真核生物および原核生物の RNA-Seq データから組み立てられた転写産物のほとんどは、タンパク質をコードしていると予想されます。コード領域の可能性が高い転写産物を同定する最も実用的な方法は、BLASTXなどによる、十分にアノテーションが付けられた近縁種の配列に対する配列相同性検索です。コード領域を予測することは、転写産物が細胞内で果たす分子的役割を決定する上で極めて重要です。しかし、新しく配列決定されたトランスクリプトームでは、そのような十分にアノテーション付けされた近縁種が利用できないことが多いです。

非モデル生物を扱う場合、参照ゲノムとトランスクリプトームが利用できないケースがあり、その場合トランスクリプトームのde-novo Assembly戦略が必要です。これらの新規トランスクリプトームは一般的に、既知のタンパク質との相同性が十分に検出できないタンパク質をコードします。このようなコーディング領域を捉えるためには、TransDecoderのような、配列組成に結びついたメトリクスに基づいてコーディング領域を予測する手法が必要です。**コーディング領域を予測する**このツールは、OmicsBoxのTranscriptomics Moduleで利用できます。

コーディング領域予測

・転写産物配列内のコーディング領域を検出

・信頼の高いORFが抽出され、機能を予測するために使用されます。

・TransDecoderを統合

TransDecoderはORFの予測を行うことで、より確実にタンパク質をコードする領域を特定します。

biosupport@filgen.jp https://filgen.jp/ 052-624-4388 052-624-4389 フィルジェン株式会社

TransDecoder によるコーディング領域の予測

TransDecoder は、Trinityに組み込まれて開発されたユーティリティです (OmicsBox 内でも使用できます)。TransDecoderは、再構成された転写産物内の潜在的なコード領域の同定を支援することを目的としています。TransDecoderは、次の基準に基づいて、コーディング領域である可能性が高い配列を識別します。

- 転写配列内の最小長のオープンリーディングフレーム (ORF)
- 対数尤度スコアは0より大きい。このスコアはGeneIDスコアと似ている。
- 上記のコーディングスコアは、ORF が最初のリードフレームにあるときに、他の2つのフォワードリードフレームのスコアと比較して最大となる。
- 候補ORFが他の候補ORFの座標に完全に包含されている場合、長い方を報告する。ただし、1つの転写産物が複数のORFを報告することもある（オペロン、キメラなどを考慮）。
- Position-Specific Scoring Matrix（PSSM）が計算、トレーニングされ、開始コドン予測を改良するために使用される。

別途コーディング尤度スコアに関係なく、機能的重要性を持つ可能性のある ORF を捕捉する感度をさらに最大化するオプションが搭載されています。ORFを既知のタンパク質との相同性についてスキャンし、共通タンパク質ドメインを同定するためにPFAMを用いてそのようなORFをすべて保持することができます。

OmicsBoxのTransDecoder

OmicsBoxを使用したTransDecoderは、Predict Coding Regionsユーティリティを介して利用できます。Trinityでショートリードをアゼンブリしたコンティグデータを使用して、 Predict Coding Regionsを簡単に適用できます。まず、ユーザーはデータと研究対象の種に応じてツールを調整できます。

ORFを見つけるための適切な遺伝暗号を選択したり、タンパク質の最小長カットオフを設定したり、ORFがどのように保持されるかを設定したりすることができます。Pfam Searchはオプションですが、強く推奨されるパラメータです。 チェックすると、ORFがPFAMに対してスキャンされ、ORF保持基準として使用されるタンパク質ドメインが特定されます。

解析後、同定されたORFは3種類の拡張子として出力されます。：

- CDS： 最終候補ORFのコード領域の塩基配列（FASTA形式）。
- Proteins： 最終候補ORFのペプチド配列（FASTA形式）。
- Coordinates: 最終的に選択された ORFのターゲット転写物内の位置情報 。(GFF形式)

トランスクリプトームのコンテキストで候補 ORF を表示するには、OmicsBox Genome Browser をお勧めします (Fig. 1)。

Figure 1. OmicsBox Genome Browser で興味深い配列を研究する

ORF Types
TransDecoderは、予測されたORFについて見つかった、長さや鎖などの詳細が得られます。さらに、開始信号と停止信号に従って予測されたORFを分類します (Fig. 2)。：

- Complete ORF:開始コドンと停止コドンが含まれている。
- 5' partial ORF:開始コドンとおそらくN末端の一部を欠いている。
- 3' partial ORF停止コドンとおそらくC末端の一部を欠いている。
- Internal ORF:5‘ と 3’ の両方で部分的である。

