平成 23 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
理研での所属研究室名：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テストラン（ベンチマークテスト）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

全体の流れは、オリジナルデータの RICC への転送、公開されているレファレンスマウスゲノム DNA 配列 30 億塩基対情報へのアラインメント/マッピング、レファレンス配列との違いの検出、検出された違いが誘発点突然変異かどうかの判定、という手順で解析を行う。そのためのプログラム群の実装と計算、データの排出と転送という手順になる。まず、商用プログラム CLC Genomic Server が RICC 上で稼働するかどうかを情報基盤センターに相談したところ、技術的に克服すべき課題が多いことを確認した。そこで、まずは、オープンソースプログラムを稼働させることに目標を切替え、第一弾として PerM と呼ばれるオープンソースプログラムを大容量メモリ計算機上で稼働させた。テストデータとして、実際の疾患モデル候補マウス系統シークエンスデータをこのプログラムに用いた。

３． 結果

バイオリソースセンターが有する次世代シークエンサーに付属する PC クラスタサーバでは、オリジナルデータを投入し、変異候補データを排出するまでに一週間程度かかっていた解析が、大容量メモリ計算機上を用いることで 3 時間程度に短縮することが分かった。

４． まとめ

次世代シークエンサーのデータ解析手法は、市販のもの、多様なオープンソースプログラムと多岐に渡り、使っているシークエンサー、生データの種類、解析目的に応じて、よりよい結果を得るためにはさまざまな組み合わせおよびそれぞれのパラメータチューニングによる試行と最適化が必要である。今までは、1 回の解析に一週間程度の解析時間がかかっていたため、条件を変えつぎつぎと再ランすることは現実的ではなかったが、一回の解析が 3 時間程度で終わることが分かっ

たので、条件設定が捗ることが期待できる。また、別途、一連の利用可能なソフトウェアが排出した変異候補を実際のゲノムに戻って実在するか実験的に検証したところ、明らかに変異として高いスコアが出ているにも拘らず、実際の解析したゲノムに存在しないことが少なからず見つかり、そういった偽陽性をもたらすそもそもの原因が最初のステップのアラインメント/マッピングに起因する可能性が極めて高いことがわかった。通常のゲノム配列解析の相同性検索は BLAST などを利用して行われるが、次世代シークセンサーデータは 1 配列データが 50bp 程度と極めて短く、かつ、1 ゲノムから 4 億タグずつ排出されているため、さらに高速なプログラムでなければ 30 億塩基からなる全マウスゲノムレファレンス配列へのアラインメント/マッピングは実質的に不可能である。その高速化実現のためのアルゴリズムによって間違ったアラインメント/マッピングが生じていると考えられる。

5． 今後の計画・展望

当研究室が有する、約 1 万系統の疾患モデルマウス候補から現在年間 50 系統ずつ次世代シーケンサーで解析する目処が今年度立った。1 系統からは 60～80GB のオリジナル配列データが産出されるので、現在の規模でも年間３～４TB のデータを RICC に転送し、保存し、解析し、RICC をもちいて効率よく行っていきたい。また、変異マウスゲノムを次世代シークエンサーで解析して得られた実際のオリジナルデータそのものを理研から即時公開し、より高速かつより精度の高いアラインメント/マッピングから変異検出まで計算できるプログラム開発のオープンコンペティションへと将来的に発展させていきたい。

6． RICC の継続利用を希望の場合は、これまで利用した状況（どの程度研究が進んだか、研究においてどこまで計算出来て、何が出来ていないか）や、継続して利用する際に行う具体的な内容

使い始めて間もないため、パラメーターのチューニングができていない。まだ試したい解析ツールもすべてテストできているわけではないため、まずは、テストデータでのチューニングを行い、当研究室が有する多数の解析データのバッチランへとつなげていきたい。

7． 一般利用で演算時間を使い切れなかった理由

（該当しない。）

8． 利用研究成果が無かった場合の理由

本報告書にもまとめているように、マウスゲノム配列のコーディング領域に生じた新規点突然変異には既存のプログラムを利用しても当チームが利用可能な PC マシーンでは 1 週間要していて、成果発表のためのデータ解析が滞っていた。そのために RICC 利用による高速化を期待して今年度利用申込をした。すでに 3 時間程度まで短縮できることがわかったので、これからパラメーターのチューニングなどを急ぎ、また、ヒトの解析では「真のリファレンス配列が欠如」（註参照）しているため、これまで全く認識されていなかった次世代シークエンシングソフトウェアの不備を量と質の両面から高速計算によって明らかに次第、成果発表をする予定である。まさに、RICC を用いることでこれまで不可能であった計算が実現可能であることがようやくわかったところで、これから実際の計算に利用していきたい。

註）公開されているマウスゲノム配列データ mm9 は標準実験系統である C57BL/6J のゲノムのものである。そして、変異マウスライブラリー構築に用いた系統もこの C57BL/6J 系統であり、mm9 として公開されているゲノム配列がそのまま「真のリファレンス配列」となっている。一方で、ヒトの解析ではそもそも「標準ゲノム」というものは存在せず実験動物的な表現を用いるとすべて「雑種」系統に相当する。ある一人のヒトのゲノム配列データはそのひとのゲノム配列そのものしか「真のリファレンス配列」とは成り得ない。（例外的にワトソン博士やベンター博士のゲノムは解読されており、両博士のゲノム解析に限っては真のレファレンス配列がある。さらに補足すると、ゲノム配列の情報解析のためのプログラム開発には、シークエンサーが実験的に排出するオリジナルデータとプログラムの精度を検証するためのリファレンス配列が必須であるが、これまで次

世代シークエンシングはヒトを中心に実施されており、個人情報保護の観点からオリジナルデータの公開が極めて限られている。加えて、開発したプログラムの精度を検証するための真のリファレンス配列がなく近似で検証するために、われわれが発見したようなアラインメント/マッピングの間違いになかなか気づかれなかったものと思われる。すなわち、理研から変異マウスゲノムの次世代シークエンシングオリジナルデータを即時公開し、解析プログラムのオープンコンペティションを開催することは、単に、理研が構築した変異マウスライブラリーのカタログ化の実現に留まらず、ヒトゲノム解析への波及効果も含めて大きな貢献をもたらすものと確信している。