Vol.2017-MPS-113 No.37
Vol.2017-BIO-50 No.37
2017/6/25

非モデル生物における条件依存的選択的スプライシングの網羅的発見手法の改良

米澤 弘毅[1,a)]　嶺井 隆平[1,b)]　小倉 淳[1,c)]

概要：選択的スプライシングは，遺伝子数が限られている中でエクソンの組み合わせを変えることで多様な遺伝子産物を発現させるメカニズムであり，非常に多くの生物における組織や発生プロセスでの異なった遺伝子発現を示す要因でもある．選択的スプライシングのメカニズムや役割は現在までモデル生物においてはよく研究されているが非モデル生物ではあまり研究が進んでいない．この原因として，選択的スプライシングを調査する方法がゲノム配列および完全なエキソン-イントロン情報を必要とすることが挙げられる．しかし、様々な生物の選択的スプライシングを概観したり，選択的スプライシングの進化的なインパクトやその役割を理解するためには，非モデル生物における選択的スプライシングの調査が不可欠である．筆者らは既に，de novo トランスクリプトームアセンブリ手法として有名な Trinity を利用した条件依存的選択的スプライシングを網羅的に発見する手法である DASE を提案しているが，本稿では各バリアントの様々な条件下における発現量の違いに関する新たな指標を計算する機能を DASE に追加し，この追加機能による条件依存的選択的スプライシングの発見能力の変化を検証した．

キーワード：条件依存的選択的スプライシング, 非モデル生物, DASE.

DASE3: Differential Alternative Splicing variants Estimation method without reference genome with improvement of its accuracy by introducing a new measure

Kouki Yonezawa[1,a)]　Ryuhei Minei[1,b)]　Atsushi Ogura[1,c)]

Abstract: Alternative splicing is a mechanism underlying gene expression diversity under the constraint of a limited number of genes and causes spatio-temporal variation in gene expression in tissues and developmental processes in most organisms. This mechanism is well studied in model organisms at present but not in non-model organisms because the current standard method requires genomic sequences as well as full annotation information on exons and introns. Nevertheless, it is necessary to uncover the landscape of alternative splicing in various organisms and to understand its evolutionary effects and roles. Previously, we have proposed the DASE method for condition-specific estimation of alternative splicing without reference genomes based on *de novo* transcriptome assembly. In this paper, we improved estimation of differentially expressed variants by means of a function evaluating how different from other mRNA variants a tuple of expression quantities of a particular variant is. The software is deposited at `https://github.com/koukiyonezawa/DASE`.

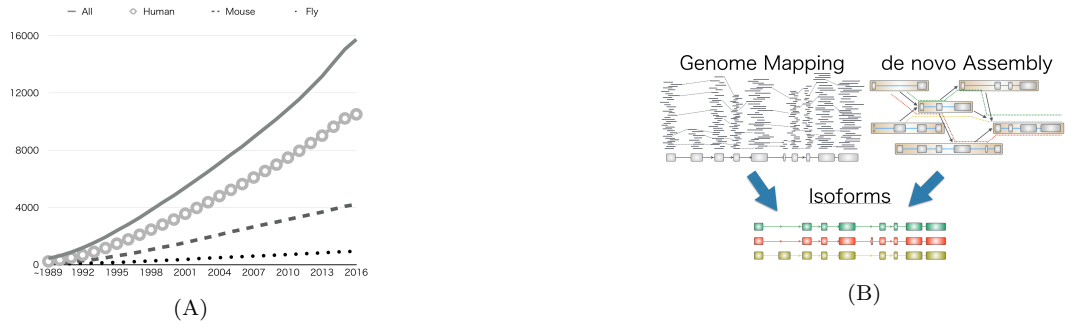
Keywords: condition-specific alternative splicing, non-model organism, DASE

1. Introduction

選択的スプライシングは，40 年前に初めて発見された 1 つの遺伝子座から複数の配列を生成するメカニズムであ

[1] 長浜バイオ大学
[a)] k_yonezawa@nagahama-i-bio.ac.jp
[b)] mineiryuhei@gmail.com
[c)] a_ogura@nagahama-i-bio.ac.jp

Vol.2017-MPS-113 No.37
Vol.2017-BIO-50 No.37
2017/6/25



(A) (B)

図 1 (A) 選択的スプライシングに関する研究論文数は急速に増加しているが，その大半はモデル生物についてのものである．(B) ゲノムマッピングと de novo アセンブリによるアイソフォーム設計．図の初出は [22].

