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）

東京大学　大学院新領域創成科学研究科

メディカルゲノム専攻

ゲノム制御医科学分野

武田　淳一

目次

第一章　ヒト完全長 cDNA 配列を用いた選択的スプライシングバリアントの同定と、タンパク機能への影響 5

1.1 緒言 6
1.1.1 選択的スプライシング（AS） 6
1.2 方法 8
1.2.1 データセットの成形 8
1.2.2 選択的スプライシング（AS）と典型的な AS パターンの判定 10
1.2.3 代表選択的スプライシングバリアント（RASV）の同定 12
1.2.4 RASV のタンパク機能アノテーション 12
1.2.5 複雑な AS パターンの判定 13
1.3 結果 15
1.3.1 ゲノムワイドに同定されたヒト AS の統計 15
1.3.2 RASV の典型的な AS パターン 17
1.3.3 RASV のタンパク機能アノテーション解析 18
1.3.4 複雑な AS パターンを有する RASV の解析 22
1.4 考察 25
1.4.1 タンパク機能に影響を与えるヒト選択的スプライシングバリアント 25

第二章　ヒトとマウスのゲノムアラインメントを用いた、選択的スプライシングの比較ゲノム解析 26

2.1 緒言 27
2.1.1 種間比較と AS 27
2.1.2 ヒト特異的なスプライシング制御 27
2.2 方法 29
2.2.1 ヒトとマウスのゲノムアラインメント 29
2.2.2 ヒト RASV の保存度判定とタンパク機能アノテーション 29
2.2.3 選択的保持イントロンの実験的検証 31
2.2.4 ヒト脳特異的 Fox 制御スプライシングの探索 31
2.3 結果 33
2.3.1 ヒトとマウスの AS の保存度 33
2.3.2 ヒトとマウスの保存 AS 遺伝子のタンパク機能アノテーション解析 33

2.3.3 選択的保持イントロンの実験的検証 36
2.3.4 非保存 AS 遺伝子のタンパク機能アノテーション解析 39
2.3.5 ヒト脳特異的 Fox 制御スプライシング候補 41
2.4 考察 43
2.4.1 ヒトとマウスの選択的スプライシングバリアントの進化的保存度とタンパク機能 43

第三章 ヒト選択的スプライシングの解析データを公開するための、データベース（H-DBAS）の開発 44

3.1 緒言 45
3.1.1 ヒト AS のデータベース 45
3.2 方法 46
3.2.1 データベースシステム 46
3.2.2 サーバーアプリケーション 46
3.2.3 Flash アプリケーション 47
3.2.4 ヒトとマウスの間で対応するエクソンの表示 47
3.3 結果 49
3.3.1 H-DBAS の構築 49
3.3.2 検索システム（簡易・詳細・BLAST） 50
3.3.3 AS Viewer 53
3.4 考察 56
3.4.1 H-DBAS の独自性 56

第四章 RNA-Seq タグを用いた、ヒト選択的スプライシングの翻訳検証 57

4.1 緒言 58
4.1.1 次世代シーケンサー（イルミナ GA）による RNA-Seq 解析 58
4.2 方法 59
4.2.1 ヒト DLD-1 細胞の細胞画分の分離と実験的検証 59
4.2.2 RNA-Seq タグの生成 60
4.2.3 RNA-Seq タグのゲノムマッピングとスプライスジャンクションの検出 60
4.2.4 翻訳する、または翻訳しない AS バリアントの同定 60
4.3 結果 61
4.3.1 細胞画分ごとの RNA-Seq タグの統計 61
4.3.2 AS バリアントへの翻訳情報のアノテーション 61
4.3.3 RNA-Seq 解析のデータベースとビューワー 63

4.4 考察 66
4.4.1 RNA-Seq 解析による AS ジャンクションの翻訳検証 66

第五章 総括 67

参考文献 70

謝辞 78

第一章　ヒト完全長 cDNA 配列を用いた選択的スプライシングバリアントの同定と、タンパク機能への影響

1.1 緒言

1.1.1 選択的スプライシング（AS）

選択的スプライシング（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(1)。AS は 1977 年にアデノウィルスで発見され、真核生物で普遍的に存在することが知られるようになった。ヒトゲノムの初のドラフト配列が公表された際、ヒト遺伝子の数が約 2 万数千個同定され、当初の予想よりかなり少ないことが明らかにされた(2,3)。これは、AS がヒト遺伝子の複雑性を産み出す最も重要なプロセスの 1 つであり、ヒトを含む高等生物にとって、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を導くのに必要なメカニズムであるためだと考えられる(4)。特に、脳における AS は神経発達などに必須であり、適切な神経回路の形成に極めて重要なダウン症候群細胞接着分子（DSCAM）は、ショウジョウバエでは AS によって生成される数千もの神経特異的 AS バリアントとしてコードされる(5)。しかし、様々な生物学的現象において見出される AS のもたらす遺伝子機能の多様性について、そのメカニズムや全体像は依然として不明な点が多い(6)。

これまで、ヒトの AS 解析は、発現配列タグ（EST）情報を用いた転写物の一部、遺伝子予測の情報を含む転写産物モデル、あるいはエクソン－エクソンジャンクションのプローブを用いることによって行われてきた(7-9)。しかし、これらは転写物の 5'末端（転写開始点）の情報が不十分なため、エクソンの数や順番に正確性を欠くという欠点がある。この欠点を克服したのが、5'末端が転写開始点である完全長 cDNA である。オリゴキャップ法(10)やキャップトラッパー法(11)等、いくつかの完全長 cDNA ライブラリーを作成する方法が開発されている。完全長 cDNA ライブラリーより単離された cDNA 配列は、高い確率で完全な個々の転写物を反映し、転写物全体（トランスクリプトーム）の特徴を調べるために必須である。2002 年に、完全長 cDNA 配列のアノテーションを行い、ヒトのトランスクリプトームを解析することを目的とした国際会議、H-Invitational(12)が開催された。この会議では、世界の 44 研究機関から 120 人以上の研究者が集まり、5 つのプロジェクト(13-17)で配列決定された計 41,118 のヒト完全長 cDNA 配列のマニュアルアノテーションが行われた(18)。2003 年には、同じプロジェクトから得られたヒト完全長 cDNA の配列数を 56,419 に拡大した H-Invitational 2 が開催された。筆者はその両方の会議に参加して、計算機アノテーションおよびマニュアルアノテーションを行った。本研究では、ヒトのトランスクリプトームにおける AS に焦点を当て、その解析には主に H-Invitational 2 でアノテーションされたヒト完全長 cDNA 配列を用いた。そして、転写物上のエクソンの位置が明確な AS バリアントを初めてゲノムワイドに同定し、その特徴を明らかにした(19)。さらに、EST、転写産物モデル、およびマイクロアレイのデータでは不可能であった、精度の高い AS バリアントのタン

パク機能アノテーション解析を行った。

1.2 方法

1.2.1 データセットの成形

解析に用いた 56,419 のヒト完全長 cDNA の内訳を表 1-1 に示す。これらは、全て 99.9%以上の sequence reliability（Phred(20)スコアが 30 以上）があり、マニュアルアノテーションによってベクターやポリ A 尾部配列が正確に切り取られたものである。ヒト完全長 cDNA 配列は、est2genome(21)によってヒトゲノム（UCSC hg16）(22)へマッピングし、同一ストランド上で 1 bp 以上共有するエクソンが 1 つ以上ある場合にクラスタリングした。タンパクの機能アノテーションは UniProt(23)と InterPro(24)を使用して行い（"1.2.4 RASV のタンパク機能アノテーション"で記述）、最終的に H-InvDB(25)へ登録した。これらのデータセットに対し、計算機による自動処理アノテーションの結果を登録した独自の AS 解析用簡易データベースと、H-Invitational で開発したアノテーションビューワーを使い、マニュアルでアノテーションを行った（図 1-1）。なお、計算機によるアノテーションは Perl バージョン 5.8.8 で作成した独自のプログラムを使用し、下記の処理を行った。まず、ゲノム上にマッピングされた完全長 cDNA 配列から、複数のエクソンで構成された転写物を選択した。このうち、5'または 3'末端がそれぞれ他の転写物の最初または最後以外のエクソン内に位置する転写物は、完全な転写物ではない（分解された mRNA に由来する cDNA である）可能性が考えられたため、これらをデータセットから取り除いた（図 1-2）(26)。5'末端が他の転写物の最初のエクソン内に位置するものを転写開始点のバリエーション、3'末端が他の転写物の最後のエクソン内に位置するものを選択的ポリアデニレーションとみなし、不完全な転写物とは判定しなかった。ゲノム再構成遺伝子群である免疫グロブリン（Ig）や T 細胞受容体（TCR）、および個体によって非常に多様性の高い遺伝子群である主要組織適合性複合体（MHC）などの免疫関連遺伝子もデータセットから取り除いた。

表 1-1 ヒト完全長 cDNA を配列決定したプロジェクト

プロジェクト/研究機関	cDNA	参照文献	URL
HUGE/KDRI	2,031	Kikuno *et al.* (13)	http://www.kazusa.or.jp/huge/
FLJ/KDRI	397	Ota *et al.* (14)	http://flj.lifesciencedb.jp/
FLJ/IMSUT	6,374	Ota *et al.* (14)	http://flj.lifesciencedb.jp/
FLJ/HRI	22,047	Ota *et al.* (14)	http://flj.lifesciencedb.jp/
MIPS/DKFZ	9,212	Wiemann *et al.* (15)	http://www.helmholtzmuenchen.de/en/mips/
MGC/NIH	15,600	Strausberg *et al.* (16)	http://mgc.nci.nih.gov/
CHGC	758	Hu *et al.* (17)	http://www.chgc.sh.cn/en/

計	56,419

HUGE: Human Unidentified Gene-Encoded Large Proteins, KDRI: Kazusa DNA Research Institute, FLJ: Full-length cDNA Japan, IMSUT: Institute of Medical Science at the University of Tokyo, HRI: Helix Research Institute, MIPS: Munich Information Center for Protein Sequences, DKFZ: German Cancer Research Center, MGC: Mammalian Gene Collection, NIH: The United States National Institutes of Health, CHGC: Chinese National Human Genome Center

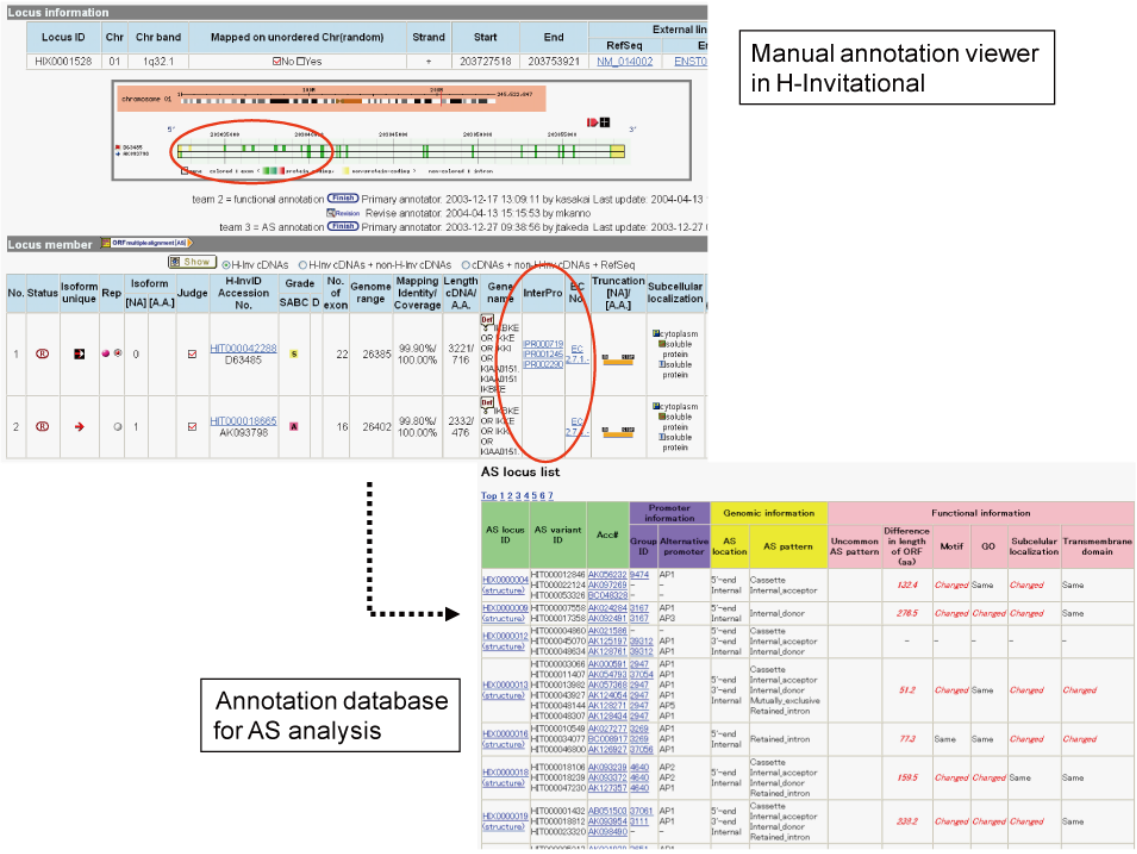


図 1-1　完全長 cDNA のマニュアルアノテーション。H-Invitational で用いられたアノテーションビューワー（上パネル）と、計算機による自動アノテーションの結果を登録した独自の AS 解析用簡易データベース（下パネル）を使ってマニュアルアノテーションを行った。赤色の丸で囲まれた部分は、マッピングされた転写物の AS イベントの場所とタンパク機能アノテーションの違いを示す。

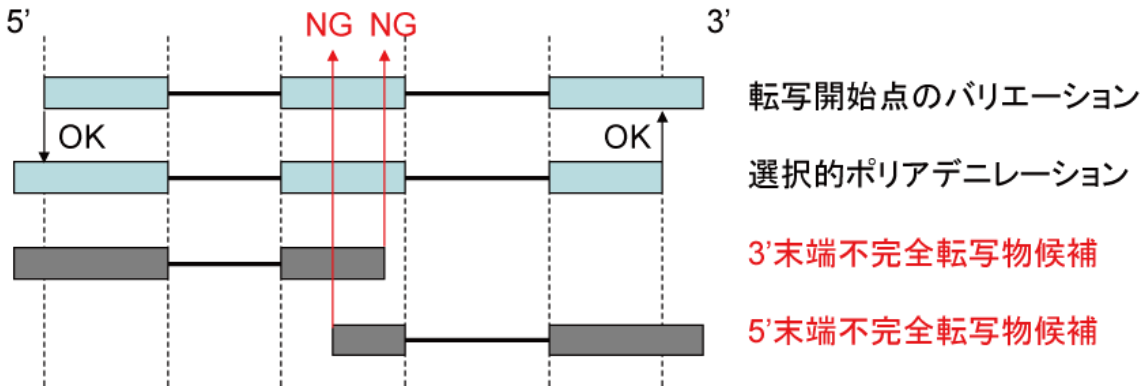


図 1-2　不完全転写物の判定。下の 2 つの転写物は、それぞれ 3’末端、または 5’末端が他の転写物の内部のエクソンに含まれており、不完全転写物と判定してデータセットから取り除いた。上の 2 つの転写物については、それぞれ転写開始点のバリエーション、または選択的ポリアデニレーションと判定し、完全長の転写物とみなした。四角はエクソン、太線はイントロンを示す。

1.2.2　選択的スプライシング（AS）と典型的な AS パターンの判定

”1.2.1　データセットの成形”で選別した転写物に対し、以下の方法で選択的スプライシング（AS）の判定を行った。判定には Perl バージョン 5.8.8 で作成した独自のプログラムを使用した。

（1）同一遺伝子に含まれる転写物に対し、エクソン―イントロン境界のゲノム位置を総当りで比較した。比較はマッピングによる不確実性を考慮して±10 bp の揺らぎを許し、10 bp までの誤差であれば同一境界とみなした。

（2）転写物のエクソンがゲノム上で他の転写物のイントロンに含まれていれば、エクソンの転写物上の位置により、5’末端 AS・内部 AS・3’末端 AS と判定した（図 1-3）。

（3）AS 判定された転写物を、5 つの典型的な AS 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 AS エクソン・選択的保持イントロン）に分類した（図 1-4）(27)。

