非モデル生物のシークエンス解析・インフォマティクス解析

中村 保一

国立遺伝学研究所 情報研究系
〒411-8540 静岡県三島市谷田 1111

Sequence and informatics analysis of non-model organisms

Yasukazu Nakamura

Department of Informatics, National Institute of Genetics
1111 Yata, Mishima, Shizuoka, 411-8540, Japan

Keywords: genome analysis, NGS, non-model species

DOI: 10.24480/bsj-review.15b2.00260

１．はじめに

この原稿では次世代シーケンサ（NGS）の出現とそれに伴う全ゲノム塩基配列決定法の変遷と現在の到達点を示し，最後にそれでもなおゲノム塩基配列を解釈する努力が必要であることを述べる。

２．全塩基配列決定プロジェクトの変遷

２−１．NGS 以前の塩基配列決定プロジェクト

まず，過去のキャピラリ型蛍光シーケンサを用いた植物のゲノム塩基配列「完全」決定手法について概観しよう。1996 年から，日欧米の研究機関が参加した国際共同プロジェクトにより，シロイヌナズナの全ゲノムの完全解読が進められた。このプロジェクトは，世界初の植物の全ゲノム構造解析としてその成果は 115 Mb, 25,498 遺伝子の決定論文として 2000 年 12 月 14 日号の Nature 誌に掲載された（AGI 2000）。このプロジェクトでは，各参加グループが若干の違いを持ちつつも，基本的には以下の手法が用いられた。まず，平均 80-100 kb 程度のインサート長をもつ P1 ないし BAC クローンライブラリを構築し，そのクローンの末端配列を使用して遺伝地図と物理地図を作成した。それらのクローンの挿入断片の塩基配列をショットガン法により短鎖で決定し，クローン単位でのアセンブル配列を完成させた。さらに，情報処理により遺伝子発見等のアノテーションを実施し，配列と解析情報を前述の地図情報に基づき染色体に沿って連結し染色体仮想的分子 pseudomolecule を再構築した。このように，全ゲノムショットガン（WGS）ではなく「クローンショットガン」の手法が必要だった理由は，現在のような全ゲノムショットガンを行うために必要な２つの要素技術，すなわち（1）全ゲノムで 10 以上の深度のリードを一気に読むことのできるハイスループットな塩基配列決定技術と（2）そのようにして得られた大量の塩基配列の断片を一気にアセンブルできる大容量かつ高速な計算機が当時存在しなかったことによる。

２−２．次世代シーケンサ（NGS）時代の塩基配列決定

２−２−１．NGS のインパクト

しかし，2007 年に出現した次世代シーケンサ（NGS）は塩基配列決定のあり方を劇的に変えた。NGS の配列決定法は，基板上での PCR による増幅とその増幅スポット上での配列合成の蛍光観察，または固相基盤上にあいたポアを核酸の単分子が通過する際の塩基により異なる電位差を検出するものであり，それまでの Sanger 法でのゲルという「液相」でサイズ分画していた配列決定法から「固相」をベースとする検出法により数桁オーダー密度を高めた配列決定を可能とし，高速かつコスト効率良く大量の塩基配列データを得ることができるようになった。筆者が事業に関わる DDBJ，ENA，GenBank からなる国際塩基配列データベース共同体（INSDC）が NGS のローデータアーカイブである SRA を始めた当初（2008 年頃）は，NGS で決定できる塩基配列長は最長で 30 塩基程度といった短いものしか得られなかったが，現在では PacBio あるいは Oxford Nanopore Technology 社による単分子シーケンサは数十キロベースオーダー以上の長鎖の解読を可能としており，とくに後者の配列決定長には理論上の限界が存在せず，サンプルの調整法でどれだけ長い DNA が得られるかが解析リードの長さを決定すると言われている。余談になるが当初，INSDC の SRA は Short Read Archive の略称をもってデータベース名としたが，後に同じ略称だが Sequence Read Archive に名称変更している。これは，解析長がかならずしも短鎖ではなくなっていたことと，受け入れる塩基配列の決定技術が NGS にほぼ完全に移行したことによるものである。

２−２−２．短鎖決定型 NGS による大量スキャフォールド時代

話を戻すと，NGS で長鎖が決定できなかった 2010 年代前半の「全塩基配列決定」プロジェクトの報告群を眺めてみると，この時代には植物・動物を問わず，計算機とアルゴリズムの進歩により一気に全ゲノムをアセンブルする WGS の試みは可能となったものの，リードとしては短鎖しか得られないため，そのアセンブルの完成度は低いものが多く，数百 Mb から数 Gb のゲノムに対し，本来の染色体の本数を遥かに上回る数万〜十万本単位のスキャフォールドからなる，まとまりの悪いアセンブル結果をもって発表されていることが多い。そこでは同じ生物から得た転写産物の 95% がそれらのスキャフォールド上にマッピングされるのであれば「コーディング領域の 95% を決定した。すなわち”ほぼ”全塩基配列が決定された」というロジックを用いており，今からみるとゲノム塩基配列プロジェクトとしては著しく完成度が低いと思える結果が散見される（もちろんこれらはテクノロジーの限界によるものでプロジェクト参加者の問題ではなく，またその生物種の遺伝子塩基配列の概観を与える等の成果によりその後の研究の発展に役だったものであり，単純に否定されるべきものではないことは言うまでもない。）