る ([1], [2])．このメカニズムは転写後遺伝子制御を行っている間の遺伝子の多様性を増加させ，異なる機能，場合によっては逆の機能を有する mRNA を生成する [3]．さらに，選択的スプライシングが組織や発生段階に寄って異なる制御をしていることも多くあることが報告されており，1 つのスプライシングアイソフォームが生命システムにおいて時空間的に特定の役割を有し得ることが示唆されている ([4], [5], [6])．最近，組織特異的な選択的エキソンが翻訳後修飾およびタンパク質間相互作用を媒介している結合モチーフのターゲットとなるサイトに多く，これによって該当タンパク質の機能を多様化させていることが報告されている ([5], [6])．スプライシングバリアントの時空間的な制御は標的となるコーディング配列の隣接した領域に内接するスプライシングコードによって調節されることが知られている [7].

このメカニズムは酵母からヒトに至るまで広い範囲の生物種に存在することが知られている ([8], [9])．それに加え，選択的スプライシングの効果は生命システムに深く関わっている．例えばヒトの遺伝子のうち 95%以上が選択的スプライシングによって制御されて 2 つ以上のスプライシングアイソフォームを産生し，それらが様々な組織や発生段階において別の機能を持つと考えられている ([4], [10])．PubMed 検索を用いると現在までに選択的スプライシングに関する論文が 16,000 本以上出されており (Fig. 1A)．この急速な増大は次世代シークエンス技術とゲノム解析のための高速な計算手法が発達してきたことが寄与している [11].

しかしながら，選択的スプライシングの研究はほぼヒトやマウス，ハエといったモデル生物を扱っている ([12], [13], [14], [15])．これには主に 2 つの理由があり，1 つは選択的スプライシングが発生における機能や疾患に深く関係しており，機能解析のためにはモデル生物が適していること [16]，もう 1 つはスプライシングバリアント発見のための標準的な手法において，リードをマッピングするためにリファレンスとなるゲノム配列やアノテーション情報を必要とすることである (Fig. 1B)．

モデル生物においては，Tophat [17]，STAR [18]，Cufflinks [19] といった通常の RNA-seq 解析の場合は最初にリファレンス配列にリードと呼ばれる短い塩基配列をマッピングし，BAM あるいは SAM フォーマットの出力を得る．Cuffdiff [19] のような選択的スプライシングの発現量解析によって，アイソフォームの複数条件間での発現量の差を計算できる．しかし，選択的スプライシングの進化的なインパクトや保存されているメカニズムを解明するためには，できるだけ多くの生物種における選択的スプライシングを広範囲に解明することが不可欠である．非モデル生物における de novo アセンブリの手法としては多くの場合 Trinity [20] が用いられる．Trinity はデータに存在する転写産物の大部分を再構成するとともに，選択的スプライシングによるアイソフォームを発見する．

通常 Trinity を使用する場合には，良い結果を得るために複数条件下で得られたリードをまとめて実行する．しかしながらこの工程により，どのスプライシングバリアントが条件特異的なのか評価が難しい．ゲノム配列を使用せずに選択的スプライシングバリアントを評価するソフトウェアは数えるほどしか存在しない．我々が調査した限りでは，ABSSeq [21]，ALDEx2 [23]，Alexa-seq [24]，ARH-seq [25]，DiffSplice [26]，MISO [27]，SpliceSeq [28]，といったソフトウェアが存在するが，そのほとんどはゲノム配列やアノテーション情報が必要である．KisSplice [29] と LEMONS [30] だけはゲノム配列がなくても動作する [21] が，ゲノム配列やアノテーションを必要としない選択的スプライシング評価用ソフトウェアはまだ数が少ない．

