

プロジェクト名(タイトル):

## Development of machine learning techniques for DNA sequencing data

利用者氏名:

〇二階堂愛(1)、尾崎遼(1)、露崎弘毅(1)、市川巧(1)、芳村美佳(1)

理研における所属研究室名:生命機能科学研究センター バイオインフォマティクス研究開発チーム

<p>1. 本プロジェクトの研究の背景、目的、関係するプロジェクトとの関係</p> <p>次世代 DNA シーケンサー(NGS)は大量のデータを出力するが、そのデータから知識を取り出すには大規模な計算が必要となる。また NGS は生命現象の様々な階層(RNA, DNA, クロマチン状態)の情報を出力する。これらの情報をいかに統合し新規知見に結びつけるかが課題となる。このような解析を実現するには多様なソフトウェアやデータベースを複雑に組み合わせて、解析ワークフローを実装・運用する必要がある。またそれぞれのデータベースやソフトウェアはバージョンアップがあり、ワークフローに改良や結果の評価が頻繁に必要となる。ゲノムデータは時間に対して定常的にデータが得られるわけではなく、生物や実験の都合によって突発的に得られるため、計算環境も突発的に用意する必要がある。このようなゲノムデータの解析環境を取り巻く問題を解決するために、データ解析ワークフローの標準化、計算環境のソフトウェアによる自動構築などに取り組む。</p>	<p>をソフトウェアから操作し、必要な計算資源を動的にオンデマンドに得るプログラムの開発を継続した。</p> <p>これまで開発してきた RamDAQ を用いて、数百検体の 1 細胞 RNA-seq データを計算したところ、ファイルシステムの inode が枯渇することが明らかになった。研究室で所有する ZFS のファイルサーバを持つ PC クラスタではこのような現象は起きていない。オミクス解析では、多数の検体がそれぞれ異なる入力ファイルとなり、そのファイルから計算に伴って、多数の出力ファイルを生成するが、このような性質は多くのオミクス解析に共通する性質である。また HOKUSAI SS と BW でファイルシステムは共通であるため BW でも同様の問題が生じると思われる。</p> <p>さらに昨年度開発した高出力型 1 細胞 RNA-seq のデータ解析パイプラインである Q2-Pipeline のアップデートを行った。特に inode が枯渇しないよう出力ファイル数を減らす機能を追加した。</p>
<p>2. 具体的な利用内容、計算方法</p> <p>ワークフローを標準ワークフロー言語 NextFlow によって実装する。ソフトウェアについてはコンテナ仮想技術を利用して実行再現性やポータビリティを担保する。これらの技術を用いて 1 細胞 RNA-seq のデータ解析ワークフローを開発する。HOKUSAI SellingShip 上に DevOps 技術を用いて自動的に計算環境を整えるソフトウェアを開発する。</p>	<p>4. まとめ</p> <p>ゲノムデータ解析環境を HOKUSAI 上に自動構築できるソフトウェアの開発を継続した。公開している 2 つのゲノムデータを対象としたワークフロー開発を継続した。HOKUSAI でたくさんのファイルを出力するオミクス解析ワークフローを実行するには、HOKUSAI のファイルシステムの inode 枯渇問題を解消する必要がある。</p>
<p>3. 結果</p> <p>1 細胞 RNA-seq のワークフローとして RamDAQ の開発を引き続き進めた。昨年度、標準ワークフロー言語 NextFlow の DSL2 によって再実装し公開したが、バグフィックスと機能追加を行った。昨年度同様に SellingShip の管理に使われている OpenStack</p>	<p>5. 今後の計画・展望</p> <p>引き続き、ワークフローの機能拡張を行う。特に inode が枯渇する問題を解消するために、ループバックファイルシステムを仮想的に構築するなどの工夫を行う。</p> <p>6. 利用がなかった場合の理由</p>