３．全塩基配列決定の現在

３−１．ロングリード時代のアセンブル

さて，本稿を執筆している現在では NGS 技術の進歩により，極めて高精度なゲノム塩基配列が低コストで得られるようになってきており，染色体 pseudomolecule の再現という意味では，大規模かつ莫大な資金により決定されたシロイヌナズナやヒトゲノムの塩基配列に比肩

しさらにそれを超える精度での配列決定が単独研究室レベルで可能となってきている。すなわち，数万以上のスキャフォールドから成る「理論上」完成した塩基配列決定の時代がおわり，再び pseudomolecule の再構築が可能になったとなると見るや，現在では染色体のテロメアからテロメアまでの真の全長を意味する「T2T 配列」が再構成されるようになり（Nurk et al. 2022），また一旦はハプロタイプの差異を解消して単一の配列にまとめ上げることを目指したアセンブル技術であったが，正確な長鎖配列を活用してアセンブルをすることにより，それぞれのハプロイドゲノムを分離して個別にアセンブルするフェージングや，また植物で過去にしばしば問題となった多倍数性のゲノムの正確なアセンブルも同様に，以前に比べ多くの資金を投入しなくても可能となりつつある。

３−２．非モデル生物のアセンブル事例

我々の最近の非モデル生物の *de novo* ゲノム塩基配列決定を，用いた技術と併せ紹介する。いずれも近縁種や同種での塩基配列決定事例はあったとはいえ，過去に大規模プロジェクトとして実施されてきたモデル生物の塩基配列決定事例ではなく，比較的小規模のグループにより，短期間に実施されたプロジェクトである。

３−２−１．イエネコゲノム

ネコは国内で 1,000 万頭近く飼育されている最も人気の高い伴侶動物であり，その健康を守るためのゲノム獣医療を推進するため，人気が高く遺伝的に多様性が高い猫種であるアメリカンショートヘア種を対象に高精度なゲノム配列の解読を目的として解読した。ゲノム塩基配列の高精度化のために当時利用可能な技術として Pacbio, illumina, Hi-C, 光学マッピングを活用したゲノムアセンブルを実施し，2020 年 5 月にプレプリントと塩基配列を公開した（Isobe et al. 2020）。塩基配列決定法としては，後述のアカシソやベンサミアナタバコで活用した Hifi 技術がまだ存在しなかったため，当時エラー率の高かった Pacbio のロングリード配列に illumina のリードを用いてエラー補正を行う手法を用いた。Hi-C と光学マッピングは共にスキャフォールドを行う手法である。Hi-C 解析は，核内で空間的に近い距離にある塩基配列同士を連結し，その DNA 断片ペアの塩基配列情報を NGS により網羅的に解読し膨大な DNA 断片ペアからゲノム上で近接関係にある配列ペアの情報を得る手法である。本来この手法はゲノムの三次元構造の研究のために開発された手法であるが，原理的に，アセンブル配列同士のスキャフォールディング（連続しない配列間の向きと距離をギャップを挟んで位置づけること）が可能となる。この手法により以前は困難であった染色体レベルの遠距離間のスキャフォールディングが可能となった。また，光学マッピングは配列特異的な制限酵素で蛍光標識した単分子のゲノム DNA をチップ上に整列させ，蛍光標識の出現パターンを画像解析することで直接観察による制限酵素地図（ゲノムマップ）を作成する手法である。これらの方法により，ほぼ全長の pseudomolecule から構成される 19 本の染色体を再現できた。解読されたゲノムの全長は 2.49 Gb であり，23,119 のタンパク質コード遺伝子を予測した。

３−２−２．アカシソゲノム

三島食品株式会社が育種を進めてきたアカシソ *Perilla frutescens* 「豊香 3 号」株について Pacbio HiFi reads を取得してアセンブルを実施，スキャフォールディングには Hi-C 解析を用い，併せて先行研究で報告された中国の青ジソ品種のゲノム配列とのシンテニーを利用することでシソゲノムの 20 本の染色体に対応した 20 本の pseudomolecule により，1.26 Gb の赤シソゲノムの 99.2% をカバーした（Tamura et al. 2022）。配列未決定領域であるギャップは 1 染色体あたり平均 4 か所以内，7 染色体はギャップの存在しない完全に再現されたと言える染色体塩基配列であった。この研究に於いて高精度な解読を可能としたポイントは PacBio 社の HiFi 技術である。HiFi はゲノム DNA を 10~20 kbp の環状 DNA とし，円環状の同一分子を繰り返し配列決定してランダムなエラーを除去することで，長鎖型シークエンサーの最大の欠点である読み取り精度の問題が解決されている。ここでは前項のネコゲノム解析で必要とした illumina リード による配列のエラー補正を実施していない。

