

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、正確な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的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(第三世代次世代シーケンサー(NGS))と呼ばれる DNA 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キロbp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確に解読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、ロングリードの配列を利用して植物バイオリソースの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を解析するには De Bruijn グラフを用いたアルゴリズムで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。ロングリードの配列を解析するために、後者のアルゴリズムを中心に利用し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 (<https://canu.readthedocs.io/en/latest/>)」と「Flye (<https://github.com/fenderglass/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い、取得したバイオリソースの解析に用いた。

3. 結果

ゲノムサイズが大きいシロイヌナズナ野生系統の配列解析

を行うためには、微生物のゲノム解析より知見を得る戦略が有効と考えられる。そこで今年度は第三世代 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い、最近当室でシロイヌナズナ培養細胞株 T87 より単離した微生物のゲノム配列解析を行った。本微生物はシロイヌナズナの細胞と共生していると考えられ、共生研究を進める上でもリソースの品質管理上も、ゲノム配列に基づいた種同定が重要である。今回得られたデータは 24,508 配列、104,761,808bp(予想されている微生物のゲノムサイズの 26.19 倍)で、一つの配列の平均長さ 4,355bp(配列長さの中央値 N50=4,786bp、最長は 81,999bp)であった。この DNA シーケンサーデータを用いて、ゲノムアセンブルを「Canu」と「Flye」で行った。利用環境は bwmpc (1 ノード; 40 コア)で実行した。「Canu」は 27 分、「Flye」は 5 分で処理を終えた。「Canu」からの最終結果は、コンティグ数 1 本(4,882,132bp)で「Flye」からはコンティグ数 1 本(4,889,922 bp)であった。

4. まとめ

シロイヌナズナ (*Arabidopsis thaliana*) 培養細胞株 T87 と共生関係が疑われる微生物のゲノム配列を第 3 世代 DNA シーケンサーで取得し、アセンブルを行った。微生物ゲノムを「Canu」と「Flye」を使ってアセンブルを行った結果、それぞれから 1 本のコンティグが出力された。これらのコンティグを比較したところコンティグの端の位置が異なっており、これら 2 本のコンティグを用いると、微生物の環状ゲノム配列になった。また、各アセンブリからは 1 本のコンティグしか出力されなかったことから、当該の微生物にはプラスミド様の構造は存在しないと予想される。一方で、今回解析に用いた配列情報が予想ゲノム配列長の約 26 倍(104Mbp)程度であったことから、プラスミドの存在を完全に否定はできないとも考えている。次年度はデータ量を増やしたア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を使って、ゲノムアセンブルの作業を実施し、環状ゲノムの情報を取得した。ミニオンのデータはポリシーケンス(同じ塩基の連続)に弱いことが報告されており、実際得られた配列には塩基の欠如が認められる部分も存在する。今後ポリシ

2019年度 利用報告書

一ケンスに強いイルミナのショートリード配列を取得して、今回得られたゲノム配列の修正を行いたい。また、コンティグ上の遺伝領域の推定などを試みたい。次世代 DNA シーケンサーの性能は向上し出力されるデータ量も増えることが予想される。今後シロイヌナズナのゲノム DNA 配列の解析にも構築中の解析環境を発展させて、バイオリソースの正確な付加情報の充実へつなげていきたいと考えている。