生物学実験 I AB
ゲノム・トランスクリプトーム・エピゲノム情報

担当者

理化学研究所 ライフサイエンス技術基盤研究センター
分子配列比較解析ユニット
ユニットリーダー
工樂 樹洋
shigehiro.kuraku@riken.jp

はじめに

生命情報科学の研究も、ある意味において「実験」である。ピペットマンやチューブを使わない作業であっても、プロトコルを読み、ポジティブコントロールやネガティブコントロールを交えて解析デザインをし、条件検討などの試行錯誤を通してはじめて信頼できるデータが得られるという点は同じなのである。

今回は、生命情報科学研究の中でも、遺伝情報の総和であるゲノム、発現する転写産物とまとめたトランスクリプトーム、そして、DNA やヒストンのメチル化を含むエピゲノムの情報に注目し、そのデータの取り扱いにおいて注意すべきこと、有用な解析ツール、そして、解析結果の解釈の仕方について紹介する。

1.　大局的な生命科学へ　－　大規模データを扱うエッセンス

1.1　NGS データ取得はどのように行われているのか？

2010 年くらいまでは、大規模データ生産は、ゲノム解析センターなど特殊な組織によっておもに行われてきた。だが、次世代シークエンス（NGS）の普及により、ごく標準的な大学や研究機関などでの大量データ取得が可能となった。より身近になったとはいえ、装置任せでよいデータが出るようになったかというと、決してそうではない。装置にかけるサンプルの調製の段階で解析の成否が決まることもある。たとえば、微量サンプルの解析においてサンプル調製における収率が悪いと、シークエンサーにかけるための必要量を得るために、過度の増幅を余儀なくされる。過度の増幅は、元来の分子の組成（RNA-seq の場合なら、遺伝子発現プロファイル）を狂わせてしまう可能性がある。また、装置が出力した結果はすぐに解釈できるようなものではなく、それを解析し生物学的な情報を得るためのバイオインフォマティクス解析の仕方によっても、得られる結論が変わることがある。

NGS 装置には、10Kbp 以上の DNA を読み取ることができるものもあるが、いわゆるショートリード（100 塩基前後の長さの DNA 配列）を得るタイプの装置が長らく主流となっている。装置の 1 回の稼働で、1Tbp のデータが取得できるようなものも存在する。出力された 100 塩基ほどの長さの配列が多数納められたファイルは、FASTQ ファイルと呼ばれている。

TASK1　FASTQC を用いて NGS 出力データの評価を行う

FASTQC は、FASTQ ファイルに含まれるデータを簡便に評価するためのプログラムであり、その出力結果をみるだけで気づくことのできる性質が色々存在する。上記の「過度の増幅」もそのひとつである。与えられたいくつかの FASTQC 結果を比較し、サンプルごとにどのような特徴がみられるか吟味してみよう。

参考 1 .1A　FASTQC プログラムの公式サイト
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

参考 1.1B　FASTQ ファイルとは？
http://bi.biopapyrus.net/perl/app/fastq.html

1.2　特定の遺伝子だけでなくゲノム全体の情報をどう理解するか？

脊椎動物の場合、ゲノムの中には、2 万個を超えるタンパク質をコードする遺伝子が存在するとされる。これだけでも相当なスケール感だが、それら遺伝子が占める領域を合計したとしても、ゲノム上のたった数%にすぎない。遺伝子以外に、転写調節を行うプロモータやエンハンサー、そして、多くは機能がわかっていない反復配列などが存在する。現在、ヒトをはじめとするさまざまな生物のゲノム情報が公開されてはいるが、それらを実際に研究に利用するのにはどうしたらよいであろうか？

TASK2　UCSC ゲノムブラウザを用いて、様々な遺伝子領域をブラウズする

ゲノムブラウザから興味のある遺伝子を探すときに、遺伝子名を入力する必要があるが、特異性の高い検索を行うにはどうしたらよいか？目的の遺伝子領域にたどり着いたら、どういった情報にアクセスできるだろうか？

参考 1.2　UCSC ゲノムブラウザ
https://genome.ucsc.edu/

1.3　多様な生物のデータをどう比較すればよいか？

原核生物から、植物や菌類、そして動物に至るまで、多数の生物のゲノム情報が公開されている。これらの情報を比較することにより、我々の祖先が、進化のどの段階で、どのような構成のゲノムを有していたのかを解析することが可能となった。当然、生物多様性がどのように生まれたのか、そのメカニズムを分子、そしてゲノムの言葉で説明その手掛かりとなる膨大な情報が、公共のデータベースの中に格納されている。