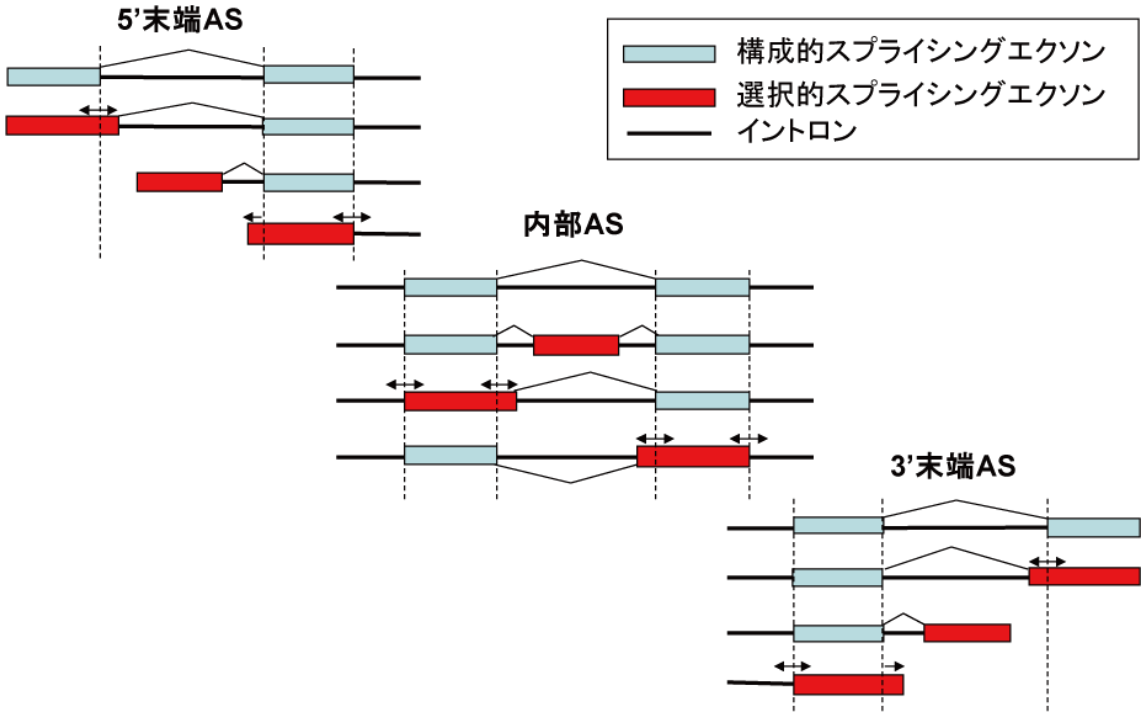


図 1-3　選択的スプライシング（AS）の判定法。AS エクソン（赤色の四角）の位置により、5’末端 AS・内部 AS・3’末端 AS に区分される。

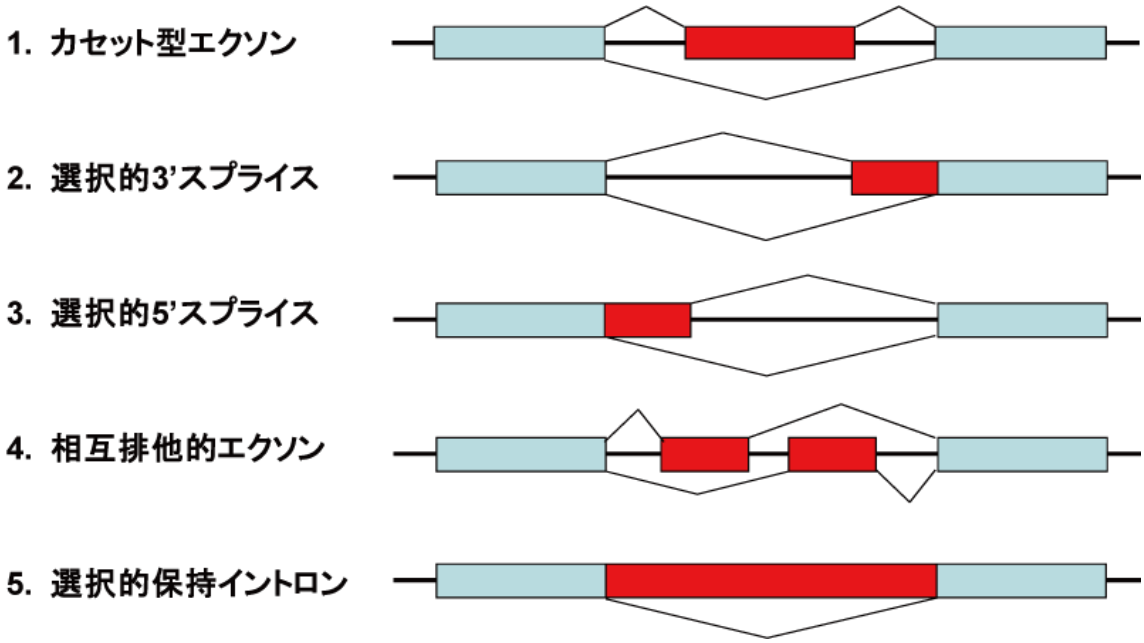


図 1-4　典型的な AS パターン。エクソンの分類およびイントロンは図 1-3 と同じ。

1.2.3 代表選択的スプライシングバリアント(RASV)の同定

選択的スプライシング(AS)によって生じたと判定された転写物(AS バリアント)について、同じエクソン構造を持つものに対しグループ分けを行なった。冗長性を省くため、その中から代表を選んで解析対象とした。これを、代表 AS バリアント(RASV)と定義した(図 1-5)。RASV は、AS グループ内でゲノムにマッピングされた範囲の長さが中央値(偶数なら長い方)の AS バリアントを選んだ。同定には Perl バージョン 5.8.8 で作成した独自のプログラムを使用した。

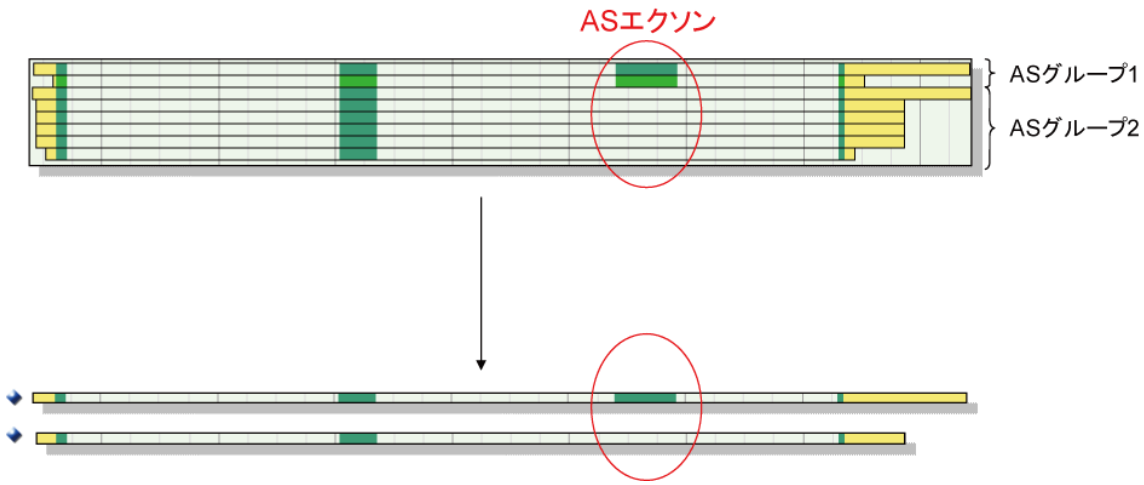


図 1-5 代表選択的スプライシングバリアント(RASV)の定義。このカセット型 AS 遺伝子の例では、AS エクソンが含まれる AS グループ 1(総エクソン数は 4)と含まれない AS グループ 2(総エクソン数は 3)に分類後、それぞれのグループ内で代表の AS バリアント(RASV)が 1 つずつ選ばれている。緑色がタンパクコード配列(CDS)、黄色が非翻訳領域(UTR)を表す。それ以外の AS バリアント内に見られる領域はイントロンを示す。

1.2.4 RASV のタンパク機能アノテーション

タンパクコード配列(CDS)(80 アミノ酸以上の ORF の中で一番長いもの)を有するヒト完全長 cDNA 配列には、タンパク機能モチーフ・遺伝子オントロジー(GO)(28)・細胞内局在化シグナル・膜タンパクドメインの 4 つのタンパク機能のアノテーションを行った。タンパク機能モチーフと GO は InterProScan によって InterPro(24)から検出された結果を、細胞内局在化シグナルは TargetP(29)と PSORT II(30)、膜タンパクドメインは SOSUI(31)と TMHMM(32)によって予測された結果を用いた。それぞれのタンパク機能アノテーションにおいて、RASV 間で異なるアノテーションがされているだけでなく、片方の RASV ではアノテーションされており、もう一方では何もアノテーションされてないものについても、”タンパク機能アノテーションに影響を与える

AS”と定義した。

1.2.5 複雑な AS パターンの判定

ヒト完全長 cDNA 配列のマニュアルアノテーションにおいて、前述した典型的な AS パターンに合致しない AS パターンを持つ遺伝子の存在を見出した。これらを、ブリッジ型・ネスト型・マルチプル CDS 型の 3 種類に分類し、”複雑な AS パターン”として解析した（図 1-6）。それぞれの AS パターンの判定基準を以下に記述する。判定には Perl バージョン 5.8.8 で作成した独自のプログラムを使用した。

(1)ブリッジ型：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。

(2)ネスト型：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、転写領域の一部を共有しているが、CDS は全く共有していないもの。

(3)マルチプル CDS 型：200 アミノ酸以上の CDS を持つ 2 つの RASV のうち、CDS を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。

なお、ブリッジ型の RASV については、RT-PCR による RNA 発現の確認を行った。そのプライマー配列を以下に示す。

primer A：5’-CGTGAGCTCGCCCGCCAGAAG-3’
primer B：5’-TCCAACTCCAGCTCCACATC-3’
primer C：5’-CGAGATGACGGGCTTTCTGC-3’
primer D：5’-GGAATGCCATCGGTGCTGG-3’
primer E：5’-CCGACTATGCAGAGGAGAAG-3’
primer F：5’-GCGTTCTGCTGCTGCTCGAG-3’
primer（GAPDH fw）：5’-TCGGAGTCAACGGATTTGGT-3’
primer（GAPDH rv）：5’-TGACGGTGCCATGGAATTTG-3’

これらは、ABI Prism 7900 Real Time PCR（ABI）を使用して標準反応条件で実行した。テンプレート RNA（PCR 当たり 50ng）は、RNA panel（BD Biosciences）を用いた。ネガティブコントロールのために、50ng のヒトゲノム DNA（Promega）をテンプレートとして用いた。

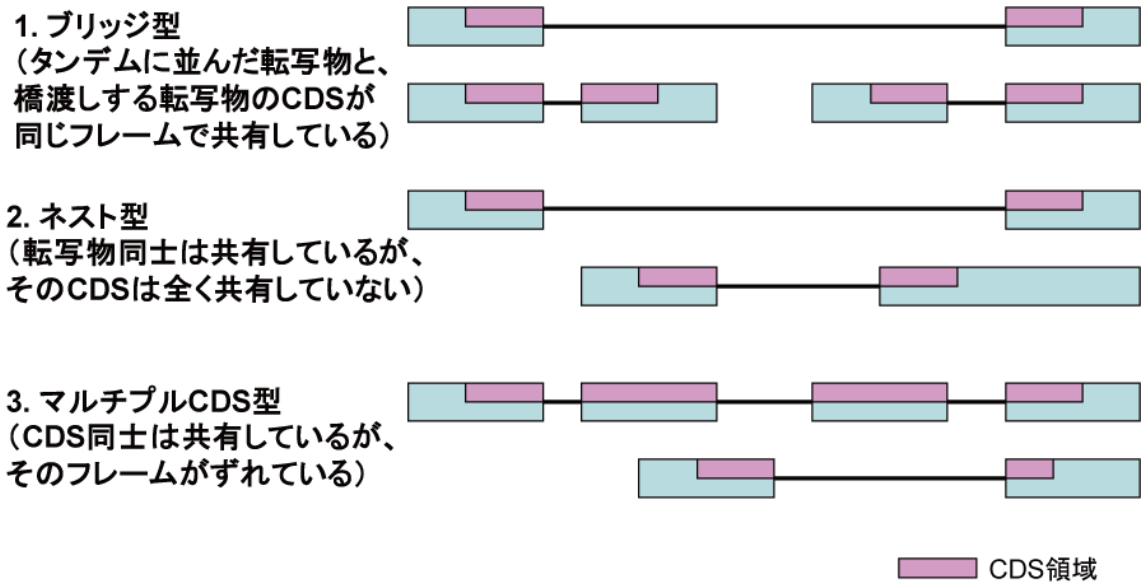


図 1-6　複雑な AS パターンの定義。四角はエクソン、太線はイントロンを示す。

1.3 結果

1.3.1 ゲノムワイドに同定されたヒト AS の統計

56,419 のヒト完全長 cDNA のうち、ゲノムにマッピングしたものが 55,036（24,425 遺伝子にクラスタリング）、データセットの成形後 1 遺伝子当たり 2 つ以上存在した cDNA は 35,030（10,127 遺伝子）であった。このデータを用いて AS の判定を行い、18,297 を RASV（6,877 AS 遺伝子）として同定した（表 1-2）。RASV の特徴は、1 AS 遺伝子当たり 2.7 RASV、1 RASV 当たり 2.1 AS エクソンであった（図 1-7）。AS イベントの位置は、5'末端、内部、3'末端のうち、エクソンの総数に対する AS エクソンの頻度が 5'末端で最も高かった（それぞれ、0.41、0.08、0.27）。また、7,494 の 5'末端 RASV のうち、47%に当たる 3,495 でそれぞれの 5'末端が 500 bp 以上離れて存在していた。これらは選択的プロモーターを用いて生成された転写産物と考えられ(33)、選択的に制御される転写と選択的に制御されるスプライシングによって、ヒトの遺伝子構造の多様化が増大していることを示唆した。

CDS に AS イベントが内包される遺伝子は 6,005（全 AS 遺伝子に対して 87%）と多く、タンパクに影響を与える AS が多いことを示唆した。タンパク機能アノテーションへの影響については、"1.3.3 RASV のタンパク機能アノテーション解析"で説明する。遺伝子単位での RASV 間のアミノ酸長の違いは平均で 123 アミノ酸であり、80%は 200 アミノ酸以下であった（図 1-8）。3%の AS 遺伝子については 500 アミノ酸以上の違いがあり、これらはタンパク機能の多様性に大きく寄与する集団であると考えられた。この集団に属するものが多いと考えられる複雑な AS パターンを持つ遺伝子については、"1.3.4 複雑な AS パターンを有する RASV の解析"で説明する。AS エクソンの生成には、レトロトランスポゾン(34)の短鎖散在反復配列（SINE）の 1 種である *Alu* が関係しているという報告がある(35)。AS エクソンと構成的スプライシングエクソンについて、RepeatMasker（Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>）を用いて検出した *Alu* を含む割合を調べたところ、それぞれ 12%と 2%であり、これまでの報告と一致していた。

表 1-2 ヒト完全長 cDNA から同定された AS の統計

	遺伝子	cDNA	総エクソン	AS エクソン	構成的スプライシングエクソン
H-Invitational 2	25,585	56,419	389,895[a]	44,727[a]	345,168[a]
ゲノムにマッピングしたもの	24,425	55,036	389,895	44,727	345,168
1 遺伝子当たり	10,127	35,030	331,924	44,727	287,197

	≧2 cDNA				
RASV	**6,877**	**18,297**	176,505	37,670	138,835
5'末端 AS	4,568	7,494	18,297	7,494	10,803
内部 AS	5,565	11,156	139,911	25,236	114,675
3'末端 AS	2,933	4,940	18,297	4,940	13,357
5' UTR AS	3,216	4,750	18,262	6,398	11,864
CDS AS	**6,005**	13,409	148,242	28,728	119,514
3' UTR AS	797	1,034	5,877	1,401	4,476

[a] マッピングされなかったエクソンについてはカウントできなかった。

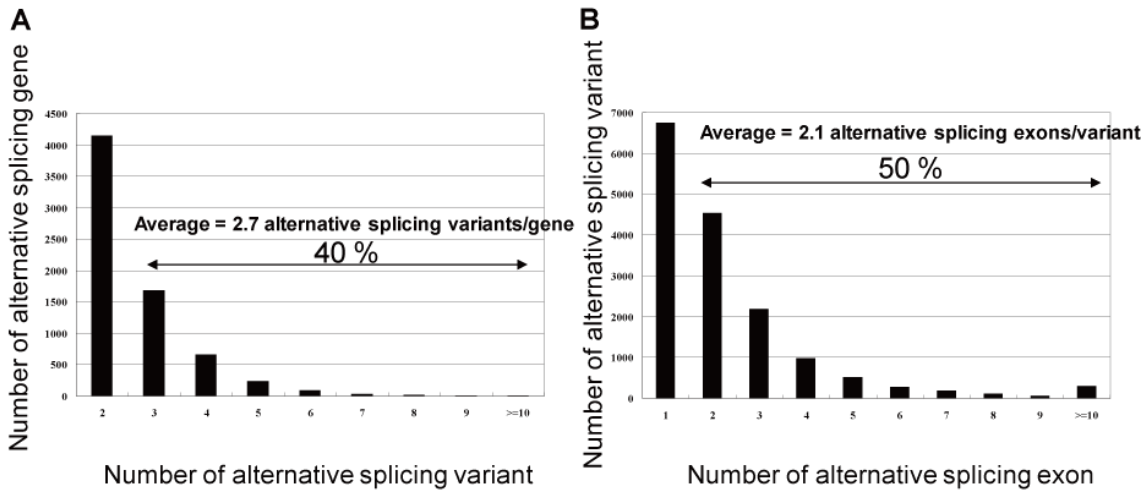


図 1-7　RASV の特徴。(A)1 AS 遺伝子当たりの RASV 数(縦軸は AS 遺伝子数、横軸は RASV 数)。平均は 2.7 で、40%の AS 遺伝子は RASV を 3 つ以上含んでいた。(B)1 RASV 当たり AS エクソン数(縦軸は RASV 数、横軸は AS エクソン数)。平均は 2.1 で、RASV の 50%は AS エクソンを 2 つ以上含んでいた。

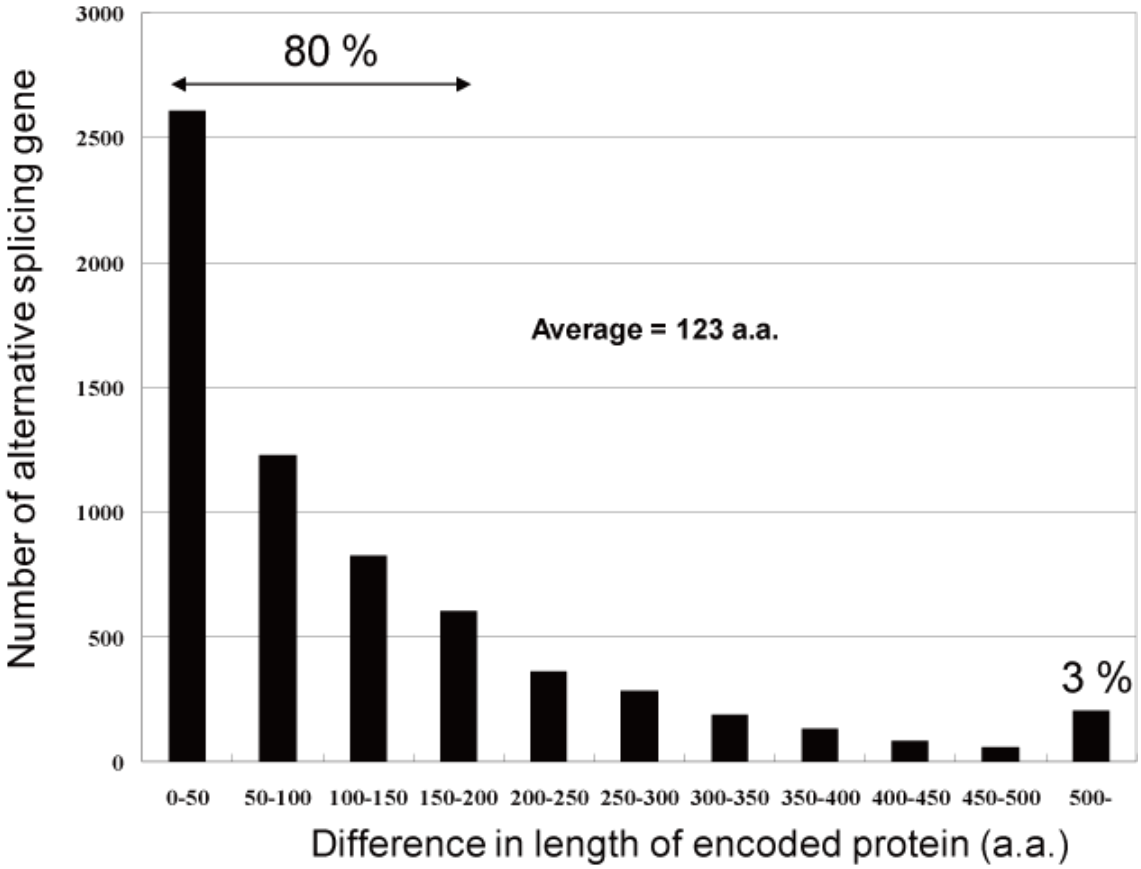


図 1-8　遺伝子単位での RASV 間のアミノ酸長の違い。縦軸は AS 遺伝子の数、横軸は RASV 間のアミノ酸長の差を示す。平均は 123 で、80%の AS 遺伝子は 200 以下、3%の AS 遺伝子は 500 以上であった。