この問題に対し，筆者らは DASE [22] と呼ばれる条件依存的選択的スプライシングバリアントの網羅的発見手法を構築した．DASE は Trinity によって得られたバリアント全てを入力とし，同じ遺伝子から産生された他のバリアントとの条件による発現量の違いおよび配列の重複度からバリアントをランク付けするアルゴリズムである．

本稿では DASE の新たなバージョンとして，発現量の違

Vol.2017-MPS-113 No.37
Vol.2017-BIO-50 No.37
2017/6/25

いに関する別の指標を実装し，種々の RNA-seq データに対してその有用性を検証した．具体的には，同一遺伝子由来のバリアントの発現量の違いの尺度としてコサイン角度を用いていたが，これをマハラノビス距離に基づいた角度を用いることで，条件依存的なバリアントに対して若干高いランクを与えることが出来た．

2. DASE アルゴリズム

フレームワーク

DASE は同じ遺伝子由来のバリアントのペアにおける発現量の差と塩基配列の重複率を使用する．全体のフレームワークは図 2 のようになっている．

de novo アセンブリ

最初に異なる条件下で得られたすべてのリードをまとめ，そのデータを入力として Trinity を実行し，バリアント[*1]の配列を得る．ここで得られたバリアントは，Trinity のルールによりクラスタに分割される．

発現量の差の計算

de novo アセンブリ実行後，すべてのバリアントに対するそれぞれの条件下での発現量を計算する．発現量の計算には Kallisto [31] がよく使用される．Kallisto は TPM (Transcripts Per Million) の計算を高速に実行する．発現量が微小なバリアントはアーティファクトである可能性があるため，ユーサが指定した発現量に満たないバリアントは無視される．より正確には，$e = \{e_1, \ldots, e_m\}$ を異なる条件下におけるバリアントの発現量の組とすると，発現量の対数の合計値 $\sum_i \log_2(e_i + 1)$ がユーザの指定した閾値 (通常は 2) 未満の場合，そのバリアントは無視される．以降において，バリアントの発現量ベクトル r を発現量の対数の組として与える．すなわち，$r = (\log_2(e_1 + 1), \ldots, \log_2(e_m + 1))$ である．

それぞれのバリアントは長さ m の発現量の対数の組を持っている．言い換えると，それぞれのバリアントは異なる条件の数を次元に持つ空間 G のある 1 点に対応する．DASE の前バージョンにおいては，同じ遺伝子由来のバリアントのペアに対して，空間 G 内の 2 ベクトルの角度 θ を計算することができる．r_1 および r_2 をそれぞれ同じ遺伝子由来の 1 対のバリアント v_1 および v_2 の発現量の対数の組とすると，G 内における 2 ベクトルの角度 θ_c は以下のようにして計算できる．

$$\theta_c(v_1, v_2) = \arccos \frac{r_1 \cdot r_2}{|r_1||r_2|}$$

なお $|r|$ はベクトル r の ℓ_2 ノルムを表す．

[*1] Trinity で得られるのはバリアントの候補であり，Trinity においてはコンティグと呼ばれるが，混乱を避けるため本稿ではバリアントという表記で統一する

本稿では，条件数が 3 以下のときにのみ適用できる発現量に関する 2 ベクトル間の角度について別の指標を用いる．実際の発現量のデータを見ると，異なる条件下における発現量には一定の相関があることが見て取れる．この相関関係を考慮するため，マハラノビス距離 [32] にヒントを得た楕円形角度を導入する．

