(P-13)

Hi-C データを含めた染色体スケール全ゲノム配列の構築検討

山口 勝司（基礎生物学研究所　技術課）

YAMAGUCHI Katsushi : Consideration of constructing a chromosome-scale whole genome assembly incorporating Hi-C data

Hi-C was originally developed as a method for analyzing the chromatin state of nuclear DNA. By isolating chromatin DNA from the nucleus while preserving its spatial structure, followed by partial fragmentation of DNA strands and proximity ligation, sequencing reads can be obtained in which sequences that are distant on the linear strand but close in the spatial structure are connected. By analyzing these reads using next-generation sequencing, the spatial organization of chromatin can be inferred. Incorporating this method into whole-genome sequencing analysis has revealed that chromosome-scale genome assembly can be dramatically improved by utilizing spatial proximity information. As a result, Hi-C has become an indispensable technology for whole-genome sequencing. In this report, based on the experience gained from applying this technology to numerous samples collected through collaborative research, we introduce examples of improvements in handling challenging samples and discuss unresolved issues.

1. 目的

Hi-C は元々、核内 DNA のクロマチン状態を解析する手法として知られる。核からクロマチン DNA を空間構造を維持した状態で単離し、部分的な DNA 鎖の断片化とそれに続く近接ライゲーションを行うことで、直鎖上では離れているが空間構造上は近接する配列同士が、繋がったリードが得られる。これを次世代シーケンサーで配列解析することで、クロマチン空間構造を推察・解析できる。この方法を全ゲノムシーケンス解析に取り入れることで、空間構造上の近接情報を活用し、染色体スケールでの全ゲノム配列のアセンブリングが、格段に向上することが 2017 年に報告された。それ以降、この手法を基盤としたキットが数社から発売され、現在では全ゲノム配列の解明において、不可欠な技術となっている。本報告では、これまでおこなってきた共同利用研究の多数のサンプルで、この技術の適用を通じて得られた経験を基に、難解なサンプルへの対処改善、さらに未解決な課題を紹介する。

2. 方法

Dovetail 社の Omni-C キットを用いて Hi-C 解析を行い、PacBio 社の HiFi リードシーケンスを hifiasm でアセンブルした contig 配列に適用した。Juicer と 3d-DNA を用いた Hi-C 解析パイプラインで処理し、染色体スケールのゲノムアセンブルを目指した。得られた結果は juicebox で可視化した。

3. 結果

解析系の中でのチェックポイントとして、作製した Hi-C のデータが良好なものかを判断する解析系が Dovetail 社から用意されている。その流れでの解析結果を示す(図 1)。PCR 重複リードや、元々直鎖上も近接しているリードは有効情報にならない。それ以外の赤字で示す割合が高いことが重要である。

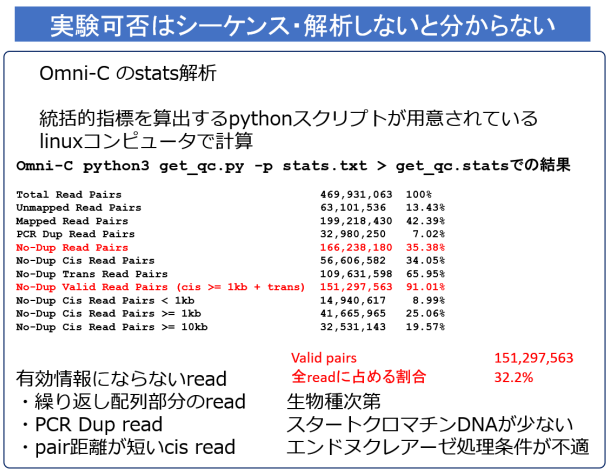

図 1.　解析結果の statistics の集計

また kit のプロトコール通り進めてもうまく行かないものも多い。図 2 の例では植物標準の葉を用いた方法では反応 buffer との馴染みが悪く、核を単離することで改善された。Juicebox での可視化で近接したリード

の組み合わせの場所に打点される。近接リードペアの直線に加え、全体的にバックが高い四角のエリアが染色体数、見られるのが理想的な結果である。このケースでは明確にスギゲノムの染色体本数が 11 本であることが示された。

