交雑個体における選択的スプライシングの解析パイプライン構築

大塚 好恵　(指導教員：瀬々 潤)

1　はじめに

植物の倍数化，特に異種の交雑による倍化である異質倍数化 (図 1) は，大型化・種無し品種などにつながり、農業的に重要な役割を果たすと同時に，倍化前に比べて広い環境適応能を示すため，環境・進化学的にも重要である．倍化した植物の例としてコーヒー，小麦，ジャガイモ，バナナが挙げられる．ところが，これらの作物に関して生物学的理解は，倍数化していないモデル生物に比較すると，それほど進んでいない．一般に，異質倍数化した異質倍数体は近縁の種が交雑するため，互いのゲノム配列が近く，どの遺伝子が働いているのかを正確に見分ける事が難しかったためである．近年，超並列シーケンサ (NGS) の発達により，安価かつ大量に遺伝子配列を観測することが可能となり，異質倍数体の理解が飛躍的に向上している．近年では，遺伝子網羅的な遺伝子発現量の解析も可能となってきた [1]．しかしながら，まだ限界も多い．一例として，1 つの遺伝子から複数の機能が発現する際に見られる選択的スプライシングは，限定的な遺伝子でのみしか観測されていない．本研究では，異質倍数体において選択的スプライシングを遺伝子網羅的に観測する計算機的な解析パイプラインを構築する．これにより，異なる親に由来する類似遺伝子群が，同一の選択的スプライシングを起こしているか，あるいは，異なるスプライシングを起こし機能分化を誘発している可能性があるのかの調査につなげる．

2　研究背景

2.1　異質倍数体の RNA-seq

NGS は，対象の塩基配列を大量に読み取ることができる．特に，NGS を利用した遺伝子網羅的な発現解析を RNA-seq と呼ぶ．RNA-seq では，サンプルから抽出した mRNA を短く切った上で大量に読み，対象種のゲノム配列のどの位置に由来するかを検索（マッピング）することで，各断片がどの遺伝子に由来するかを調べる．各断片をリードと言う．この操作を数千万本のリードに対し行うことで，全遺伝子の発現量を観測する．RNA-seq では細胞中の RNA を一緒くたに解析するため，異質倍数体の RNA-seq ではリードの親種を区別した由来遺伝子の定量的な観測は，困難であると考えられていた．しかし，親種の配列に存在する配列の差異を考慮し，どちらの親の遺伝子から転写されたものかを分類することが可能となっている [1]．

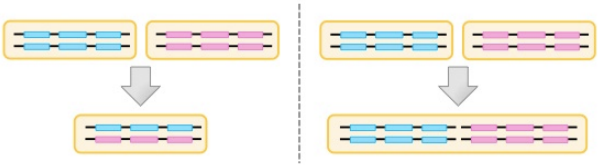

図 1: 倍数化の説明．左はヒトのような二倍体，右は二倍体が交雑した四倍体である．

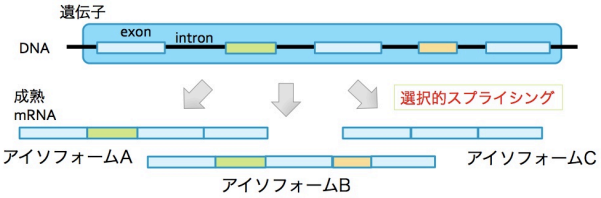

図 2: 選択的スプライシング．生成された mRNA をアイソフォームと呼ぶ．

2.2　選択的スプライシング

生物が環境適応能を獲得する過程で重要な現象として，倍数化を起こす事の他に選択的スプライシングの利用が挙げられる．遺伝子領域には遺伝情報を持つエクソン部分と，それらをつなぐイントロン部分がある．前駆体 RNA から成熟 RNA になる際にイントロンを除きエクソン同士をつなぎ合わせることをスプライシングという．殆どの場合全てのエクソンを利用するが，いくつかの遺伝子では限られたエクソンをつなぎ合わせる選択的スプライシングが観測される (図 2)．また，１つの遺伝子から複数の種類のスプライシングが生成されることもあり，１つの遺伝子から複数のタンパク質を生成することが可能となる．異質倍数体においてもこの選択的スプライシングが起こっていると考えられるが，これまでに遺伝子網羅的な解析は行われてきていない．RNA-seq を用いてスプライシングを観測するソフトウエアとして，Cufflinks[2] を始めとしたソフトウエアが開発されているが，異質倍数体に適用すると，相同遺伝子同士を混同し，正しくスプライシングパターンを同定できない．本研究では，この問題を解決し，異質倍数体から取られた RNA-seq よりスプライシングパターンを同定する．

3　手法

3.1　対象とする植物

提案手法は，どの異質倍数体に対しても適用可能な汎用的な手法であるが，本研究では植物のモデル生物であるシロイヌナズナの近縁種である *Arabidopsis halleri*(ハクサンハタザオ；2 倍体)，*Arabidopsis lyrata*（西洋ミヤマハタザオ；2 倍体）の 2 種を両親にもつ *Arabidopsis kamchatica*（オウシュウミヤマハタザオ；異質 4 倍体）の解析を行った．解析に際して，3 サン

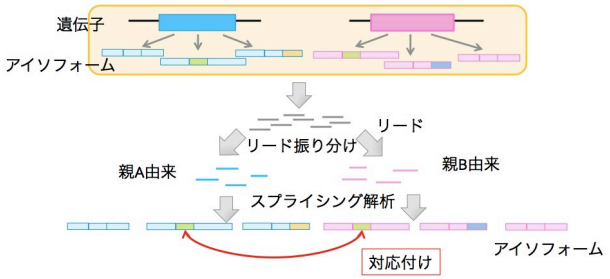

図 3: 提案手法の流れ．

プルから取得した RNA-seq，計 119,123,927 リードを使用した．

3.2 提案手法

本研究で提案する解析は，以下のような構成である (図 3)．

由来の特定： *A. kamchatica* から抽出された各リードの由来する親種を推定する．マッピングのソフトとして STAR[3]，リードの由来判定ソフトとして HomeoRoq[1] を用いた．