1.3.2　RASV の典型的な AS パターン

18,297 の RASV（6,877 AS 遺伝子）について、5 つの典型的な AS パターンの数を表 1-3 に示す。最も多いのはカセット型エクソンであり、これはヒト 22 番染色体で報告された結果の通りであった(36)。カセット型エクソンには、組織によって異なるスプライシングの制御（エクソンのインクルージョン・エクスクルージョン）に関わるシス制御モチーフが多く存在することが知られているため(37)、このパターンを持つ AS バリアントが多いのだと考えられる。選択的保持イントロンも多数見出されたが、これにはイントロンがスプライス除去されていない pre-mRNA の形のままのものが含まれている可能性が考えられた。このパターンを示す AS について、ナンセンス変異介在的 mRNA 分解（NMD）(38)を引き起こす可能性がある RASV を調べると、1 つしか存在しなかった。よって、本研究で見出されたほぼ全ての選択的保持イントロン RASV は、ヒトのタンパク生成の多様性に寄与すると考えている。

表 1-3 RASV の典型的な AS パターンの統計

	AS 遺伝子	RASV
カセット型エクソン	**3,020**	**8,166**
選択的 3'スプライス	1,758	4,896
選択的 5'スプライス	1,686	4,537
相互排他的 AS エクソン	210	636
選択的保持イントロン	1,970	2,803

1.3.3 RASV のタンパク機能アノテーション解析

解析に用いたタンパク機能アノテーション(タンパク機能モチーフ・細胞内局在化シグナル・GO・膜タンパクドメイン)について、これらに影響を与える AS の数を表 1-4 に示す。タンパク機能モチーフに影響を与える AS 遺伝子は合計で 3,015 であった。図 1-9A に示すように、Iκ B kinase epsilon gene(Iκ B キナーゼε 遺伝子(IKKε))(NM_014002)内に、キナーゼドメインが欠失している新しい AS バリアント(AK093798)を同定した。IKK 複合体は、転写因子 NF-κ B へシグナルを伝達することによって NF-κ B を活性化させるため、免疫や炎症反応の際に重要な役割を演じる(IKK 複合体が NF-κ B と結合していた Iκ B をリン酸化して分解することにより、NF-κ B が核内に移行して標的遺伝子の転写を活性化させる)(39)。IKKε のキナーゼ欠失変異体は、ホルボールエステル(PMA)や T 細胞受容体によって引き起こされる NF-κ B の誘導を阻害するが、腫瘍壊死因子(TNF)-α やインターロイキン(IL)-1 によって引き起こされるその活性化については効力がないことが報告されている(40)。AK093798 に代表される IKKε のキナーゼ欠失バリアントは、これら 2 つのシグナル伝達経路の間の変調器(モジュレーター)として働き、細胞が 2 つの経路から生じるシグナルの相対的な量を、AS を利用することによって 1 つの遺伝子で調節できるようにするために存在しているのかもしれない。次に、AS 遺伝子とタンパク機能モチーフの関係を表 1-5 に示す。フィッシャーの正確確率検定により、AS 遺伝子に有意($p < 10^{-16}$)にタンパク機能モチーフが濃縮されていることが示された。また、タンパク機能モチーフを含むエクソンの特徴として、平均 1.6 AS エクソンにタンパク機能モチーフが含まれるのに対し、構成的スプライシングエクソンは平均 3.0 であった。AS エクソンにタンパク機能モチーフが含まれる頻度が高いという傾向は、EST を基にした研究でも報告されている(41)。

図 1-9B に細胞内局在が異なる RASV のペアを示す。最も多かった細胞内局在の違いは、図の例のように nuclear(核)と cytoplasm(細胞質)で、数は 2,455 であった。2 番目に多かったのは secretory(分泌小胞)と plasma membrane(細胞膜)で、数は 1,145 であった。AS の結果によって同じ遺伝子の転写産物間でこのような違いが生じるのは、タンパクの細胞内局在を

変化させるために AS が使われることが多いからだと考えている。タンパク配列を用いた先行研究からも、タンパクアイソフォーム間で細胞内局在の異なるものが多いという報告がある(42)。

図 1-9C には、膜タンパクドメインが異なる RASV のペアを示す。選択的保持イントロンの内部に膜タンパクドメインを持つ RASV と、その部分がイントロンとしてスプライス除去された RASV を同定した。RASV 間で GO の異なる例は、タンパク機能モチーフの異なる例と同じく図 1-9A に示す。また、AS 遺伝子に濃縮して観察されるタンパク機能モチーフと GO を表 1-6 に示す。この結果は、シグナル伝達や転写制御に関わるタンパク機能モチーフと GO が、AS によって影響を受ける可能性が高いことを示唆している。

表 1-4 タンパク機能アノテーションに影響を与える AS の統計

	AS 遺伝子	RASV
タンパク機能アノテーションに変化が認められる数	**4,481**	**12,542**
タンパク機能モチーフ	3,015	8,727
細胞内局在化シグナル	2,982	8,624
GO	1,779	5,179
膜タンパクドメイン	1,348	3,933

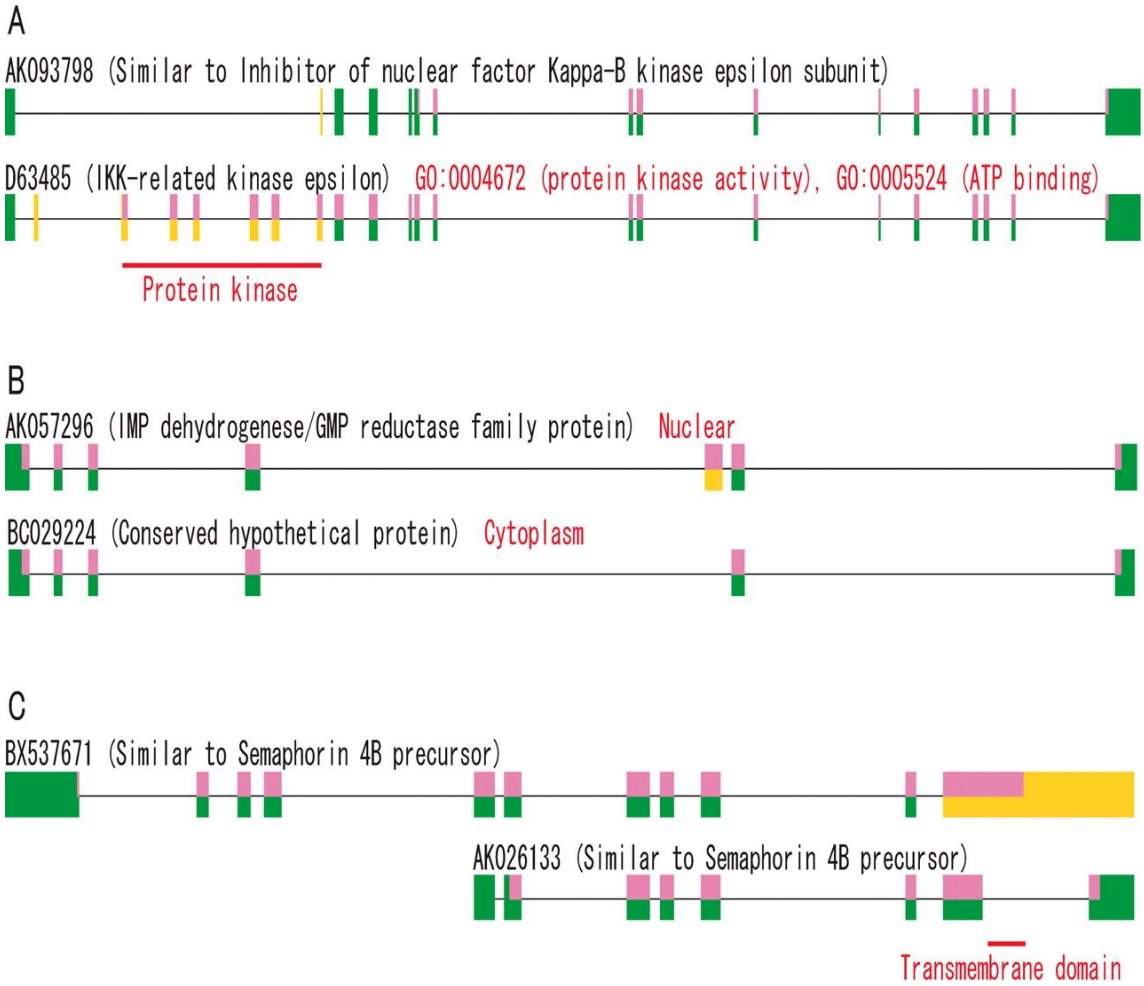


図 1-9　タンパク機能アノテーションに影響を与える RASV の例。(A)タンパク機能モチーフと GO が異なる例。Iκ B キナーゼε 遺伝子の中で、AK093798 がキナーゼドメインのタンパク機能モチーフを欠失した新しい AS バリアントとして同定された。(B)細胞内局在が異なる例。一つの RASV は核、もう一つの RASV は細胞質に存在すると予測された。(C)膜タンパクドメインが異なる例。片方の RASV ではイントロンとしてスプライス除去される選択的保持イントロンの中で、膜タンパクドメインが予測された。黄色い四角は AS エクソンを示す。桃色が CDS、緑色が UTR を表す。

表 1-5　AS 遺伝子とタンパク機能モチーフの関係

	タンパク機能モチーフを含む遺伝子	タンパク機能モチーフを含まない遺伝子	計
AS 遺伝子	5,523	1,354	6,877
非 AS 遺伝子	7,241	10,307	17,548
計	127,64	11,661	24,425

表 1-6　AS 遺伝子に濃縮して観察されるタンパク機能モチーフと GO

InterPro ID	AS 遺伝子に現れるタンパク機能モチーフ	全遺伝子に現れるタンパク機能モチーフ	*p* 値 [a]	Definition
IPR003598	417	495	$<10^{-16}$	Immunoglobulin C-2 type
IPR000867	73	79	$<10^{-16}$	Insulin-like growth factor-binding protein (IGFBP)
IPR003088	114	211	$<10^{-15}$	Cytochrome c, class I
IPR008957	55	78	10^{-15}	Fibronectin, type III domain
IPR002017	56	88	10^{-12}	Spectrin repeat
IPR002472	62	103	10^{-11}	Palmitoyl protein thioesterase
IPR002035	42	60	10^{-11}	von Willebrand factor, type A
IPR000595	22	25	10^{-10}	Cyclic nucleotide-binding domain
IPR003034	31	42	10^{-9}	DNA-binding SAP
GO ID	**AS 遺伝子に現れる GO**	**全遺伝子に現れる GO**	***p* 値 [a]**	**GO term**
GO:0003676	451	1,112	$<10^{-16}$	Nucleic acid binding
GO:0003700	327	518	$<10^{-16}$	Transcription factor activity
GO:0003677	276	603	$<10^{-16}$	DNA-binding
GO:0004713	164	318	$<10^{-16}$	Protein tyrosine kinase activity
GO:0005215	164	299	$<10^{-16}$	Transporter activity
GO:0008270	148	276	$<10^{-16}$	Zinc ion binding
GO:0005520	73	79	$<10^{-16}$	Insulin-like growth factor-binding
GO:0005524	379	967	10^{-14}	ATP binding
GO:0003824	190	429	10^{-13}	Catalytic activity
GO:0016491	116	237	10^{-11}	Oxidoreductase activity

[a]6,877 AS 遺伝子および 24,425 全遺伝子を用いて、フィッシャーの正確確率検定により求めた。

1.3.4 複雑な AS パターンを有する RASV の解析

典型的な AS パターンに該当しないが、タンパクの多様性に貢献していると考えられる複雑な AS パターンを 3 種類見出した（表 1-7）。これらのうち、ブリッジ型と定義した AS の例を図 1-10A 上パネルに示す。RASV である AK000438 が 2 つの遺伝子、SERF2（NM_005770）と HYPK（NM_016400）を橋渡ししている。これらの遺伝子は神経筋疾患に関連しており、それぞれ脊髄性筋萎縮症とハンチントン病の変更遺伝子候補として同定された(43,44)。AK000438 は、SERF2 と HYPK の転写物のプライマーを用いた RT-PCR（”1.2.5 複雑な AS パターンの判定”を参照）によって、正常組織での RNA 発現を確認した（図 1-10A 下パネル）。

図 1-10B に、ネスト型と定義した AS の例を 2 つ示す。それぞれ、タンパク機能モチーフを含む RASV と含まない RASV がある。タンパク機能モチーフを含まない RASV は、共に Conserved hypothetical protein（機能未知だが種を超えて保存されているタンパク）であった。さらに、ゲノムに対し 15 Kbp 以上に渡って identity と coverage がほぼ 100%であり、マッピングエラーとは考えられないため、細胞内に存在する転写物を反映していると考えられた。図 1-10B 下パネルの例では、ネスト型の RASV ペアが 5’末端エクソンを共有している。先行研究からも、少数例についてこの RASV ペアと同様 1 つのプロモーターから共発現し、組織特異的または細胞タイプ特異的に発現する独立した 2 つの遺伝子の存在が報告されている(45)。

図 1-10C に、マルチプル CDS と定義した AS の例を 2 つ示す。図 1-10C 上パネルの AS では、AK097244 の最後のエクソン内の CDS と、AK000272 の 2 番目のエクソン内の CDS が共有しているが、それぞれの CDS のフレームは異なっている。同様に、図 1-10C 下パネルは AK096258 の 2 番目のエクソンと BC029781 の 5 番目のエクソンのそれぞれの CDS が異なったフレームで共有している。興味深いことに、この遺伝子のもう 1 つの RASV である BC043484 は、前半のタンパクが BC029781 のフレームと等しく、後半のタンパクが AK096258 のフレームと等しかった。BC043484 では、6 番目のエクソンがフレームをスイッチする役割をしていた。この遺伝子の機能は未知だが、本研究で新しく発見された AS バリアント（”1.3.3 RASV のタンパク機能アノテーション解析”を参照）と同様、モジュレーターとして働いているのかもしれない(46)。

表 1-7 複雑な AS パターンの統計

	AS 遺伝子	RASV
複雑な AS パターンの数	**316**	**1,033**
ブリッジ型	129	604
ネスト型	172	390
マルチプル CDS 型	27	56

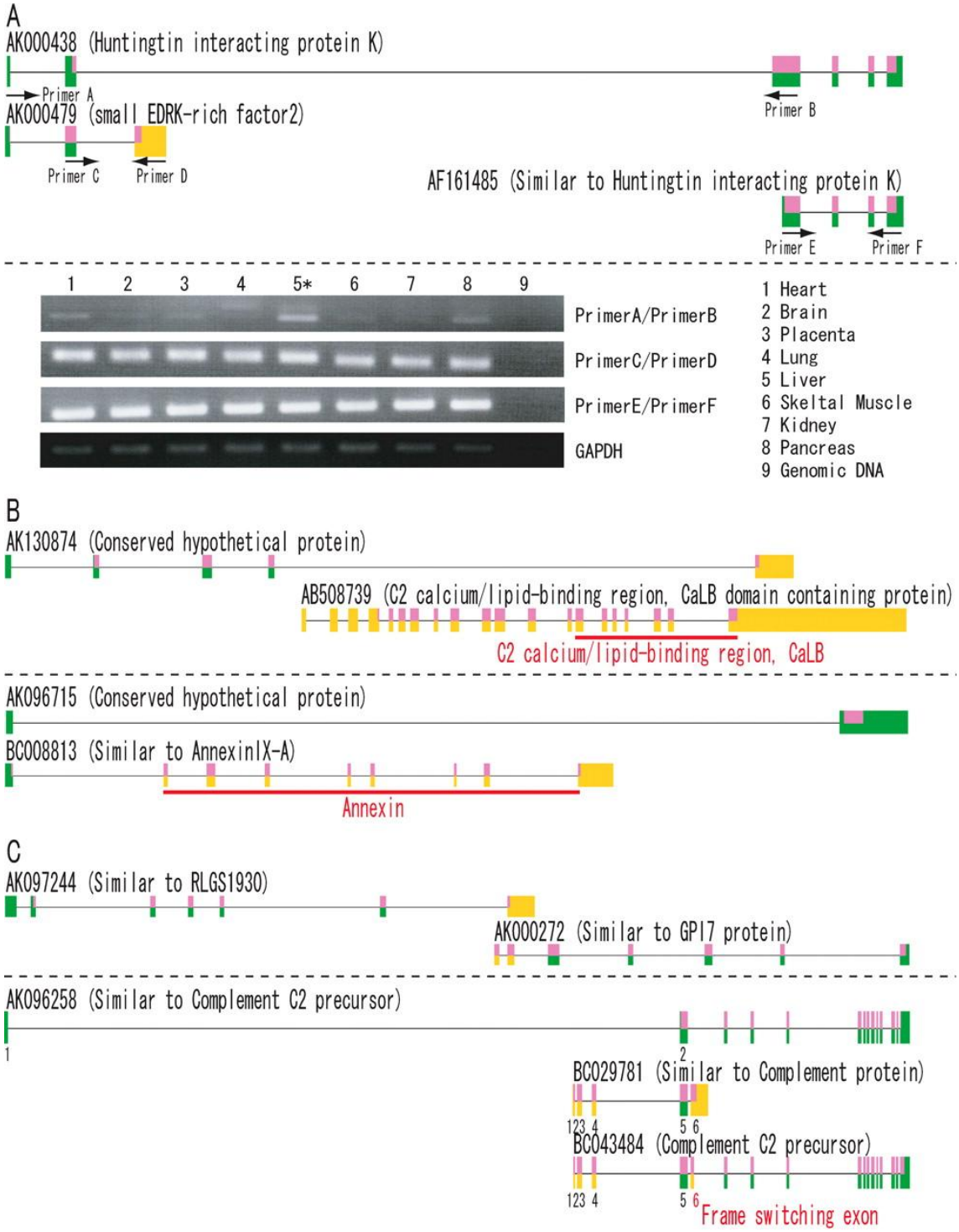


図 1-10　複雑な AS パターンの例。(A)上パネルはブリッジ型の例。AK000438 が 2 つの遺伝子を橋渡ししている。下パネルは RT-PCR の結果。5 レーンの正常組織(肝臓)で、ここに挙げた 3 つ全ての RASV の RNA 発現を確認した。(B)ネスト型の例。上下パネル共にタンパク機能モチーフのない RASV が Conserved hypothetical protein であった。下パネルは同じプロモーターを共有しながら、全く異なるタンパクを生成する例でもある。(C)マルチプル CDS 型の