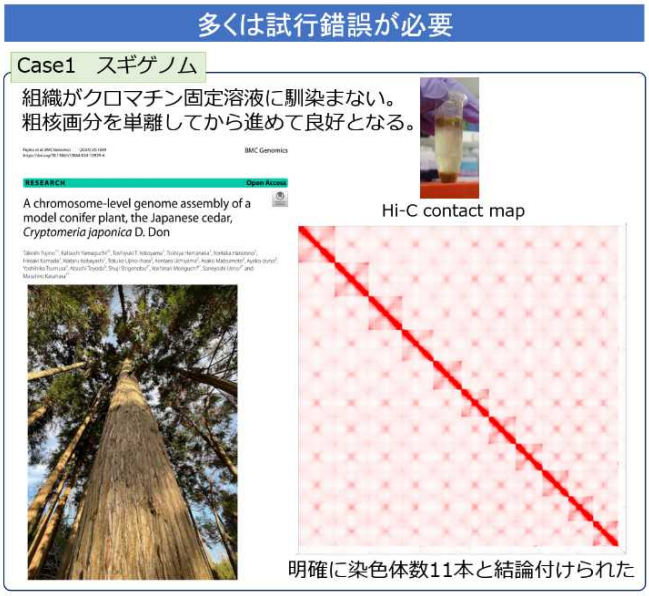

図 2.　核単離を先におこなうことでうまくいったケース

状況に応じ、部分的にプロトコールを改変し、途中段階での残存量を評価し、またクロマチン DNA の初期状況を確認することにした。図 3　case2 は、うまく行かなかった 1 例で、単離時点でクロマチン DNA が分解している。単離クロマチンは固定単離後にヌクレアーゼ処理をして近接ライゲーションに進む必要があり、この時点ですでにクロマチン DNA が分解しているのは好ましくない。これについては一般的なヌクレアーゼ阻害剤である EDTA を添加することで、初期のクロマチン DNA 分解を抑制でき、最終的に良好な結果となった。また同様にサンプルによっては途中のチェックポイントで、残存する DNA 量が不足している場合がある。その場合でも、最終的な PCR サイクル数を増やすことで、PCR による重複リード割合が増えたとしても、そうでない有効なリード自体も増やすことができ、目的を達せられたケースもある（図 3　Case3）。一方、未だうまく行っていないものも多い。以下図 4　Case4 では一見良好な結果に見えるが、推定される染色体本数の倍を示している。すなわち染色体の短腕と長腕を繋ぐリードが得られていない。図 4　Case5 では内生ヌクレアーゼによる分解が EDTA 添加によっても改善しないもので、最終結果も不良であった。

4. 考察

上記のように、現状はどの生物種においても安定的に良好な結果が得られるには到達していない。しかし、個々の対処法を遂行することで、新しいサンプルに対する、対応の順序は整備され、より成功率は上がっている。現在、新 kit もメーカーから販売に至っており、順次それらでの状況も検討する予定である。

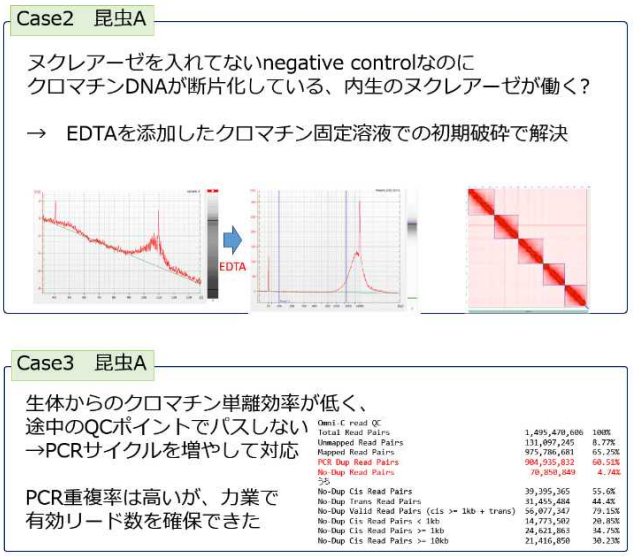

図 3.　検討によりうまくいったケース

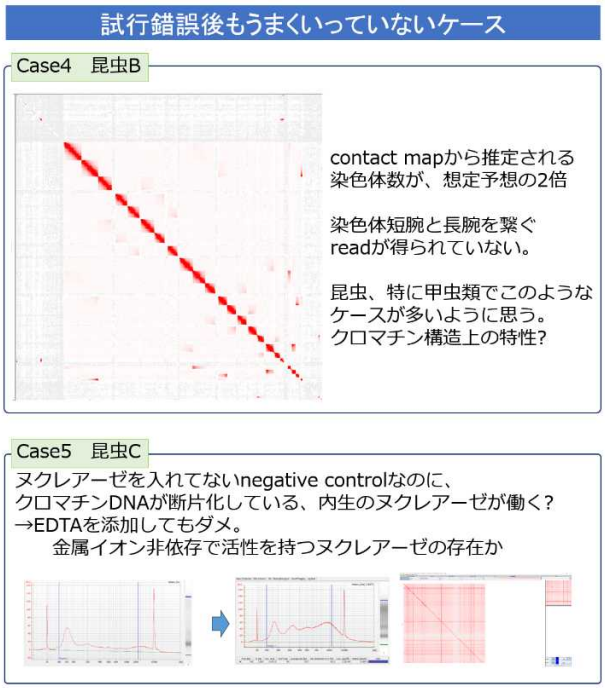

図 4.　試行錯誤後もうまくいっていないケース

謝辞

本発表は NIBB の統合ゲノミクス共同利用研究の複数課題のサンプルを用いておこなっている。共同利用研究の関係諸氏、トランスオミクス解析室の重信秀治教授、技術支援員の松本美和子氏、池田弥華氏に感謝申し上げる。