ゲノム研究の歴史と技術革新

兼崎　友

古より生物学者たちは生命というメカニズムを解き明かそうと躍起になってきた．これまでの生物学上の研究トレンドの転換点は，大科学者たちの卓越したアイデアに加え，新しい技術との融合によって引き起こされることが度々あり，本稿では特にゲノム研究にまつわる数々の発見と技術の関わり，そして現在の生物学において不可欠の研究機器となった次世代シーケンサーの現状について，主に実験研究者側の視点から紹介してみたい．

DNAの発見とその構造解明に寄与した人々と技術

1866年にMendelが遺伝の法則を発表した少し後の1869年，チュービンゲン大学のMieschерは細胞核の大量分離法を研究する過程で「核酸」を発見した[1]．彼はサケの精子中にこの物質が多量に存在すること，精子の中に炭素，酸素，窒素，リンが含まれることを発見する．しかし，この時代はまだタンパク質こそが遺伝物質という意見が根強く，それ以上の概念の構築には至らなかった．その後，1944年，ロックフェラー研究所のAveryによる肺炎レンサ球菌を用いた形質転換実験[2]や1952年のHersheyとChaseによるT2ファージの感染実験[3]から，遺伝情報を担う本体がDNAであることが示されると，DNAの構造解析が生物学における大きな研究課題として認識されるに至った．1906年にTswettにより発明されたカラムクロマトグラフィーの技術[4]を改良し，1949年，ChargaffがDNAにおいてアデニンとチミンの総量が等しいこと，グアニンとシトシンの総量が等しいことを発見する[5]．また，1914年にLaueによって発見されたX線回折現象を応用したX線結晶構造解析の手法がDNA構造解析にも投入され，1950年頃よりアメリカのPaulingとCorey, イギリスのWilkinsらのグループがDNA分子の構造解析を開始する．そして1953年，ケンブリッジ大のCrickはハーバード大から訪れていたWatsonとともに，かの有名なDNAの二重螺旋構造モデルを発表した[6]. Crickが博士論文を執筆している最中に，Watsonは針金とブリキ板でDNAの構造模型を組み立てたとされている．よく誤解されるが，彼ら自身はX線結晶構造解析を行っておらず，あくまで当事のX線回折データとChargaffの規則を満たす構造モデルを提唱したに過ぎない．またWatsonとCrickの二重螺旋構造モデルは，ケンブリッジ大の女性研究者Franklinが撮った未発表のX線回折写真を無断で入手したことで完成した疑惑があることは余りにも有名である．話は逸れるが，生物学の研究に電子計算機が使用され始めるのもこの頃からで，微分方程式やタンパク質のX線結晶構造解析などに使われていた．

ゲノムの概念と塩基配列解読手法の整備

直径20ÅのDNA二重螺旋構造が明かされた後，生物学者たちはDNAの構造と組成の解明から，配列と機能の解明へ, すなわち「ゲノム解析」へと興味を移していく．ゲノム（genome）とは，遺伝子（gene）と「全体」を意味する接尾語（ome）を合わせた造語であり，1920年にWinklerによって配偶子が持つ染色体の一組として定義され[7]，1930年に木原が「ある生物をその生物足らしめるのに必須な遺伝情報」と再定義した概念である．DNA二重螺旋構造の発見以後，1956年のKornbergによるDNAポリメラーゼの単離[8]，1960年代のNirenbergらによるコドン表の完成，1977年に発表されたSanger法[9]，Maxam & Gilbert法によるDNA配列解析決定法[10]や1983年のMullisによるPCR法の確立など，いずれもノーベル賞が授与されたこれらの偉大な研究成果に加え，1987年にApplied Biosystems社が世界初の自動DNAシーケンサーABI 370を発売したことで，ついに大規模なゲノム解析の時代が始まることになる．特にSanger法の改良法である蛍光色素を用いたダイターミネーター法は，その後20年にわたり大規模ゲノム解析を牽引する手法となり，今も日常的に使用されている．また技術的な面からは，1930年代のTiseliusによる電気泳動法の発見[11]も，その後の改良を経てゲノム解析の歴史に大きな貢献を果たした．

