論文内容の要旨

論文題目　完全長 cDNA 情報を用いたヒトの選択的スプライシング解析

（Large-scale identification and characterization of human alternative splicing variants based on full-length cDNAs information）

氏名　武田　淳一

序論

選択的スプライシング（alternative splicing; AS）は、転写された mRNA 前駆体（pre-mRNA）上から様々なパターンで pre-mRNA のイントロンがスプライス除去され、複数の異なるエクソンによって構成される成熟 mRNA を生成する現象である。AS は、ヒトを含む高等真核生物において、複雑な細胞内遺伝子システムを構成するためのタンパク機能の多様化などに貢献していると考えられているが、そのメカニズムや全体像は不明である。これまでヒトの AS 解析は、EST 情報を用いた転写物の一部あるいは遺伝子予測の情報を含む転写産物モデルを用いて行われてきた。これらは 5'末端の情報が不十分のため、検出される AS イベントの位置にバイアスがかっていたり、アミノ酸配列の情報が欠けることによって正しいタンパク機能アノテーションを行うことができない。本研究で、筆者はヒトの転写産物情報として 5'末端が転写開始点である完全長 cDNA を用い、さらにそれらを計算機プログラムおよびマニュアルでアノテーションすることによって、EST や遺伝子予測のデータでは不可能であった精度の高い AS バリアントのゲノムワイドな同定と、そのタンパク機能アノテーションの解析を行った。また、共通祖先遺伝子であるオルソログについて、ヒトとマウスで AS バリアント単位での比較を行い、その保存度についての解析を行った。その結果、タンパク機能アノテーションに影響を与える AS バリアントは多数認められるものの（1 章）、ヒトとマウスで進化的に保存された AS バリアントは相対的に非常に少ないこと、そしてヒト特異的な AS バリアントを持つ遺伝子には、生殖に関係し精巣で発現するタンパク機能が濃縮して観察されたこと（2 章）を明らかにした。また、これらの解析結果を一般に公開すべく、データベース（H-DBAS; http://www.h-invitational.jp/h-dbas/）を構築した（3 章）。さらに、次世代シークエンサーによるヒト細胞内のポリソーム画分、あるいは核画分から精製した RNA を用いた RNA-Seq タグの解析から、翻訳に用いられていると思われる転写産物と、核内に留まり翻訳には用いられていないと思われる AS バリアントを区分して解析した（4 章）。

47-087338：武田淳一

材料と方法

ヒトの完全長 cDNA は、主にヒト完全長 cDNA アノテーション会議（H-Invitational 2）でアノテーションされた 56,419 本の配列を用いた。同一遺伝子内の cDNA 配列について、ゲノム上のエクソンーイントロン境界を総当たりで比較し、AS バリアントを同定した。同じ構造を持つ AS バリアントのグループから cDNA を 1 つ選び､それを代表 AS バリアント（representative AS variant; RASV）と定義した。この RASV を用い、CDS やタンパク機能アノテーションの違いを比較した。また、従来の典型的な概念と合致しない複雑な AS パターン、すなわち、タンデムにマッピングされた 2 つの遺伝子を橋渡しする RASV が存在するが､リードスルーとは異なり CDS を同じフレームで共有しているもの（ブリッジ型)、同一遺伝子内で入れ子状にマッピングされた 2 つの RASV のうち、転写領域の一部は共有しているが CDS は全く共有していないもの（ネスト型)、200 アミノ酸以上の CDS を持つ 2 つの RASV のうち、CDS は共有しているもののフレームがずれてアミノ酸配列が異なるもの（マルチプル CDS 型）についても区分し、解析に用いた。

ヒト完全長 cDNA から同定した RASV の進化的保存度を調べる比較ゲノム解析は､FANTOM3 および Mammalian Gene Collection（MGC）から収集したマウス完全長 cDNA を対象として行った。ヒトとマウスのゲノムアラインメントは、BLASTZ でアセンブルしたものを用いた。種間保存度の判定はエクソン単位で行い（閾値は coverage=70%かつ identity=60%)、エクソンアラインメントの結果をもとに RASV 単位での保存度を判定した。ヒトの RASV と対応するマウスの cDNA が全て保存されたエクソンで構成されていた場合、それらを同一スプライシングバリアント（equally-splicing variant; ESV）と定義した。遺伝子に 2 つ以上の ESV が存在していた場合、すなわち、同一遺伝子内で AS イベントを含む 2 つ以上の RASV が保存されていた場合、それらを保存 AS 遺伝子と定義し、タンパク機能アノテーションの統計解析に用いた。

次世代シークエンサー（Illumina GA）による RNA-Seq 解析には、ヒト DLD-1 細胞の細胞質・核・ポリソームの各画分から精製された RNA を用いた。その RNA-Seq 解析の結果から、計 148,376,598 本の 36bp シングルエンドタグ（RNA-Seq タグ）配列を得て、これらをショートリードのアセンブルプログラムである Bowtie を使ってヒトゲノム上にマッピングし、スプライスジャンクション検出プログラムである TopHat を使ってスプライスジャンクションを検出した。これらのスプライスジャンクションから、既知転写物（RefSeq）の AS ジャンクションと同じゲノム上の位置にあるものを同定し、さらに核に特異的なものを選択した。ポリソーム由来の RNA-Seq タグについては、RefSeq にマッピングした。