Σ をすべてのデータ D に対する共分散行列とすると，2 つのベクトル r_1 と r_2 間のマハラノビス距離は以下のように定義される．

$$d_M(r_1, r_2) = \sqrt{(r_1 - r_2)^T \Sigma^{-1} (r_1 - r_2)}$$

ここで r^T はベクトル r の転置行列を，M^{-1} は行列 M のぎゃキュ行列を意味する．I を単位行列とした時，$\Sigma = I$ であればマハラノビス距離はユークリッド距離と一致する．見方を変えると，r_1 と r_2 のユークリッド距離は，r_1 を中心とする円が r_2 に届いたときの半径であり，マハラノビス距離は r_1 を中心とし，共分散行列 Σ で規定される楕円が r_2 に届いたときの径の長さである．

2 ベクトル r_1 および r_2 のマハラノビス距離を考慮した角度は以下のように定義される．単純化のため，2 条件下で発現量の差を考える．E_O を原点を中心とした共分散行列 Σ で規定される楕円とし，その長軸を e_1 とする．a および b をそれぞれ楕円 E_O の長軸および短軸の長さとする．通常，e_1 は xy-平面における直線 $y = x$ とほぼ一致する．最初に，e_1 が x 軸に一致するように D 内のすべてのデータ点を回転させる．この回転行列を R とし，$i = 1, 2$ に対して $u_i = Rr_i$ とする．また，$\arg(u)$ をベクトル u と x 軸との角度とする ($\arg(u) \in [-\pi/2, \pi/2]$)．上記を用いて，発現量ベクトル r_1 と r_2 を持つバリアント v_1 および v_2 の間の新たな角度 $\theta_M(r_1, r_2)$ を以下のように定義する．

$$\theta_M(v_1, v_2) = \begin{cases} \left| aE\left(\frac{\arg(u_1)}{a}, k\right) - aE\left(\frac{\arg(u_2)}{a}, k\right) \right| \\ \qquad \text{if } \arg(u_1)\arg(u_2) \geq 0 \\ \left| aE\left(\frac{\arg(u_1)}{a}, k\right) \right| + \left| aE\left(\frac{\arg(u_2)}{a}, k\right) \right| \\ \qquad \text{if } \arg(u_1)\arg(u_2) < 0 \end{cases}$$

ここで $k = \sqrt{1 - b^2/a^2}$ であり，$E(\phi, m)$ は第二種楕円積分を意味し，$E(\phi, m) = \int_0^\phi [1 - m\sin(t)]^{1/2}\, dt$ と定義される．また，$\Sigma = I$ のとき，$\theta_M(v_1, v_2) = \theta_c(v_1, v_2)$ である．別の見方をすると，通常の角度が原点 O を中心とした単位円に 2 ベクトル r_1 および r_2 を射影した 2 点間の弧の長さに対応し，$\theta_M(v_1, v_2)$ は原点 O が中心かつ共分散行列 Σ で規定される楕円に 2 ベクトルを射影した 2 点間の弧の長さに対応することになる．なお上記のどちらの角度 θ に関しても，$\theta \in [0, \pi/2]$ である．

実験の条件数が 3 の場合は，2 ベクトル r_1 および r_2 から規定される平面 P が楕円体 E を 2 分割し，その切り口 P_c 上に O が存在する．この P_c が xy-平面と一致するように

