2022年度　利用報告書

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:

○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

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

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。

2.　具体的な利用内容、計算方法

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。

3.　結果

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。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。

4.　まとめ

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。

5.　今後の計画・展望

引き続き、ワークフローの機能拡張を行う。BDRを始め全理研でゲノムデータ解析を実施している研究センターや研究室に技術提供を行う。またほかのシーケンス技術のワークフローに実装をすすめる。

6.　利用がなかった場合の理由