

スギゲノムの解読と解析

Genome sequencing and analysis of Japanese cedar

東京大学新領域創成科学研究科メディカル情報生命専攻

藤野 健

研究成果概要

本研究では、京都大学化学研究所スーパーコンピュータシステムを利用し、スギゲノムの解読と解析に取り組んだ。

スギのゲノムサイズはフローサイトメトリーを用いた既存の解析で約 11 Gb という推定が得られており、これはヒトゲノム(約 3 Gb)に比べてはるかに巨大である。また、スギゲノムは極めて頻度の高い反復配列を持つことが予備解析から判明している。このため、スギゲノムの解析には大きな計算資源を必要とする。

我々はスギから高分子 DNA を抽出し、PacBio HiFi リードを得ることでゲノムアセンブリを行った。この結果、N50 で約 12Mb、合計で約 9 Gb のコンティグを得ることができた。さらに Omni-C 技術によって構築した Illumina ライブラリをシーケンシングした結果と合わせることで、コンティグの 97 % を 11 本の染色体上に配置することができた。

こうして得られたアセンブリを検証しアノテーションを行うにあたり、京都大学化学研究所スーパーコンピュータシステムを利用した。遺伝学的地図の作成に使用したマーカー配列をゲノムアセンブリ上にマップし対応関係を調べることでアセンブリの正確性を評価した。アノテーションは、RepeatModeler による反復配列の同定、HISAT2 や minimap2 による RNA-seq データのマッピング、BRAKER による遺伝子予測、StringTie や PASA によるゲノム上での transcriptome アセンブリ、EVIDENCEModeler によるコンセンサスアノテーションの作成という一連の手順で行われた。さらに transcriptome アセンブリ、反復配列ライブラリおよび既存配列データベース (SwissProt, TrEMBL, RefSeq, OrthoDB) をもとにしたアノテーションのフィルタリングを行った。BUSCO によりアノテーションに含まれる core gene set の割合を調べたところ約 91 % の完全性を持つという評価が得られた。

発表論文(謝辞あり)

Fujino, Takeshi, et al. "A chromosome-level genome assembly of a model conifer plant, the Japanese cedar, *Cryptomeria japonica* D. Don." *bioRxiv* (2023): 2023-02.