Vol.2017-MPS-113 No.37
Vol.2017-BIO-50 No.37
2017/6/25

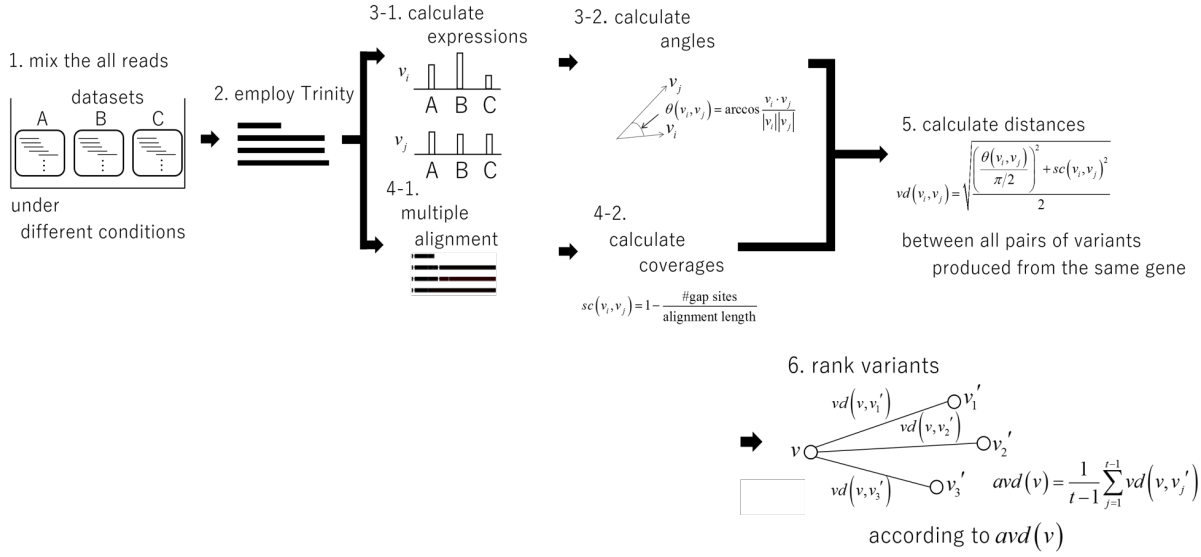


図 2 DASE のフレームワーク.

回転させれば，2 条件の場合と同様の手続きが適用できる．

配列の重複度の計算

いくつかの遺伝子は，異なる条件下では発現量だけでなくエキソンスキッピングやイントロンリテンション (e.g., [33]) といったエキソン構造の大きな変化を伴うことがある．そこで，2 つのバリアント間の塩基配列の重複度を考慮に入れる．この目的のため，同じ遺伝子由来のバリアントの塩基配列に対して MAFFT [34] を実行する．各塩基の違いではなく塩基配列のブロックを見るため，MAFFT におけるギャップペナルティをデフォルトの 1.53 よりかなり大きい 10.0 に設定する．

MAFFT によるマルチプルアラインメントにより，同じ遺伝子由来のバリアントのペア全てに対して重複度を以下のように計算する．s_1 および s_2 をそれぞれ，同じ遺伝子由来のバリアントのペアのアラインメント後の塩基配列とし，$s_{i,j}$ を塩基配列 i の j 番目の塩基とする．また $\ell = |s_i|$ とする．ブロック内のあるサイトに着目すると，そのサイトにある塩基はギャップでないときに重複度を高くすべきである．そのようなサイトの数を ℓ_{nuc} とし，$s_{1,j}$ と $s_{2,j}$ の片方がギャップであるサイトの数を ℓ_{gap} とする．さらに，$s_{1,j}$ と $s_{2,j}$ の両方ともギャップであるサイトの数を ℓ_{neg} とする．(ペアワイズアラインメントをしていないため，2 配列のあるサイトが両方ともギャップとなり得る．) なお，$\ell = \ell_{\mathrm{nuc}} + \ell_{\mathrm{gap}} + \ell_{\mathrm{neg}}$ である．このとき，配列 s_1 および s_2 を持つバリアント v_1 および v_2 の配列重複度は以下のように計算される．

$$sc(v_1, v_2) = \frac{\ell_{\mathrm{nuc}}}{\ell - \ell_{\mathrm{neg}}}.$$

なお，$sc(v_1, v_2) \in [0, 1]$ である．

2 つのバリアント間の距離の計算

2 つのバリアント v_1 および v_2 の距離を，発現量ベクトルの角度 $\theta(v_1, v_2)$ と配列の重複度 $sc(v_1, v_2)$ から計算する．これは上記 2 つの距離を $[0, 1]$ に正規化した後，2 乗和の平均の平方根と定義する．

$$vd(v_1, v_2) = \sqrt{\frac{(\theta(v_1, v_2)/(\pi/2))^2 + sc(v_1, v_2)^2}{2}}.$$

