平成 30 年度　利用報告書

課題名（タイトル）：次世代シーケンサーを用いたバイオリソースの特性解析

利用者氏名：○井内聖（1）

理研における所属研究室名：（1）バイオリソース研究センター実験植物開発室

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

バイオリソース研究センター（BRC）実験植物開発室が保有しているバイオリソースの特性解析のために、ゲノム情報の取得を目指している。近年、ゲノム配列取得技術の急速な発展がおきており、シロイヌナズナ 1001 ゲノムプロジェクトや中国を中心とした多くの植物種の全ゲノム配列の解析が進行中である。また、ゲノム編集技術の登場に伴い、BRC が保存・提供するシロイヌナズナ野生系統を中心とした植物リソースの信頼性、及び付加価値を向上するために、全ゲノム配列情報が不可欠となっている。これまで、ショートリードと呼ばれる 500bp 以下の配列情報を用いた解析を研究室の PC でどうにか行ってきた。近年ロングリード次世代シーケンサーと呼ばれるシーケンサーの普及が進んできた。このロングリードの配列は、長いものでは100、000bpを超えるものが現れている。ロングリードの配列はゲノム情報を正しいものにするためには非常に重要であると考えている。当研究室では、バイオリソースのロングリードの配列を取得してより良いゲノム情報を整備しようと作業を進めている。ロングリードの配列を含む次世代シーケンサーデータの解析には、大きなコンピューター資源を必要とすることから、これら配列の解析を行うためにスーパーコンピュータシステムの利用が必須であると考えて解析を進めている。

2.　具体的な利用内容、計算方法

ショートリードの配列データーを解析するには De Bruijn グラフを用いたアセンブルを行うのが一般的である。一方ロングリードは overlap-layout-consensus というアプローチでアセンブルを行うのが一般的である。今回、ロングリードの配列を解析するために、後者の手法を用いてプログラムされているオープンソースソフトウェアである「Canu」と「Flye」を使った。これらソフトウェアを実行できるように、スーパーコンピュータシステムに設定を行った。

3.　結果

第三世代 DNA シーケンサーであるオックスフォードナノポア社のミニオンを使って、シロイヌナズナ（Arabidopsis thaliana）培養細胞株 T87 のゲノム配列を取得した。今回得られたデーターは 220,673 配列、1,473,846,473bp（シロイヌナズナゲノムの 12.28 倍）で、一つの配列の平均長さ 4,355 塩基（配列長さの中央値 N50=10,547ｂｐ、最長は 96,710bp）であった。この DNA シーケンスデーターを用いて、ゲノムアセンブルを「Canu」と「Flye」でを行った。利用環境は bwmpc（1 ノード；40 コア）で実行した。「Canu」は 35 時間、「Flye」は 30 分で処理を終えた。「Canu」からの最終結果は、コンティグ数 1,382 本（N50=102kｂｐ）で「Flye」からはコンティグ数 1,370 本（N50＝167ｋｂ）であった。

4.　まとめ

シロイヌナズナ（Arabidopsis thaliana）培養細胞株 T87 のゲノム配列を第 3 世代 DNA シーケンサーで取得し、アセンブルを行った。今年度は、「Canu」と「Flye」を使って、アセンブルができる環境をHOKUSAIアカウントに構築する目的を達成できた。同じデーターを使って、「Canu」と「Flye」で解析を実施した結果、出力されたコンティグ数には大差ないように見られたが、解析時間が 70 倍ほど異なっていた。出力されたコンティグ数は約 1300 本得られたが、理想的にはシロイヌナズナは 5+2 本である。今回のデーターでは、ゲノムの平均カバレッジが 12 倍程度と低かったためにアセンブルの効率が下がったのではないかと考えている。次年度はデーター量を増やしたアッセンブルを実施したい。

5.　今後の計画・展望

ロングリードの次世代 DNA シーケンサーのデーターを使って、アッセンブルの作業を実施できる環境を整えたが、今後、得られたアセンブルの結果としてのコンティグが正しいのか検証を行いたい。コンティグ上の遺伝領域の推定などを試みたい。第 3 世代 DNA シーケンサーの性能も日進月歩で向上し出力されるデーター量も増えることが予想される。次年度取得する予定のロングリードの DNA 配列の解析にも今回構築した解析環境を利用して、バイオリソースの付加情報の充実へつなげていきたいと考えている。