例。上パネルは、共有している CDS のフレームが異なる RASV ペア。下パネルの AK096258 と BC029781 も、同様に異なるフレームで CDS を共有していた。BC043484 の 6 番目のエクソンは CDS のフレームを変えるスイッチの役目を有し、BC043484 の前半の CDS は BC029781 と共通のフレーム、後半（6 番目のエクソン以降）の CDS は AK096258 と共通のフレームであった。色は図 1-9 と同じ。

1.4 考察

1.4.1 タンパク機能に影響を与えるヒト選択的スプライシングバリアント

56,419 のヒト完全長 cDNA（25,585 遺伝子）を用い、18,297 の代表選択的スプライシングバリアント（RASV）（6,877 AS 遺伝子）を初めてゲノムワイドに同定した。この RASV について、典型的な AS パターン・タンパク機能アノテーションに影響を与える AS・複雑な AS パターンなどについての解析を行った。結果として、タンパク機能アノテーションに影響を与える AS 遺伝子が 4,481（全 AS 遺伝子に対して 65%）と多く、タンパク機能の多様性に AS が大きく寄与していることを示した。また、既知の AS バリアントに含まれるタンパク機能モチーフを含まない新しい AS バリアントを多数同定した。これらの AS バリアントは、タンパクの機能の一部を失うことによって、細胞内におけるシグナル伝達に際しモジュレーターとして働いているのかもしれない。なお、AS によって影響を受けるタンパク機能のほとんどは、シグナル伝達や転写制御に関わるものであった。また、典型的な AS パターンに該当しないが、タンパクの多様性に貢献していると考えられる複雑な AS パターンの解析では、ブリッジ型・ネスト型・マルチプル CDS 型の計 316 遺伝子（全 AS 遺伝子に対して 5%）を同定した。これらは、タンパクの多様性を必要とする細胞内遺伝子システムのさらなる複雑さに寄与するものとして、興味深い例だと考えている。

本研究で行った解析は、2 つの利点を有している。第一に、計算機プログラムを用いた自動処理に加えて、マニュアルアノテーションを併用した点である。計算機上、あるいは人為的に明らかなエラーや例外的な事例に対し、プログラムの修正・変更等のフィードバックを行い、自動処理を改良することが可能となった。第二に、解析に完全長 cDNA を用いた点である。完全長 cDNA ではない転写物は、エクソンの完全な形を反映していないため、多くのエクソンにまたがるタンパク機能モチーフなどが検出されない可能性がある。さらに、完全長 cDNA ではない転写物は、5'末端が必ずしも転写開始点とは限らないため、細胞内局在を予測するシグナルペプチドのような位置依存の配列（N 末端アミノ酸）に対し、誤ったアノテーションが成される可能性がある。本研究では、全て完全長 cDNA 配列を用いており、それらを計算機プログラムとマニュアルの両方で相互補完しながらアノテーションしたことによって、精度の高いヒト AS バリアントの同定と解析結果を得ることができたと考えている。

第二章　ヒトとマウスのゲノムアラインメントを用いた、選択的スプライシングの比較ゲノム解析

2.1 緒言

2.1.1 種間比較と AS

異なる種間でゲノム配列をアラインメントすることにより、その配列相同性から種間の進化的保存度を調べることが定法となっている。共通の祖先を持つ遺伝子はオルソログ遺伝子と呼ばれ、一般的に高い相同性を示す。ヒトとマウスのオルソログ遺伝子については、これまでの研究からその全体的構造の保存度は高いが(47)、その内包する選択的スプライシング（AS）については、保存度は相対的に低いと報告されている(48,49)。ただし、これらは EST を用いた部分的配列の解析結果であり、エクソンの数や順番が考慮された転写物配列全体像における AS の保存性については依然として不明であった。本研究では、ヒトとマウスの完全長 cDNA を用いて、第一章で同定した代表選択的スプライシングバリアント（RASV）単位での比較ゲノム解析を行い、RASV に種間保存度のアノテーションを加えることによって、ヒト RASV の進化的観点から見たタンパク機能アノテーションの特徴を解析した(50)。

2.1.2 ヒト特異的なスプライシング制御

ヒトとマウスの比較ゲノム解析を行うことにより、ヒト特異的な AS 遺伝子を同定した。この AS 遺伝子のうち、ヒト特異的に制御されるスプライシングの解析を試みた。スプライシングは、pre-mRNA のイントロンに含まれるスプライスサイト配列（主に GT-AG、まれなケースとして GC-AG と AT-AC がある）を、U2 タイプのスプライソソーム（U1・U2・U4・U5・U6 の 5 つ の核内低分子リボ核タンパク（snRNP）と様々なスプライシング因子を含むタンパク複合体）によって認識されることから始まる（AT-AC と GT-AG を認識する U12 タイプのスプライソソーム（U11・U12・U4atac・U5・U6atac の 5 つの snRNP と様々なスプライシング因子を含むタンパク複合体）も少ないながら存在する）(51)。このスプライソソームが、pre-mRNA のイントロンのスプライス除去を段階的に行い、最終的に成熟 mRNA を生成する(52)。AS は、エクソン内に含まれるスプライシングエンハンサー（ESE）(53)とスプライシングサイレンサー（ESS）(54)それぞれに結合する SR タンパクと hnRNP が、snRNP に作用することによって引き起こされる(27)。イントロンに含まれるスプライシングシス制御モチーフもいくつか報告されている。特に、組織および位置特異的に結合して AS を引き起こす Fox や Nova などのトランス作用スプライシング因子については、実験的・情報学的解析によって詳細な標的 RNA マップが作成されている(55,56)。今回、ヒトとマウスの完全長 cDNA およびゲノムアラインメントを用いた比較ゲノム解析から得た AS の種間保存度の結果を用い、ヒト独

自の脳進化に関わるASの機能を調査するため、ヒト脳特異的にFoxがスプライシング制御を行うと考えられる例の探索を行った。

2.2 方法

2.2.1 ヒトとマウスのゲノムアラインメント

ヒトとマウスのゲノムアラインメント配列は、比較ゲノムブラウザーの G-compass(57)とオルソログデータベースの Evola(58)で使われたものを用いた。ヒトとマウスのゲノム（UCSC hg18、mm8）(22)を、BLASTZ(59)（オプションパラメーターとして C=2）で対合し、スコアが 3000 以上のゲノムアラインメントを用いた。アラインメントされた領域が重複していた場合は、相互ベストヒット法によりヒトとマウスのゲノム領域を必ず 1 対 1（オルソログ領域）に対応させた。

2.2.2 ヒト RASV の保存度判定とタンパク機能アノテーション

ヒトの完全長 cDNA 配列は、H-Invitational 2 に Mammalian Gene Collection（MGC）(16)から追加した計 64,034 本を用いた。マウスの完全長 cDNA 配列は、FANTOM 3(60)と MGC を合わせた計 118,775 本を用いた（表 2-1）。ヒト AS の判定は、”1.2.2　選択的スプライシング（AS）と AS パターンの判定”と同様の方法で行い、20,392 の RASV（7,601 AS 遺伝子）を同定した。保存度の判定は、エクソン単位・RASV 単位・同一スプライシングバリアント（ESV）単位で行った（図 2-1）。それぞれの判定基準を以下に記述する。判定には Perl バージョン 5.8.8 で作成した独自のプログラムを使用した。

（1）エクソン単位：保存度の判定は、エクソン単位を基準として行った。ヒト RASV エクソンの全長または CDS 領域が、ヒトーマウスのゲノムアラインメントにマップされなかった、もしくはマップされても閾値（coverage=70%かつ identity=60%）に満たなかった場合、非保存エクソンとした。ゲノムアラインメントに閾値以上でマップされたが、対応するマウス転写産物にエクソンとして見出されたなかった場合、ゲノム保存エクソンとした。対応するマウスエクソンとも閾値以上でアラインメントされた場合は、転写物保存エクソンとした（図 2-1A）。

（2）RASV 単位：（1）で同定されたエクソンの保存度から、RASV 単位での保存度を判定した。RASV を構成するエクソンのうち、1つでも非保存エクソンが存在していれば、非保存 RASV とした。非保存エクソンが存在せず、ゲノム保存エクソンが存在していれば、ゲノム保存 RASV とした。転写物保存エクソンのみで構成されていれば、転写物保存 RASV とした（図 2-1B）。

（3）ESV 単位：（2）で同定された転写物保存 RASV のうち、転写物全長または CDS 領域にお

いて、対応するマウスの完全長 cDNA も全て転写物保存エクソンで構成され、エクソンの数も順番も等しかった場合、それらを同一スプライシングバリアント（ESV）と定義した。遺伝子内に 2 つ以上の ESV が存在していた場合、すなわち、AS イベントを含む 2 つ以上の RASV が保存されていた場合は、その遺伝子を保存 AS と定義した（図 2-1C）。

保存 AS として判定された遺伝子に対しては、第一章同様 AS によるタンパク機能モチーフ・GO・細胞内局在化シグナル（TargetP と WoLF PSORT(61)）・膜タンパクドメインへの影響について解析した（”1.2.4　RASV のタンパク機能アノテーション解析”を参照）。

表 2-1　マウス完全長 cDNA を配列決定したプロジェクト

プロジェクト/研究機関	cDNA	参考文献	URL
FANTOM3/RIKEN	101,789	Carninci *et al.* (60)	http://fantom3.gsc.riken.jp/
MGC/NIH	16,986	Strausberg *et al.* (16)	http://mgc.nci.nih.gov/
計	**118,775**		

FANTOM3: Functional Annotation of Mouse 3, RIKEN: RIKEN Institute, MGC: Mammalian Gene Collection, NIH: The United States National Institutes of Health

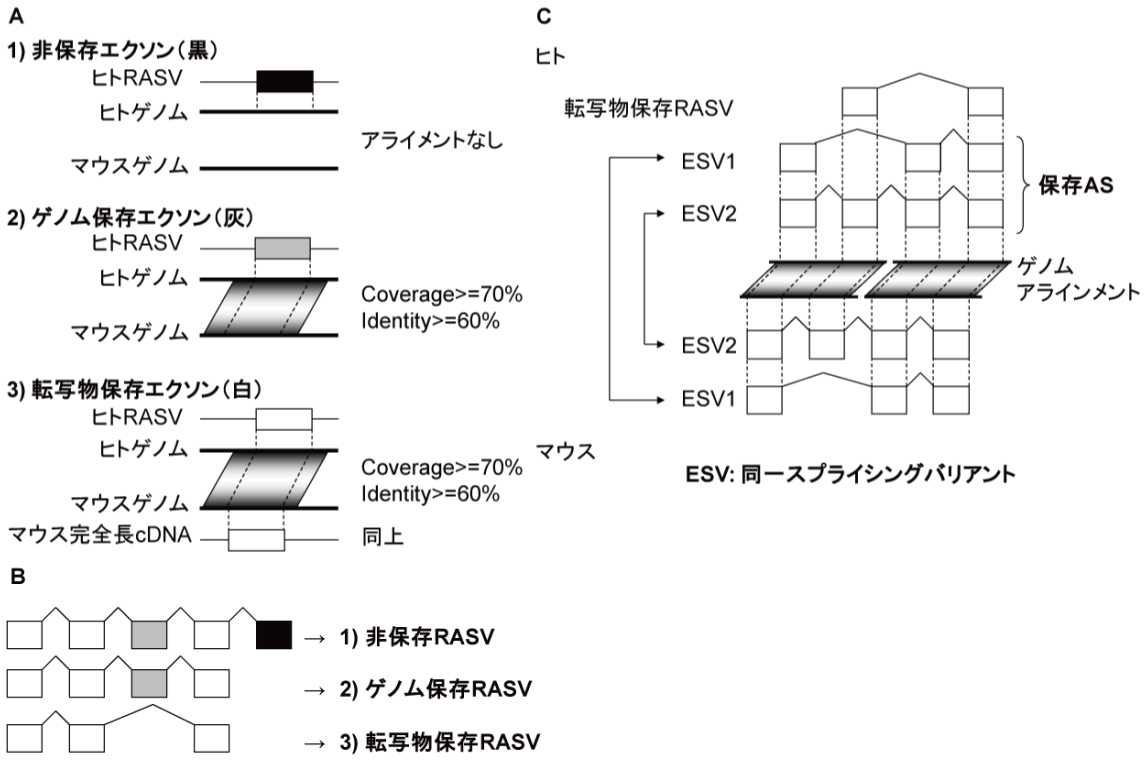


図 2-1　ヒト RASV の保存度判定法。（A）エクソン単位での判定。ヒト RASV のエクソンが、ヒトーマウスのゲノムアラインメントを介し、マウスゲノムと全くアラインメントしていない、または閾

値（coverage=70%かつ identity=60%）以下でアラインメントされていれば非保存エクソン、マウスゲノムと閾値以上でアラインメントされているが、対応するマウス完全長 cDNA のエクソンが存在しない、または閾値以下でアラインメントされていればゲノム保存エクソン、対応するマウス完全長 cDNA のエクソンとも閾値以上でアラインメントされていれば転写物保存と判定した。（B）エクソンの判定結果を用いた RASV 単位での判定。1 つでも非保存エクソンがあれば非保存 RASV、非保存エクソンがなく 1 つでもゲノム保存エクソンがあればゲノム保存 RASV、全て転写物保存エクソンで構成されていれば転写物保存 RASV と判定した。（C）種間で保存された RASV と保存 AS の同定。転写物全長および CDS 領域において、ヒトの転写物保存 RASV が対応するマウスの転写物とエクソン数およびその順序ともに等しければ、同一スプライシングバリアント（ESV）と判定した。ESV が 2 つ以上ある遺伝子を、保存 AS 遺伝子と判定した。

2.2.3 選択的保持イントロンの実験的検証

ポリソーム画分は、(62)に記述されているように、3 × 10^7 以上のヒト前骨髄球性白血病細胞株（HL60）の細胞を使って精製した。細胞の沈殿物は、100 ユニットの RNase 阻害剤を含む 1ml の溶解バッファー（20mM Tris-HCl（pH 7.5）・10mM NaCl・3mM $MgCl_2$・0.04M sucrose・0.5% NP40・1mM dithiothreitol）でけん濁した。沈殿物は 10 分間氷上で溶解し、核と細胞残骸は、4℃で 10 分間、1,000g で遠心分離することによって取り除いた。溶解物は 11ml の 15%/50% (w/v)ショ糖密度勾配の最上部に重ね、4℃で 135 分間、Beckman SW41Ti ローターの中で 36,000g で遠心分離した。密度勾配分留装置（Towa Labo, Japan, Model 152-001）は、勾配を 11 の等画分に分けるために用いた。吸光度は 260nm で測定し、それぞれの画分はプロテイナーゼ K で処理した。RNA はフェノールとクロロホルム（$CHCl_3$）を使って抽出し、エタノールで沈殿した後、それぞれの mRNA に対して解析を行った。First-strand cDNA 合成は、メーカーによって指示された 17 bp dT プライマーと SuperScript II を用いて行った。得られた cDNA は定量化し、定量 RT-PCR 解析に用いた。解析結果は、グリセロアルデヒド-3-リン酸デヒドロゲナーゼ（GAPDH）PCR 産物に対して標準化を行った。PCR サイクルのパラメーターは、50℃で 2 分間、95℃で 10 分間、続いて 95℃で 30 秒を 35 サイクル、57℃で 1 分間、72℃で 1 分間とした。ABI PRISM HT7000 Sequence Detection System（Applied Biosystems）を PCR 産物の検出に使用した。最終的な PCR 産物の大きさと完全性（不鮮明な増幅産物とプライマーダイマーの混入率）は、アガロースゲル電気泳動で確認した。

