

論壇

次世代型シーケンサーの最近の進歩と植物ゲノム解析の現状

東京大学大学院 農学生命科学研究科教授

磯部祥子

植物ゲノム解析における技術の進化は、驚異的なスピードで進み続けている。2000年に高等植物で初めてゲノムが解読されたシロイヌナズナ、2004年に作物で初めてゲノム解読が達成されたイネにおいて、ゲノム解読プロジェクトは全世界、あるいは国家レベルの大きな事業であった。一方、そこから約四半世紀が経過した現在、植物の全ゲノム解読は一研究室単位で実施可能な解析となりつつある。

この大きな進展をもたらしたのは、次世代型シーケンシング (NGS: Next Generation Sequencing) 技術である。従来のサンガーシーケンシング法は、シーケンスの正確性には優れていたものの、解析速度やコストの面で大きな制約があった。しかし、2000年代初頭に登場した NGS 技術により、膨大な量の DNA 断片を並行して解析することが可能となり、ゲノム全体の解読速度とコスト効率を飛躍的に向上した。最初に商業化された NGS 技術は、ゲノム配列の読み取り長が短い「ショートリード」と呼ばれるデータを生成するタイプだった。この技術は高スループットと高精度な配列解析を特徴としており、ゲノム全体を迅速に解析することができる。NGS を利用した植物ゲノム解読の最初の報告は、2012年に発表されたトマトゲノムの解析だと考えられる¹。この研究では、NGS 技術の先駆けである Roche/454 シーケンシングプラットフォームによって得られたショットガンシーケンス配列と、サンガー法で得られた BAC および Fosmid ライブラリの配列を組み合わせることでゲノム解読が行われた。それ以降の約 20 年間で、およそ 800 種類近くの植物ゲノム解読が報告されており、その数は現在も増え続けている²。特に、2015 年頃から解析コストの低下に伴い普及が進んだロングリードシーケンサーが、ゲノム解読の爆発的な増加を後押しした。また、Hi-C ライブラリなど、ゲノム配列を結合 (スキャフォールドリング) するために適したライブラリ構築技術の進歩とも相まって、多くの植物種で染色体レベルのゲノム配列を構築できる時代へと突入した。

さて、「ショートリード」「ロングリード」に大別される NGS 技術であるが、それぞれのカテゴリで技術が革新し続けている。ショートリード解析においては、算出するデータ量の増大とそれに伴う解析コストの削減が 2020 年ごろまで続いた*。ショートリード配列は、主にゲノム配列変異の検出、遺伝子発現量解析 (RNA-Seq) のほか、作成するライブラリを工夫することで、ゲノム上のメチル化領域を検出する Chip-Seq やオープンクロマチン領域を特定する ATAC-Seq など、エピジェネティック解析にも用いられている。最近のショートリード解析

技術のトレンドとしては配列の高精度化があり、シングルセル解析など微細なサンプル由来の変異検出等で利用が進んでいる。現在はヒトの癌細胞における変異解析などで用いられることが多く、植物での利用は進んでいない状況であるが、今後、植物においてもシングルセル解析や Skim-Seq 解析（全ゲノム配列を低い重複度で取得し、インピュテーション解析等でゲノム全域にわたる変異を検出する技術）など、高精度ショートリード配列を用いた新たな解析が広がることが期待される。

一方、ロングリード解析は、Pacific Biosciences (PacBio) 社と Oxford Nanopore Technologies (ONT) 社が提供する解析プラットフォームを基盤として発展してきた。ロングリード配列は、主に全ゲノムアセンブル解析に大きく貢献している。技術が登場した当初は、得られる配列の品質が不十分でエラー率が高く、ショートリードを用いた配列補正が必要だった。しかし、PacBio 社が 2019 年に販売を開始した Sequel II システム以降、高精度なロングリード配列 (Hifi リード) の取得が可能となり、ショートリードによる補正が不要となった。

ロングリード技術の導入により、アセンブル配列の長さが大幅に拡張され、ハプロイドゲノムの構築が実現した。すなわち、ショートリードは通常 100~150 塩基長、最大でも 300 塩基長と非常に短いため、異なるハプロイドゲノムが混在している場合、それらを識別することができない。このため、NGS 解析の黎明期には、ゲノムのアセンブルにはホモ個体やハプロイド個体を用いなければ、アセンブル解析の高精度化は困難だった。一方で、ロングリード技術の導入により、ハプロイドゲノムの識別が可能となったことで、ヘテロ個体や高次倍数体など、異なるハプロイドゲノムが混在している個体の解析が可能となった。

さらに、バイオインフォマティクスの進歩も植物ゲノム解析において重要な役割を果たしている。例えば、2022 年頃より Hifi リードをアセンブルする際にしばしば用いられるツール「Hifiasm」が、ハプロタイプごとの配列を構築する機能を備えるようになったことで、ハプロタイプを識別してゲノムを構築する haplotype-resolved assembly が広く行われるようになった。このような技術進歩の結果、2024 年にはゲノム構造が非常に複雑である異数体である栽培サトウキビのゲノム解読が発表された。これは、アセンブル技術が一つの到達点に達したことを示す重要な成果であり、非常に感慨深いものがある³。

バイオインフォマティクス技術の進歩は解析にかかる計算コストの飛躍的な低下ももたらした。配列解析コストと計算コストの両者が低下したことで、ゲノム配列の解読は多くの研究機関が協力して行う大規模なプロジェクトではなく、個々の研究者が保有する材料について小規模なプロジェクトベースで実施することが可能になった。これにより、対象種において代表的な一つの参照ゲノムを構築するのではなく、品種・系統ごとにゲノムアセンブルを行い、それらを比較する pan-genome 解析の時代に突入している。さらに、以前はゲノム配列上の遺伝子予測が難しく、複数の手法を組み合わせるなど試行錯誤を必要とするケースも多かったが、最近では深層学習を活用したツールが登場し、高い精度で迅速に遺伝子を予測することが可能となっている。

一方、植物ゲノム解析には未解決の謎も多く残されている。例えば、ゲノムサイズが非常に大きい植物や、高度に複雑な遺伝子重複を持つ植物の解析は依然として難しい。高次倍数性種も、特に同質倍数体において同祖染色体の識別が難しいため、解析の難しさが残っている。さらに、エピゲノムの変化がどのように植物の形質や環境適応に影響を与えるかについての理解もまだ完全には得られていない。また、細胞壁の存在からシングルセル解析によるアプローチもヒトや動物に比べて遅れている。知れば知るほど次の謎が生じ、ゲノム解析の深化はとどまることがないが、それぞれの時点で得られたゲノム配列情報をもとに、例えば、特定の環境条件下での植物の適応メカニズムや、病害抵抗性を持つ遺伝子の特定など、実際の応用に直結する研究が進んでいる。これらは気候変動や環境問題に対処するための持続可能な農業の推進、バイオエネルギーの開発、新しい医薬品の創出など、幅広い分野での応用が期待されており、植物ゲノム解析の進展は、これからも私たちの生活に多大な影響を与える可能性を秘めている。

*<https://www.genome.gov/sequencingcosts/>

参考資料

- 1 Tomato Genome Consortium. 2012, The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485, 635-41.
- 2 Marks, R. A., Hotaling, S., Frandsen, P. B., and VanBuren, R. 2021, Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, 7, 1571-8.
- 3 Healey, A. L., Garsmeur, O., Lovell, J. T., et al. 2024, The complex polyploid genome architecture of sugarcane. *Nature*, 628, 804-10.