審 査 の 結 果 の 要 旨

氏　名　　桑　飞

本論文は「A global approach for identification and quantification of splicing variants using RNA-Seq data（RNA シーケンシングデータからのスプライシング変異産物の検出と定量に関する包括的アプローチ)」と題し、5 つの章から構成されている。

第 1 章「Introduction」では本論文の背景と目的及び構成について述べている。ゲノム DNA から RNA が転写されるプロセスにおいて、選択的スプライシングはトランスクリプトームやプロテオームの多様性の生成に主要な役割を担っており、構造、機能的に異なる mRNA とタンパク質変異は、発生、疾患に影響を与えている。ヒト遺伝子の転写産物の約 95%は複数のスプライスパターンを持ち、細胞、組織間で異なった発現を示す。しかしながら、哺乳類の選択的転写産物のレパートリーとその制御機構に関する知見は現在では乏しい。次世代シーケンシング技術は膨大な量のトランスクリプトーム情報をもたらし、その解析に道筋を示したと言える。

膨大な量の転写産物の配列決定とそれらのゲノムへのマッピングにより、RNA-Seq 技術は新規転写産物の発見とその発現量の推定を可能とした。これまでの RNA-seq データ解析手法は mRNA のアセンブリとアノテーション、新規エクソンや遺伝子の発見、発現レベルの推定などに焦点が当てられていた。現在RNA-Seq解析で頻繁に使用されるCufflinks,やScripture等の手法は*de novo*アノテーションが行えるが、転写産物の正確な定量は難しく大きな課題として未だに残っている。

第 2 章「Data resource」および第 3 章「methodology」では本研究で用いたデータおよび開発した解析手法について述べている。RNA-Seq リードおよびベースカウントによるスプライスバリアントの同定と定量が可能な網羅的アプローチを開発した。既知遺伝子アノテーション若しくはアセンブル後の転写産物構造の情報を用いた同定と相対的定量が可能な統計モデルである。本手法の検証は、転写産物の量、種類を事前に得ることができるように、FluxSimulator により生成した人工データを解析することにより行った。1 億 2000 万の 75 塩基のペアエンドリードから、18,000 のトランスクリプトバリアントを得た。アノテーションレファレンスの利用の有無で得られた 2 種類の結果を Cufflinks の結

果と比較した。

第 4 章「Result」では開発した手法を用いた解析結果について述べている。アノテーションを利用した場合には、新規手法ではリード数の異なる 3 つのデータセットで、8,090， 8,928， 9,452 の転写産物を同定した。 一方 Cufflinks はそれぞれ 12,448， 13,317，13,579 個を同定し、偽陽性を含むと考えられた。Cufflinks と比較して真の発現レベルと高い相関が得られた。発現比においては 88%以上が適正に推定されたのに対して、Cufflinks では 51%が良好に推定された。さらに発現比の推定だけでなく、発現レベルにおいても Cufflinks より正確であった。

de novo assembly 解析も新手法として適用したところ、57%正確であったのに対し Cufflinks は 9%であった。Poisson model がこの様な過少推定を行っていると考えられる。シーケンシング過程が単純なランダムサンプリングと仮定すれば、サンプル中の全塩基から一様且つ独立に全てのリードが得られるはずである。しかしながら、リードは真にランダムではなく一様でもない場合も多く、これは RNA-Seq 実験の技術的なバイアスによるものと考えられる。他の理由として、アイソフォームの発現量推定にリードカウントを使用するのは短いエクソンのカウントにおいて過大推定になるかもしれない。

第 5 章「Conclusion and Discussion」では以上の成果を要約し，議論している。本研究では既存の方法より正確に速く RNA アイソフォームの同定と定量を正確におこなうために幾つかの独創的な開発を行った。第一に、リードカウントの代わりにベースカウントを適用した。第二に、Poisson モデルに変わる新規の確率的アルゴリズムを開発した。第三に、更に正確に速く推定するためのアルゴリズムを追加した。Subexon Matrix、スプライシングリード情報を用いることなど。本研究で新規アプローチを開発したが、その推定には自身の制限による不正確さが含まれている。一つ目の不利点として、仮定が含まれる同定フェイズ（Identifying Phase）が厳密過ぎると発現が低いスプライシングバリアントを検出できない。しかし、これらの２つの仮定を除けばより多くの偽陽性を含むことになる。もう一つの不利点として遺伝子構造が複雑すぎる場合にはアイソフォーム検出に失敗することが多い。しかし、複雑な構造の場合、現行の RNA-seq データからはどんな手法を使っても難しい。

以上のことを勘案して審査員 5 人の合議による最終結論は、本論文の提出者が自立して研究活動を行い、高度な専門的業務に従事するために必要な能力およびその基礎となる豊かな学識を有していることを示すものであると判断した。

よって本論文は博士（工学）の学位請求論文として合格と認められる。