平成 24 年度 RICC 利用報告書

課題名（タイトル）：

次世代シーケンサーのバイオインフォマティクス解析

利用者氏名：二階堂 愛
所属：神戸研究所 発生・再生科学総合研究センター 先端技術支援・開発プログラム ゲノミクス解析室 機能ゲノミクスユニット

１． 本課題の研究の背景、目的、関係するプロジェクトとの関係

本課題では、RICC 上に次世代シーケンサーのデータ解析パイプラインを構築することを目的とした。次世代シーケンサーは一度の実験で、数百 Gb のデータを出力する。このデータを高速にデータ解析し、生物学的な知識に結び付けるには、大量の計算リソースが必要とされる。次世代シーケンサーのデータは生体サンプルや遺伝子単位など細かいユニットに分割することができるため、計算も RICC のように大規模な CPU 数を誇るコンピュータで分散計算しやすいデータである。これらのことから、私は RICC を利用し、次世代シーケンサーデータに対して一連の解析が行えるようデータ解析パイプラインを構築することを目指した。特に、タンパク質がゲノム上のどの位置に結合しているかを調べることができる ChIP-seq のデータ解析パイプラインを RICC に構築する。さらに 1 細胞 RNA-Seq のデータ解析パイプラインの研究開発を行う

２． 具体的な利用内容、計算方法今年度は、1 細胞 RNA-Seq のデータ解析パイプラインを以下の 3 つのステップに分けて構築した。(1)前処理（データクオリティ評価、シーケンスデータのゲノムへのマッピング、データフィルタリングなど)、(2)発現量の定量化（転写産物の量をマッピングデータから推定する)、(3)転写量がゆらいでいる遺伝子を、機械学習を用いて取り出す。このステップを約 20 個以上の 1 細胞に対し Java や Ruby, C/C++, R で書かれたプログラムを Rake で繋ぎ合せ、Grid Engine によるジョブスケジューラを利用し、分散計算ができるように工夫した。ただし開発は RICC ではなく、ラボの PC クラスタを利用した。

３． 結果

約 20 細胞分の RNA-Seq データ（約 600 GB 相当）のデータ解析を分散して行うことができた。ただし、RICC がとても混雑しており、さらにデータが大量なので RICC に送る時間がかかることも問題となったため、開発や実行のテストは主に研究室の PC クラスタで行なわざるをえなかった。ただし原理的には RICC で実行可能なシステムになっている。

４． まとめ

1 細胞 RNA-Seq データ解析パイプラインを構築し運用した。

５． 今後の計画・展望

今年度は、ChIP-seq , RNA-Seq のデータ解析パイプラインを中心に実装・運用してきたが、今後は、エピジェネティクスを測定する ChIP-seq のデータ解析のパイプラインを開発する。また RICC の混雑状況をみて、テストも行いたい。

６． 利用がなかった場合の理由

RICC がとても混雑しており、さらにデータが大量なので RICC に送る時間がかかることも問題となったため、開発や実行のテストは主に研究室の PC クラスタで行なわざるをえなかった。