2.2.4 ヒト脳特異的 Fox 制御スプライシングの探索

上述のヒトのゲノム保存エクソンを用い、ヒトの脳で特異的にスプライシングを制御している

可能性のあるFox結合サイトを探索した。Foxタンパクが結合するRNA上の配列はUGCAUGで、脊椎動物の間で保存されていることが知られている(63)。また、Foxタンパクはそのシス制御モチーフに組織特異的(脳(特に小脳)・心臓・骨格筋)に結合すること、および位置特異的(スプライシング対象エクソンに対し、上流イントロンに位置すれば抑制、下流イントロンに位置すれば促進に働く)に結合することが知られている(55)。本解析では、ヒトRASVのゲノム保存エクソンのうち、両隣を転写物保存エクソンで挟まれたカセット型エクソンを対象とした。なお、ヒト完全長cDNAのRASVは、DDBJ(64)リリース73のフラットファイルフォーマット中(FEATURESのsourceの/tissue_type)に、脳のみで発現したと記述されているものを用いた。解析対象のヒトASエクソンが、対応するマウスのゲノム上で、マウスの転写物(完全長cDNAを含むmRNA(DDBJ・RefSeq(65)リリース23・Ensembl(66)リリース44の計310,028本)、EST(計4,366,565本)、およびマウスの脳で発現したRNA-Seq(67)タグ(ELANDのマッピング結果ファイルをSAMtools(68)でSAMファイルに変換後、不要な配列を取り除いた計12,599,212本)と共有せず、マウスに見出されないこと(ゲノム保存であること)をプログラムによる確認およびGBrowse2(69)上で目視した。そして、そのエクソンの下流300 bp以内のゲノム配列からTGCATG (Fox結合サイト)を探索し、マウスゲノム上で対応するFox結合サイトに変異があるものを、ヒト脳特異的Fox制御スプライシング候補とした(図2-2)。本解析で行った探索や確認には、Perlバージョン5.8.8で作成した独自のプログラムを使用した。

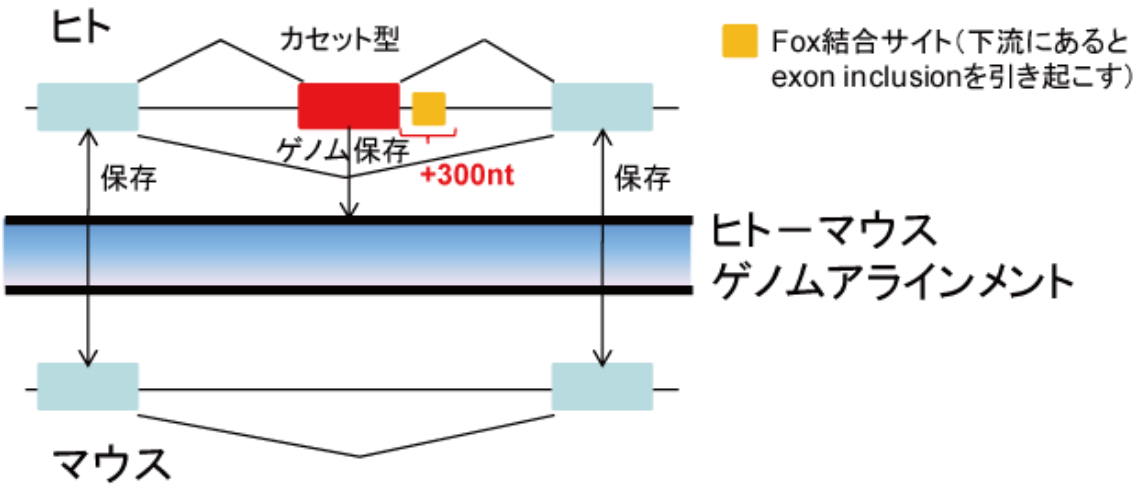


図2-2 ヒト脳特異的Fox制御スプライシングの概念図。本解析では、ヒトのRASVは脳のみで発現した完全長cDNAから同定されたものを用いた。解析対象のヒト脳特異的Foxスプライシングの対象エクソンは、両隣のエクソンがヒトとマウスで保存されたカセット型ASのゲノム保存エクソン(赤色の四角)で、下流にあるFox結合サイトがマウス側で変異を起こしているものとした。

2.3 結果

2.3.1 ヒトとマウスの AS の保存度

ヒトとマウスの AS の保存度の統計を表 2-2 に示す。エクソン単位（図 2-1A）では 74%が保存されていたが、RASV 単位（図 2-1B）では 38%であった。この中で、ESV と保存 AS 遺伝子（図 2-1C）と判定されたのは、それぞれ 23%と 3%であった（表 2-3）。3%の保存 AS 遺伝子は、189 個と非常に少数であった。これらについて、ヒト RASV の進化的特徴を知る目的で、タンパク機能アノテーション解析を行った。合わせて、非保存 AS 遺伝子に関しても、タンパク機能アノテーション解析を行った。

表 2-2 ヒトとマウスの AS の保存度の統計

	計	非保存	ゲノム保存	転写物保存
全エクソン	199,426	27,879 (14%)	23,412 (12%)	**148,135 (74%)**
AS エクソン	49,842	12,196 (25%)	9,064 (18%)	28,582 (57%)
RASV	20,392	8,296 (41%)	4,410 (21%)	**7,686 (38%)**
AS 遺伝子	7,601	1,716 (23%)	1,241 (16%)	4,644 (61%)

表 2-3 同一スプライシングバリアント（ESV）と保存 AS の統計

	ESV	保存 AS
RASV	**4,624 (23%)**	431 (2%)
AS 遺伝子	3,570 (47%)	**189 (3%)**

2.3.2 ヒトとマウスの保存 AS 遺伝子のタンパク機能アノテーション解析

図 2-3 に保存 AS 遺伝子の例を示す。図 2-3A に示す phosphoinositide-3-kinase regulatory subunit（ホスホイノシチド-3-キナーゼ（PI3 キナーゼ）調節サブユニット）は、5'末端が AS の RASV ペアを有する。ヒトの BC094795 とマウスの BC026146、ヒトの BC030815 とマウスの BC051106 がそれぞれ ESV として対応し、全てのエクソンの数と順番が種間で保存されている。この遺伝子には、p85-α ・p55-α ・p50-α の 3 つのタンパクアイソフォームがすでに同定されており(70)、BC094795 と BC030815 はそれぞれ p85-α と p55-α に対応している。p85-α は、p55-α にはない N 末端領域に Rho GTPase-activating protein（Rho GTP アーゼ活性化タンパク（RhoGAP））（IPR008936）と Src homology-3（Src ホモロジー3（SH3））（IPR001452）のタン

パク機能モチーフを含んでいる。p85-α とp55-α は、insulin receptor substrate protein（インスリン受容体基質（IRS）タンパク)から PI3 キナーゼの p110 -kDa catalytic subunit（p110 触媒サブユニット）へ異なる効率でシグナルを伝達しているという報告がある(71)。この遺伝子の AS が媒介するシグナル伝達経路は進化的に古くから存在し、生物の生存に重要なものであると考えられる。

図 2-3B は、AS の及ぼすタンパク機能への影響が未知の遺伝子の例である。cysteinyl-tRNA synthetase（システイニル tRNA シンセターゼ（CARS））としてアノテーションされたこの遺伝子の RASV のうち、BX647906 のカセット型エクソンにのみ、glutathione S-transferase C-terminal-like（グルタチオン-S-トランスフェラーゼ（GST）C 末端様）（IPR010987）が含まれていた。GST の C 末端領域は基質活性化に重要な役割を演じるが、このタンパク機能モチーフは GST だけでなく、多くのタイプの aminoacyl-tRNA synthetases（アミノアシル tRNA シンセターゼ（aaRSs））にも現れる(72)。いくつかのタイプの哺乳類の aaRSs は、GST の C 末端領域を介して translational elongation factor -1（翻訳伸長因子（EF）-1）と結合する。この結合は、アミノアシル化された tRNA をリボソームの A 部位へ移動させるのを促進すると考えられている。新しく同定された RASV である BX647906 と BC002880 の機能的役割は、CARS から EF 複合体へ同族（すなわち、システイニル）tRNA の配送に関わっている可能性がある。これにより、細胞内の特別な環境下で翻訳効率を制御しているのかもしれない(73)。

保存 AS 遺伝子の特徴は、全 AS 遺伝子よりもカセット型エクソンが相対的に多く見られ、逆に選択的保持イントロンが相対的に少なかった（表 2-4）。また、RASV 間で異なるタンパク長の差の平均は、全 AS 遺伝子が 87 アミノ酸に対し、保存 AS 遺伝子は 52 アミノ酸と短いにも関わらず、タンパク機能アノテーションへの影響は保存 AS 遺伝子の方が相対的に大きかった。より多くのカセット型エクソンを用いることにより、保存 AS 遺伝子はより過激にタンパクを変化させているのかもしれない。

表 2-5 には、保存 AS 遺伝子に濃縮して観察される GO term とタンパク機能モチーフを示す。GO term では、DNA binding（DNA 結合）（GO:0003677）・Peroxidase activity（ペルオキシダーゼ活性）（GO:0004601）・Response to oxidative stress（酸化的ストレス応答）（GO:0006979）が、フィッシャーの正確確率検定により保存 AS 遺伝子で有意（$p < 0.01$）に濃縮して観察された。これらの機能を持つ遺伝子が種間で保存されているのは、主に細胞の恒常性を維持する基礎的な役割を担い、進化的に不変な AS による制御を必要とするからだと考えている。タンパク機能モチーフに関しては、nucleotide-binding（ヌクレオチド結合）・protein kinase（タンパクキナーゼ）・WD40 protein（WD40 タンパク）が、保存 AS 遺伝子において AS の影響を受けたものの中で多数見出されたが、これらは全 AS 遺伝子でも共通して多く見出された。一方、脊椎動物において転写制御だけではなく、胚発生時に複数の細胞外シグナルを統合するエフェクターとしても働く transforming growth factor -β -stimulated clone-22（トランスフォーミング成長因子（TGF）-β 誘導クローン 22（TSC-22））（哺乳類での別名は DSIP-immunoreactive

peptide（Dip）・ショウジョウバエでの別名は Bun）（IPR000580）（74)や、真核生物で、紫外線損傷や浸透圧ストレスに対して細胞応答を開始する Basic-leucine zipper（bZIP）transcription factor（bZIP 転写因子）（IPR004827）（75)といったタンパク機能モチーフは、保存 AS 遺伝子で同じく有意（$p < 0.01$）に濃縮して観察された。環境変化などによってタンパクの機能モチーフを切り替えることは、基本的な細胞機能に必須であるため、この機能に対応する AS 配列も種間で保存されたと考えられる。これらの結果から、遺伝子機能の違い（切り替え）を保存した AS 配列のコアデータセットから成る保存 AS バリアントと、それ以外の AS バリアントでは、AS の機能が異なると考えられる。タンパク機能モチーフのサブセットは、保存 AS 遺伝子と種特異的 AS 遺伝子で機能的に異なるのかもしれない。

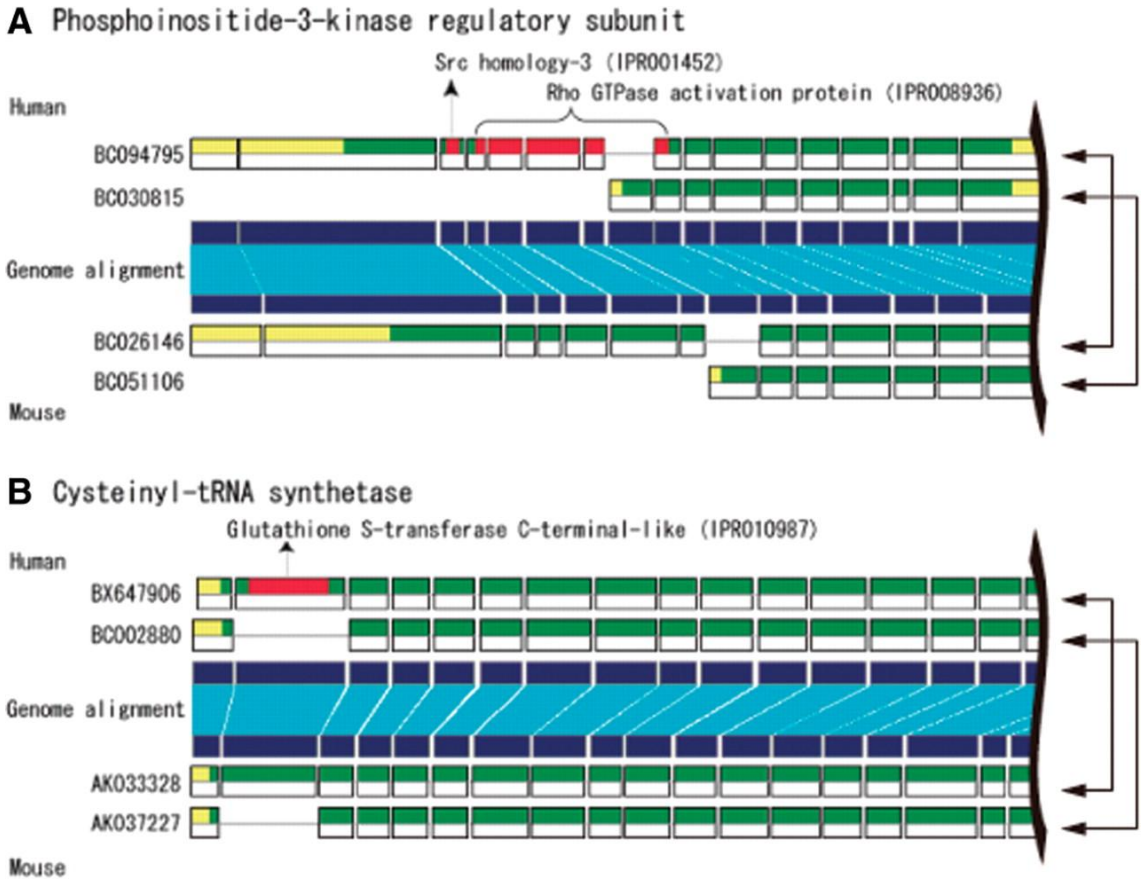


図 2-3　保存 AS 遺伝子の例。（A）phosphoinositide-3-kinase regulatory subunit（ホスホイノシチド-3-キナーゼ（PI3 キナーゼ）調節サブユニット）。ヒトとマウスで、N 末端領域にタンパク機能モチーフが含まれる RASV と含まれない RASV が保存されている。（B）cysteinyl-tRNA synthetase（システイニル tRNA シンセターゼ（CARS））。ヒトとマウスで、カセット型エクソン内にタンパク機能モチーフが含まれる RASV と含まれない RASV が保存されている。矢印で結ばれた転写物が、ヒトとマウスの同一スプライシングバリアント（ESV）を示す。緑色が CDS、黄色が UTR、赤色がタンパク機能モチーフ領域を表す。

表 2-4　全 AS および保存 AS 遺伝子に含まれる典型的な AS パターン

	全 AS 遺伝子	保存 AS 遺伝子
カセット型エクソン	3,584 (35%)	**66 (42%)**
選択的 3'スプライス	1,988 (19%)	30 (19%)
選択的 5'スプライス	1,990 (20%)	33 (21%)
相互排他的 AS エクソン	237 (2%)	4 (2%)
選択的保持イントロン	2,477 (24%)	**26 (16%)**

表 2-5　保存 AS 遺伝子に濃縮して観察される GO term とタンパク機能モチーフ

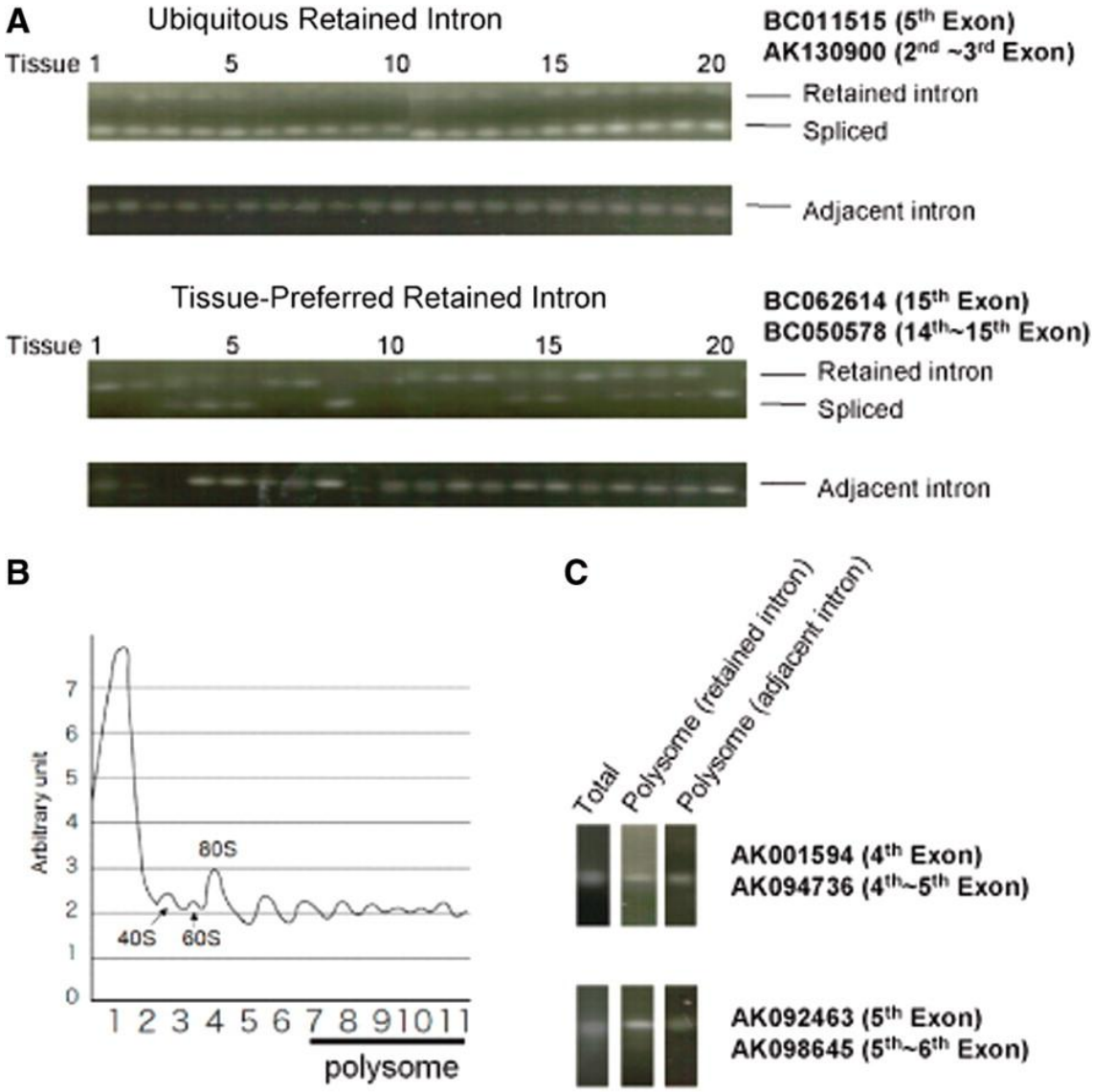
GO ID	GO term	保存 AS 遺伝子	全 AS 遺伝子	*p* 値 [a]
GO:0003677	DNA binding	17	333	0.0085
GO:0004601	Peroxidase activity	3	14	0.0077
GO:0006979	Response to oxidative stress	3	14	0.0077
InterPro ID	**Definition**	**保存 AS 遺伝子**	**全 AS 遺伝子**	***p* 値 [a]**
IPR004827	Basic-leucine zipper (bZIP) transcription factor	3	4	0.0005
IPR000580	TSC-22/Dip/Bun	2	3	0.0057