ゲノム解析時代の幕開けと生物情報科学

初めて全ゲノム配列が決定されたのは，1976年のRNAウイルスのバクテリオファージMS2[12]，1997–1978年のDNAウイルスのバクテリオファージφX174である[13,14]．この頃の放射性同位体を用いたゲノム解析

著者紹介　東京農業大学生物資源ゲノム解析センター（研究員）

法では，被曝上限のため，1年間に1人の人間が決定できる塩基長は1 kb以下程度であったようである．1977年頃より，早くも「バイオインフォマティクス（生物情報科学）」という語が使われ始めたようで，1982年には塩基配列公開用のデータベース，EMBL data libraryとGenBankが相次いで整備され，最初は約600種類の塩基配列が収納された．その後，1988年にはNCBIが設立されている[15]．1995年，生物で初めて全ゲノムが解読されたのはゲノムサイズ1.83 Mbの真正細菌*Haemophilus influenzae*で，その後1997年末までに，大腸菌，枯草菌，シアノバクテリアなど10種類以上の細菌と出芽酵母のゲノムが決定された．そして1990年から2003年までに世界各国の協力により3000億円以上の予算を投じたヒトゲノム計画が実施され，多数のギャップ領域は残るものの，ついに約30億塩基対のヒトゲノムが解読された[16]．ヒトゲノム計画の支援のため，1998年にワシントン大のGreenとEwingにより配列断片の大規模アセンブル用プログラム「Phred」が開発された[17]．現在のゲノム解析機器の主流である次世代シーケンサーが検出する塩基の正確度（QV値：quality value）は今でもPhred quality scoreと呼ばれる数値で表現されることが多い．

ポストゲノム解析の時代

1990年代半ば以降は，ゲノム配列が決定されたモデル生物に対する興味が集中し，これらのゲノム上の個々の遺伝子の破壊や機能解析を行う「ポストゲノム解析」や「逆遺伝学的解析」というフレーズが一世を風靡した．ゲノム情報が整備されたモデル生物では，1995年のスタンフォード大のBrownによるマイクロアレイ解析系の確立[18]に端を発する網羅的な転写産物解析（トランスクリプトーム解析）をはじめ，プロテオーム解析，メタボローム解析，フェノーム解析といった細胞全体を相手にするさまざまなオミクス解析の手法が一斉に整備されていった．その一方，非モデル生物は実験系の整備・開発が進まず，モデル生物との情報の差が開いた時代でもある．

次世代シーケンス技術の登場と急速な淘汰

2005–2007年，Sanger法も電気泳動法も使わない新型シーケンサーが相次いで発表された．新しい塩基配列解読法よりもさら衝撃的だったのは，その解析速度と解読塩基量であった．2007年当時の性能でも，それまでのキャピラリー式シーケンサーの300倍以上の解読塩基量を誇ったことから，これらの新型シーケンサー群はまとめて「次世代シーケンサー」と呼ばれた．その後も改良や新型機種の登場により解読塩基量が年々増加し，2016年現在，単体で解読塩基量のもっとも多い機種であるHiSeq X（Illumina社）は，1回の稼働で最大1.8 Tbもの塩基配列を解読する（フローセル2枚分）．実に，ヒトのゲノム（3 GB）600人分もの塩基量である．一方で，各社の次世代シーケンサー間の競争と淘汰のスピードも凄まじく，1億円近い価格の機種がわずか7年程度で性能が時代遅れになり，生産停止に追い込まれるほどである．

