The Society for Bioscience and Bioengineering, Japan

DDBJの新型シークエンサーのデータアーカイブと解析支援系の提供

中村　保一

大規模塩基配列情報とアーカイブ

新型シークエンサー　新型シークエンサーとは，従来型のサンガー法による塩基配列決定技術を用いた蛍光キャピラリーシークエンサーとの違いに着目して用いられている呼称である．一般には次世代シークエンサーと呼ばれることが多いが，幅広く利用されるようになって久しく，すでに「現世代」のシークエンサーであることから，筆者は新型シークエンサーと呼ぶことをすすめている．Next generation sequencerの略称でNGSと記載されていることが多いが，new generation sequencerでも同じ略称であるNGSが利用できるため混乱が少ないという理由もある．現在主流となっている新型シークエンサーは，1）固層のチップもしくは微小なビーズ上にDNAを固定，2）その上でPCRによる同一配列を持つDNAの増幅をおこない，3）DNAポリメラーゼまたはリガーゼによる逐次的DNA合成を行い，塩基伸長ごとの蛍光発光を高密度の色別スポット画像群として撮影し，4）同一スポットの時系列の蛍光色を解析することで，並列で大量の配列決定を行うものである．よく利用されている機材としては，illumina社のGenome Analyzer System，Roche社の454 GS FLX System，Life Technologies社のSOLiD Systemがある．

日本DNAデータバンク　日本DNAデータバンク（DDBJ, DNA Data Bank of Japan）は，国際塩基配列データベース協力体制（INSDC，International Nucleotide Sequence Databank Collaboration）[1]の一員である．研究者はINSDCを構成する三機関，日本のDDBJ，欧州EMBL-Bank（EBI），および米国GenBank（NCBI）のいずれかを通じて塩基配列データを登録することで，上記すべての塩基配列データバンクから配列情報を公開することが可能となっている（図1）．DDBJは，文部科学省からの運営費により，静岡県三島市の国立遺伝学研究所に設置された共同利用事業センターとして運営されており，データバンクの維持と改善に取り組んでいる．DDBJは，後述する大学共同利用施設である「国立遺伝学研究所スーパーコンピュータシステム」の運用もあわせて担当している．

Sequence Read Archive（SRA）の発足　新型シークエンサーの生データを共有し再利用性を高めるための新たなデータアーカイブがSRA[2]である．DDBJにおいても，以下に紹介するとおりINSDCの一員として欧米のパートナーとの協調のもとで新たなサービスとして構築・展開を始めている．

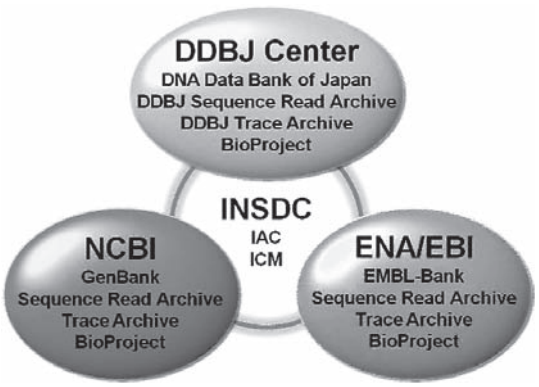

図1．国際塩基配列データベース協力体制図．INSDCは日本のDDBJ米国NCBIと欧州EBIのEuropean Nucleotide Archive（ENA）の三機関から構成されている．DDBJ，GenBank，EMBL-Bankは伝統的な完成配列の塩基配列データバンクである．SRAは新型シークエンサーの，TraceArchiveはサンガー法によるトレースデータを格納するローデータのアーカイブである．さらにプロジェクト情報を集中管理するBioProjectデータベースのデータ交換体制が始まっている．図中のIACは国際諮問委員会，ICMは国際実務者会議を示している．