[a]189 保存 AS 遺伝子および 7,601 全 AS 遺伝子を用いて、フィッシャーの正確確率検定により求めた。

2.3.3　選択的保持イントロンの実験的検証

ヒトとマウスの比較ゲノム解析結果より、ヒトの全 RASV のうち非保存 RASV が大きな割合を占めた。非保存 RASV が、cDNA クローニングから生じたアーティファクト（不完全にスプライスされた mRNA や、ゲノム DNA 混入物質から作られた cDNA）である可能性が危惧された。配列情報からだけでは、実際に存在する選択的保持イントロン（保存 AS 遺伝子では相対的に減少してる AS パターン）かどうかを見分けることは難しい。CDS を有するとアノテーションされた選択的保持イントロン RASV から、転写物保存、ゲノム保存、非保存の各 RASV ペアを、それぞれ 14、15、5 組選んで実験的検証を行った。半定量リアルタイム RT-PCR によって、20 種類のヒト正常組織での発現パターンを測定した。少なくとも 1 つ以上の組織で観察された選

択的保持イントロン RASV に対応する PCR 産物は、転写物保存、ゲノム保存、非保存の各 RASV ペアで、それぞれ 14、3、3 組であった。隣接するエクソンについても対照実験を行ったが、どれも完全にスプライスされていた。組織に対して多様な発現パターンが観察され、両方の RASV が普遍的に発現しているもの（図 2-4A 上パネル）や、相互排他的に発現しているもの（図 2-4A 下パネル）が存在した。これらの結果は、選択的保持イントロンの AS バリアントを生じるほとんどの cDNA が実際に存在する転写物であることを示唆する。さらに、選択的保持イントロン RASV のタンパクへの翻訳について実験的検証を行った。ヒト前骨髄球性白血病細胞株である HL60 において、タンパク合成を行うリボソーム画分（ポリソーム画分）から、ショ糖密度勾配遠心法で RNA を回収した（図 2-4B）。図 2-4B に示す 7～10 のポリソーム画分から精製した RNA を、リアルタイム RT-PCR で解析した。HL60 において、選択的保持イントロンである全ての転写物保存 RASV の発現が観察された。これらのうち 9 つについては、適切なサイズの明快なバンドをポリソーム画分の RT-PCR で確認した（図 2-4C、D）。この実験的検証では、全ての保存カテゴリーで選択的保持イントロン RASV の翻訳の証拠を観察した。更なる解析は必要ではあるが、非保存の選択的保持イントロン RASV においても、その多くがタンパクへの翻訳に用いられている可能性が示唆された。



D

	transcript-conserved	genome-conserved	non-conserved
primer set	14	15	5
20 tissues	14	3	3
HL60 total RNA	14	3	3
HL60 polysome	9	0	1

図 2-4　選択的保持イントロン RASV の RT-PCR による実験的検証。(A)ヒトの 20 種類の正常組織に対し、普遍的に発現する選択的保持イントロン RASV ペア(上パネル)と、相互排他的に発現する選択的保持イントロン RASV ペア(下パネル)。各レーンの組織名は、1. 副腎、

2. 骨髄、3. 小脳、4. 脳全体、5. 胎児の脳、6. 胎児の肝臓、7. 心臓、8. 腎臓、9. 肝臓、10. 肺全体、11. 胎盤、12. 前立腺、13. 唾液腺、14. 骨格筋、15. 精巣、16. 胸腺、17. 甲状腺、18. 気管、19. 子宮、20. 脊髄。(B)RT-PCR に用いた、ヒト前骨髄球性白血病細胞株(HL60)から分離したポリソーム画分。(C)HL60 のタンパク合成中のリボソーム(ポリソーム画分)と混合した選択的保持イントロン RASV の、選択的保持イントロンと隣のイントロンの発現結果。(D)転写物保存・ゲノム保存・非保存の各カテゴリーにおいて、20 種類の組織・HL60 の total RNA・HL60 のポリソーム画分で RNA 発現を確認した選択的保持イントロン RASV ペアの数。

2.3.4 非保存 AS 遺伝子のタンパク機能アノテーション解析

ヒト RASV の非保存エクソンの特徴を解析するために、末端エクソンにおけるマウスとの保存度を調べた。表 2-6 に示すように、RASV の 5'末端に位置する約半数の非保存エクソンに AS が見出された。これは、選択的プロモーターから生じた選択的 5'末端エクソンはヒトに豊富に存在するが、進化的には保存されていないという先行研究の結果と一致する(76)。逆に、RASV の非保存エクソンに含まれる CDS は、構成的スプライシングエクソンより AS エクソンの方が観察される頻度が少なかった(表 2-7)。これは、多くの 5'末端エクソンが UTR に位置するエクソンのためだと考えられる。また、非保存の AS エクソンはタンパク機能モチーフを含む割合が小さく(表 2-7)、そのタンパク機能モチーフは保存 AS で濃縮して観察されたものと種類が大きく異なっていた。

非保存の AS エクソンを持つ遺伝子に最も高頻度に観察されたタンパク機能モチーフは、Krueppel-associated box(KRAB)(IPR001909)であった(表 2-8)(77)。このタンパク機能モチーフは、AS によって機能的に制御されていることが示されているタンパク相互作用ドメインである(78)。KRAB タンパクは、哺乳類の精巣の決定と分化に必要な sex-determining region Y (*SRY*/*Sry*) gene(SRY 性決定遺伝子)の転写を制御すると考えられている(79)。KRAB が非保存の AS エクソンに存在しやすいという事実は、生殖系におけるヒトの転写制御ネットワークの独自進化を示唆するのかもしれない。

他に 4 つのタンパク機能モチーフ、GAGE(IPR008625)・nuclear-pore-complex-interacting (核膜孔複合体(NPC)相互作用)(IPR009443)・phospholipase A2 active site(ホスホリパーゼ A2 活性部位)(IPR013090)・T-complex 11(IPR008862)が、非保存の AS エクソンを持つ遺伝子のみで観察された(表 2-8)。GAGE と nuclear-pore-complex-interacting は、共に機能未知であるが、それぞれヒト特異的、類人猿特異的であることが知られている(80,81)。ヒト遺伝子の発現データベースである H-ANGEL(82)データから、GAGE と T-complex 11 を持つ遺伝子は精巣で特異的に発現していることを観察した(図 2-5)。GAGE を持つ遺伝子のタンパク機能は非常に興味深いが、GAGE はヒト特異的であり、種間比較による解析は不可能である。一方、T-complex 11(TCP11)を持つ遺伝子はヒトとマウスの両種に存在していた。TCP-11 は、

fertilization-promoting peptide（受精促進ペプチド（FPP））の受容体をコードする遺伝子で、ヒトとマウスにおいて生殖能力のある成体の精巣のみで発現するため、精子の機能と受精能力に重要であることが報告されている(83)。非保存の AS エクソンに濃縮された GAGE と T-complex-11 が、生殖系に含まれる種特異的因子に関わっているのは興味深い。非保存の AS エクソンはまた、レトロトランスポゾンと高い割合で関連していた（表 2-7）。ヒトのレトロトランスポゾンは類人猿特異的（*Alu*）であり、マウスのレトロトランスポゾンはげっ歯類特異的であることが主要な原因である(84)。エクソン内に含まれるスプライシングエンハンサー（ESE）はヒト RASV のエクソンに散在しており、非保存の AS エクソンと保存 AS エクソンの間で明確な差異を見つけることはできなかった（表 2-7）。

表 2-6 ヒト RASV の末端エクソンにおける保存度の統計

	計	非保存	ゲノム保存	転写物保存
全 5'末端エクソン	20,392	7,687 (38%)	4,410 (21%)	8,295 (41%)
AS 5'末端エクソン	8,748	**4,201 (48%)**	2,350 (27%)	2,197 (25%)
全 3'末端エクソン	20,392	5,497 (27%)	2,983 (15%)	11,912 (58%)
AS 3'末端エクソン	6,636	2,482 (37%)	1,383 (21%)	2,771 (42%)
全末端エクソン	40,784	13,184 (32%)	7,393 (18%)	20,207 (50%)
AS 末端エクソン	15,384	6,683 (44%)	3,733 (24%)	4,968 (32%)

表 2-7 ヒト RASV エクソンの保存度とスプライシングの関係

保存度/スプライシング [a]	計	CDS	タンパク機能モチーフ	レトロトランスポゾン	ESE
C/CS エクソン	133,901	95,583 (71%)	27,805 (21%)	2,523 (2%)	130,104 (97%)
C/AS エクソン	37,646	20,805 (55%)	6,192 (16%)	2,308 (6%)	35,702 (95%)
NC/CS エクソン	15,683	7,030 (45%)	1,898 (12%)	2,549 (16%)	14,961 (95%)
NC/AS エクソン	12,196	**3,516 (29%)**	**812 (7%)**	**4,544 (37%)**	11,701 (96%)
全エクソン	199,426	126,934 (64%)	36,707 (18%)	11,924 (6%)	192,468 (97%)

[a]C：保存、NC：非保存、CS：構成的スプライシング、AS：選択的スプライシング。

表 2-8 非保存かつ AS 遺伝子で頻繁に観察されたタンパク機能モチーフ

InterPro ID	Definition	タンパク機能モチーフを含む非保存の AS エクソンを持つ AS 遺伝子	タンパク機能モチーフを含む AS エクソンを持つ AS 遺伝子	*p* 値 [a]
IPR001909	Krueppel-associated box（KRAB）	26	44	0.0003
IPR008625	GAGE	3	3	0.0806

IPR008862	T-complex 11	3	3	0.0806
IPR009443	Nuclear-pore-complex interacting	3	3	0.0806
IPR013090	Phospholipase A2 active site	3	3	0.0806

[a]1,716 非保存 AS 遺伝子および 7,601 全 AS 遺伝子を用いて、フィッシャーの正確確率検定により求めた。

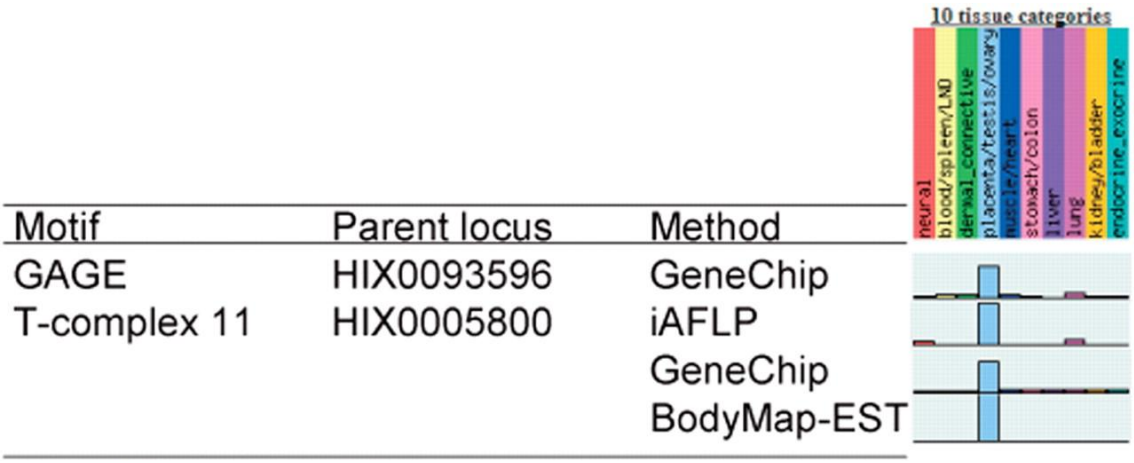


図 2-5　GAGE または T-complex 11 を持つ遺伝子の発現パターン（H-ANGEL(82)のデータより）。水色のバーが精巣での発現を示す（バーの高さは相対的で、トータルで 100%になる）。

2.3.5 ヒト脳特異的 Fox 制御スプライシング候補

脳で発現する完全長 cDNA を用いたヒトの AS 解析およびヒトとマウスの比較ゲノム解析の結果より、両隣のエクソンがマウスのエクソンと保存（転写物保存）されているヒトのカセット型 AS かつゲノム保存エクソンは 228 個であった。そのうち、エクソンインクルージョンを導く TGCATG（Fox 結合サイト）が、このエクソンの下流 300 bp 以内のイントロン上で見いだされたものは 10 個であった。このエクソンが、マウスゲノムに見出されず（マッピングされたマウスの完全長 cDNA を含む全ての mRNA・RefSeq・Ensembl 転写物、全ての EST、およびマウスの脳で発現した RNA-Seq タグとゲノム上の対応する場所で共有しなかった）、ヒトの Fox 結合サイトと対応するマウスゲノムの配列に変異が見つかったものは 4 個であった（図 2-1）。これらは、ヒト脳特異的に Fox にスプライシング制御されると考えられる候補である。そのうちの 1 つのゲノムアラインメント配列を図 2-6 に示す。このエクソンを含むヒト遺伝子にアノテーションされた、糖輸送に関する GO term（phosphoenolpyruvate-dependent sugar phosphotransferase system（ホスホエノールピルビン酸塩依存糖ホスホトランスフェラーゼシステム）（GO:0009401）、sugar:hydrogen symporter activity（糖水素共輸送体活性）（GO:0005351））は、

全 AS 遺伝子に対して濃縮して観察されたため、ヒト独自の脳進化に AS によって制御される糖輸送が関わっているかもしれない。本解析は脳での解析に限定したが、Fox は心臓と骨格筋でも特異的に発現している。マウスによる Fox の解析結果と比較することによって、脳を含むこれらの組織におけるヒト独自のスプライシングのメカニズムを明らかにできるかもしれない。

図 2-6 ヒト脳特異的 Fox 制御スプライシング候補のゲノムアラインメント配列。上がヒト、下がマウスのゲノム配列を示す。下線部の配列が、ヒトのみに存在する AS エクソンを示す。青字はヒトのスプライスサイト配列、赤字はヒトの Fox 結合サイトを表す。

2.4 考察

2.4.1 ヒトとマウスの選択的スプライシングバリアントの進化的保存度とタンパク機能

本研究では、ヒトとマウスの完全長 cDNA を用いた AS 配列の進化的起源の解析を行った。ヒトとマウスのゲノムアラインメントを介した比較ゲノム解析により、同一スプライシングバリアント（ESV）は 4,624（ヒト RASV に対して 23%）、保存 AS 遺伝子は 189（ヒト AS 遺伝子に対して 3%）と、種間での保存度が低いことが明らかとなった。進化を通じて、高等真核生物のゲノムはその長さを増し、紫外線のような環境要因やトランスポゾンのような内部要因によって生じる変異に晒される頻度を増してきたと思われる。その結果、哺乳類や他の高等真核生物は、数多くの新しい AS バリアントを蓄積してきたが、細胞の恒常性の維持などに必要な一部のコアとなる AS 配列を含むバリアントを除き、多くは種特異的に生じてきたのかもしれない。本研究では、進化的保存度とタンパク機能アノテーションの情報を統合することによって、非保存の AS エクソンが CDS に影響を及ぼす頻度が低いことを明らかにした（表 2-7）。種分化後に AS によって容易にエクソンが作られるが、それらが CDS 上に存在した場合は、子孫に残りにくい異常なタンパクとして生成される可能性が高いことを反映しているのかもしれない。

また、ヒトの AS バリアントが大きな割合で脳や精巣で生成されることも興味深い(85)。これらの組織は、数多くのニューロンや精子細胞で構成され、1 つの細胞が独自の機能を持ちながらも集団として相互補完することによって機能的に働いていると考えられる。高度に特殊化された細胞による多様性自体がこれらの組織で重要なため、AS などによって引き起こされる個々のタンパクの偶発的な異常に対して比較的耐性があるのかもしれない。AS によって生みだされるこれらの組織の多様性が進化の原動力として働き、種分化のプロセスを加速しているとすれば興味深い。

第三章　ヒト選択的スプライシングの解析データを公開するための、データベース（H-DBAS）の開発

3.1 緒言

3.1.1 ヒト AS のデータベース

選択的スプライシング（AS）のデータベースとしては、ASD(86)や ASAP(87)などがあるが（後にそれぞれ ASTD(88)、ASAP II(89)にアップデートされた）、これらは主に EST や転写産物モデルを用いて解析されているため、転写物の完全性が重要となる AS とタンパク機能の関係についての情報はほとんどない。筆者は、ヒト完全長 cDNA 配列を用いて解析した結果をデータベース化することにより、これまでゲノムワイドでは明らかにされていなかったタンパク機能アノテーションに影響を与える AS バリアントの情報を公開した。このデータベースを H-DBAS（Human-transcriptome DataBase for Alternative Splicing、URL：http://www.h-invitational.jp/h-dbas/）と名付け、2006 年にバージョン 1 を初めて公開した(90)。これは、H-InvDB(25)を中心として様々な目的に応じた生物学データベースを開発するという、H-Invitational(12)の活動の一環として行った。その後、ヒトとマウスにおける AS バリアントの種間比較などの新しいアノテーション情報の追加や、転写物・ゲノムなどのアップデートを適宜行い、2010 年にはバージョン 6 まで更新している。H-DBAS を用いることにより、ヒトとマウスで保存された AS 遺伝子や、ヒト特異的 AS 遺伝子における AS バリアント間のタンパク機能アノテーションの違いなどを詳細に調べることができる。