結果と考察

1 章：ヒト完全長 cDNA を用いた選択的スプライシングのゲノムワイドな同定と、タンパク機能アノテーションに影響を与える選択的スプライシング

全ヒトゲノム中で、18,297 の RASV（6,877 AS 遺伝子）を同定した。AS には典型的な 5 つのパターン（カセット型エクソン、選択的 3'スプライス、選択的 5'スプライス、相互排他的エクソン、選択的保持イントロン）があり、それぞれ、3,020、1,758、1,686、210、1,970 の遺伝子において見出した。CDS に影響を及ぼす AS 遺伝子は 6,005（87%）であり、さらに 4 つのタンパク機能アノテーション（タンパク機能モチーフ・GO・細胞内局在化シグナル・膜タンパクドメ

イン）に影響を与える AS 遺伝子は 4,481（65%）であった。加えて、AS 遺伝子は非 AS 遺伝子に対してタンパク機能モチーフを含む頻度が有意に高いことを確認した。これらの結果は、AS がタンパク機能の多様化に大きく寄与していることを示唆する。タンパク機能モチーフを用いた解析では、I κ B kinase- ε (IKBKE)内に protein kinase を含まない新規の RASV を同定した。この RASV は、protein kinase を含む RASV の機能の一部を失っていると考えられるため、細胞内における NF- κ B のシグナル伝達に際し、モジュレーターの役目を果たしていると考えられる。なお、AS 遺伝子に濃縮して現れる GO とタンパク機能モチーフは、ほとんどがシグナル伝達と転写制御であった。また、RASV 間でほとんど CDS が異なる 3 種類の複雑な AS パターン（ブリッジ型・ネスト型・マルチプル CDS 型）を有する遺伝子は、316（5%）存在した。これらは、1 つの遺伝子から明らかに異なるタンパクを複数生成するため、細胞内遺伝子システムの更なる複雑化に寄与する興味深い例だと考えている。

2 章：ヒトとマウスの完全長 cDNA を用いた、ゲノムアラインメントを介した種間保存およびヒト特異的選択的スプライシングのタンパク機能アノテーション解析

マウスの cDNA に対するヒト RASV のエクソン自体の保存度は高い（74%）が、ESV では 23%、保存 AS 遺伝子になると 3%しか存在しなかった。GO とタンパク機能モチーフ解析の結果、保存 AS 遺伝子に濃縮して現れる RASV の機能は、peroxidase activity・bZIP transcription factor・TSC-22 など、主に細胞の恒常性を維持するためのものであることが明らかとなった。これは、種に関係なく細胞の基本的な機能のため、進化的に不変な AS 配列を必要とするからだと考えられる。保存 AS 遺伝子の例として、phosphoinositide-3-kinase (PI3 kinase) regulatory subunit を示す（Figure 1）。この 2 つの RASV は、insulin receptor substrate (IRS) protein から PI3-kinase の p110-kDa catalytic subunit へ、異なる効率でシグナルを伝達していることが知られている。一方、ヒト特異的な AS 遺伝子に有意に現れるタンパク機能モチーフは、GAGE や T-complex 11 など、精巣で発現するものが多かった。精巣は、脳と同様従来から AS が多数見出される組織として知られ、種特異的機能獲得あるいは種分化を解析する上で重要な組織だと考えられている。ヒトとマウスの ESV や保存 AS 遺伝子の割合が少ないことと合わせると、多くの AS バリアントは種ごとに独自に進化してきたことを示唆する。

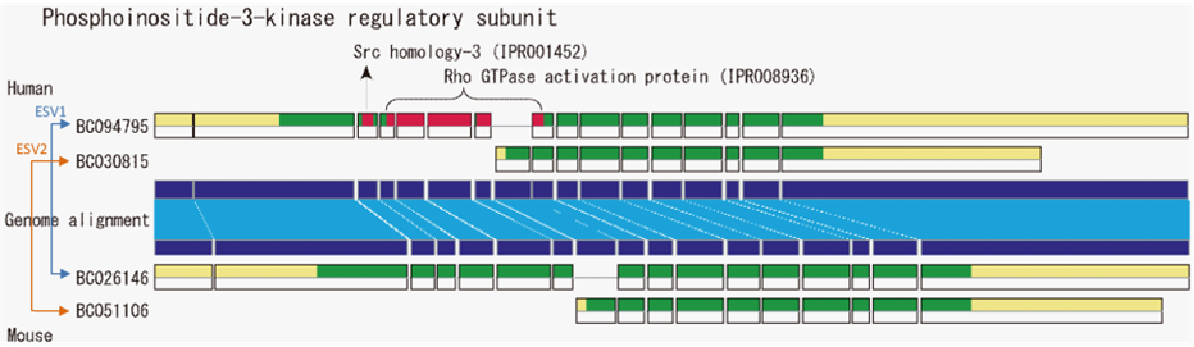

Figure 1　保存 AS 遺伝子の例（PI3-kinase regulatory subunit）。ヒト BC094795 とマウス BC026146、ヒト BC030815 とマウス BC051106 がそれぞれ ESV である。緑が CDS、黄が UTR、赤がタンパク機能モチーフを表す。この図では、構成的スプライシングイントロンを短くして表示している。

