

プロジェクト名(タイトル):

Linked-short reads を用いた de novo ゲノムアセンブリ

利用者氏名:

○安岡 有理(1)

理研における所属研究室名:

(1)生命医科学研究センター応用ゲノム解析技術研究チーム

1. 本プロジェクトの研究の背景、目的、関係するプロジェクトとの関係

linked-short reads とは、長鎖 DNA に分子バーコードを付与してからライブラリ作成をすることで、ショートリードでありながら疑似的なロングリードシーケンスを実現する技術である。この技術を用いることで、単一のショートリードライブラリに由来するシーケンスデータから高品質な *de novo* ゲノムアセンブリや、ハプロタイプレベルのゲノム解析が可能となる。本研究では、シロアゴガエル、アフリカツメガエル、ネッタイツメガエル、ヒトの linked-short reads データを用いて *de novo* ゲノムアセンブリを行う。それによって、リファレンスゲノムのないシロアゴガエルはまずはリファレンスゲノムの作成、リファレンスゲノムのある他の動物はハプロタイプの同定を行う。アフリカツメガエルは性染色体領域がすでに同定されているので、その領域が linked-short reads で同定できるかによって、本手法の有効性を検証する。また、ネッタイツメガエルは性染色体領域が未同定なので、本手法による同定を試みる。ヒトについては、すでに実施済みのロングリードシーケンスの結果と比較することで、構造変異の同定に本手法が有効であるかを確認する。これらの計算のためには、1TB に及ぶ大容量メモリが必要とされ、一時ファイルの大きさも数 TB に及ぶため、スーパーコンピュータの利用が必要不可欠である。

2. 具体的な利用内容、計算方法

各動物の細胞・組織から抽出した長鎖 DNA を用いて linked-short read ライブラリ(TELL-seq を使用)を作成し、Illumina NovaSeq にてシーケンスデータを得る。専用解析ソフト(TELL-READ, TELL-SORT, TELL-LINK)を用いて、シーケンスデータから *de novo* ゲノムアセンブリや構造多型検出、フェーズ解析を行う。

3. 結果

本年度は利用しなかった。

4. 今後の計画・展望

データ量を少なくした test データで解析の感触をつかみながら、HOKUSAI BW の有効な利用方法を模索する。新たな知見が得られれば随時学会・論文報告を行う。

5. 利用がなかった場合の理由

大容量メモリ演算サーバ(ACSL)の最大経過時間が 48h と短いため、数日～数週間単位の計算が予想される *de novo* ゲノムアセンブリ解析には用いることができなかった。