公共NGSデータ再利用のためのデータ整備

Quanto / ChIP-Atlas / wPGSA online

Database Center for Life Science, Japan
○大田達郎
仲里猛留
坊農秀雅

Dept. of Dev. Bio., Grad. Sch. of Med. Sci., Kyushu Univ., Japan
沖真弥

Med. Sci. Innov. Hub Prog., RIKEN, Japan
川上英良

公共DBの一次配列データから二次的DBを構築する

様々な研究目的、様々な対象生物種に対して新型シーケンサー (NGS) を利用することで得られた大規模な配列データは、DDBJ などの公的データレポジトリにその一次塩基配列データ が登録される。DBCLS では、国内外の大学や研究機関と協力し、大量に登録された配列データに対して目的に応じたソフトウェア・ワークフローを実行することで、二次的なDBを構築し、それに基づくデータ解析技術を国内の提供している (Fig. 1)。

以下に、現在 DBCLS で実運用を行っている3つのサービスを紹介する。

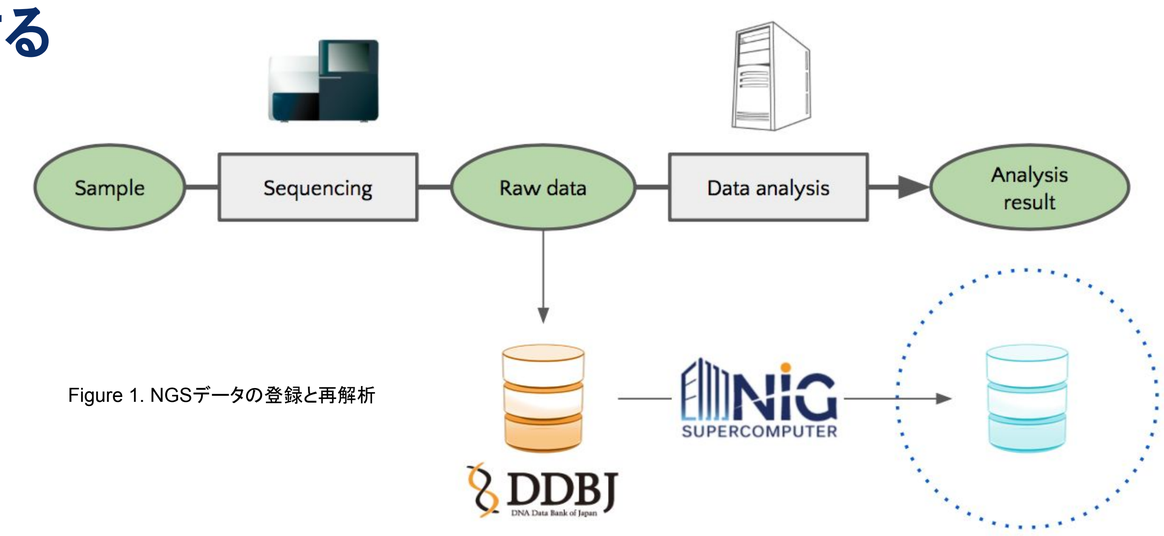

Figure 1. NGSデータの登録と再解析

Quanto
公共NGSデータのクオリティ情報のDB

「研究に用いられたデータはインターネットを通じて公開し、第三者によるアクセスを可能とすることでデータの再利用を促すべき」というオープンデータ・オープンサイエンスの考えが広まるに従って、データの共有と再利用が盛んになっている。しかし、NGS から得られるデータはデータサイズ、データ数ともに増大しており、必要なデータをレポジトリから取捨選択することは困難となっている。

Quanto は、公共データレポジトリに登録された全ての NGS データを対象に QC (Quality Control) 処理を実施することで、配列データに定量的なメタ情報を付加し、登録されたデータ同士の比較やデータセットの絞込を可能にするためのDBである。サンプル数にして20万以上のデータを迅速に処理するため、遺伝研スーパーコンピューターシステムを活用している。Quanto は DBCLS SRA (http://sra.dbcls.jp/) [1] 内でも提供されている(Fig. 2)。