SRAの現状　新型シークエンサー由来データの保存と共有を目的としたアーカイブであるSequence Read Archive（SRA）は2007にNCBIによって開始された．引き続きEBIとDDBJが参加することにより，INSDCとして国際的な協調体制にあるSRA事業が展開されてきている．DDBJは当初NCBI SRAへの代理登録を実施する形でSRA事業に参加してきたが，2009年に自らのアーカイブとしてDDBJ Sequence Read Archive（DRA）を立ち上げ，運営を始めた．SRAの公開登録データはINSDCの伝統的な塩基配列データ同様に，INSDCに参加している三機関でそれぞれ査定したのち登録を受け付け，相互に共有・公開している．図2に生物分類群ごとの，SRA登録データ量比の推移を示す．初期はヒト（Primates＝霊長類）の情報に限定されていたが，次第にさまざまな生物種の研究に新型シークエンサーが活用されるようになってきていることがわかる．とくに，オレンジ色で示されるEnvironmental samplesすなわち環境由来材料を培養することなく直接塩基配列を決定し，微生物群集のゲノムを偏りなく決定する研究が広く可能となっており，一定の大きさを占めてきていることに注

著者紹介　国立遺伝学研究所・総合研究大学院大学（教授）　E-mail: yn@nig.ac.jp

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

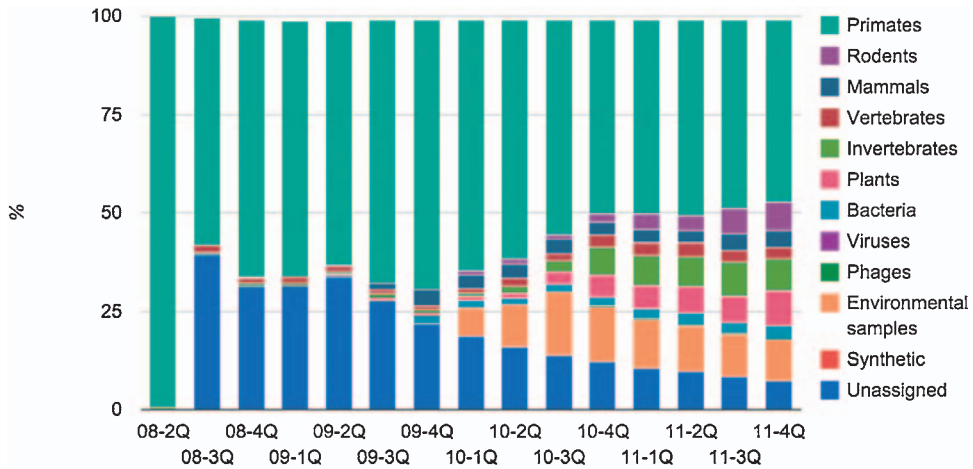

図2．生物分類群ごとのSRAデータ量比の推移．SRAが始まった当初はヒト由来配列を中心とした霊長類のデータが殆どを占めているが，2010年頃から，さまざまな生物種や環境由来サンプルの登録配列の比率が増えて来ている．現在は，幅広い生物群を対象に新型シークエンサーの配列決定技術が研究に利用されるようになってきていることがわかる．

目されたい．

SRAへのデータ登録　SRAは古典的なINSDCの塩基配列データであるフラットファイル形式とは異なり，複数のファイルから構成されるデータパッケージの形で登録をうけ，保管・公開されている．配列に関わるデータ本体と，そのデータについて記述するメタデータから構成されている．メタデータの規格はINSDC共通であり，5種類のXMLファイルにより記述するシステムとなっている．プロジェクトやサンプル，個々のランなどのメタ情報が階層的になることが多く，それぞれにデータに一対一で対応させるよりも，異なるファイルのセットとして持っておき，相互に参照する形をとる方が便利であるという理由がある．このため，登録と公開データの利用にあたってはこのデータ構造に沿ったXMLファイルを用意する必要がある．これは一般の生物学者には困難な作業であるため，DRAでは，データ登録を支援するために，グラフィクスユーザインターフェイスにより要求された情報を順次入力して行くことでXMLについての知識がなくても必要なファイルが作成されるウェブアプリケーションを作成し提供している．新型シークエンサー由来の登録データはときにテラバイトオーダーの莫大なものになるため，インターネット経由でのデータ転送だけでなく，ポータブルハードディスクに書き込んだデータを郵送などの方法で物理的に送付する方式も利用可能である．

