

[ 研究区分 : 学際的・先端的研究 (A) ]

研究テーマ : 次世代シーケンスデータの新規解析法の開発とその環境ゲノミクスへの活用	
研究代表者 : 生命環境学部 生命科学科 准教授・菅 裕	連絡先 : hsuga@pu-hiroshima.ac.jp
共同研究者 :	
<b>【研究概要】</b> 次世代シーケンス(NGS)は生命科学の分野のみならず、環境科学においても重要な手法となりつつある。しかしながら本学ではその導入は遅れており、一部の教員の高価な外注による利用にとどまる。本学内における次世代シーケンス法の定着と、得られる大規模データを学内で解析可能とすることをめざし、H27年度は、NGSの手法の一つである Mate Pair reads が本来持つ情報を効率よくゲノムアセンブリに落とし込むためのソフトウェア(DIMP)を開発した。DIMPにより、カキゲノムの品質を大幅に高めることができた。	

### 【研究内容・成果】

次世代シーケンスの手法の一つである Mate Pair reads は、ある長さ(例えば 5kb)をもったゲノム断片の両端を大量に読む技術であり、繰り返し構造などのためアセンブリ(断片のつなぎ合わせ)が困難なゲノムを「解きほぐす」ために使われる。真核生物の巨大なゲノムを正確につなぎ合わせる上で不可欠なテクニックとなりつつある。ところが、これまでのアセンブリソフトウェアでは、読み取った配列(Read)をうまくゲノムに貼り付けて並べられないことが原因で、本来活用されるべき Read 同士の距離情報が大きく失われてしまうという問題があった。そこで、距離情報を含めた形で Read をゲノムに mapping (並べて貼り付ける作業)することで、ゲノムの繋ぎ合わせの誤りを視覚的に検出できれば、Mate Pair reads が本来持つ情報を最大限に利用できるのではないかと考えた。

アイデアは ruby により実装され、ソフトウェアの名称は DIMP

(Draft genome Improvement by Mate Pair library) とした。

次に、リアルデータで DIMP の有用性を確かめるため、カキ(マガキ/*Crassostrea gigas*)のドラ

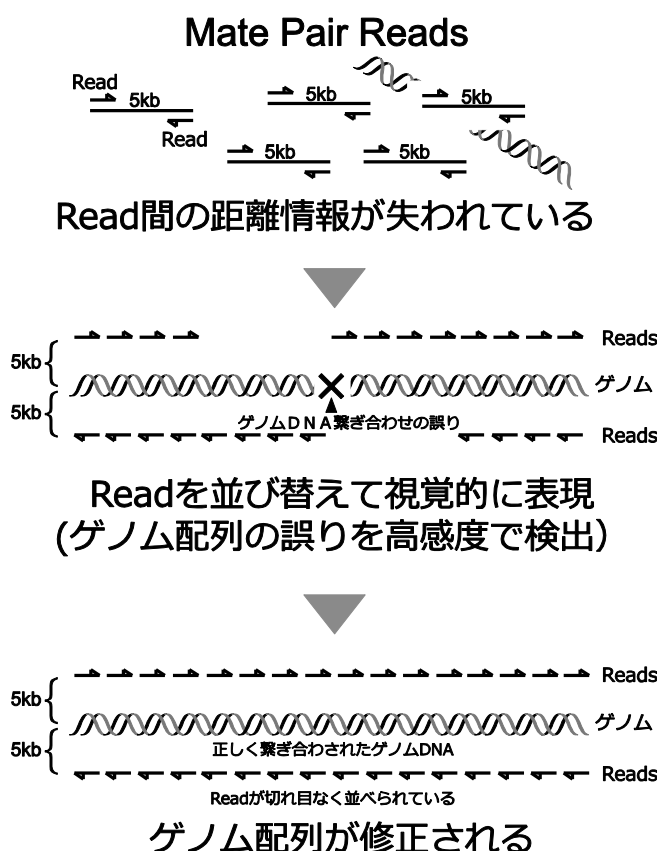


図1 DIMP のアイデア

---

フトゲノム(Zhang et al. 2012)の改善を試みた。カキゲノムは巨大(559Mbase)で、その全てを DIMP により改善することは困難である。そこで、カキドラフトゲノム中でもっとも大きな Scaffold19 (約 2Mbase) に DIMP を適用した。その結果、驚くべきことに、既に Nature 誌に発表されているこのデータには 4 1 か所に及ぶ mis-scaffolding (つなぎ合わせの誤り) が含まれていることが明らかになった。また、DIMP の適用により、それまで Scaffold19 に含まれていた全長 137kbase に及ぶ 67 個の gap (情報がうまく得られず、空白で放置された部分) は 11 個 (総計 5kbase) に減少した。更に、この過程で gap を埋めた塩基配列を詳細に解析した結果、このうち 52%にも及ぶ領域が、オリジナルのドラフトゲノムには含まれていない配列であることが判明した。すなわち、DIMP はカキドラフトゲノムの誤りを修正したばかりでなく、アセンブリの過程で消失した情報を復活させることができたことになる。いうまでもなく、比較ゲノム学 (複数種のゲノム情報を比較解析することで生物種間の形態や性質の違い、またその進化の過程を明らかにしようとする分野) においては、ゲノムから特定の情報が失われることは、その解析結果に致命的な影響を与える。それどころか、既にゲノム (生物の全遺伝情報) として発表されているこのデータは、定義上、"gen - ome"ですらないことになる。

次に、既に発表されている類似ソフト、REAPR 及び GapFiller との比較を行った。全自動で配列を修正するこれらのソフトと、基本的に手作業での修正の補助をめざした DIMP とではフェアな比較とはならない。それでも DIMP では 11 個に減少した、オリジナルアセンブリに含まれる 67 個のギャップは、REAPR+GapFiller だと 63 個にしか減少せず、また総ギャップ長はむしろ 203kbase に増加するなど、DIMP の優位性を示す結果となった。

ソフトウェア自体の論文は現在投稿準備中であるが、DIMP を使用して品質を高めたゲノムは既にいくつかのプロジェクトで使用されており、その一つは最近 eLife 誌 (IF 9.322)に掲載された。

#### 参考文献

- Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biol* 13, R56.
- Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. (2013). REAPR: a universal tool for genome assembly evaluation. *Genome Biol* 14, R47.
- Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., et al. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490, 49-54.

#### 【発表論文】

- De Mendoza, A., Suga, H., Permanyer, J., Irimia, M., and Ruiz-Trillo, I. (2015). Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife* 4, e08904.