なお vd についても $vd(v_1, v_2) \in [0, 1]$ が成り立つ．

各バリアントへのランク付け

同じ遺伝子由来の他のバリアントと大きく異なる発現パターンを持つバリアントを収集するため，それぞれのバリアントは同じ遺伝子由来の他のバリアントとの距離の平均に従ってランク付けされる．より正確な記述のため，ある遺伝子 g から産出されたバリアント v を考える．g から産出されたバリアントの数を t とし，v 以外のバリアントを $v'_1, \ldots, v'_{t-1}$ ($i = 1, 2$ に対して $v'_i \neq v$) と表記する．このとき，v の平均距離 $avd(v)$ は以下のように計算される．

$$avd(v) = \frac{1}{t-1} \sum_{j=1}^{t-1} vd(v, v'_j)$$

すべてのバリアントは $avd(v)$ の降順でランク付けされる．

3. ヒト胎児の RNA-seq データセットによる検証結果

3.1 データセットと前処理

本稿で用いるデータセットは，ヒトの胎児の脳と心臓の RNA-seq データである．このデータは生後 12 週の胎児から分離され，HiSeq 2000 でシークエンスしたものであり，データサイズは脳のデータが 2.3G，心臓のデータが 2.7G

Vol.2017-MPS-113 No.37
Vol.2017-BIO-50 No.37
2017/6/25

である．なお，これらの SRA ID は SRR1663122 および SRR1663124 である．(詳細は [35] を参照のこと)

Trinity を実行する前に，十分なクオリティがないリードは FASTX-Toolkit [36] を用いて除去した．その結果，ヒトの胎児の脳からのリード数は約 450 万，心臓からのリード数は約 530 万となった．これらのデータセットを混ぜて Trinity に与え，179,997 個のバリアントが得られた．

発現量の計算には Kallisto [31] を使用した．

3.2 性能評価の指標

θ_c および θ_M を用いた DASE の性能を比較するため，以下のような手続きを行った．最初に，Trinity のエントリのうち DASE でランクの高い順に n 個のエントリ ($n \in \{500, 600, 800, 1000, 1500, 2000, 2500\}$) を取得し，これらのエントリを脳と心臓のどちらで大きな発現量を持つかによって 2 つのグループに分割した．脳と心臓のそれぞれのグループに属しているバリアントに対する双方向 BLAST([37] 等を参照) によって，バリアントと対応するヒトの Ensembl の遺伝子エントリを取得した．ここで取得したエントリが各組織とどのくらい関連しているかを調べるため，Web サービスとして公開されている TopAnat [38] を使用した．TopAnat は Ensembl の遺伝子 ID を入力として，様々な組織を表すオントロジーである UBERON エントリ [39] を返す．TopAnat による出力のうち，E-value が `1e-10` 以下のエントリを採用することとした．

DASE の性能評価として，それぞれの組織に関連すると思われる UBERON エントリの数を比較した．ある組織に関連している UBERON エントリの数が多ければ，DASE がその組織と関連していると思われるバリアントに比較的高いランクを与えていることが示唆される．

3.3 検証結果

θ_M と θ_c を用いた DASE と各組織に対応する UBERON エントリ数の関係を図 3 に示す．図 3 より，θ_M を用いた DASE の方が全般的に性能が良いことがわかる．

図 3 のようになった要因の 1 つとして，θ_M を用いた DASE の場合，発現量による座標平面において共分散行列 Σ で規定される楕円の長軸方向に近い場所での差異をより大きく見積もることが挙げられる．一方，θ_c を用いた DASE の場合は，発現量による座標平面のどこであっても差異を同じように評価する．このため θ_M を用いた場合，発現量そのものの値がそれほど大きくなく，脳および心臓のどちらかにおいてのみ発現したバリアントのランクが下がる．