SRA登録情報の利用　SRAに登録されたデータを利用する場合，DRA検索サイトDRASearch[3]での検索が便利でかつ高速である．DDBJでは2012年3月，スーパーコンピュータのリプレイスによって合計約5ペタバイトの記憶容量のディスクシステムを用意しているが，新型シークエンサーの解析コストは年々下がっており，その利用はますます増大することが想定されるため，SRAとして現在の形でどこまでアーカイブしつづけられるかは，INSDCとしても大きな課題として議論されてきている．

DDBJ Omics aRchive（DOR）の開発計画　新型シークエンサーの出力は上述のDRAを通じてSRAデータベースに登録でき，それらをアセンブルしゲノム塩基配列を再構築した場合には，完成配列として伝統的なDDBJのデータバンクに登録することができる．しかし，RNAseqのような定量的なNGS由来解析情報をアーカイブするデータベースはDDBJには存在しない．そのため，そのような実験情報を公開するためには，日本の生物学者は，生データはDDBJのDRAに登録し，定量的解析情報はNCBIのGEOもしくはEBIのArrayExpressに登録し公開するといった，登録の二度手間が必要になる．この問題を克服するため，2010年にDDBJはDDBJ Omics aRchive（DOR）の立ち上げを企画し，ArrayExpressと共通のMAGE-TABという記載方法で定量データをアーカイブし，データ交換を行う国際協力体制をつくることを決めた．残念ながらこのアーカイブは主として予算不足による運営上の問題により実際の立ち上げが大きく遅れたが，2012年度に新運営体制のもと開発を開始，登録受付システムの骨格を構築する方向で計画を再始動している．

大規模塩基配列解析支援系の提供

遺伝研スーパーコンピュータ　2012年3月に更新された遺伝研スーパーコンピュータシステムは，約5600コアから構成されLINPACK Rmax: 82.9 TFlopsのスペックを有しており，2012年6月時点のスーパーコンピュータTop500ランキングで世界280位，国内21位[4]

NII-Electronic Library Service

The Society for Bioscience and Bioengineering, Japan

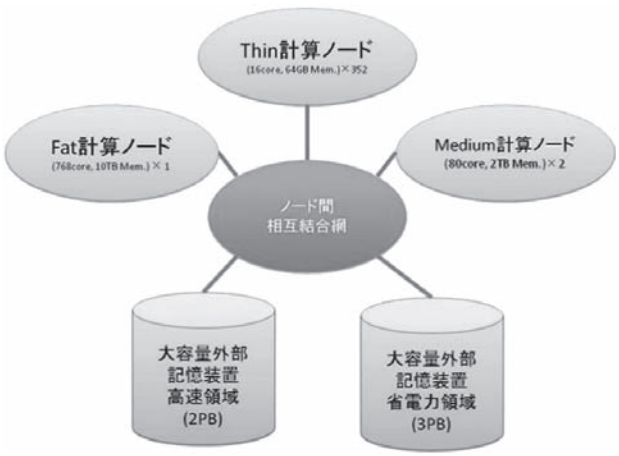

図3．国立遺伝学研究所スーパーコンピュータ概要．システムは大別して三種類の計算ノードにて構成される．Thin計算ノードは64 GBメモリを搭載した通常のPCクラスタであり，Medium計算ノード（2台）は2 TB，Fat計算ノード（1台）は10 TBのメインメモリを搭載しており，大量のメモリを要求するゲノムアセンブリなどの大規模塩基配列処理に特化した構成となっている．

