科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　２　日現在

機関番号：　１３３０１
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５７０６３

研究課題名（和文）次世代シーケンサーによる系統解析の革新

研究課題名（英文）Revolution of Phylogenetic Analysis with Next Generation Sequencers

研究代表者
西山　智明（Nishiyama, Tomoaki）

金沢大学・学際科学実験センター・助教

研究者番号：５０３９０６８８

交付決定額（研究期間全体）：（直接経費）　　2,900,000円、（間接経費）　　870,000円

研究成果の概要（和文）：次世代シーケンサーを用いて、系統学的に枢要な位置にある生物の転写産物とゲノム両方のシーケンシングを行う事で、これまで困難であった古い分岐順序を統計的に高い信頼性をもって推定する事ができるようにすることを目指して研究を進めた。ゲノムサイズが400 Mb以下程度の生物については核ゲノムまで決定する事が現実的に可能であることがわかった。ゲノム配列を決定する事は、発現量の低い遺伝子、全長の長い遺伝子までデータを取得する事を考えるとかえって安価であるかもしれない。

研究成果の概要（英文）：This research aimed at sequencing both the genome and the transcriptome of organisms that occupy pivotal position in the phylogeny to allow resolution of difficult phylogenetic problems such as deep branches with high statistical confidence. A liverwort species, Jungermannia infusca, with an estimated genome size of 386 Mb was sequenced and assembled, indicating that sequencing nuclear genome of less than 400 Mb is realistically doable, and could be less expensive than focusing RNA-seq when one want to obtain low-expression level genes and very long transcripts.

研究分野：　生物学

科研費の分科・細目：　生物多様性・分類

キーワード：　ゲノム　トランスクリプトーム　系統解析

１．研究開始当初の背景

陸上植物の基部系統関係については、申請者らの葉緑体に共通にコードされている全タンパク質のアミノ酸配列の解析からコケ植物３群(セン類、タイ類、ツノゴケ類)が単系統であるという結果が得られていた(Nishiyama et al. Mol Biol Evol. 2004 21:1813-1819)。しかし、新たに小葉類の配列が加わり色素体の遺伝子だけでは十分な解像度が得られない事がわかって来た。また、Qiu ら(PNAS 2006 103:15511-15516)はGC含量の違いがあるにもかかわらず、核酸レベルの解析に基づいてタイ類が陸上植物のもっとも基部で分岐し、ツノゴケ類が維管束植物にもっとも近縁であるという系統関係を提唱しており、かなり広く受け入れられている。

陸上植物及び近縁の生物のゲノム解析は被子植物で複数の種類の植物ゲノムが決定されている他､セン類のヒメツリガネゴケ、小葉類のイヌカタヒバのゲノムが解読済みであり、タイ類のゼニゴケのゲノム解読が進行中である。しかし、外群のシャジクモについては EST 解析にとどまっており、ツノゴケ類とシダ類についてはゲノム解析がなされていない。申請者はこの問題を解決するには多数の核コード遺伝子を用いた系統解析を行うことが必要であると考え、実際に複数の核コード遺伝子の cDNA クローニング及び配列決定を進めつつあった。

そうした中、新型シーケンサーが市場に登場し､国内でも使えるようになってきた。新型シーケンサーはシーケンスのコストパフォーマンスに優れるが、シーケンスサンプルの調整法及びデータ解析について従来のシーケンシングとは異なる取り扱いが必要となる。特にデータ解析法においては、単一種のゲノム解析を目指しての研究は進展しているが、多様性研究という観点での開発はほとんどなされていないのが現状である。

２．研究の目的

本研究では、一種の生物を徹底的に解析するゲノム科学と、一部特定領域だけを多種間比較する系統学を融合し、ゲノムワイド系統解析手法の確立を目指す。新型シークエンサーを用いて、画期的に高スループット低コストで、系統解析に用いる事ができるデータを取得する方法を開発し、陸上植物の基部系統という難題の解決を目指す。

３．研究の方法

（１)データ取得

代表的な陸上植物（裸子植物のクロマツ、イチョウ；シダ類のリチャードミズワラビ；コケ植物タイ類のオオホウキゴケ・コマチゴケ；コケ植物ツノゴケ類のホウライツノゴケ）および緑藻類メソスティグマについて、イルミナの Paired-End ライブラリーを調整し、次世代シーケンサーによる解読を行った。

さらに、オオホオキゴケについて Mate Pair ライブラリーを作成し解読を行った。

また、RNA-seq ライブラリーの調整法について標準的には真核型 mRNA の 3'末端に存在する polyA を利用して精製する方法が従来用いられていたが、この方法だと 5'側領域の出現頻度が相対的に低下するおそれがある。そこで、oligo dT を利用した polyA RNA の精製に代わり、rRNA を除去する方法についても比較検討することとした。このため、オオホウキゴケの単一の RNA より両方の方法での RNA-seq ライブラリーを調整し塩基配列を解読した。