TASK3　Ensembl（「アンサンブル」と読む）ゲノムデータベース内をブラウズしどういった生物のゲノム情報が利用できるか把握する

ゲノム情報が公開されている生物には、何か共通点があるだろうか？また、個々の生物のゲノム情報それぞれにアクセスできること以上に、種間の比較や進化学的視点からの解析を手助けする情報は含まれているか？

参考 1.3　Ensembl ゲノムデータベース
http://www.ensembl.org/

1.4　大規模データ解析に適した環境とは？

次世代シークエンサーの出力データや、それに基づいたゲノム規模の情報は、すべての生物学者から手の届くものとなった。しかし、正確かつ効率的にそれらを用いた解析を進めるためには、それに適した解析環境を準備することが望ましい。これに適しているのは、文字・数値の扱いを簡便にするようなプログラミングや、それを可能にする既存の解析プログラムをコマンドラインから実行するような解析の進め方である。具体的には、Mac の「ターミナル」というアプリケーションや **Linux オペレーティングシステム**がこれを可能にする。より大規模な計算を行うにはとくに後者が適しているといえる。

2. ゲノム情報を利用する

2.1 ゲノム塩基配列はどの程度信頼できるだろうか？

ゲノムの中には、クロマチンに強固に保護されて DNA が露出されにくい部分や、DNA としては単離できるが配列決定の難しい単純繰り返し配列も存在する。ゲノム情報の読み取りには、もともとこのような問題があった。次世代シークエンスの登場によってゲノム配列情報取得の敷居は確実に下がったが、世に出る情報の質が上がったかという観点では実は疑問がある。やはり、ショートリードに依存しているために、配列の繋ぎ合わせ（アセンブリ）の精度が如実に結果に影響するということと、長い反復配列に弱い、ということがおもな理由である。本来、得たゲノム情報を綿密に評価してから公開できれば理想であるが、そもそも新規の情報を得るために行う作業であるために、誰も正解を知らないというのが現実である。また、脊椎動物くらいの大きなゲノムとなれば、アセンブリに数週間かかってしまう。リーズナブルな時間経過の中で評価を行い、その結果を受けて再度改善を図る、という手順を踏みにくいということが、クオリティの評価を難しくしている要因でもある。

TASK4　ゲノムアセンブリ配列の評価を行う(assemblathon_stats.pl)

全ゲノム配列は、ヒトゲノムでは、染色体レベルの情報にまとめ上げられているが、公開されているそれ以外のゲノム配列の多くは、単純な配列のオーバーラップをもとに繋がれたコンティグ（contig）と、コンティグのあいだの間隔の情報を基にしてさらに繋ぎ合わされたスキャフォールド（scaffold）に分けられる。それらの配列の長さや中身を調べることにより、ゲノム情報のクオリティを評価することができる。この種の評価を行うには、具体的に何を指標にしたらよいだろうか？

参考 2.1　Assemblathon2 プロジェクト
https://gigascience.biomedcentral.com/articles/10.1186/2047-217X-2-10

2.2　公開されているゲノム塩基配列はゲノム情報を網羅できているのか？

上の項では、単純に配列を構成する塩基の数や組成に注目して、ゲノム配列のクオリティを評価した。しかし、アセンブリプログラムの中には、無理に配列を長くつないでしまう性質（'over-assembly'）をもつものもある。その場合、単に長く繋がったからといってクオリティが高いと評価してよいか甚だ疑問である。そこで、多用されているのが、ゲノム配列中に存在するはずの遺伝子の構造がどの程度復元されているかを定量化し、それを完成度のスコアとする評価法である。この目的のために最初に作られた CEGMA というプログラムは、近年、開発者がサポート終了を宣言し、BUSCO というプログラムにとって代わられた。

TASK5　ゲノムアセンブリの完成度を評価する（CEGMA & BUSCO）

この方法では、「存在するはずの」遺伝子をどう選ぶのか、そして、それらを長大なゲノム配列中にどう探すのか、という点で検討を要する。これらを考慮したうえで、CEGMA と BUSCO の良い部分を取り入れて整えられた、ゲノムやトランスクリプトームアセンブリの評価を行うためのインターフェースが gVolante（「ジーボランチ」と読む）である。このサイトでは、自身の持つアセンブリ配列ファイルをアップロードして評価できるだけでなく、公開されているゲノムアセンブリの完成度のスコアを閲覧することもできる。

参考 2.2A　ウェブサーバーgVolante
http://gvolante.riken.jp/

参考 2.2 B
原 雄一郎「どのアセンブリを使うか?: 分子系統学的観点に基づくアセンブリの評価」日本進化学会ニュース 2016. 17(1): 23-29.

