Don’t sleep through the assembly!

桑原　ひかる（筑波大学 生物学類）　　指導教員：徳永　幸彦（筑波大学 生命環境系）

[導入]

次世代シーケンサー(NGS)の登場をきっかけに、ゲノム解析の技術は日々進歩しており、今やゲノム情報は医学系の分野だけでなく、保全や進化に関する分野にも活用されている。非モデル生物を対象とした解析も増え、数多くの結果や生データが、インターネット上のデータベースに登録されている。

NGS によるシーケンシングの手法としてショットガン・シーケンシングがある。DNA 鎖の断片化をランダムに行うもので、効率的ではあるものの断片(リード)をつなぎ合わせる過程(アセンブリ)が複雑になってしまう。ゲノム配列の中に存在する重複や反復を短いリードで再構築するほか、シーケンシングのエラーを考慮する必要もある。リードのゲノム上の位置に偏りが生じた場合、アセンブリ結果の配列が 1 本にまとまらないことも多い。

こうした課題を克服するため、サンプル調整といったウェットな段階の手法だけでなく、アセンブリのアルゴリズムの段階など、ドライな段階においても改良が進んできた。しかし一方で、アセンブリのためのソフトウェア(アセンブラー)が多様化しているため、研究者がその中から使用するものを選ばなくてはならず、異なるアセンブラーで作り出した結果を比較する必要があるが、評価に用いられる指標も煩雑な状態である。指標として広く採用されているものとしては、アセンブリで再構築された断片群(コンティグ)の長さや数、存在が示唆される遺伝子の復元率、リファレンスにアライメントした際のミスマッチ数、などがある。ただしこれらの値から「良い」アセンブラーを選ぶための基準は研究者に委ねられるところが多く、明確ではない。また、これらの指標それ自体は、それぞれの結果の中身、すなわち塩基配列そのものを直接的に比較するものではない。さらに、アセンブラー選択や結果の評価が問題なく進んだとしても、そこまでの過程を繰り返して再現性を検証した例は極めて少ない。

本研究ではこうした現状を踏まえ、塩基配列そのものに注目したアセンブラー間の比較と、シーケンスとアセンブリを繰り返すことを想定したシミュレーションを行った。

[手法]

一連の実験には、大腸菌ファージ phiX-174 の全ゲノム配列(NCBI accession number: NC_001422)と phiX-174 の全ゲノムをシーケンスしたリードデータ(https://github.com/gigascience/galaxy-bgisoap/tree/master/test-data/phiX174 より入手)を用いた。

まず 13 種類のアセンブラーを用いてデータをアセンブルし、各結果の類似度を、レーベンシュタイン編集距離を用いて計算した。配列の長さと中身それぞれについて、リファレンス配列を含めてクラスター分析を行いデンドログラムを作成した。

次に全ゲノム配列から、入手したデータと同じ構造のリードをシミュレーションで 50 セット作成し、13 のアセンブラーでそれぞれについてアセンブルを行い、同様にデンドログラムの結果を比較した。また、phiX-174 のコーディング領域の復元率を相同性検索によって調べた。

[結果・考察]

元データを使用したアセンブラー間の比較に関しては、13 種類すべてのソフトウェアにおいて、全ゲノム配列と長さは近いものの、それぞれ固有な配列が再構築された。長さの類似度のクラスターと塩基配列の類似度のクラスターは樹形が似ており、長さがリファレンス配列に近ければ配列の中身も近い、という傾向がみられた。また、編集距離によるクラスター分析によって、リファレンス配列の長さによらずアセンブラー間の結果の配列の比較を行うことができた。

シミュレーションによる再現性の検証においては、全アセンブラーのすべての結果でコーディング領域が 100%復元された。しかし、同一のアセンブラーを使った 50 回分の結果をクラスタリングすると、長さ、配列の中身ともに同じ結果が得られることは皆無であることが示された(図 1)。リファレンス配列が存在しない非モデル生物の場合、このような結果のばらつきはドラフトアセンブリーの作成や個体識別に大きく影響すると考えられる。またリファレンス配列が存在する生物の場合も、それがばらつく結果のうちの一つであることを認識する必要がある。

さらに、シミュレーションで作成した同一リードデータのアセンブリを比較することで、前半で比較したアセンブラーごとの塩基配列のデンドログラムは、リードデータによっては同じ樹形を示さないことがわかった。アセンブラーを選択するための比較の結果が、アセンブラーのパフォーマンスだけでなくリードデータにも影響を受けていることが示唆された。シーケンシングが手軽になり、ゲノム情報が幅広い分野に利用されている今、アセンブリにミスがあった場合の影響も様々な研究に及びかねない。本研究で明らかになった結果の変動や再現性の検証に着目した、新たな評価指標を作り出す必要がある。

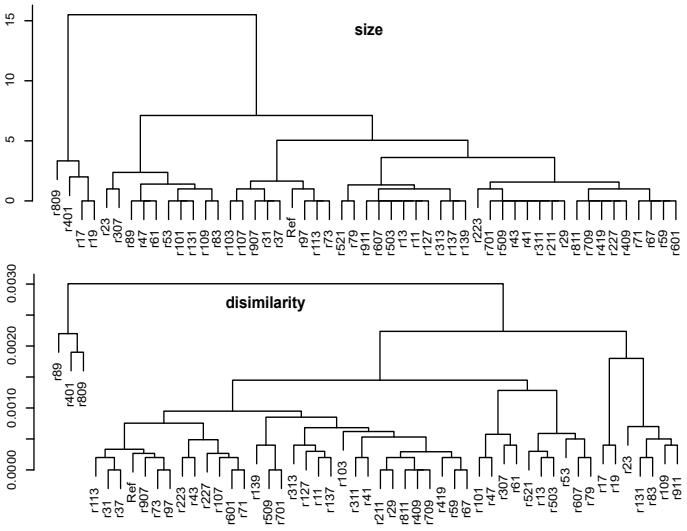

図1. シミュレーションによる比較(アセンブラーとしてSPAdes使用)。上部で配列の長さ、下部で中身を比較している。各数字はそれぞれリードのデータセットの番号で、Ref はリファレンス配列である。