タイ類の基部で分岐したオオホウキゴケ*Jungermannia infusca*、ストレプトファイツ類の基部で分岐したと考えられる*Mesostigma viride* について Paired-end library, Mate-pair ライブラリーを作製しシーケンシングを行った。

（２）アセンブリー

オオホウキゴケの paired end library, Mate pair library 双方のデータを ALLPATHS-LG を用いてアセンブルした。

（３）アノテーション

得られたオオホウキゴケのゲノムデータを CEGMA を用いて解析し、真核生物全般で非常に保存的な遺伝子セットがどれだけ見つかるかを評価した。また、ここで見つかった超保存的遺伝子の構造にもとづいて ab initio アノテーションプログラム Augustus のトレーニングを行った。

多重リピート配列を同定するため RepeatModeler を用いてオオホウキゴケの種特異的反復配列のモデルを作成した後、RepeatModeler を用いて反復配列を同定し、mask したゲノム配列データを用意した。

得られたオオホウキゴケ RNA-seq データを、ALLPATHS-LG による反復配列マスク済みゲノムに TOPHAT を用いてマッピングした。TOPHAT で得られた、スプライシングを考慮したアラインメントからイントロン情報を抽出し、Augustus で処理し、Augustus を再トレーニングした。

さらに、シロイヌナズナとヒメツリガネゴケのアミノ酸配列データセットを query に exonerate を用いて相同性が保存されている領域を同定し、これも Augustus 用のヒントとして、RNA-seq からのイントロン情報、ヒメツリガネゴケおよびシロイヌナズナの遺伝子との相同性情報を統合した遺伝子予測を実行した。このときは、マスクしていない元のゲノム配列を用い、推定多重配列領域は、exon ではないであろうというヒント情報としての推定を行った。

４．研究成果

ペアエンドライブラリーのみにもとづく解析は、当初予定より困難であったが、メイトペアライブラリーを作成することで良好な解析結果が得られる事がわかった。

Mesostigma viride については、Paired-end については期待された量のデータ(43.5 Gb)が得られたが、Mate-pair については、非常に duplicate のレベルの高い(90%)データとなり、更なるシーケンスの追加が必要であった。この原因については、DNA 標品が酵素反応に抵抗性を示す事から DNA の品質に原因があると考えており、今後、DNA 抽出法の検討が必要である。

オオホウキゴケについては、ペアンドライブラリー、メイトペアライブラリーともに良好なものが得られ、Allpaths-LG を用いたアセンブリーの結果、Total scaffold 383 Mb, Scaffold N50 244 kb, Contig N50 14.2 kb の比較的良好な概要ゲノム配列が得られた。CEGMA による超保存的 248 遺伝子の評価でも、complete 91%, partial 95%に達した。

Total Contig は 287 Mb であり、96 Mb はアセンブリーギャップであり、その多くはリピート配列と考えられる。さらに、RepeatModeler, RepeatMasker を用いた解析により、65 Mb の LTR element, Unclassified のリピート 76 Mb 等計 155 Mb が散在反復配列として認識され、マスク後の配列としては 130 Mb が残った。

RNA-seq のマッピング結果、polyA RNA の精製による方法では 75～85%程度のマッピング効率であったのに対し、rRNA を除去する方法では 40～50%程度の効率に留まった。

RNA-seq のデータをイントロンのヒントにした探査では 13,703 遺伝子、15,345 モデルが見つかり、さらに、シロイヌナズナ、ヒメツリガネゴケとの類似性をヒントに加えて、25,105 遺伝子、27,157 モデルを推定した。