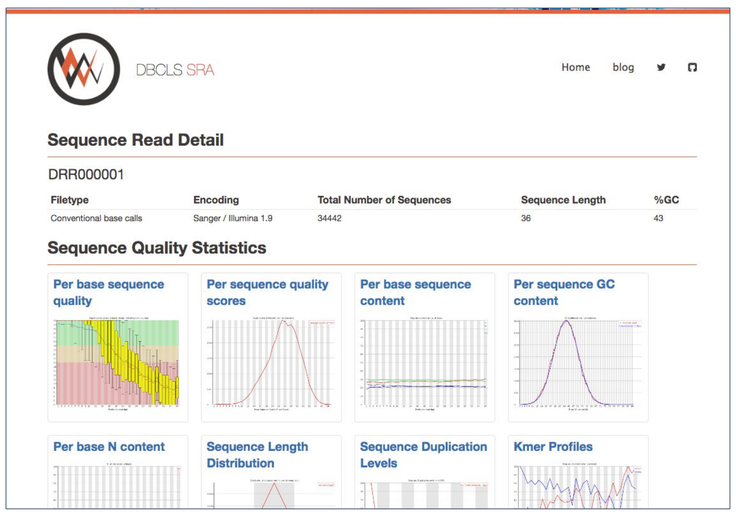

Figure 2. DBCLS SRA におけるQC情報の可視化

クオリティ統計値のRDF化による他DBとの統合

シーケンシング技術の発展は目覚ましく、機械だけでなく試薬の改良、DNA 抽出技術の改良などによって、NGS から得られるデータは質・量ともに年々向上している。シーケンス出力の増大は様々な生物学的応用を可能にし、公共データレポジトリには多様な種類のデータが登録されている。従って、NGS データの再利用にあたって最適なデータを選択するためには、配列断片の量や長さ、ベースコール精度や GC 含量などの情報が必須である。

公共 NGS データの再解析に基づくデータベースを構築する際に、Quanto の情報をより柔軟に扱えるよう、データごとのクオリティ統計値を RDF 形式で提供している (https://integbio.jp/rdf/)。これにより、データ検索だけでなく、NGSデータを元にしたデータベースとQC情報の統合が可能となる。

ChIP-Atlas
既報ChIP-Seqデータの網羅的再解析DB

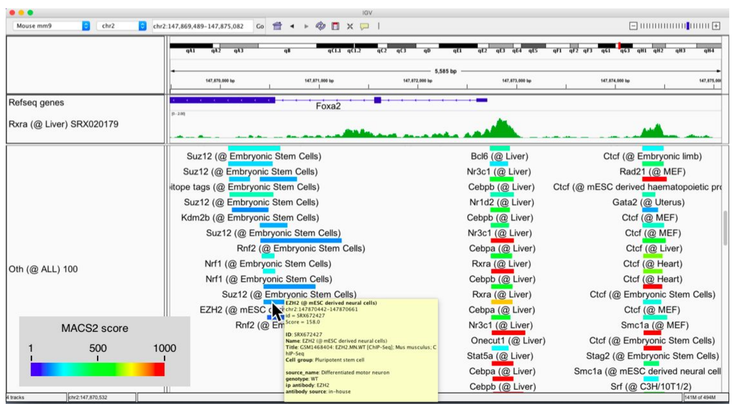

Figure 3. Peak Browser による転写因子結合サイトの可視化

既報ChIP-Seq/DNase-Seqデータの再利用

クロマチン免疫沈降法 (ChIP) によって転写因子に結合したゲノム上の DNA 配列を網羅的にシーケンスする ChIP-Seq, DNase によってオープンクロマチン領域をゲノムワイドに計測する DNase-Seq は、共に遺伝子発現制御機構の解明において重要な実験手法である。世界中の研究者によって NGS を用いたデータが多く発表されているにも関わらず、そのデータを参照、もしくは解析に再利用するためには、配列データをその都度ダウンロードし、解析ソフトウェアを用いて再計算する必要があったため、**特に計算機科学の技術を持たない実験研究者にとってデータの再利用が非常に困難**であった。

そこで九州大学 沖真弥助教の協力の元、公共データレポジトリに登録されたデータを再解析し、実験研究者が容易に既存研究のデータを再利用できる DB である ChIP-Atlas (http://chip-atlas.org) を開発した (Fig. 3)。ChIP-Atlas では再解析されたデータをゲノムブラウザで閲覧する Peak Browser, 転写因子のターゲット遺伝子を探索する Target Genes (Fig. 4), 転写因子の共局在解析を行う Colocalization, ピーク情報に対して Enrichment 解析を行う *in silico* ChIP の4つの機能を提供しており、データは毎月更新している。