H-DBAS のシステムは、Tomcat 上で動作する Web アプリケーションである。サーバーサイドで動作する Servlet は、H-DBAS バージョン 1 からデータの追加以外は基本的に変わっていない。現行のビューワーは、H-DBAS バージョン 5 から 2 つの Flash アプリケーション（AS Viewer とアノテーションコントローラー）を使用している。H-DBAS バージョン 4 までのビューワーは Java アプレットを使用していたが、1 つの Web ページ内で表示されない、クライアントの PC へ Java をインストールしなければならない、Java アプレットの起動が遅いなどの理由により、Flash への切り替えを行った。H-DBAS のトップページから RNA-Seq 解析のデータベースへアクセスすることも可能であるが、これについては第四章で詳しく説明する。

3.2 方法

3.2.1 データベースシステム

データベースには、フリーのリレーショナルデータベースマネージメントシステム（RDBMS）である MySQL バージョン 5.0.19 を用いた。ヒトでは、ヒトでしか解析していないデータ（タンパク機能アノテーション（”1.2.4 RASV のタンパク機能アノテーション解析”を参照）・ESE/ESS・Fox/Nova 結合サイト・レトロトランスポゾン・SNP(91)）を含む 27 のテーブルを、マウスについては 13 のテーブルを作成した。MySQL のテーブルデータと Java（Java SDK バージョン 5.0）オブジェクトとの対応付け（O/R マッピング）には、Hibernate バージョン 3.0 を使用した。O/R マッピングツールのフレームワークとして Hibernate を選んだ理由は、他の O/R マッピングツールに比べて柔軟性が高く多機能であること、そしてオブジェクトクエリ言語の Hibernate Query Language（HQL）を使えることである。RDBMS で使われる Structured Query Language（SQL）には一部独自拡張による差異（方言）があるが、HQL を使うことにより共通の言語で記述することができる。また、ヒトとマウスのゲノム配列と、ヒトーマウスのゲノムアラインメント配列は、ファイル I/O でデータを呼び出す方式を採用している。

3.2.2 サーバーアプリケーション

Web コンテナとして Apache Tomcat バージョン 5.5 を使用した。Java Servlet を主な処理に使用し、JavaServer Pages（JSP）の表示と画面遷移は Struts バージョン 1.2 で制御している。フレームワークとして Struts を用いているため、開発モデルである Model-View-Controller（MVC）に準拠しており、作業分担・画面デザインの変更・国際化対応・柔軟な画面遷移などを容易に行うことができる。Struts による画面遷移には原則 5 つのファイル（Form クラス・Action クラス・入力用 JSP・処理成功時 JSP・エラー検出時 JSP）が必要で、それらは WEB-INF/struts-config.xml ファイルで管理される。表 3-1 に H-DBAS の各機能と Struts の構成を示す。ユーザビリティーの向上（検索項目の動的な変化等）のために所々 Asynchronous JavaScript + XML（Ajax）を使用し、その際の通信は JavaScript Object Notation（JSON）を使用している。

表 3-1　H-DBAS の各機能と Struts の構成

	Form クラス [a]	Action クラス [b]	処理成功時 JSP	エラー検出時 JSP

簡易検索	SearchAction Form	SimpleSearch Action	asLocus.jsp	fetchLocusError.jsp
詳細検索	SearchAction Form	SearchAction	asLocus.jsp	fetchLocusError.jsp
検索結果（前/次）	SearchAction Form	ResultPageAction[c]	asLocus.jsp	fetchLocusError.jsp
検索結果（ダウンロード）	なし	DownloadHixListAction	なし[d]	downloadError.jsp
AS ローカス構造図	なし	LocusOverviewAction	ASOverview.jsp	fetchLocusError.jsp
AS ローカス構造図（ダウンロード）	DownloadForm	DownloadHixAction	なし[d]	downloadError.jsp
ゲノムアラインメント	ShowAlignmentForm	ShowAlignmentAction	alignmentInformation.jsp	fetchLocusError.jsp

[a] 入力を保存する Java クラス。

[b] 入力を実行する（データベース検索など）Java クラス。処理内容は、通常 execute メソッド内に記載される。

[c] execute 以外のメソッドを使用している。前は backPage()、次は nextPage()。

[d] 成功時は画面の表示ではなく、データがダウンロードされる。

3.2.3 Flash アプリケーション

AS Viewer と、AS Viewer 内で動作するアノテーションコントローラーがある。共に、フレームワークに Adobe Flex Framework バージョン 3.5 を使用した。AS Viewer とアノテーションコントローラー間の通信は、ローカルコネクションを使用している。ローカルコネクションは同時に開いている SWF（Flash ファイル）同士が連携し合う Flash 独自の通信方法である。AS Viewer とサーバーとの通信は、Actionscript Message Format（AMF）通信を使用してる。AMF により、RemoteObject（HTTPService を拡張した HTTP/HTTPS 通信用クラス）を使用してサーバーとクライアント間を容易に通信させることができる。また、AS Viewer は、Flash 起動時に必要なデータを全てロードする仕組みにしている。

3.2.4 ヒトとマウスの間で対応するエクソンの表示

比較ゲノム解析によってヒトとマウスでの対応を確認したエクソンは、AS Viewer（”3.3.3 AS

Viewer”で記述）のエクソン中心（デフォルト）表示形式で種間での対応を分かりやすく表示するため、縦に一列に並べて表示させるようにした。この表示方法を実現するために、ヒトとマウスで対応したエクソンの配列同士を BLAST(92)の bl2seq でペアワイズアラインメントし、その結果からマウスのエクソンの位置をヒトのエクソンのゲノム位置に合わせた。AS Viewer の全体表示形式では、ゲノム位置を合わせていないのでヒトとマウスで対応するエクソンは一列に揃っていないが、それらを線でつなげることにより対応関係を明示するようにしている。

3.3 結果

3.3.1 H-DBAS の構築

H-DBAS のトップページの上部を図 3-1 に示す。右上には簡易検索・ヘルプ・言語切り替え（日本語と英語）がある。H-DBAS 内の全てのページでこれらの機能を使用することができる。その下のタブからは、AS の機構・データと解析方法・統計・ダウンロード・用語集・リンクのページを表示することができる。トップページ内の”検索と解析ページ”のカテゴリーには、H-DBAS の機能を集約した 5 つのリンクがあり、以下に短く記述する。

（1）AS Viewer への直接リンク：HIX（H-Invitational の遺伝子 ID）をクリックすることにより、AS Viewer の組み込まれた画面（AS ローカス構造図）を直接参照することができる。AS Viewer については、”3.3.3 AS Viewer”で説明する。キーワードをクリックすると、検索結果リスト画面が表示される。

（2）詳細検索へのリンク：詳細検索については、” 3.3.2 検索システム（簡易・詳細・BLAST）”で説明する。

（3）BLAST 検索へのリンク：BLAST 検索については、” 3.3.2 検索システム（簡易・詳細・BLAST）”で説明する。

（4）比較ゲノム解析ページへのリンク：ヒトとマウスで保存された AS 遺伝子、およびヒト特異的な AS 遺伝子のリストを表示する。

（5）RNA-Seq 解析ページへのリンク：RNA-Seq 解析およびそのデータベースについては、第四章で説明する。

図 3-1　H-DBAS のトップページの上部。各番号の内容については本文を参照。

3.3.2　検索システム（簡易・詳細・BLAST）

H-DBAS は、3 種類の検索システムを有する。1 つ目は前述した簡易検索で、常に画面の右上に位置している。生物種の選択とキーワードの入力のみという簡単な構造である。2 つ目は詳細検索で、6 カテゴリー計 22 の検索項目から AS 遺伝子の絞り込みを行うことができる（図 3-2）。それぞれのカテゴリーの内容の詳細を以下に記述する。

（1）基本情報：簡易検索とほぼ同じであるが、キーワードだけでなく、各種 ID（HIX（H-Invitational の遺伝子 ID）・HIT（H-Invitational の転写物 ID）・アクセッションナンバー・RefSeq ID・Ensembl ID・HUGO 遺伝子シンボル・Entrez Gene ID・OMIM ID・EC ナンバー）ごとに入力できる。

（2）ゲノム構造：染色体番号、ストランド、ゲノム位置などのゲノムの基本情報の他に、スプライスサイト配列（GT-AG・GC-AG・AT-AC）、NAGNAG（エクソン－イントロン境界の 3’末端（AG）側でスプライスサイト配列がタンデムに現れるモチーフ）（93）、レトロトランスポゾン（LTR・LINE・SINE・*Alu* のみの SINE）の選択ができる。

（3）AS ゲノム構造：第一章の方法で同定された RASV の数と、AS の位置（5’末端・内部・3’末端）、AS 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）に加え、転写開始または終結に関係する選択的第一エクソンと選択的最終エクソンの AS パターンを選択できる。

(4)タンパク機能アノテーション:RASV の ORF 長、タンパク機能モチーフの ID(InterPro ID)と名前、GO の ID と term、細胞内局在化シグナル(WoLF PSORT および TargetP での予測結果)、膜タンパクドメイン(TMHMM および SOSUI での予測結果)のタンパクに関する情報を選択できる。また、ナンセンス変異介在的 mRNA 分解(NMD)を導く中途終止コドン(PTC)を含む ORF も選択できる。SR タンパクが NMD 結合 AS によって制御されているという報告があり(94)、AS のメカニズム解明に重要なものもあると考えられるからである。

(5)AS タンパク機能アノテーション:第一章の方法で解析された RASV のアミノ酸長の差、UTR または CDS に存在する AS、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、複雑な AS パターン(ブリッジ型・ネスト型・マルチプル CDS 型)について選択できる。

(6)AS 比較ゲノム解析:第二章の方法で解析された、ヒトーマウスのゲノムアラインメントを介したヒト RASV とマウス転写物との保存度(非保存・ゲノム保存・転写物保存・同一スプライシングバリアント(ESV)・保存 AS)を選択できる。

3 つ目は BLAST 検索で、FASTA 形式の塩基またはアミノ酸配列をコピーアンドペーストで直接、あるいはファイルで入力することにより、ヒト完全長 cDNA データセットから同定された RASV の中から、クエリー配列と配列相同性の高いものを検索することができる。

図 3-2　詳細検索画面。6 カテゴリー計 22 の検索項目があり、自由に組み合わせて AS 遺伝子の絞り込み検索を行うことができる。

3.3.3 AS Viewer

AS Viewer は、ユーザーが AS 解析の結果をインタラクティブに操作して閲覧できるシステムである（図 3-3）。主な操作可能機能は 4 つあり、詳細を以下に記述する。なお、AS Viewer は AS ローカス構造図画面の一機能である。AS Viewer に表示された遺伝子やそこに含まれる転写物の様々な情報（由来細胞・発現組織・発生段階・性別など）を、AS ローカス構造図画面から得ることができる。

（1）メインビュー：RASV とゲノムが表示される。以降で説明する（2）～（4）で操作した結果は、全てこの場所に反映される。ゲノムには常にスプライスサイト配列と SNP が表示されている。スプライスサイト配列と SNP を関連付けた解析は、ヒト集団を理解する上で重要だと考えているからである(95)。また、AS グループの中から代表（RASV）に選ばれなかった残り全ての AS バリアント（Other ASV）や、参考転写物配列として同じ遺伝子にマッピングされた RefSeq と Ensembl を表示することもできる。

（2）スケールビュー：RASV の表示形式の選択と拡大縮小を行う。デフォルトでは RASV の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が、全てのイントロンをマッピングデータ通りに表示することもできる。また、RASV の塩基やアミノ酸配列を表示する場合もこのビューの機能を使う。

（3）アノテーションコントローラーボタン：第一章で解析した結果を表示するために用いる機能である。クリックするとページ内で移動可能な Flash の画面として現れる。ヒト RASV 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）を表示するだけでなく、これらタンパク機能アノテーションに影響を与える RASV ペアを調べることもできる。また、複雑な AS パターンを示す RASV や、レトロトランスポゾン、ESE/ESS、Fox/Nova 結合サイトを表示できる。図 3-4 は、アノテーションコントローラーを表示し、タンパク機能アノテーションに影響を与える RASV ペアを選択した状態である。

（4）種間比較ボタン：第二章で解析した結果を表示するために用いる機能である。比較ゲノム解析の結果より、ヒトの AS 遺伝子に対応するマウスの遺伝子が存在していれば、プルダウンメニューにそれがリストアップされる。マウスの遺伝子を選択すると、その遺伝子がマウスゲノムとともにメインビューの下に現れる。ヒトとマウスのオルソログ遺伝子間に同一スプライシングバリアント（ESV）が存在する場合、同じ色と数字（ESV 1、ESV 2 など）で対応付けて表示される。右隣のチェックボタンをチェックすることにより、ヒト RASV のエクソンの保存度（非保存・

ゲノム保存・転写物保存）が色分けされて表示される。図 3-4 は、種間比較ボタンからマウスの遺伝子を選択し、さらにエクソンの保存度チェックボタンをチェックした状態である。

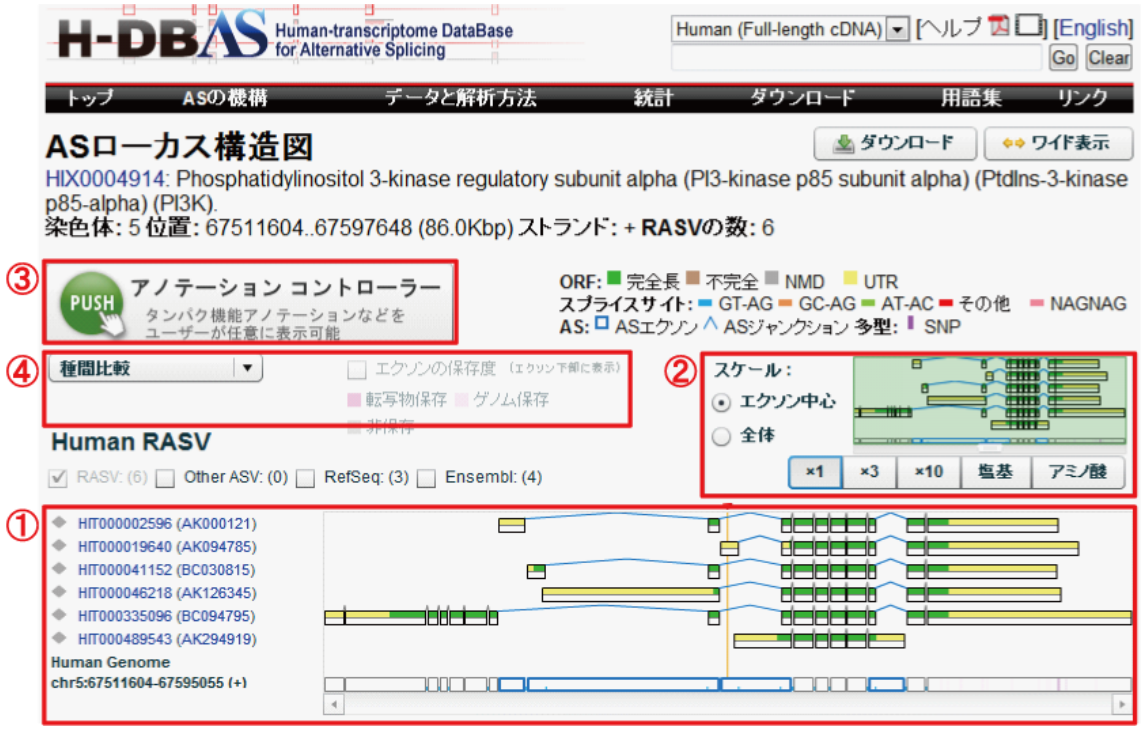


図 3-3　AS Viewer のデフォルト表示例（サンプルとして、図 2-3A の phosphoinositide-3-kinase regulatory subunit（ホスホイノシチド-3-キナーゼ（PI3 キナーゼ）調節サブユニット）を使用）。（1）メインビュー。RASV とゲノムの情報が表示される。（2）スケールビュー。RASV の表示形式の切替え、塩基・アミノ酸配列の表示を行う。（3）アノテーションコントローラーボタン。ヒト RASV のタンパク機能アノテーション解析結果の表示に用いる。（4）種間比較ボタン。ヒト RASV とマウス転写物との比較ゲノム解析結果の表示に用いる。

図 3-4　AS Viewer のアノテーションコントローラーと種間比較ボタンを使用した状態。メインビューで、タンパク機能モチーフに影響を与える RASV ペア（背景をハイライト表示）や、ESV（ヒトとマウスの転写物 ID の背景を同色でハイライトし、同色の線でつなげて表示）を同時に観察することができる。ヒト RASV エクソンの保存度も色分けして表示されている。なお、エクソン中心（デフォルト）表示形式では、ヒトとマウスで対応しているエクソンは縦に一列に並べて表示されている。

3.4 考察

3.4.1 H-DBAS の独自性

H-DBAS に登録されているヒトの AS バリアントは、完全長 cDNA から同定されており、ゲノムにマッピングされた転写物配列上のエクソンの数や順番が明確である。それゆえ、転写物上の AS イベントの位置や、アミノ酸配列の位置依存タンパク機能（タンパク機能モチーフや細胞内局在化シグナル）を正しくアノテーションすることができる。このような精度の高いデータを用いた AS バリアントの解析結果を、充実した検索システムとインタラクティブに操作できて直感的に分かりやすいビューワーを開発することにより、ユーザーに有益な形で提供している。

公開している情報は、第一章の AS のタンパク機能アノテーション解析と、第二章の AS の比較ゲノム解析の結果である。AS バリアント単位、そして、同じ遺伝子構造を持つ AS バリアントグループの代表である RASV 単位でこれらの情報を公開しているデータベースは、世界的にみてもユニークである。この利点を活かし、さらに様々なアノテーションを RASV に追加することによって、より利用価値の高いデータベースへ発展させることができると考えている。

第四章　RNA-Seq タグを用いた、ヒト選択的スプライシングの翻訳検証

4.1 緒言

4.1.1 次世代シーケンサー（イルミナ GA）による RNA-Seq 解析

第一章および第二章の解析結果により、完全長 cDNA から同定されたヒトの選択的スプライシング（AS）バリアントに、タンパク機能アノテーションに影響を及ぼすものが数多く見出された。一方で、マウスとの進化的保存度は相対的に非常に低いことが明らかとなった。進化的に保存していないヒトの AS バリアントは、タンパク機能モチーフを用いた解析により、一部は精巣に濃縮して機能していることが確認されたが、タンパク機能モチーフを持たないものも多かった。生物学的な意味はなく、単にゲノムに生じる内因性の転写のノイズとして現れたという可能性も依然として否定できない。ヒト AS バリアントの更なるアノテーションとして、配列情報による解析だけではなく、それらが翻訳されている証拠を実験的に確認する必要があると考えられた。

