平成 22 年度 RICC 利用報告書

課題名（タイトル）：

次世代シーケンサ出力の de novo 解析

利用者氏名：

〇吉田　拓広*
櫻井　哲也*
持田　恵一**

所属　　：

***　横浜研究所　植物科学研究センター**
メタボローム基盤研究グループ/ゲノム情報統合化ユニット
****　横浜研究所　植物科学研究センター**
メタボローム基盤研究グループ/ゲノム/機能開発研究グループ

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

次世代シーケンサが相次いで実用化され、出力される配列データが飛躍的に増加し、ゲノム研究の潮流となりつつある。これら配列データを統合し、それぞれの生物がもつゲノムの全体像を再構成する de novo アセンブル品質は、有用遺伝子探索などゲノム情報を用いた下流の研究の正否を決定する。高等生物ゲノムの de novo アセンブルには、大量の RAM をもつ計算環境が必要不可欠である。

２．　具体的な利用内容、計算方法

シークエンスアセンブルソフトウェア MIRA を導入し、次世代シーケンサから出力された約 80 ギガバイトの配列データを用いて de novo アセンブルを行った。

３．　結果

MIRA を用いて結果を得ることは出来なかった。

４．　今後の計画・展望

MIRA 以外の de novo アセンブルソフトを試す予定。具体的には、Velvet、SOAP 等。

５．　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

大量データを扱う場合、簡易利用でのメモリ使用容量最大値 120GB ではメモリが不足のため、暫定的にメモリを 480GB まで使用できるように設定変更をしていただいた。また計算時間も最長 72 時間まで延長してもらいジョブを投入したが、72 時間経過したところで Elapse 超過となりジョブが終了した。途中までしか実行できなかったテスト結果から、メモリは 150GB あれば収まることが判明したが、大量データをクエリとする場合どのくらいの経過時間設定が必要かが不明である。経過時間を設定内で終了させるためには、今後、大量データを規定時間内で処理が終了する最適なデータ量に分割するなどの工夫が必要である。また、MIRA 以外のアセンブルソフトが多数存在するので、これらのソフトの導入を試み、大容量メモリ計算機に適したアセンブルソフトの検証も行う予定である。

６．　利用研究成果が無かった場合の理由

簡易利用の最大経過時間 24 時間以内に計算が終わらなかった為。また、72 時間まで計算時間を延長してもらったが、この場合でも計算時間が足らなかった。