参考 2.2 C
原 雄一郎「脊椎動物ゲノム・トランスクリプトームアセンブリ完全度を評価する」バイオサイエンスとインダストリー 毎日学術フォーラム　2016. 74(3): 228-230.

2.3　「扱っている生物のゲノム情報が公開されていない！」といった場合には？

かつては、ゲノムシークエンスが行われている生物の情報を集約し、その検索ができるサイトが存在したが、単独研究室での新規ゲノム配列取得が身近になった現在、多数のプロジェクトが乱立し、情報の整理・集積もままならない。自分が扱っている生物のゲノム情報を利用したいが公開されていない、という場合には、まず立ち止まって、ゲノム情報が本当に必要だろうかと自問自答することをお薦めする。身近になったとはいえ、高精度なゲノム情報を自信を持って世に送り出すためには多大な費用そして時間がかかる。また、NGS 装置やバイオインフォマティクスのスキルを持った人材など限られたリソースを有効に活用するためにも、本当に誰も手を付けていないのか、そして、本当にリソースをつぎ込む価値の十分ある対象なのかをまず吟味することが強く推奨される。仮に、ゲノムシークエンスは必要ないと判断した結果、トランスクリプトーム解析だけで済むということなら、難易度や労力は確実に下がるといえるだろう。

いっぽう、どうしてもゲノム情報が必要、ということになれば、まずは、めぼしいゲノム解析拠点など、関係筋へのコンタクトを直接試みるとよい。ゲノム情報が公開されていなくても、すでに情報は整備され、それを用いて解析中あるいは出版準備中、という段階のプロジェクトは非常に多いので、尋ねてみると教えてもらえることがあるかもしれない。ゲノム解析はトップジャーナルに採択されているプロジェクトが目につくが、年々その敷居が上がっており、スムーズに出版できていないだけかもしれない。もし、誰もゲノムシークエンスを手掛けていないのなら、その時点ではじめて自分が着手することを考えるべきである。ゲノムシークエンスの方法、なかでも高次スキャフォールディングの方法は、次から次へと新しい手法が登場しており、その組み合わせ方によっても結果が大きく変わりうる。DNA 試料の調製やシークエンスのプランだけでなく、シークエンス後のアセンブリをはじめとするインシリコ解析の部分についても必ず着手する前に「誰がどう手を動かすのか」を綿密にプランしておくべきである。

参考 2.3　JGI GOLD データベース
https://gold.jgi.doe.gov/

3.　トランスクリプトーム情報を利用する

3.1　遺伝子発現から何がわかるのか？それらはどのように可視化できるか？

ゲノム情報とは違って、トランスクリプトーム情報（‘transcriptome’は転写産物全部という意味）は、注目する細胞・時期によって異なる様相を呈する。血液ではヘモグロビンが、そして網膜ではロドプシンが高いレベルで発現し、いっぽうヒストンはどの組織の細胞でも発現しているという具合である。網羅的な解析手法を用いれば、ゲノム中に存在する２～３万個の遺伝子それぞれの発現レベルを一度に調べることができる。こういった解析は、古くはサンガーシークエンスを用いた EST（Expressed Sequence Tags)、そしてより近年ではマイクロアレイ（microarray)、そしてより最近では RNA-seq という方法によって行われてきた。

TASK6　EST データ(NCBI UniGene)やマイクロアレイデータ（GEO）を閲覧する

公共データベースには、論文投稿の際などに登録されたデータが多数格納されている。どのデータベースにどのような情報が、どのような形で保管されているか確認しよう。また、そのデータを取得するのにどのような試料をつかってどのような実験を行ったのかという詳細情報はどの程度辿ることができるであろうか？

参考 3.1 A　NCBI UniGene データベース
https://www.ncbi.nlm.nih.gov/unigene

参考 3.1 B　NCBI Geo データベース
https://www.ncbi.nlm.nih.gov/geo/

3.2　RNA-seq 実験はどのようにデザインするのがよいか？

次世代シークエンスを利用した RNA-seq（RNA シークエンス）という手法を用いれば、興味を持っている組織や細胞に発現する遺伝子の転写産物を一網打尽に検出できる。ただし、発現量の低い遺伝子や転写産物の長さや塩基組成が極端な遺伝子は、検出から漏れる可能性があることに注意が必要である。

