012): 2.

[12] Roger D Peng. "Reproducible research and Biostatistics." Biostatistics 10.3 (2009): 405-408.

[13] Antony T. Vincent, and Steve J. Charette. "Who qualifies to be a bioinformatician?." Frontiers in genetics 6 (2015).

[14] Sudhir Kumar, and Joel Dudley. "Bioinformatics software for biologists

in the genomics era." Bioinformatics 23.14 (2007): 1713-1717.

[15] Allegra Via, et al. "Best practices in bioinformatics training for life scientists." Briefings in bioinformatics (2013): bbt043.

[16] Takeru Nakazato, Tazro Ohta, and Hidemasa Bono. "Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive." (2013): e77910.

[17] "HiSeq 2500 Specifications." Illumina. Web. 23 Aug. 2015. <http://www.illumina.com/systems/hiseq_2500_1500/performance_specifications.html>.

[18] Xosé M. Fernández-Suárez, Daniel J. Rigden, and Michael Y. Galperin. "The 2014 nucleic acids research database issue and an updated NAR online molecular biology database collection." Nucleic acids research 42.D1 (2014): D1-D6.

[19] Eleni Giannoulatou, et al. "Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie." BMC bioinformatics 15.Suppl 16 (2014): S15.

[20] Jason O'Rawe, et al. "Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing." Genome med 5.3 (2013): 28.

[21] Andy Rimmer, et al. "Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications." Nature genetics 46.8 (2014): 912-918.

[22] Lincoln D. Stein. "The case for cloud computing in genome informatics." Genome Biol 11.5 (2010): 207.

[23] Joel T. Dudley, and Atul J. Butte. "In silico research in the era of cloud computing." Nature biotechnology 28.11 (2010): 1181-1185.

[24] Bill Howe. " Reproducibility, Virtual Appliances, and Cloud Computing." Implementing Reproducible Research. CRC/Taylor and Francis, 2014. Print.

[25] Carole A. Goble, et al. "myExperiment: a repository and social network for the sharing of bioinformatics workflows." Nucleic acids research 38.suppl 2 (2010): W677-W682.

[26] Khalid Belhajjame, et al. "Using a suite of ontologies for preserving workflow-centric research objects." Web Semantics: Science, Services and Agents on the World Wide Web (2015).

[27] Belinda Giardine, et al. "Galaxy: a platform for interactive large-scale genome analysis." Genome research 15.10 (2005): 1451-1455.

[28] Kate R. Rosenbloom, et al. "The UCSC genome browser database: 2015 update." Nucleic acids research 43.D1 (2015): D670-D681.

[29] Kai Wang, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic acids research 38.16 (2010): e164-e164.

[30] Davis J. McCarthy, et al. "Choice of transcripts and software has a large effect on variant annotation." Genome medicine 6.3 (2014): 1-16.

[31] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols 4.1 (2008): 44-57.

[32] Junjun Zhang, et al. "International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data." Database 2011 (2011): bar026.

[33] Syed Haider, et al. "BioMart Central Portal—unified access to biological data." Nucleic acids research 37.suppl 2 (2009): W23-W27.

[34] Chris Stark, et al. "BioGRID: a general repository for interaction datasets." Nucleic acids research 34.suppl 1 (2006): D535-D539.

[35] Richard N. Smith, et al. "InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data." Bioinformatics 28.23 (2012): 3163-3165.

[36] Xiaoshu Wang, Robert Gorlitsky, and Jonas S. Almeida. "From XML to RDF: how semantic web technologies will change the design of'omic'standards." Nature biotechnology 23.9 (2005): 1099-1103.

[37] Natalya F. Noy, et al. "BioPortal: ontologies and integrated data resources at the click of a mouse." Nucleic acids research (2009): gkp440.

[38] Richard Côté, et al. "The ontology lookup service: bigger and better." Nucleic acids research 38.suppl 2 (2010): W155-W160.

