平成 24 年度 RICC 利用報告書

課題名（タイトル）：

次世代シークエンサーデータの解析

利用者氏名：○権藤 洋一, 村田 卓也, 福村 龍太郎
所属：筑波研究所 バイオリソースセンター バイオリソース関連研究開発プログラム
新規変異マウス研究開発チーム

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

ゲノム解析の先端分野において、次世代シークエンサーを用いた解析が行われはじめて来ている。当研究室では、ヒト疾患モデルマウスの開発を目指し、RIKEN ENU-based Gene-driven Mutagenesis を展開し、疾患モデルマウスリソースを開発している。すでに 10,000 系統を超えるマウスリソースを凍結精子およびゲノム DNA としてアーカイブ化した。1 系統は約 5000 の点突然変異をゲノム全体にランダムに持っており、総数 5 千万の点突然変異を蓄積したライブラリーとなっている。遺伝子をコードする領域はゲノムの 1～2%なのでコード領域だけでも約百万の変異が総数 3 万といわれる遺伝子に誘発されている。すなわち、1 遺伝子あたり平均 30 を超える点突然変異を、10000 系統の変異マウスライブラリーとして利用公開している。これまで、ユーザーの要望する遺伝子ごとに変異を持つ系統をこのライブラリーから個別にスクリーンして提供してきたが、次世代シークエンサーを用いて一気に各系統がもつ変異を検出しカタログ化することを目標としている。まずは、各ゲノムのコーディング領域 5 千万塩基対上に誘発されている点突然変異の検出から始めた。次世代シークエンサーが産出するデータは莫大であり、その解析には RICC がふさわしいと考え、2011 年 8 月に情報基盤センターに相談するところから始めた。実際、11 月に利用登録し、まずは解析プログラムの評価を行うことを当年度の目標とした。

２． 具体的な利用内容、計算方法

筑波研究所で稼働している次世代シークエンサー、SOLiD4(ライフテクノロジー社)に付属する、解析プログラム BioScope のライセンスが無償提供されるとの知らせを受け、RICC での稼働の可能性を調査した。また、CLC 社は、様々な次世代シークエンサーの機種とその産出データのフォーマットの違いを意識することなく、同一のプラットフォームで解析可能なプログラムをリリースしている（CLC Genomics Server）。この解析プログラムが RICC で稼働するかどうかも併せて調査した。

３． 結果

BioScope は、Linux ベースの PC クラスタで稼働するソフトウェアではあったが、RICC のジョブ・スケジューラーのもとで稼働させるためには膨大なプログラム変更が必要であることが判明したため、RICC への導入を断念した。また、CLC 社の CLC Genomics Server には、RICC での可動性の余地が残されたため、今後の調査継続案件となった。

４． まとめ

現状の次世代シークエンサーが産出するデータは、まだまだ「ジャンク」なものが多い。日常的に解析データを取り扱うこととは別の次元で、パラメーターを見直すなど、よりよい解析チューニングは必要と思われる。この目的においては、RICC の高速演算環境は極めて魅力的であり、RICC 上で稼働可能な解析パイプラインの構築を目指していくことは重要であると思われる。オープンソースのプログラムを含め、幅広い選択肢の中から最高の解析環境を構築できるよう、継続的な挑戦が必要である。

５． 今後の計画・展望

情報基盤センターとの定期的な情報交換の結果、理研内部に多数導入されている次世代シークエ

ンサーの存在と、次世代シークエンサーが生み出すデータの再解析を希望する潜在ユーザーの存在を共有することが出来たと思われる。今後とも RICC 上で稼働可能な再解析パイプラインの構築に向けて努力する所存である。

6． 利用がなかった場合の理由

現状は予備調査の段階であり、RICC を用いて具体的に再解析をさせるタイミングではなかったため。