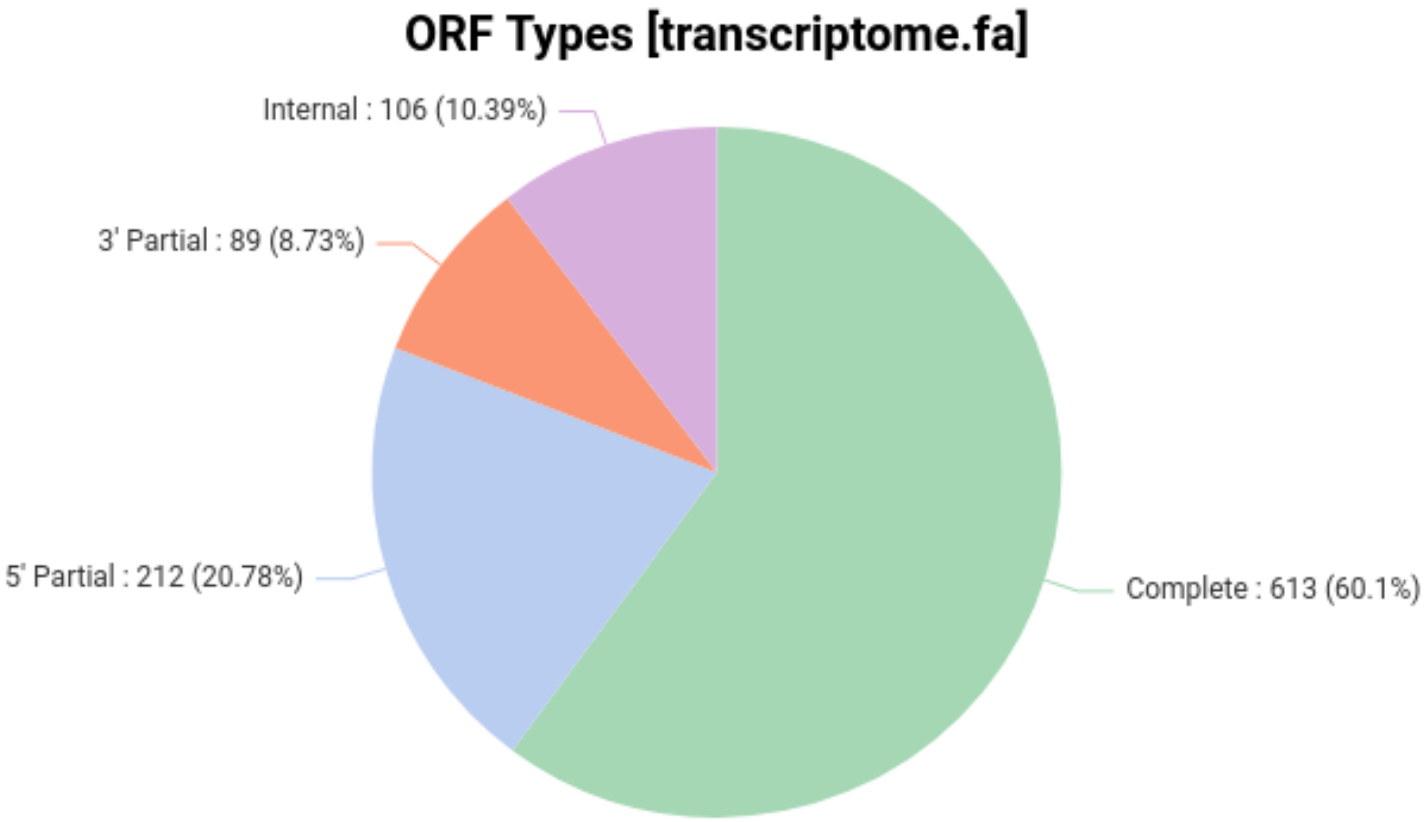

Figure 2. ORF分類の円グラフ

OmicsBox はタンパク質の機能を予測するために最も信頼性の高いORFを抽出します。実際には、この戦略を適用した後、 OmicsBox では、このステップに広く知られている Blas2GO methodologyを使用する相同性ベースの機能アノテーションパイプラインに直接リンクできます (Fig. 3)。

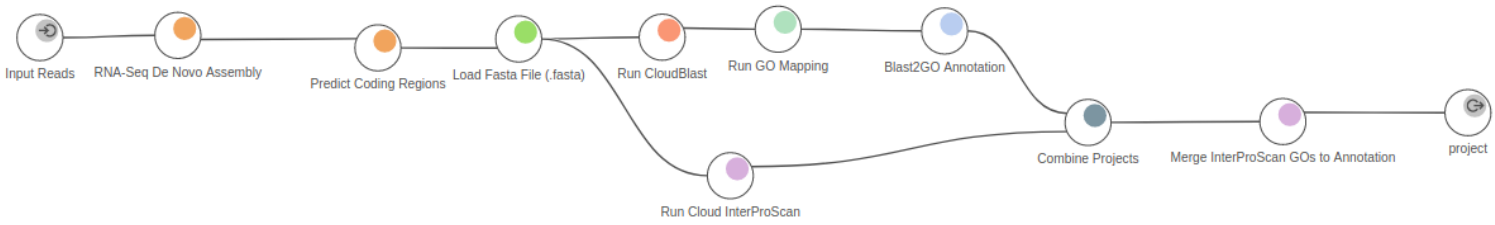

Figure 3.解析ワークフローの例

References
•Predict Coding Regions User Manual.
•Haas BJ et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), 1494-512.
•TransDecoder 5.5.0. Haas, BJ. and Papanicolaou, A. 2019. https://github.com/TransDecoder/TransDecoder/wiki.

OmicsBox

実績は高いがコマンドライン型であったりOSに制限がある
オープンソフトウェアを多数搭載

それらの解析をマウス操作で簡単に解析できる

- 本稿で使用したTransDecoderツールの他、このツールの前工程であるツール（ Trinityによるde novo assembly、BUSCOによるコンティグデータ品質評価、CD-HITによるクラスタリング）やその後の工程のためのツール（BlastやBlast2GOアノテーション、コンティグデータを使用したRNA-Seq定量解析）を搭載
- OmicsBoxの紹介ページはこちら