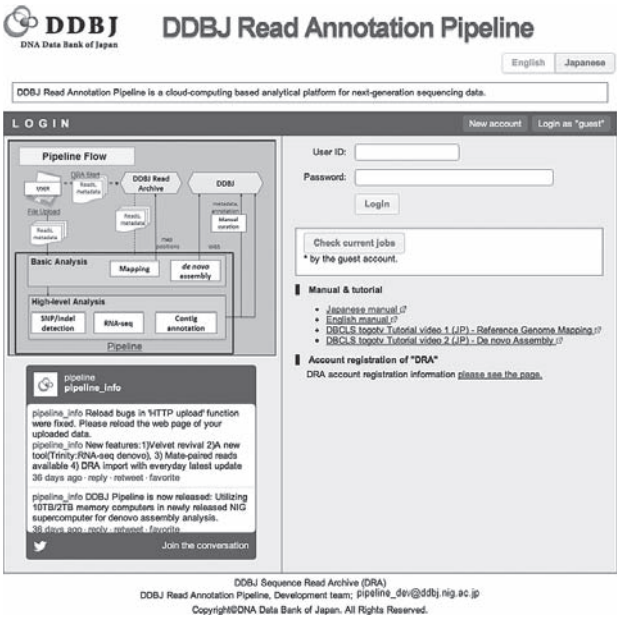

図4．DDBJリードアノテーションパイプライン．DDBJリードアノテーションパイプラインは，新型シークエンサー由来の大量配列の解析を支援することを目的としたウェブベースの解析ツールである．このサービスは参照ゲノム配列に対するリードのマッピングや*de novo*アセンブリを行なう「基礎処理部」と，基礎処理の後の一塩基置換（SNPs）の検出や発現解析などの解析を行う「高次処理部」の二段階で構成されている．

の処理能力をもつシステムである（図3）．このシステムの特徴は，汎用スパコンの主流となっているGPGPUを利用し浮動小数点演算の高速化を目指すものではなく，塩基配列情報処理の効率を追求するためCPUに重点を置いた設計がなされていることである．高速演算を可能とした350台余のマシンは一般的なPCクラスタの構成をとっているが，Fat nodeと呼ばれるサーバ（1台）は10 TBの共有メインメモリを利用可能なNUMA（Non-Uniform Memory Access）システムをもっており，Medium nodesと呼ばれるサーバ（2台）は2 TBのメインメモリを搭載しており，これらは大容量メモリが必要となる処理に供するための特別な機器である．このような非常に大規模なメモリを要する演算としては，大量の短い塩基配列を組み合わせ，一本の長大なゲノム塩基配列を再構築する「アセンブル」と呼ばれる処理がある．データを保管するストレイジとしては，同時に大量のノードからアクセスをしても高速にI/O処理のできるラスタ型と，利用していないときにはハードディスクをスピンダウンして消費電力を抑えるMAID型を合わせて5 PB（ペタバイト）の容量を導入している．このシステムは現在，学術解析用途のユーザには無償で利用できるよう提供されており，遺伝研ホームページからオンラインで申込可能である．

DDBJ Read Archive Pipeline　遺伝研スーパーコンピュータは上記のとおり新型シークエンサーの解析用途にフォーカスしたシステムであるが，その利用方法は対話的なコマンドの実行ではなくUniva Grid Engine（UGE，過去にはSun Grid Engineとして知られていたジョブキューイングシステム）でのバッチジョブ投入システムの利用が要求されるため，ベンチワークを行っている一般の生物学者の利用には敷居が高い．そこで，新型シーケンサー由来の塩基配列データに対して頻繁に行われる解析を容易に実行できるようにWWWのグラフィクスインタフェイスから解析を実行できるサービスを開発し提供している（図4）[5]．このようなサービスの展開によって，スーパーコンピュータの解析パワーを一般の生物学者にも容易に使っていただけるように配慮している．

終わりに

本稿では，INSDCとDDBJが展開している新型シークエンサーのアーカイブとその周辺サービスを紹介した．塩基配列決定技術の大きな進歩とともに配列アーカイブが提供するサービスも刻々と変化してきておりいる．サービスの利用手順などの最新情報はDDBJホームページ[6]から確認していただきたい．

文　献

1) Karsch-Mizrachi, I. *et al.*: *Nucl. Acids Res.*, **40**, D33 (2012).
2) Kodama, Y. *et al.*: *Nucl. Acids Res.*, **40**, D38 (2012).
3) http://trace.ddbj.nig.ac.jp/DRASearch/
4) http://i.top500.org/site/48477
5) http://p.ddbj.nig.ac.jp
6) http://www.ddbj.nig.ac.jp

NII-Electronic Library Service