様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３１　日現在

機関番号：　１５３０１
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１４８５４

研究課題名（和文）果実発育・成熟関連遺伝子の完全長cDNA配列と選択的スプライシングの網羅的解析

研究課題名（英文）Comprehensive analysis of full length cDNA and alternative splicing for the genes related to fruit development and ripening

研究代表者

牛島　幸一郎（Ushijima, Koichiro）

岡山大学・環境生命科学研究科・准教授

研究者番号：２０３７９７２０

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：第３世代の次世代シークエンサーの普及により，ゲノム解読だけではなく転写産物の解読に利用されている．本研究ではメロン果実のIso-seq解析を行い，完全長cDNA配列と選択的スプライシングについて解析を行った．３つの異なる方法で完全長cDNA配列ライブラリーを構築し，PacBioとNanoporeを用いて配列解読を行った．得られたIsoformをメロンゲノムにマッピングして比較した結果，PacBioの解読では 約6,500遺伝子座，Nanoporeの解読では12,000の遺伝子座からの転写産物が得られた．そのうち，509遺伝子座が新規の遺伝子座であった．

研究成果の概要（英文）：3rd generation sequences promotes not only genome sequencing but also transcript analysis. Iso-seq analysis using PacBio are very useful method for reading full length cDNA and detecting alternative splicing. In this analysis, we carried out Iso-seq analysis for melon fruits. We used three different methods for construction of full length cDNA library. The libraries were sequenced by PacBio sequel and Nanopore MinION. Isoforms were mapped on the reference genome of melon and compared them. PacBio and Nanopore isoforms fell into ~ 6,500 and 12,000 loci respectively. 509 loci of them were located at intergenic region and predicted novel loci.

研究分野：　園芸学

キーワード：　Iso-seq

１．研究開始当初の背景

ゲノムプロジェクトの進展により多くの生物で全ゲノム配列，全遺伝配列が明らかとなってきている．この過程で，その複雑さに対して高等生物の遺伝子の数が少ないことが古くから指摘されている．実際にシロイヌナズナやヒトは線虫と同程度の数の遺伝子しか保持していないが，線虫と比べ非常に複雑な器官を形成し，高度な生活環を営んでいる．これは，限られた数の遺伝子でそれらを実現するために選択的スプライシングが利用されており，実質的な遺伝子の数を増やし，細かな機能分化・調節を可能としていると考えられている．

近年の Illumina HiSeq に代表される高速シークエンサーの出現によりトランスクリプトーム解析が容易になっている．しかし，Illumina リードは１つが 100 bp 程度と短いため，長さやゲノムでの重複，選択的スプライシングなどにより，全長配列やスプライスバリアントが正確に特定できていないケースが多い．近年，PacBio から平均 10 kbp のロングリードの解読が可能なシークエンサーが登場し，また Illumina も同程度の長さの配列の解読を可能にする Synthetic Long Read のシステムを販売している．これらを利用すれば１リードで１つの転写産物の完全長配列を決めることが出来るので，ヒトなどでは全スプライシングを特定する Iso-seq 解析が盛んになってきており，植物などの他の生物にも応用可能であると考えられる．

２．研究の目的

植物でも選択的スプライシングの重要性は指摘されており，特定の遺伝子に限った研究であるが選択的スプライシングと機能・形質との関連が明らかとなっている．果実の形成や発育，成熟は非常に複雑な現象で有り，多くの遺伝子が協調的に働いていると考えられる．選択的スプライシングも関わっていると考えられるが，その網羅的な解析は成されていない．本研究では，クライマクテリック型果実の代表の一つであるメロンに関して Iso-seq 解析を行い，遺伝子の完全長配列を決定・カタログ化し，研究基盤となるデータベースを作成することを目的とする．

３．研究の方法

(1) 材料および RNA 抽出

メロン品種 ”春３“ の葉および複数の成熟段階の果肉をサンプリングした．フェノール抽出により total RNA を抽出した．

(2) ライブラリーの構築

本研究では Clontech 社の SMARTer kit，Lexogen 社の Teloprime kit，また Oligocaping 法の３つの方法を採用した．各々の方法について，解読後に demultiplex の操作をできるように，逆転写用プライマーにバーコード配列を付加した．cDNA の合成と共にアダプター配列を付加して，TaKaRa PrimeGLX で増幅した．得られた断片は AMPureXP で増幅後に，３つのライブラリーを混合してシークエンシングに供試した．