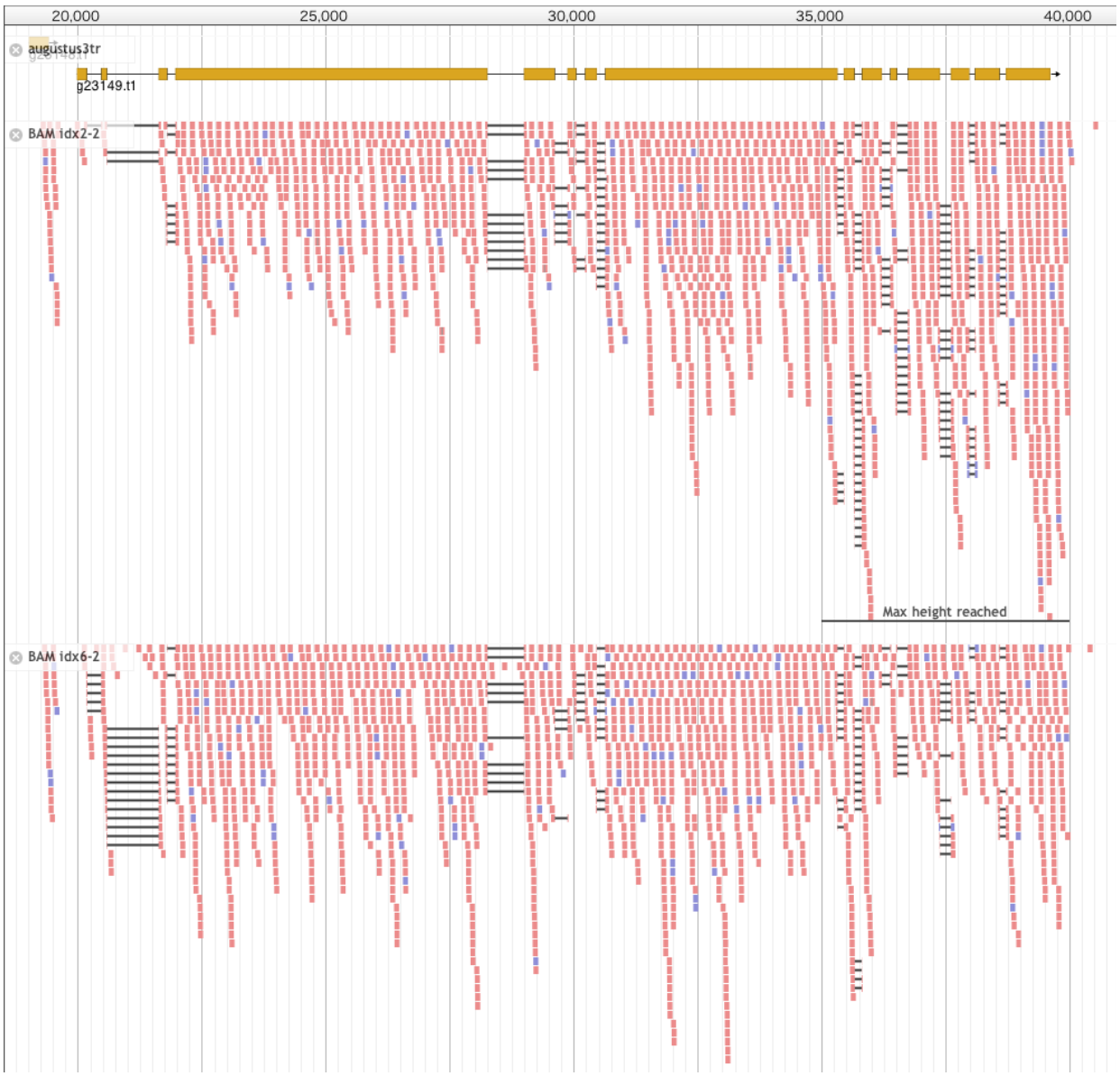
シロイヌナズナで最長の遺伝子 DOC1 の相同遺伝子が正しく再構成できているか検討した。NP_186875.2 の配列を query にして BLASTP 検索を行い 37 番目のアミノ酸から 5095 番目にわたるアラインメントが得られた。ヒット領域のアラインメントの 5295 座位中 48%が一致、65%が類似、5%が Gap とほぼ全長にわたりもっともらしい配列が得られていると評価した。こういう長い mRNA の 5’端側シーケンスは、RNA-seq をしてもなかなかつながらず、5’RACE を幾度も繰り返してもなかなか到達できないものであり、ゲノムの配列を読む事と組み合わせることでかえって容易に有効にデータが取得できるということが判明した。

ゲノムシーケンシングは、発現量が低いなどの理由により、RNA-seq で配列取得するのが困難な遺伝子についても、もともと同程度の頻度で存在するので安定した網羅性が期待される。本研究により、オオホウキゴケで実際にこの戦略がうまく行く事が示された。今後、さらにゲノムサイズが大きな生物で、効率的にデータを取得する方法についての研究開発が望まれる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
西山　智明　　（NISHIYAMA, Tomoaki）
金沢大学・学際科学実験センター・助教
研究者番号：50390688



オオホウキゴケ DOC1 相同遺伝子領域
推定された遺伝子構造と、RNA-seq のデータが良く一致している事を示している。また polyA RNA を用いた RNA-seq (中段)より、rRNA 除去による RNA-seq の方が、5’末端付近まで均一な厚みのデータが得られている。