TASK7　提示された生物学的な問いに答えるための実験デザインを考えてみよう

与えられた問いに答えるための正しい実験デザインの仕方を検討しよう。機械ではなく生物であるからこそ存在する生命現象のゆらぎを考慮するために、「反復試行（replication）」がどのような役割を果たすであろうか？大規模な実験ほど考慮を要する「バッチ効果(batch effect)」とは何か？

さらに、遺伝子発現プロファイルを見分ける際、より高い解像度での解析を実現するために、シークエンスの段階でも気を付けるべきことはあるだろうか？

4.　エピゲノム情報を利用する

4.1　エピゲノム情報から何がわかるのか？それらはどのように可視化できるか？

「エピゲノム（epigenome）」という語の「エピ」という部分単独では「～の上の」を意味する。すなわち DNA 配列に表れないゲノムワイドな情報のことだが、具体的には、DNA メチル化とヒストンタンパク質の修飾を指している。DNA メチル化の検出には、バイサルファイトシークエンス法が、また、ヒストン修飾の検出には、クロマチン免疫沈降シークエンス（ChIP-seq）法が多用される。

ゲノム DNA そのものの情報ではないとはいえ、ゲノムの位置依存的にエピゲノム情報は存在しているため、ゲノムブラウザの一部がエピゲノム情報を表示するなど、表示上はゲノム情報が基準となっている。どの細胞・時期を見るかによって、DNA やヒストンの修飾の状況が大きく異なることもあるため、トランスクリプトームの解析と同様、試料の用意の段階から綿密な見極めが必要である。

TASK8　既存のエピゲノム情報を閲覧してみよう（UCSC ＆ ENCODE）

古くからゲノム情報を提供している UCSC のゲノムブラウザは、代表的なエピゲノム情報を表示することもできる。ENCODE（エンコード）は、‘Encyclopedia of DNA Elements’という国際プロジェクトの略称であり、このプロジェクトは、5 年かけてヒトゲノム中の機能因子の大量検出を目標に掲げて行われたものである。その一部として、エピゲノム情報の大量取得が行われ、独自のプロジェクトサイトにて、公開されている。

参考 4.1 A　UCSC ゲノムブラウザ
https://genome.ucsc.edu/

参考 4.1 B　ENCODE プロジェクトサイト
https://www.encodeproject.org/

参考 4.1 C
「ヒストン修飾や転写因子の結合領域を同定するコツやポイント」
門田満隆、蓑田亜希子『次世代シークエンス解析スタンダード』　羊土社　2014 年. 105-121.

参考 4.1 D

「ChIP-seq 実験を成功に導くための秘訣：ChIP 条件の最適化とサンプル QC について」

田中かおり、門田満隆『実験医学 2016 年 10 月号』 Vol.34 No.16.

5. 大規模情報に惑わされないために

5.1　データ解析スキルを身につける

DNA シークエンス技術の革新によってデータ生産のスピードが向上したいま、いわゆるビッグデータを扱う他分野と同様、生物学においても「データサイエンス」の重要性が高まっている。大量データを扱う際、より汎用的な解析アプローチを求めるなら、Linux システムにおけるプログラミングのスキルを習得していることが望ましい。多くの大規模データ解析プログラムは Linux 環境で作られ、Linux 環境での利用を想定されていることが多いためである。

5.2　既存の情報を効率的に利用する

大がかりな実験になるほど費用も手間もかかる。ゲノム情報の準備について前述したように (項目 2.3)、自らがすべてのデータ生産を行うべきかについては常に検討の余地がある。自らがデータを生産しなくても、公共のデータベースに既に多くのデータが格納されている。特定の細胞・時期に注目する場合には、その目的を満たす公共データが存在する可能性は低いかもしれないが、もし、自分でデータ生産を行わなくてよいと判断した場合は、費用も時間も節約できる。

5.3　賢い NGS ユーザーになる

次世代シークエンスを行っただけで生物学の問いが解けるなどという簡単なものではないが、この技術の登場によって、実験を行う規模、タイムスパン、そして予算の割り当て方が大きく変化した。研究者の考え方、そして時間の使い方を大きく変える技術革新だったともいえる。とはいえ、次世代シークエンス自体は、研究プロジェクトを進めるうえでの 1 ステップにすぎず、使用するサンプルの綿密な準備や、シークエンサから出力されたデータの詳細な解析が重要であることには変わりはない。むしろ、それら前後のステップが最適化されていない場合には、容易に無駄な出費が発生してしまう、という望ましくない状況を招いているのが現実である。次世代シークエンスは決して魔法ではなく、単なる「実験」にすぎない。すべてを人任せにせず、次世代シークエンス後の解析デザインを含めて、事前に全体的な構想を立てて、計画し実行する必要がある。