(3) シークエンシング

同一のライブリーミックスを用いて PacBio の Sequel と Nanopore の MinION の２つの方法で解読した． Sequel についてはサイズセレクションを行わずに解読している．MinION についてもサイズセレクションは行っていないが，1D と 2D の異なった解読法で解読した．

(4) 完全長配列の取得

① PacBio のリード

PacBio が提供している SMRT link には Iso-seq のパイプラインが組み込まれている．そこで，ライブラリごとに demultiplex を行い，各構築手法ごとに完全長のリードを得た．

② Nanopore のリード

PacBioのCCSの様な明確なリード補正の手法がない．そこで，ゲノムのアセンブリーソフトである CANU のエラー補正機能を利用してリードの補正を行った. PacBio とはことなり，断片を完全に解読しているとは考えにくかったので，両端にアダプター配列がある断片のみを抽出して，ゲノム配列にマッピングした.

(5) 配列解析

PacBio リードについては STAR，GMAP，minimap2 の３つの aligner を用いて，メロンゲノム（3.6.1）に mapping した．得られた sam ファイルをカスタムスクリプトで gff フォーマットに変換した. gffcompare などで既存のゲノムアノテーションと比較し，mapping について評価した．Nanopore については minimap2 でのみ mapping を行った.

４．研究成果

(1) Mapping ソフトの選択

第３世代の長鎖を大量に Mapping するソフトウェアは複数開発されているが，発展途上の感がある．そこで，PacBio で解読したリードを用いて Minimap2，STAR，GMAP といった代表的な Mapping ソフトを試すことにした．リファレンスにはメロンのゲノム配列（3.6.1）を使用した.

Map 率については STAR が最も劣っていた（表１）. GMAP や minimap２は 99%と高い Map 率をしめした．しかし，実際にはクオリティが低い mapping（ミスマッチが多い，短い）を多く含む. そこで, ソフトウェアが算出する Map クオリティ，リードのカバレージを元にフィルタリングをおこなった． 3 者とも 50%を切るマップ率となったが，最も良かったのは minimap2 でそのマップ率は 49%であった.

表１　Mapping率

ソフト	フィルター前		フィルター後	
	マップ数	率 (%)	マップ数	率 (%)
minimap2	218,529	99.86	107,720	49.22
STAR	170,734	78.02	89,727	41.00
GMAP	217,200	99.25	103,417	47.26

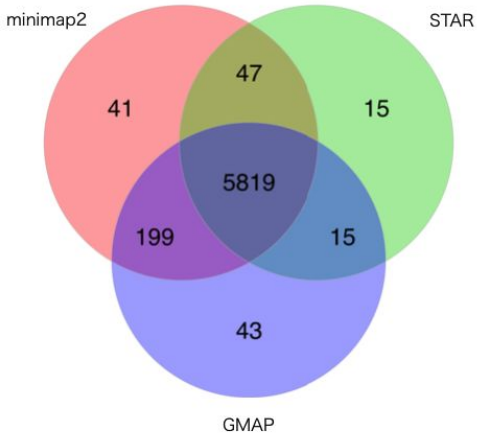


図１ マッピングソフトの比較

３者のソフトエア間での遺伝座の重なり具合を解析した（図１）． ３者を合わせると 6,179 遺伝子座であり，そのうち 94%の 5,810 遺伝子が３者間で共通しており，どのソフトウェアを選択しても結果に大幅な違いが出ないと考えられた.

既存の CM4.0 アノテーションと比較を行い，転写開始点（TSS），転写終結点（TES）と Isoform の比較を行ったが，minimap2 がもっともよい結果となった．そこで，以降は Minimap2 を用いた解析を行うことにした.