2016年末現在，今後の発展性も考慮したうえで把握しておくべき代表的な次世代シーケンサーとしては，HiSeqシリーズ，MiSeq（Illumina社）などのショートリード型（1本のリード塩基長は300 base以下だが，大量の塩基配列を同時に解読可能）と，PacBio RSII，Sequel（Pacific Biosciences社），MinION，PromethION（Oxford Nanopore Technologies社）などのロングリード型（1本のリード塩基長が10 kb以上）の2タイプがある．表1に各次世代シーケンサーの特徴をまとめた．一塩基伸長法は4つの塩基に別々の蛍光標識をつけておき，一塩基の伸張反応を行うごとにレーザーで蛍光を検出する手法で，検出する塩基の精度が非常に高いのが特徴である．一分子シーケンス法は，ミクロレベルの孔内にDNA polymeraseが固定されており，その孔に1分子の長鎖DNA断片を入れて合成反応を行わせ，取り込まれた蛍光標識dNTPをリアルタイムで検出する手法である．長いDNAを読めるが，検出する塩基配列の精度は低い．イオン電流検知法は，ナノレベルの細孔を持つチャンネルタンパク質中を長鎖DNA分子が通過する際の極微小なイオン電流の変化から塩基を特定する手法である．この手法も長いDNAを読めるが，検出する塩基配列の精度は低い．ロングリード型を使う場合は十分なリード量を確保するか，あるいはショートリード型のデータで配列を校正するかといった工夫が必要となる．また，ロン

表1．主要な次世代シーケンサーとその解析原理・性能

機種	塩基解読手法	解読塩基長	解読塩基量
HiSeqX	一塩基伸長法	～150 × 2	～900 Gb
MiSeq	一塩基伸長法	～300 × 2	～20 Gb
PacBio RSII	一分子シーケンス法	10～50 kb	～800 Mb
Sequel	一分子シーケンス法	5～50 kb	～5 Gb
MinION	イオン電流検知	～200 kb	～5 Gb
サンガー法	Dyeターミネーター法	700	

*使用する試薬キットにより塩基長，解読塩基量ともに異なる．各社の改良により各数値は定期的に向上中である．

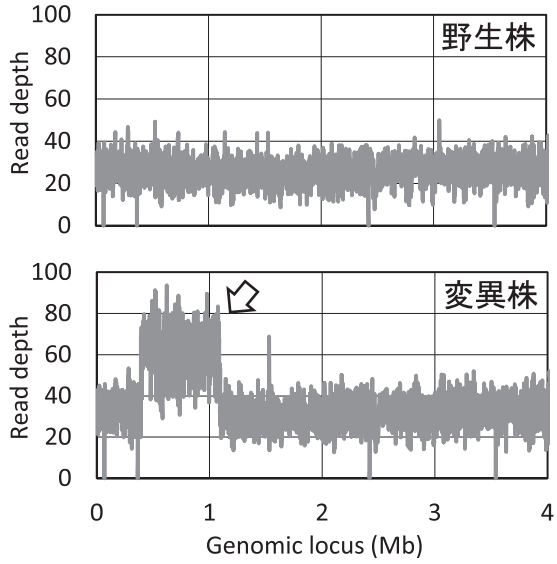

図1．ゲノム中に生じた重複領域の検出例．ゲノム情報既知の細菌の野生株と突然変異株についてMiSeqでリシーケンス解析した例．リードデータをゲノム情報にマッピングすると，リード深度の変化から矢印の領域が重複を起こしたことが分かる．重複領域末端のリード配列を詳細に解析すると，重複配列がゲノム上のどこに転位したかも分かる．

グリード型は新規のゲノム解析においてショートリード型のシーケンサーを圧倒する成績を示す．一方，ショートリード型は同時に解読できるリード本数が数十億本にも達するため，トランスクリプトーム解析やコピー数多型解析などの定量性を求める解析用途にきわめて有用である．図1にショートリード型シーケンサーでの解析例を示す．この全ゲノムレベルの変異解析が今やわずか1週間で完了する．

次世代シーケンサーという機器は，機器本体と維持管理費，保守契約費，解析を行うサーバーの維持費，高価な試薬代と人件費など，莫大なコストの問題を抜きにしては語れない．たとえばIllumina社のHiSeqシステムの場合，年間1000万円程度の保守費用を要求されるが，機器の故障頻度も高いため必要経費という印象である．正直に言って，個々の研究室レベルで維持するのは困難な機器であるため，設備と人材が揃った大規模なゲノム解析拠点を全国にいくつか整備し，設備更新しながら多様な研究分野を支援することが国策レベルで重要と思われる．