[39] Hongyan Wu, and Atsuko Yamaguchi. "Semantic Web technologies for the big data in life sciences." Bioscience trends 8.4 (2014): 192-201.

[40] Eric Jain, et al. "Infrastructure for the life sciences: design and implementation of the UniProt website." BMC bioinformatics 10.1 (2009): 136.

[41] Shin Kawano, et al. "TogoTable: cross-database annotation system using the Resource Description Framework (RDF) data model." Nucleic acids research (2014): gku403.

[42] "List of RNA-Seq Bioinformatics Tools." Wikipedia. Web. 21 Aug. 2015. <https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools>.

[43] Galaxy Tool Shed. Web. 21 Aug. 2015. <https://toolshed.g2.bx.psu.edu/>.

[44] Kent Beck. Extreme Programming EXplained: Embrace Change. Reading, MA: Addison-Wesley, 2000. Print.

[45] Tim Berners-Lee. "Linked Data." 27 July 2006. Web. 20 Aug. 2015. <http://www.w3.org/DesignIssues/LinkedData.html>.

[46] K. Jarrod Millman, and Fernando Perez. "Developing Open Source Scientific Practice." Implementing Reproducible Research. CRC/Taylor and Francis, 2014. Print.

[47] Michael R. Berthold, et al. "KNIME: The Konstanz information miner." Data analysis, machine learning and applications. Springer Berlin Heidelberg, 2008. 319-326.

[48] Tom Oinn, et al. "Taverna: a tool for the composition and enactment of bioinformatics workflows." Bioinformatics 20.17 (2004): 3045-3054.

[49] Jeremy Goecks, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol 11.8 (2010): R86.

[50] Daniel Blankenberg, et al. "Galaxy: a web ‐ based genome analysis tool for experimentalists." Current protocols in molecular biology (2010): 19-10.

[51] Samik Ghosh, et al. "Software for systems biology: from tools to integrated platforms." Nature Reviews Genetics 12.12 (2011): 821-832.

[52] Alexandre G. de Brevern, et al. "Trends in IT Innovation to Build a Next Generation Bioinformatics Solution to Manage and Analyse Biological Big Data Produced by NGS Technologies." BioMed Research International 2015 (2015).

[53] "Galaxy Wiki: Statistics about the Galaxy Project." Web. 20 Aug. 2015. <https://wiki.galaxyproject.org/GalaxyProject/Statistics>.

[54] James Taylor. "Adventures in Scaling Galaxy" Web. 20 Aug. 2015. <https://wiki.galaxyproject.org/Documents/Presentations?action=AttachFile&do=view&target=2014_Taylor_BioData_Scaling.pdf>

[55] "Galaxy Wiki: Virtual Appliances." Web. 20 Aug. 2015. <https://wiki.galaxyproject.org/VirtualAppliances>.

[56] "DBCLS Galaxy." Web. 20 Aug. 2015. <http://dbcls.rois.ac.jp/services/sequencing/dbclsgalaxy>.

[57] "ChEMBL Virtual Machine." The ChEMBL-og. Web. 22 Aug. 2015. <http://chembl.blogspot.co.uk/2013/10/chembl-virtual-machine-aka-mychembl.html>.

[58] Robert C. Gentleman, et al. "Bioconductor: open software development for computational biology and bioinformatics." Genome biology 5.10 (2004): R80.

[59] Raymond, Eric S. The Cathedral & the Bazaar. 2nd ed. Sebastopol: O'Reilly Media, 2008. Print.

[60] Richard Stallman. "The GNU Operating System and the Free Software Movement." Open Sources : Voices from the Open Source Revolution. O'Reilly, 1999. Print.

[61] "Genomics Virtual Lab." Web. 20 Aug. 2015. <https://genome.edu.au/>.

[62] DDBJ P-Galaxy. Web. 20 Aug. 2015. <https://p-galaxy.ddbj.nig.ac.jp/>.

