

スギゲノムの解読と解析

Genome sequencing and analysis of Japanese cedar

東京大学 大学院新領域創成科学研究科 メディカル情報生命専攻 笠原研究室 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。

スギ花粉症は日本で大きな社会問題となっており、その解決は喫緊の課題である。特に、高度成長期に植えられ、花粉を大量に放出している(伐採にはちょうど適した時期の)スギ人工林を伐採し新たに植林を行う(条件が悪い土地ではあるいは植えない)ことが重要である。しかし、輸入材と比べた場合に日本林業の商業的な採算性は厳しく、特に私有地のスギ人工林の植え替えを促進するには多額の補助金を注ぎ込む必要があり、財政事情が厳しい政府あるいは地方自治体では二の足を踏んでいた。

また、スギ花粉を出さないいわゆる無花粉スギ(雄性不稔スギ)はスギ花粉症への抜本的対策として期待されているが、自然変異体を利用しているため原因遺伝子は不明であり、各地の気候に適し商業的に優れた特性を持つ様々な雄性不稔スギを大量生産することはできない。これらの課題を解決するための基礎研究として我々はスギゲノム配列を解読している。スギゲノム(約 11 Gb)はヒトゲノム(約 3 Gb)の 3.8 倍の大きさを持ち、予備解析の結果、ヒトゲノムより遙かに解読が難しいゲノム構造(極めて頻度の高い反復配列)を持っていることが分かっており、ゲノム解読には予測不能な極めて大きな計算量を必要とする。

このため、スーパーコンピュータを用いてスギゲノム概要配列の構築を目的としたゲノムアセンブリを行なった。アセンブリにおいては Oxford Nanopore 社のシーケンシング技術で得られたデータを用いた。まず、ゲノム配列の読み取りによって得られた信号データから、ゲノムアセンブラの入力となる塩基配列を得るにあたり GPU ノードを利用した計算の高速化を図った。続いて、本年度得られたデータと昨年度までに得られていたデータを合わせて入力とし、wtdbg2 アセンブラでゲノム概要配列の構築を行なった。この際、数十パターンにわたる様々な条件で並列に計算を行った。

Oxford Nanopore 社のシーケンシングデータはエラー率が高いことが知られているため、ゲノム概要配列に対して Illumina 社のシーケンシングデータをマップし、HyPo によってエラー訂正を行なった。このようにして得られたゲノムアセンブリに対する評価として BUSCO による core gene set の検出数の計算や、別の実験で得られたマーカー配列のデータをもとにした遺伝地図とゲノムアセンブリとの整合性の調査などを行っており、これらの評価を通じて今後もゲノムアセンブリの高精度化、コンティグ長の増大を目指す予定である。