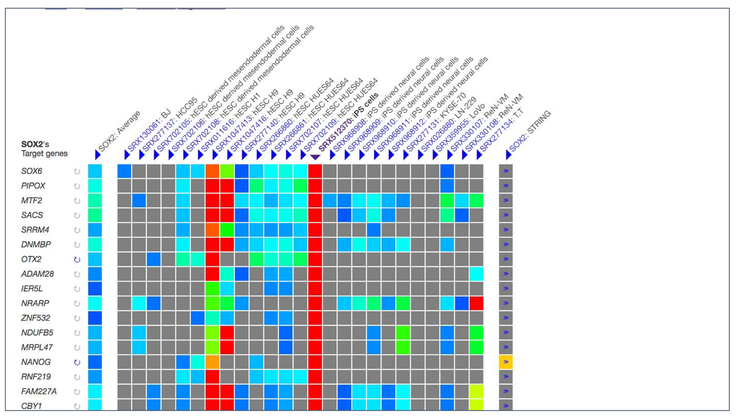

Figure 4. Target Genes の結果ページの例

wPGSA online
ChIP-Seqデータに基づく遺伝子発現制御予測

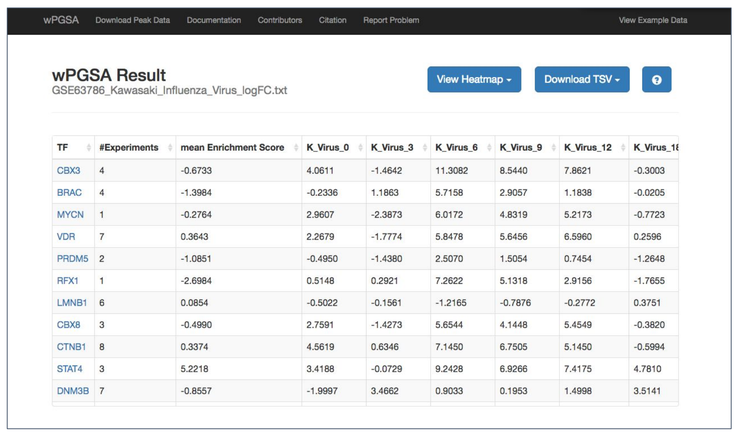

Figure 5. wPGSA online での解析結果の例

公共NGSデータに基づくデータ解析手法の開発

NGSの性能が向上し、より網羅的な計測が安価に実施できるようになった結果として、実験データの蓄積が急速に進んでいる。これまで、塩基配列の類似性などを用いて行われていた様々な生体分子の相互作用の予測も、NGSデータを用いることによって、**実験的に確かめられたデータに基づく予測**が可能となった。

理化学研究所 川上英良上級研究員らの開発した手法 wPGSA [2] は、公共DBに登録された ChIP-Seq データの再解析データをベースに、網羅的な遺伝子発現データから転写因子を予測するものである。wPGSA を用いた発現制御の解析を誰でも簡単に行えるよう、川上上級研究員の元、オンラインでデータ解析を行うことのできる wPGSA online をDBCLSで開発、運用している (Fig. 5)。

本手法に限らず、公共NGSデータレポジトリに登録されたデータに基づく手法には**新たなデータが追加される度に解析の精度向上や対象生物種などバリエーションの増加が見込める**というメリットがある。そのため、公共NGSデータベースの整備やデータ検索性の向上、データの質の担保、計算機やソフトウェアなどデータ再利用のためのインフラ・技術を高い質で維持することが今後の医学・生物学の発展にとって非常に重要であるといえる。

参考文献

1. Nakazato, T., Ohta, T., & Bono, H. (2013). Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. *PLoS One*, *8*(10), e77910.
2. Kawakami, E., Nakaoka, S., Ohta, T., & Kitano, H. (2016). Weighted enrichment method for prediction of transcription regulators from transcriptome and global chromatin immunoprecipitation data. *Nucleic acids research*, gkw355.