なお，θ_M と θ_c を用いた DASE によるランキングの違いを評価するため Spearman 相関係数を計算したところ 0.769 であった．この相関は強いものであるが，ランキング同士が類似しているとまでは言い切れない数値となっている．

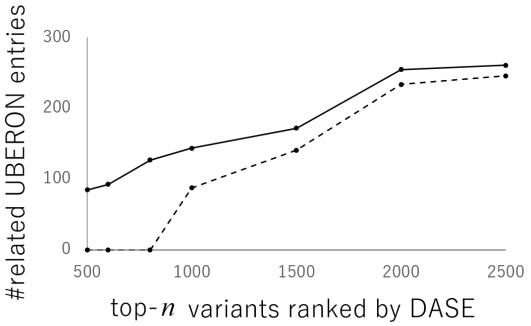


図 3 性能比較．実線と点線はそれぞれ，θ_M と θ_c を用いた DASE における性能を示している．

4. 結論と今後の課題

本稿では，マハラノビス距離にヒントを得た楕円上の角度と従来のコサイン角度を用いた DASE の性能評価を行い，楕円上の角度を用いた DASE の方が高い性能を有することが示された．

DASE の有用性を示すために，より多くの性能評価実験を行うことが今後の課題として挙げられる．そのためにはまず，RNA-seq 実験のデータセットのうち，実験で同一個体の別組織あるいは異条件下において，スプライシングバリアントの発現量が複数の遺伝子について調査されているデータセットを収集することが必要である．しかしながら現在までに調査した限りでは，論文において異条件下でスプライシングバリアントの発現量を確認した遺伝子の数は多くても 20 個に届かない．

次に DASE の改良のために，実験条件数が 3 を超える場合にもマハラノビス距離に基づいた角度を用いた DASE が実行できるように拡張する必要がある．ただこの場合，回転行列のみでは動作しないことが考えられるため，別のアイディアが必要となる．

また DASE の機能拡張に関する今後の予定として，以下の 2 点が挙げられる．1 点目は条件依存的なエクソンの網羅的発見手法の構築である．同一遺伝子から算出されたバリアントのマルチプルアラインメント実行結果を見ると，異なる条件下で発現量が大きく異なると思われるエキソンが存在することがわかる．さらに，ゴリラとオランウータンのような近縁種間の RNA-seq データを DASE に入力し，特定の生物種に特徴的な選択的スプライシングを発見する手法を構築することも予定している．これを実装するためには，異なる生物種の RNA-seq データを混合したことによって起こるアーティファクトの除去機能が必要となる．

謝辞 本研究は日本学術振興会 (No. 90360560; 小倉)，創薬プラットフォーム推進事業 (No. 16am0101042j0005; 米澤)，および科研費基盤 (C) (No. 17K00419; 米澤) によ

る支援を受けている．

参考文献

[1] Chow, L.T. et al. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, Vol. 12, No. 1, pp. 1-8 (1977).

[2] Berget S.M. et al. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA*, Vol. 74, No. 8, pp. 3171-3175 (1977).

[3] Modrek, B., and Lee, C.J. A genomic view of alternative splicing. *Nat. Genet.*, Vol. 30, No. 1, pp. 13-19 (2002).

[4] Pan, Q. et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, Vol. 40, No. 12, pp. 1413-1415 (2008).

[5] Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*, Vol. 456, No. 7221, pp. 470-476 (2008).

[6] Nilsen, T.W., and Graveley, B.R. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, Vol. 463, No. 7280, pp. 457-463 (2010).

[7] Barash, Y. et al. Deciphering the splicing code. *Nature*, Vol. 465, No. 56, pp. 53-59 (2010).

[8] Ellis, J.D. et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell.* Vol. 46, No. 6, pp. 884-892 (2012).

[9] Buljan, M. et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell*, Vol. 46, No. 6, pp. 871-883 (2012).

[10] Sultan, M. et al., A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, Vol. 321, No. 5891, pp. 956-960 (2008).