次世代シーケンサーで何ができるのか

次世代シーケンサーは一台の機械によりゲノム解読だけでなく，さまざまな核酸オミクス解析を可能にする点がきわめて優れている．これまでのDNAマイクロアレイなどの技術のように，生物種ごとに固有の解析用プラットフォームをカスタム製造する必要がないため，従来では不可能であったゲノム情報のない非モデル生物において，いきなりトランスクリプトーム解析をはじめとするオミクス解析を実施可能になった．次世代シーケンサーを用いた主要な解析手法としては，新規ゲノム解析，メタゲノム解析，リシーケンス解析，コピー数変異解析，トランスクリプトーム解析（RNA-seq），転写開始点解析（TSS-seq），DNA結合タンパク質やRNA結合タンパク質の核酸結合領域解析（ChIP-seq），エピジェネティクス（メチローム，ヒストン修飾）解析，特定遺伝子の多型解析（amplicon-seq），などがある．これらの解析手法の実用的な使用目的としては，たとえば，産業的に有用な微生物のゲノム情報の特許化，防疫や外来微生物の混入などの検査，作物の品種改良や連鎖地図の作製，がん関連などの特定遺伝子の変異やコピー数多型の検出，土壌や動物の腸内などの微生物叢に含まれる生物種の網羅的解析と有用微生物の探索，などがある．研究目的に応じ，wet，dryそれぞれの解析手法の改良法や解析プログラムが日々開発・発表されており，すでに膨大な数になっている．そのすべての解析法を網羅した経験のある研究者はおそらくいないため，新たに次世代シーケンサーを用いた解析を始める研究者は，それぞれの解析法に関するノウハウの情報収集が必須となる．たとえば，新規ゲノム解析時の*de novo* assemblerだけでも，Platanus，SOAPdenovo2，HGAP，Spade，MaSuRCA，Newbler，Celera Assembler，Velvet，ALLPATHS-LG，他多数のプログラムに加え，有償の汎用ソフトウェアとしてCLC genomics workbench, Strand-NGSなどがある．しかし，基本的にはシーケンス用のDNAライブラリ長よりも長い重複配列領域はアセンブル困難なので，異常によくつながった場合はミスアセンブルの可能性も考慮する必要がある．そのような場合，アセンブルして作った配列断片（contig）にリードデータをマッピングし，異常なマッピングパターンの有無や平均リード深度などの数値をチェックすると良い．またリード全体のk-mer値，あるいはGC含量の頻度分布グラフから，コンタミした生物やオルガネラ由来のリードの割合などを見積もることができる．RNA-seqやChIP-seqといった他の解析用途についても同様に多数のプログラムやノウハウが存在するため，それぞれ検討が必要である．

次世代シーケンサーの登場によって著しい進歩が見られた研究分野はいくつもあるが，特筆すべき分野として以下の二つをあげておきたい．第一は，進化と種内多様性（パンゲノム）解析，およびその応用分野である．次世代シーケンサー登場以前はコストパフォーマンスの問

題から，ゲノム解読された生物種の亜種や変異株の全ゲノム解析は重要視されなかったが，今では同属亜種のゲノム情報が大量に整備されることで，その生物属のアイデンティティを構成する遺伝子セットである「コアゲノム」や「最小ゲノム」が解明されつつあり，それを応用した最小ゲノム微生物の人工合成まで始まっている[19]．また，25年以上にわたる大腸菌の継代培養実験により集団内での遺伝的多様性の変遷や再現的な突然変異の発生現象が確認されたり[20]，病原性結核菌1687検体の全ゲノム解析から患者間での感染ネットワークが推定されたり[21]と，従来ではありえない規模でのゲノム解析結果が基礎と応用の両面で突然変異に関する新しい知見を生み出している．またアメリカで確立された実験用マウス系統（ヨーロッパ原産）のゲノム配列の10%が，日本の江戸時代の愛玩用マウス系統由来[22]である可能性が高いことが判明するなど，同一属内や同一種内のゲノムデータが増えることが文化や歴史的な発見にも貢献している．