(2) ライブラリー構築法の評価

３つの方法を利用してライブラリーを構築したが Oligocapping 法は断片の増幅が少なく，他の 2 者より得られたリードが少なかった． ３つのライブラリーについて，得られた完全長配列を minimap2 にてマッピングして，gffcompare によるゲノムアノテーションとの比較や TSS，TES の解析を行った．TSS や TES の解析に関しては Teloprime が最も良い結果となった．ただ，相互に比較すると，リードの母数にかかわらず一定数の作成法に特異的な遺伝子座が存在していた（図２）. ライブラリー作成法の違いが解読遺伝子の種類に影響を及ぼすと考えられたので，３つの手法を混ぜた状態で解析を行った.

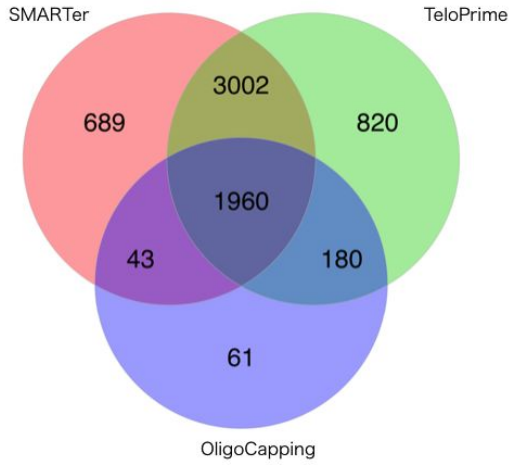
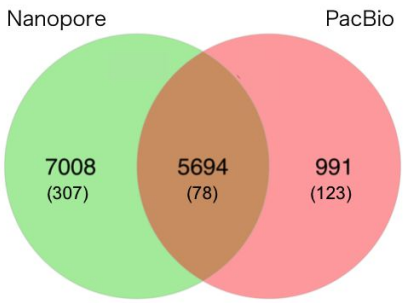


図２ ライブラリー構築法の比較

(3) Nanopore 利用の検討

Nanopore は PacBio と同様に解読精度は 90%程度であり，CCS の様な補正も出来ない．たとえば，STAR でリードデータを補正無しで Map すると数%程度しか mapping 出来ない．そこで，CANU でリードを補正した後に，Minimap2 で mapping を行った．gffcompare を用いて転写産物のマージや比較を行った（図３）．その結果，約 12,000 遺伝子座にマップされていることが明らかになった．PacBio の解析で特定されたのは，6500 遺伝子座であり，Nanopore の方が約２倍の遺伝子座を特定できた．２つの解読手法の間で共有する遺伝子座は 5694 であった．



(4) 新規遺伝子座の特定

今回の解析で PacBio，Nanopore 合わせて 13,693 の遺伝子座が特定された．大半が既存のアノテーションデータの遺伝子座とオーバーラップしていたが，509 遺伝子座は遺伝子領域以外にアノテーションされた新規の遺伝子座であった．

(5) まとめ

非モデル作物における完全長 cDNA の解析をライブラリー構築法とシークエンシング法について複数条件で行い比較した．その結果，ライブラリー構築の方法によって増幅する断片種が異なると考えられ，可能であれば複数の方法を併せて解析する方が好ましいと考えられた．シークエンス法については一般に用いる PacBio に比べて，Nanopore からは２倍の Isoform がえられた．これは，リード数が PacBio の装置は ZMW の数に制限されるのに対して，常に新規の DNA が通過し解読される Nanopore の方が有利であると考えられる．しかし，解読精度が低いため，今回のようなクオリティの高いゲノム配列が手に入る生物ならば問題ないが，参照配列の無い生物種では依然として PacBio を用いた Iso-seq 解析が有利であると考えられた．

５．主な発表論文等

研究期間中に発表当行っていないが，H30 年度中に論文投稿，研究発表を行う予定である．

６．研究組織

(1)研究代表者

牛島幸一郎（USHIJIMA，Koichiro）
岡山大学・大学院環境生命科学研究科・准教授
研究者番号：20379720

(2)研究分担者

門田 有希（MONDEN，Yuki）
岡山大学・大学院環境生命科学研究科・准教授
研究者番号：30646089

赤木 剛士（AKAGI，Takeshi）
京都大学・大学院農学研究科・助教
研究者番号：50611919