３−２−３．ベンサミアナタバコゲノム

ベンサミアナタバコ（*Nicotiana benthamiana*）はナス科タバコ属に属し実験植物として長年利用されてきた。近年、とくに「接ぎ木」を成立させる機構の研究で注目されている（Notaguchi et al. 2020）。タバコ属植物は近縁種との交雑が繰り返され複雑なゲノム構造を持つため，それまで断片的なゲノム情報しか得られていなかった。本研究では，前項のアカシソゲノム決定同様，Pacbio Hifi リードを用いて *de novo* 全ゲノムアセンブリを行い，1,668 コンティグ，全長 3.1 Gb を構築した （Kurotani et al. 2023）。染色体の pseudomolecule と考えられる最長から 21 本目までのスキャフォールドに 2.8 Gb の配列が含まれ，アセンブルされたゲノム長の 95.6%を占めた。

３−３．それでも必要な地道な遺伝子「機能」予測の努力

上記の３プロジェクトはここ数年の間に行われたものであるが，その間も配列決定のコストは下がり続けており，リード取得だけのコストで言えば，これらの生物種の配列決定の規模感はすでに 100 万円を切っており，情報解析を担う共同研究先さえあれば，非モデル生物のゲノム塩基配列完全決定は研究室単位で実施できる研究となりつつある。昨今機械学習・人工知能技術を用いた遺伝子構造予測技術や機能予測技術も利用可能となってきており，より多くのゲノム塩基配列が低コストで基盤情報として得られるようになったことで，塩基配列の解釈という行為であるアノテーションも次第に簡便にできるようになってきた。しかし，配列決定後の一次アノテーションは，代表的かつ生物種間で保存性の高い遺伝子は網羅しているが，限られた組織や発生のステージでのみ発現する遺伝子は記載されていないことがままある。我々はフタホシコオロギ *Gryllus bimaculatus* の神経ペプチドの発現解析を実施する過程で，そのゲノムドラフト塩基配列上に，我々が解析したい遺伝子群のうち 2 遺伝子しかアノテーションされていないことに気づいた。そこで，昆虫で既知の 43 神経ペプチドをシードとして精度の高い配列のマッピングとその結果のアラインメントをもとに，長期間に渡る継続的なウェット研究者との検討・確認によるマニュアルキュレーションを実施することにより 41 の未記載の神経ペプチド遺伝子をアノテーションし，そのセットを昆虫の神経ペプ

チド研究に寄与する基盤データとして手順のワークフローと共に報告した（Mochizuki et al. 2023）。神経ペプチドのように時期・組織特異的に発現し，また最終産物が数ペプチドであるような発見しにくい（しかし興味深い現象を司るような）遺伝子群については，こうした地道な情報処理とその確認という手順を経て初めて，網羅的な実験のための基礎データが得られるのである。

４．おわりに

本稿では近年極めて短期間かつ低コストで実現できるようになった全ゲノム塩基配列決定により，これまで未確定であった非モデル生物のゲノムに光をあてることができるようになったことを紹介し，しかしそれでもなお「泥くさい」しかし実はそこが「面白い」生物学者と情報生命学者との密接な連携による，塩基配列上の意味を解釈する努力の必要性を述べた。本学会の非モデル植物研究者の皆様にゲノムを決定しまたそれを活用することを考えていただく一助となれば幸いである。

引用文献

The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408: 796–815. doi: 10.1038/35048692

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A et al. (2022) The complete sequence of a human genome. Science 376: 44–53. doi: 10.1126/science.abj6987

Isobe S, Matsumoto Y, Chung C, Sakamoto M, Chan T-F, Hirakawa H, Ishihara G, Lam H-M, Nakayama S, Sasamoto S et al. (2020) AnAms1.0: A high-quality chromosome-scale assembly of a domestic cat *Felis catus* of American Shorthair breed. BioRxiv, doi: 10.1101/2020.05.19.103788

Tamura K, Sakamoto M, Tanizawa Y, Mochizuki T, Matsushita S, Kato Y, Ishikawa T, Okuhara K, Nakamura Y, Bono H. (2022) A highly contiguous genome assembly of red perilla (*Perilla frutescens*) domesticated in Japan. DNA Res 30: 1–8. doi: 10.1093/dnares/dsac044

Kurotani K, Hirakawa H, Shirasawa K, Tanizawa Y, Nakamura Y, Isobe S, Notaguchi M. (2023) Genome Sequence and Analysis of *Nicotiana benthamiana,* the Model Plant for Interactions between Organisms. Plant and Cell Phys 64: 248–257. doi: 10.1093/pcp/pcac168

Mochizuki T, Sakamoto M, Tanizawa Y, Seike H, Zhu Z, Zhou YJ, Fukumura K, Nagata S, Nakamura Y. (2023) Best Practices for Comprehensive Annotation of Neuropeptides of *Gryllus bimaculatus.* Insects 14: 121–121. doi: 10.3390/insects14020121

Notaguchi M, Kurotani K , Sato Y, Tabata R, Kawakatsu Y, Okayasu K, Sawai Y, Okada R, Asahina M, Ichihashi Y et al. (2020) Cell-cell adhesion in plant grafting is facilitated by β-1,4-glucanases. Science 7: 698-702. doi: 10.1126/science.abc3710