3 章：ヒト選択的スプライシングの解析データを盛り込んだデータベースの開発と公開

完全長 cDNA に依拠し、AS を転写産物バリアント単位で解析することが可能で、さらに種間比較も行えるデータベースは世界的に見てもユニークである。これらの情報を一般に向けて発信するため、独自のヒト AS のデータベース、H-DBAS（Human-transcriptome DataBase for Alternative Splicing）を開発し、公開した（URL: http://www.h-invitational.jp/h-dbas/）。

H-DBAS は 2006 年にバージョン 1 が公開され、2010 年の 9 月にはバージョン 6 まで更新されている。H-DBAS の特徴は、ユーザーがビューワー上で RASV をインタラクティブに操作できることである。特に、前述した解析の結果であるタンパク機能アノテーションに影響を与えるヒトの RASV や、ヒトとマウスの ESV・保存 AS を直感的に操作して観察することができる。また、6 カテゴリー計 22 の検索項目から目的の RASV を探すための詳細検索や、配列相同性の高い RASV を探すための BLAST 検索など、有用な検索機能も実装している。

4 章：RNA-Seq タグを用いた、選択的スプライシングの翻訳検証

ヒト細胞の細胞質・核・ポリソームの各画分の RNA-Seq 解析により検出したスプライスジャンクションのうち、既知遺伝子として採用したモデル転写産物である RefSeq のスプライスジャンクションとゲノム上の位置が一致したものは、それぞれ 47,615、47,260、51,041 であった。同一遺伝子内でこれらが AS を構成するものは、1,067、1,021、1,114 であった。このうち、254 の AS ジャンクションを核で特異的に検出した。これらの AS バリアントは核内に留まり、タンパクに翻訳されないと考えられる。この例として、caspase 4, apoptosis-related cysteine peptidase (CASP4)を示す（Figure 2）。一方、ポリソーム由来の RNA-Seq タグがマッピングされた 8,440 の RefSeq AS バリアントについては、タンパクへ翻訳されると考えられる。本解析では 1 つの細胞のみを用いたが、将来このような翻訳情報を様々な細胞から得ることにより、ヒト特異的に多数生じた AS バリアントは生物学的に意義があるのか、あるいはゲノムに内在する転写のノイズとして生じたのか、についてのアノテーションを付加することができると考えている。

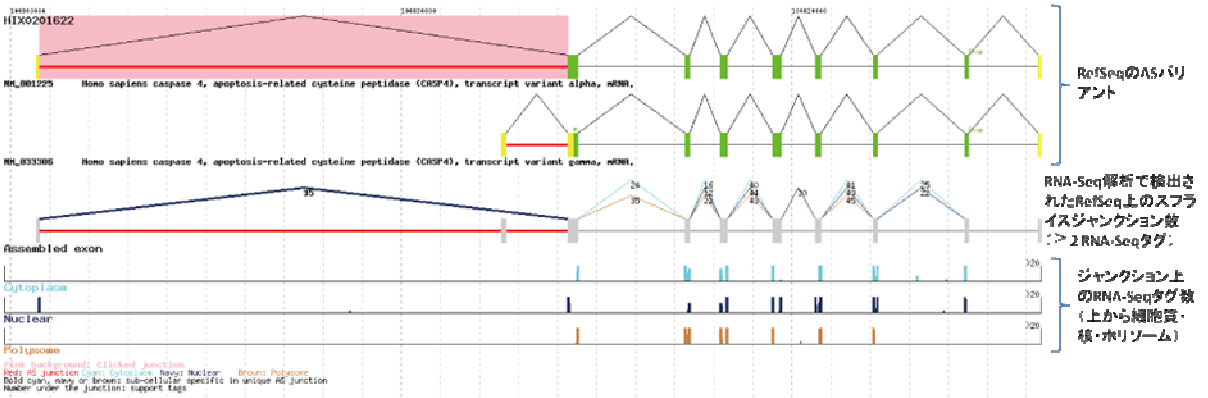

Figure 2　タンパクに翻訳されないと考えられる AS バリアントの例（CASP4）。上段の桃色の背景は、RNA-Seq 解析によって核特異的に検出され、かつ RefSeq の AS ジャンクション（赤色の太線）とゲノム上の位置が一致したスプライスジャンクションである。中段の水色の線は細胞質、紺色の線は核、茶色の線はポリソーム由来のスプライスジャンクションを表し、その下の数字は RNA-Seq タグの数を示す。太線は細胞画分特異的であることを示す。下段のスプライスジャンクション上の RNA-Seq タグの数を示すバーの色は、中段と同じ。

47-087338：武田淳一