第二は，メタゲノム解析である．従来の主力解析法であったDGGE法に対し，データ量と検出感度で勝る次世代シーケンサーの登場により，河川，海洋，土壌，空中，動物の腸内，糞便，醸造発酵環境など，あらゆる環境下での生物叢の解明が爆発的に進んでいる．その配列情報は，公共データベースに登録されることで次回の解析の精度を向上させる正のフィードバックループを形成しており，きわめて生物情報科学と相性の良い研究分野ともいえる．またQIIME，MG-RASTなどの配列解析パイプラインも便利である．たとえば，世界各国ごとの人々の腸内細菌叢の特徴や食物との関連などの研究は，さまざまな産業との関連性からもきわめて重要である[23]．また，近年注目されている「環境DNA解析」は，メタゲノム解析データに含まれる痕跡レベルの配列から，観測対象の生物種に出会えなくてもその種が周辺環境に生息しているどうかを判別する手法であり，今後の生態学の分野をより発展させそうである．加えて，環境中には単独培養不可能な未知の微生物群が圧倒的多数存在することは以前から知られていたが，それらが塩基配列レベルで次々と明らかにされつつある．ただし，サンプルの保存方法やDNA抽出法，DNA増幅法により結果に多大な影響が出る点はDGGE法の頃から変わっておらず，この点には注意を要する．

技術が拓くゲノム研究の今後

次世代シーケンサーとその周辺機器は未だ猛烈な勢いで開発が進んでおり，今後も性能が向上すると思われる．たとえば，10x Genomics社が開発したChromiumシステムのように，1細胞の単離からシーケンス用ライブラリ調製までを一つの機器で完結でき，かつショートリード型シーケンサーでも*de novo* assembly結果を改善できる周辺機器なども登場している．今後も技術革新により今の実験生物学者がイメージする研究計画の限界が次々と打ち破られることが予想される．いつの日か，「先生，ゲノム情報が未知の生物など地球上にいるのですか？」と学生に聞かれる日が来るのかもしれない．

文　献

1) Miescher, F.: *Medicinisch-chemische Untersuchungen*, **4**, 441 (1871).
2) Avery, O. T. *et al.*: *J. Exp. Med.*, **79**, 137 (1944).
3) Hershey, A. and Chase, M.: *J. Gen. Physiol.*, **36**, 39 (1952).
4) Tswett, M.: *Berichte der Deutschen botanischen Gesellschaft*, **24**, 316 (1906).
5) Chargaff, E. *et al.*: *J. Biol. Chem.*, **195**, 155 (1952).
6) Watoson, J. D. and Crick, F. H.: *Nature*, **171**, 737 (1953).
7) Winkler, H.: *Verbreitung und Ursache der Parthenogenesis im Pflanzen - und Tierreiche*, p. 166 (1920).
8) Kornberg, A. *et al.*: *Biochim. Biophys. Acta*, **21**, 197 (1956).
9) Sanger, F. *et al.*: *Proc. Natl. Acad. Sci. USA*, **74**, 5463 (1977).
10) Maxam, A. M. and Gilbert, W.: *Proc. Natl. Acad. Sci. USA*, **74**, 560 (1977).
11) Tiselius, A.: *Transactions of the Faraday Society*, **33**, 524 (1937).
12) Fiers, W. *et al.*: *Nature*, **260**, 500 (1976).
13) Sanger, F. *et al.*: *Nature*, **265**, 687 (1977).
14) Sanger, F. *et al.*: *J. Mol. Biol.*, 1**25**, 225 (1978).
15) Attwood T. K. *et al.*: *Bioinformatics—Trends and Methodologies*, Chapter 1, In Tech (2011).
16) International Human Genome Sequencing Consortium: *Nature*, **431**, 931 (2004).
17) Ewing, B. *et al.*: *Genome Res.*, **8**, 175 (1998).
18) Schena, M. *et al.*: *Science*, **270**, 467 (1995).
19) Huchison, C. A. *et al.*: *Science*, **351**, 1414 (2016).
20) Tanaillon, O. *et al.*: *Nature*, **536**, 165 (2016).
21) Guerra-Assunção, J. A. *et al.*: *eLife*, DOI: 10.7554/eLife.05166. (2015).
22) Takada, T. *et al.*: *Genome Res.*, **23**, 1329 (2013).
23) Nishijima, S. *et al.*: *DNA Res.*, **23**, 125 (2016).