[63] "Galaxy Series: Data Intensive and Reproducible Research" Giga Science Article Collection. Web. 20 Aug. 2015. <http://www.gigasciencejournal.com/series/Galaxy>.

[64] Thomas J. Hudson, et al. "International network of cancer genome projects." Nature 464.7291 (2010): 993-998.

[65] John N. Weinstein, et al. "The cancer genome atlas pan-cancer analysis project." Nature genetics 45.10 (2013): 1113-1120.

[66] Tim Berners-Lee, James Hendler, and Ora Lassila. "The semantic web." Scientific american 284.5 (2001): 28-37.

[67] François Belleau, et al. "Bio2RDF: towards a mashup to build bioinformatics knowledge systems." Journal of biomedical informatics 41.5 (2008): 706-716.

[68] Simon Jupp, et al. "The EBI RDF platform: linked open data for the life sciences." Bioinformatics 30.9 (2014): 1338-1339.

[69] Toshiaki Katayama, et al. "BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains." Journal of biomedical semantics 5.1 (2014): 1-13.

[70] "Semantic Web Stack." Wikipedia. Web. 22 Aug. 2015. <https://en.wikipedia.org/wiki/Semantic_Web_Stack>.

[71] UniProt SPARQL Endpoint. Web. 22 Aug. 2015. <http://beta.sparql.uniprot.org/>.

[72] Natalya F. Noy, et al. "Creating semantic web contents with protege-2000." IEEE intelligent systems 2 (2001): 60-71.

[73] Bernadette Hyland, et al."Best practices for publishing linked data." W3C Working Group Note 09 January 2014.

[74] BioHackathon 2014. Web. 22 Aug. 2015. <http://2014.biohackathon.org/>.

[75] "RDFizing Database Guideline" Togo Wiki. Web. 22 Aug. 2015. <http://wiki.lifesciencedb.jp/mw/RDFizingDatabaseGuideline>.

[76] Matthias Hert, Gerald Reif, and Harald C. Gall. "A comparison of RDB-to-RDF mapping languages." Proceedings of the 7th International Conference on Semantic Systems. ACM, 2011.

[77] Christian Bizer, and Andy Seaborne. "D2RQ-treating non-RDF databases as virtual RDF graphs." Proceedings of the 3rd international semantic web conference (ISWC2004). Vol. 2004. Hiroshima: Citeseer, 2004.

[78] Marcelo Arenas, et al. "A Direct Mapping of Relational Data to RDF." W3C Recommendation 27 September 2012.

[79] Souripriya Das, Seema Sundara, and Richard Cyganiak. "R2RML: RDB to RDF Mapping Language." W3C Recommendation 27 September 2012.

[80] D2RQ Mapper. Web. 22 Aug. 2015. <http://d2rq.dbcls.jp/>.

[81] Tina A. Eyre, et al. "The HUGO gene nomenclature database, 2006 updates." Nucleic acids research 34.suppl 1 (2006): D319-D321.

[82] David Croft, et al. "The Reactome pathway knowledgebase." Nucleic acids research 42.D1 (2014): D472-D477.

[83] Daniel Barrell, et al. "The GOA database in 2009—an integrated Gene Ontology Annotation resource." Nucleic acids research 37.suppl 1 (2009): D396-D403.

[84] Toshiaki Katayama. "D3SPARQL: JavaScript Library for Visualization of SPARQL Results." CEUR-WS 2014, Vol-1320, paper 39.

[85] https://github.com/ryotas/sparql-gateway

[86] "Serializing SPARQL Query Results in JSON." Web. 21 Aug. 2015. <http://www.w3.org/TR/rdf-sparql-json-res/>.

[87] Peter Amstutz, Nebojša Tijanić "Common Workflow Language, Draft 2" Web. 21 Aug. 2015. <http://common-workflow-language.github.io/draft-2/>.