[11] Martin, J.A., and Wang, Z. Next-generation transcriptome assembly *Nat. Rev. Genet.*, Vol. 12, No. 10, pp. 671-682 (2011).

[12] Barbosa-Morais, N.L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, Vol. 338, No. 6114, pp. 1587-1593 (2012).

[13] Johnson, J.M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, Vol. 302, No. 5653, pp. 2141-2144 (2003).

[14] Modrek, B., and Lee, C.J. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, Vol. 34, No. 2, pp. 177-180 (2003).

[15] Gan, Q. et al. Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell Res.*, Vol. 20, No. 7, pp. 763-783 (2010).

[16] Fu, R.H. et al. Aberrant alternative splicing events in Parkinson's disease. *Cell Transplant.*, Vol. 22, No. 4, pp.653-661 (2013).

[17] Trapnell, C. et al. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, Vol. 25, No. 9, pp. 1105-1111 (2009).

[18] Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, Vol. 29, No. 1, pp. 15-21 (2013).

[19] Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, Vol. 7, No. 3, pp. 562-578 (2012).

[20] Grabherr, M.G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, Vol. 29, No. 7, pp. 644-652 (2011).

[21] Yang, W. et al. ABSSeq: a new RNA-Seq analysis method based on absolute expression differences and generalized Poisson model. *BMC Genomics*, Vol. 17, p. 541 (2016).

[22] Yonezawa, K. et al. DASE: Conditional-specific differential alternative splicing variants estimation method without reference genome sequence, and its application to non-model organisms. *2016 IEEE International Conference on Bioinformatic and Biomedicine (BIBM)*, pp. 324-329 (2016).

[23] Gloor. G. ALDEx2: ANOVA-Like Differential Expression tool for compositional data. ALDEx manual modular (2015).

[24] Griffith, M. et al. Alternative expression analysis by RNA sequencing. *Nat. Methods*, Vol. 7, No. 10, pp. 843-847 (2010).

[25] Rasche, A. et al. ARH-seq: identification of differential splicing in RNA-seq data. *Nucleic Acids Res.*, Vol. 42, No. 14, p. e110 (2014).

[26] Hu, Y. et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.*, Vol. 41, No. 2, p. e39 (2013).

[27] Katz, Y. et al. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, Vol. 7, No. 12, pp. 1009-1015 (2010).

[28] Ryan, M.C. et al. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, Vol. 28, No. 18, pp. 2385-2387 (2012).

[29] Sacomoto, G.A. et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. *BMC Bioinformatics*, Vol. 13, No. 6, pp. 1-12 (2012).

[30] Levin, L. et al. LEMONS-A tool for the identification of splice junctions in transcriptomes of organisms lacking reference genomes. *PLoS ONE*, Vol. 10, No. 11, p. e0143329 (2015).

[31] Bray, N.L. et al. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, Vol. 34, pp. 525-527 (2016).

[32] Mahalanobis, P.C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, Vol. 2, pp. 49-55 (1936).

[33] Karen, H. et al. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, Vol. 11, pp. 345-355 (2010).

[34] Katoh, K. et al. MAFFT: a novel method for rapid muitiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, Vol. 30, No. 14, pp. 3059-3066 (2002).

[35] Yan, L. et al. Epigenomic landscape of human fetal brain, heart, and liver. *J. Biol. Chem.*, Vol. 291, No. 9, pp. 4386-4398 (2016).

[36] Gordon, A. and Hannon, G.J. FASTX-Toolkit: FASTQ/A short-reads pre-processing tools. Available from `http://hannonlab.cshl.edu/fastx_toolkit/`.

[37] Jordan, I.K. et al. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.*, Vol. 12, pp. 962-962 (2002).

[38] Bastian, F.B. Bgee: database and R package for retrieving the preferred anotomical expression localization of a list of genes, or of a single gene, in animals. *Plant and Animal Genome XXV Conference* (2017).

[39] Mungall, C.J. et al. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, Vol. 13 (2012).