構造推定： 親種を同定したリードについて，各親種において構造の推定を行う．そのために Cufflinks[2] を利用する．Cufflinks は，マッピングの結果を入力することで，遺伝子のスプライシングパターンを同定する．エクソンをまたぐリードがあった時，そのエクソンが隣同士であれば通常のスプライシングであるが，中間にあるエクソンを飛ばして対応付けられるようであれば，選択的スプライシングが起こっていると考えられる．この情報から，遺伝子から発現している選択的スプライシングのパターンを抽出する．

アイソフォーム同士の対応付け： Cufflinks の結果から，アイソフォームの塩基配列について，対応付けを Reciprocal Best Hit(RBH) 法を用いて行う．対応付けには，遺伝子配列検索における標準手法である BLAST[4] を用いる．*A. halleri* 由来のすべてのアイソフォーム中から，*A. lyrata* 由来の個々のアイソフォームに対し最も E-value の低いアイソフォームを検索する．同様に *A. halleri* からの検索も行う．この結果から，*A. halleri* と *A. lyrata* で双方において 1 番類似していると判定されたアイソフォーム同士を対応関係にあるとした．

4 結果と考察

4.1 構造推定の結果

A. halleri 由来，*A. lyrata* 由来の発現した遺伝子数がそれぞれ 28,200 個，26,562 個に対し，発現したアイソフォーム数はそれぞれ 50,715 個，49,274 個であった．このことから，選択的スプライシングにより 1 つの遺伝子から平均 1.8 個のアイソフォームが生成されていると考えられる．次に，各遺伝子から生成されたアイソフォーム数を解析した．*A. halleri* 由来,*A. lyrata* 由来ともに，1 つの遺伝子から 1 つのアイソフォームのみ発現している割合が 60 %であった．一方で，選択的スプライシングを起こしたと考えられる遺伝子が 40 %存在した．最も多くのアイソフォームが観測された例としては，1 つの遺伝子から 15 個のアイソフォームが生成されている例もあった．

4.2 対応付けの結果

Akama らの研究で同定された *A. kamchatica* の重複遺伝子のペアは 24,880 個である．これらの遺伝子が生成するアイソフォーム数の合計は，*A. halleri* 側が 48,768 個，*A. lyrata* 側が 49,024 個であった．これらに対し RBH 法による対応付けを行い，さらにその中から対応遺伝子間のアイソフォームのみの結果に絞った．図 4（a）は，対応する遺伝子がもつアイソフォームの数の分布である．対応する遺伝子同士が生成するアイソフォーム数は，必ずしも非常に高い相関を示している訳ではないが，十分高い相関が見られ，アイソフォームを生成しやすい遺伝子は，いずれの親でもアイソフォームを生成しやすい傾向が見られた．

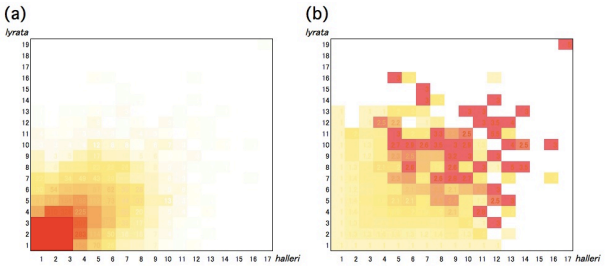

図 4: （a）対応遺伝子のアイソフォーム数の分布（b）アイソフォーム数の組み合わせごとのペア数の平均数.

図 4（b）は，各対応遺伝子内における相同的なアイソフォームの数の組み合わせをまとめたものである．最大のペア数は *A. halleri* が 12 個と *A. lyrata* が 11 個のアイソフォームを持つとき，8 個が対応付けられる場合であった．全体的に，アイソフォーム数に比べ対応付けられるペアのアイソフォーム数が少ない．確かに，アイソフォーム数が多いほど平均ペア数も多くなっているが，アイソフォーム数が 10 個以上でも平均ペア数が高々 2 ～ 5 程度で，予想よりペア数が少ない結果が得られた．これより，異質倍数体におけるアイソフォームの生成は，重複遺伝子間で共通するのではなく，独立に働いていることが示唆される．

5 まとめと今後の課題

異質倍数体におけるアイソフォームの解析パイプラインを構築した．今後は対応遺伝子間でのスプライシングの相違や，環境による対応遺伝子間でのスプライシングの変化を調べると共に，それらの発現量を考慮することで遺伝子網羅的な解析を行い，本パイプラインの有効性の検証を行っていきたい．

謝辞

産業技術総合研究所・ゲノム情報研究センター　赤間悟研究員より大変有用なアドバイスを頂きました．

参考文献

[1] Akama, S., Shimizu-Inatsugi, R., Shimizu, KK. and Sese, J.: Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucl. Acids Res.*, 42(6),e46(2014).

[2] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, DR., Pimentel, H., Salzberg, SL., Rinn, JL. and Pachter, L.: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*,7,562-578(2012).

[3] Dobin,A., Davis,CA., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,TR.: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*,29(1),15-21(2013).

[4] Altschul, SF., Madden, TL., Schäffer, AA., Zhang, J., Zhang, Z., Miller, W and Lipman, DJ.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*,25(17),3389-3402(1997).