

**【目的】** 従来、真核生物において、mRNA 中の一番長い open reading frame (ORF) 配列である coding sequence (CDS) のみがタンパク質として翻訳されると考えられてきた。しかし、リボソームプロファイリング法により翻訳されている RNA の領域を調べたところ、CDS 以外の非翻訳と考えられてきた領域も翻訳されていることが示唆されている。質量分析により細胞で発現するタンパク質の解析を行うと、CDS 由来の既知タンパク質と一致するスペクトルは約 40%である。これまで、残りの 60%はデータとして高品質で信頼性が高いものであっても、CDS がコードする既知タンパク質と一致しないという理由で無視されていた。本研究では、これら既知のタンパク質をコードしない小さな ORF を一括りにして unannotated ORF (uaORF) と定義し、研究対象とする。uaORF がコードするタンパク質はほとんどが未同定・未解析であり、細胞や生物を説明する上で「足りないもの」の極めて有力な候補である。現在までに同定されている既知タンパク質に uaORF がコードする新規タンパク質を加えれば、現行のプロテオームデータが大幅に拡張でき、細胞機能の実行役を担うプレイヤーが出そろおうだろう。そこで本研究では、ヒト多能性幹細胞における uaORF の同定とマッピングを目的とする。

**【方法】** 二つの独立したヒト多能性幹細胞を用いて、直鎖 RNA (mRNA および長鎖非翻訳 RNA) のマッピング、環状 RNA のマッピングおよびリボソームプロファイリングによる翻訳領域の同定を行った。

**【結果】** 直鎖 RNA (mRNA および長鎖非翻訳 RNA) は通常の RNA-seq を実施することでデータを得た。次に環状 RNA を濃縮する目的で、直鎖 RNA を特異的に消化する RNase R でトータル RNA を処理し、環状 RNA を RNA-seq により効率的に検出した。結果として、二つの独立したヒト多能性幹細胞株でそれぞれ 10,000 種類の環状 RNA が検出され、共通して発現していた環状 RNA は 6,788 個であった。さらに、リボソームプロファイリングを実施し、データベース上に存在しない新規翻訳領域を塩基レベルで見出した。uaORF マッピングデータとリボソームプロファイリングデータと組み合わせることにより、uaORF のうち、実際に翻訳されている領域を一塩基レベルで正確に決定した。その結果、検出した全 ORF のうち 62%が既知の CDS であった。一方で、新規の uaORF は 4,500 個以上同定され、その割合は全 ORF の 19%であった。また、報告されている翻訳開始点とは異なる位置から翻訳が開始されている CDS のバリエーションは全 ORF の 19%であった。

これまで見過ごされてきたアノテーションされていないタンパク質