実験的検証には、サンガー法を利用した従来のシークエンサーではなく、低コスト・高スループットの次世代シークエンサーを用いた。次世代シークエンサーは、イルミナ・ABI・ロシュの各企業から製品化され、性能の向上に伴い、近年活発に用いられている。ロシュの製品を除き、数 10 から 100 bp 程度の短いタグが 1 ラン当たり 100 Gb 以上配列決定されるため、解析には高性能のコンピューターと特殊なアルゴリズムを用いたプログラムを必要とする（ELAND・SOAP2(96)・Bowtie(97)などのマッピングソフトウェアが存在する）。次世代シークエンサーを用いたトランスクリプトーム解析では、ポリ A を持つ RNA をショットガンで断片化して配列決定する RNA-Seq(67)が遺伝子発現解析として用いられており、本研究でもこの方法を用いた。RNA-Seq を用いた AS 解析の先行研究により、複数エクソンを有する転写物を含むヒトの遺伝子が AS 遺伝子である割合は、92-94%であることが推定されている(98)（EST の解析では 40-60%、マイクロアレイの解析では 74%と推定されていた(7,9)）。RNA-Seq のメリットは、EST と比べて分画できる、実験精度が一定、カバレッジが高いこと、マイクロアレイと比べてプローブによる制限がない、相対ではなく絶対発現量が分かる、発現量の小さい RNA も検出できることなどである。

本研究ではこの RNA-Seq により、ヒト細胞内の特定画分に存在する mRNA を解析し、既知遺伝子に対してポリソームに存在してタンパクへの翻訳が行われていると考えられる AS バリアント、および核内にのみ存在して翻訳に用いられていないと思われる AS バリアントの同定を行った(99)。

4.2 方法

4.2.1 ヒト DLD-1 細胞の細胞画分の分離と実験的検証

ヒトの大腸ガン細胞株である DLD-1 細胞は様々な実験で用いられており、解析データを豊富に有する細胞である。この DLD-1 細胞から、ショ糖密度勾配遠心法（SDG）によって細胞質・核・ポリソーム（タンパク合成中のリボソーム）の各画分を分離した。それぞれの細胞画分は、細胞質タンパクであるグリセロアルデヒド-3-リン酸デヒドロゲナーゼ（GAPDH）と核タンパクであるラミン A/C のウェスタンブロッティング、そして、核 RNA である sno/scaRNA のリアルタイム RT-PCR 解析によって確認した（図 4-1）。sno/scaRNA のプライマー配列を以下に示す。

scaRNA2：ACGCGTGAGTGTGTGAGTGT　　GCAGGAGGAGAGCTTTTCATT
scaRNA12：TGATGAGACTAAGGCGAATGC　GCACCAGAAATGAAGGCAAG
scaRNA10：AATCTTGGTGGGCGATACAG　　CCCTGATACCCTGAACATGC
scaRNA13：GTAGTCTTGGAGCCGCACAG　　GTGGCAACAGTGACCAGAAA
snoRNA73：CTCTGTCCAAGTGGCGTAGG　　GACAGGACTCTGGGAAGCTG

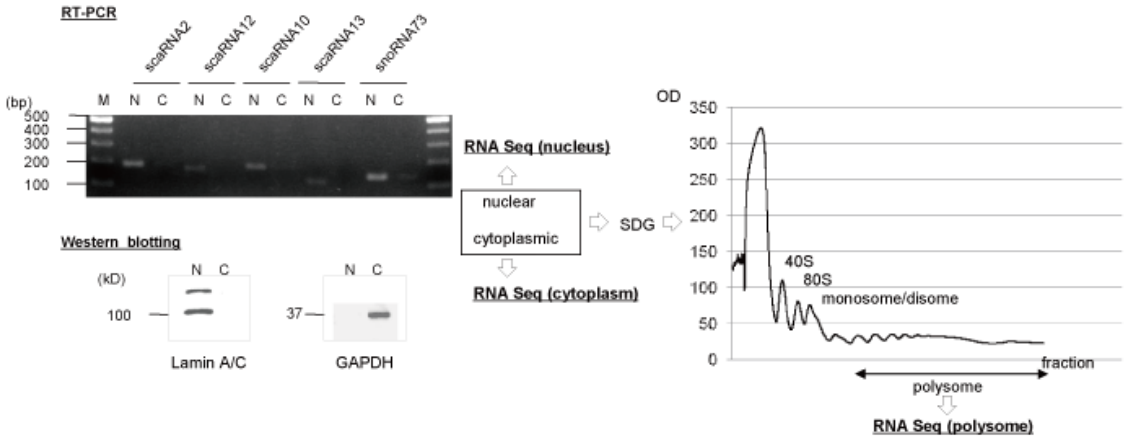


図 4-1　RT-PCR とウェスタンブロッティングによる分離した細胞画分の確認。sno/scaRNA の RT-PCR とラミン A/C のウェスタンブロッティングにより核（N）であること、GAPDH のウェスタンブロッティングにより細胞質（C）であること、SDG 後の吸光度（OD）からポリソームであることを確認した。

4.2.2 RNA-Seq タグの生成

ヒトDLD-1 細胞から分離した細胞質・核・ポリソームの各細胞画分より、ポリAを持つ mRNA を抽出した。抽出の際 DNase 処理を行い、ゲノム DNA を取り除いた。次にこれらの cDNA を作成し、ショットガンで断片化した後、Illumina Genome Analyzer（GA）でシークエンシングを行った。最終的に、細胞質、核、ポリソームで、それぞれ 46,354,139、47,120,831、54,901,628 のシングルエンド 36 bp RNA-Seq タグを生成した。

4.2.3 RNA-Seq タグのゲノムマッピングとスプライスジャンクションの検出

RNA-Seq タグは、ヒトゲノム（UCSC hg18）(22)に対し、ショートリード用のアラインメントプログラムである ELAND および Bowtie バージョン 0.10.0.2（デフォルトパラメーター。ミスマッチは 2 bp まで許容）を用いてマッピングした。スプライスジャンクションは、ショートリード用のスプライスジャンクション検出プログラムである TopHat(100)バージョン 1.0.9（デフォルトパラメーター。スプライスサイト配列は GT-AG のみサポート）を用いて検出した。全ての検出されたスプライスジャンクションに対し、RNA-Seq タグが 2 つ以上マッピングされたものを、確度の高いスプライスジャンクションとして以降の解析に用いた。

4.2.4 翻訳する、または翻訳しない AS バリアントの同定

ヒトの既知転写物として、RefSeq(65)リリース 23 転写物のマッピングデータを用いた。さらに、”1.2.2 選択的スプライシング（AS）と AS パターンの判定”と同じ方法で AS 判定を行い、これをヒトの既知 AS バリアントとした。翻訳する AS バリアントを同定するために、RefSeq 転写物のエクソンとポリソーム画分由来の RNA-Seq タグのゲノム位置情報を比較した。翻訳しない AS バリアントの同定は、TopHat によって検出した細胞質・核・ポリソームの各細胞画分由来のスプライスジャンクションのうち、核画分特異的スプライスジャンクションを用いて、そのゲノム位置情報を RefSeq 転写物のスプライスジャンクションのものと比較することによって行った。比較には Perl バージョン 5.8.8 で作成した独自のプログラムを使用した。RNA-Seq タグの数は膨大であり、そのゲノム上の配列位置情報を総当りで調べるには時間がかかりすぎる。そのため、以下の方法を行った。まず、RNA-Seq タグとエクソンの位置情報データファイルを連結し、ゲノム上の開始位置を昇順で並べ替えた。次に、このファイルの先頭から順に RNA-Seq タグとエクソンのゲノム位置情報を比較した。これにより、無駄な比較を省いて処理の高速化を実現した。

4.3 結果

4.3.1 細胞画分ごとの RNA-Seq タグの統計

RNA-Seq 解析の統計を表 4-1 に示す。計 148,376,598 本の RNA-Seq タグを用い、7,396,342 本のスプライスジャンクションを検出した。また、ポリソーム由来の RNA-Seq タグがマッピングされた既知（RefSeq）エクソンは 90,900 個存在していた。

表 4-1 RNA-Seq タグによる解析の統計

	RNA-Seq タグ	ゲノムにマッピングされた RNA-Seq タグ	RNA-Seq タグによって検出されたスプライスジャンクション	RNA-Seq タグがマッピングされた既知（RefSeq）エクソン
細胞質	46,354,139	28,906,833	1,312,273	81,547
核	47,120,831	28,939,028	1,952,803	85,923
ポリソーム	54,901,628	29,720,537	4,131,266	**90,900**
計	148,376,598	87,566,398	7,396,342	258,370

4.3.2 AS バリアントへの翻訳情報のアノテーション

表 4-2 に、RefSeq の遺伝子および転写物と、その中で 90,900 個のポリソーム由来の RNA-Seq タグのいずれかがマッピングされたエクソンを有するものの数を示す。10,923 の AS バリアントのうち、8,440（77%）でタンパクに翻訳している可能性が高いことを示した。

RefSeq 転写物のスプライスジャンクションと、それらと同じゲノムに位置する細胞質・核・ポリソームの各細胞画分由来のスプライスジャンクションの数を表 4-3 に示す。254 本の AS ジャンクションが核特異的に検出され、このスプライスジャンクションを持つ転写物は核内に留まりタンパク合成をしないと考えられた。例として、Caspase 4, apoptosis-related cysteine peptidase（カスパーゼ 4 アポトーシス関連システインペプチダーゼ（CASP4））を図 4-1 に示す。この遺伝子には 2 つの AS バリアントがあり、最初のエクソンのゲノム位置が異なる選択的第一エクソン型の AS パターンを有している。AS バリアントの一方である NM_001225 は、最初のイントロンが AS ジャンクションであり、さらに核由来の RNA-Seq タグのみが 35 本マッピングされた核特異的スプライスジャンクションでもあった。カスパーゼは主にシグナル伝達によってアポトーシスを引き起こすが、近年多様な非アポトーシス機能が発見された(101)。カスパー

ゼファミリーの 1 つであるカスパーゼ 4 は、炎症にも関わるグループに分類され、古典的な LPS 刺激誘導 Toll 様受容体（TLR）4 シグナル伝達に重要な役割を演じ、NF-κ B の活性化と結果としての IL-8 と CCL4（共に前炎症性メディエーター）の上方調節と分泌を導く(102)。本研究で用いたヒト DLD-1 細胞では、カスパーゼ 4 の 1 つの AS バリアントは mRNA に転写されてスプライスを受けた後、核内に留まっていた。

表 4-2　RefSeq の遺伝子および転写物と、その中でタンパクに翻訳するものの統計

	遺伝子	転写物	ポリソーム由来の RNA-Seq タグがマッピングされたエクソンを有する転写物を含む遺伝子	ポリソーム由来の RNA-Seq タグがマッピングされたエクソンを有する転写物
全	19,181	26,814	13,870（72%）	19,226（72%）
AS	4,090	10,923	**3,388（83%）**	**8,440（77%）**

表 4-3　RefSeq のスプライスジャンクションと、その中で各細胞画分から検出されたスプライスジャンクションとゲノム位置が同じであったものの統計

	SJC[a]	細胞質由来 SJC[a]	核由来 SJC[a]	ポリソーム由来 SJC[a]	細胞質特異的 SJC[a]	核特異的 SJC[a]	ポリソーム特異的 SJC[a]
全	186,137	47,615	47,260	51,041	6,649	6,980	9,799
AS	14,491	1,067	1,021	1,114	260	**254**[b]	299

[a]SJC：スプライスジャンクション。

[b]AS ジャンクションかつ核特異的に検出されたスプライスジャンクションを持つ AS バリアントは、タンパクに翻訳されないと考えられる。

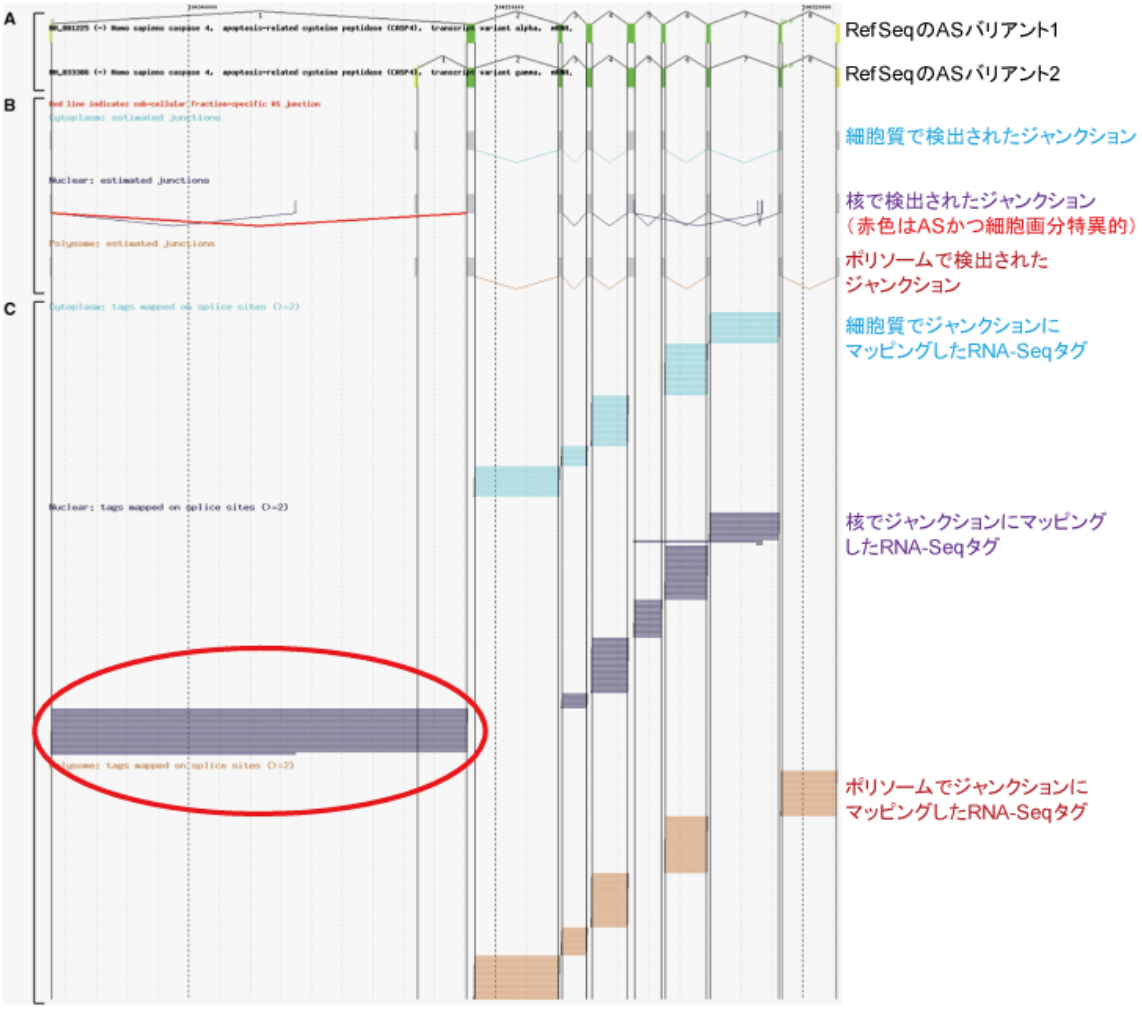


図 4-1　核特異的 AS ジャンクションを有する AS バリアントを含む遺伝子の例（Caspase 4, apoptosis-related cysteine peptidase（カスパーゼ 4 アポトーシス関連システインペプチダーゼ（CASP4））。（A）RefSeq の AS バリアントを示す。緑色が CDS、黄色が UTR を表す。数字はイントロンの番号を示す。（B）マッピングされた 2 つ以上の RNA-Seq タグを基に検出されたスプライスジャンクションを示す。細胞質、核、ポリソーム由来のスプライスジャンクションは、それぞれ水色、紺色、茶色の線で表されている。赤色の太線は、AS ジャンクションかつ細胞画分（ここでは核）特異的スプライスジャンクションを表す。灰色の四角は、（A）に示した全 AS バリアントのエクソンのゲノム位置をまとめたもの（Assembled exon）を表す。（C）実際にスプライスジャンクションにマッピングした RNA-Seq タグを示す。赤丸で囲まれた部分は、AS かつ核特異的ジャンクションにマッピングした 35 本の RNA-Seq タグを表す。色の内訳は（B）と同じ。

4.3.3　RNA-Seq 解析のデータベースとビューワー

RNA-Seq 解析の結果は、データベースに格納して検索・表示を可能にした（URL：

http://www.h-invitational.jp/rnaseq4hdbas/)。データベースには、フリーのリレーショナルデータベースマネージメントシステム(RDBMS)である PostgreSQL バージョン 7.3.10 を用い、データの表示には Perl バージョン 5.8.8 で作成した独自の CGI プログラムを使用した。検索項目は 3 種類あり、それぞれ以下に記述する。

(1)HIX(H-Invitational の遺伝子 ID)・RefSeq 転写物・HUGO 遺伝子シンボルの各 ID 検索と、タンパク名のキーワード検索。

(2)AS ジャンクションの AS 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・選択的第一エクソン・選択的最終エクソン)の選択。

(3)RNA-Seq タグのマッピング情報から検出されたスプライスジャンクションの由来細胞画分(細胞質・核・ポリソーム)の選択。

解析結果は、図 4-1 のビューワー、または図 4-1 の RNA-Seq タグのマッピング情報をまとめた図 4-2 のビューワーとして表示される。図 4-2 では、スプライスジャンクションにマッピングされた RNA-Seq タグを数字とバーで表している。RNA-Seq タグを詳細に調べたい場合は、本解析データを登録した次世代シークエンサーのショートリード表示用ゲノムブラウザーである GBrowse2(69)から、RNA-Seq タグの配列などを閲覧することができる(URL: http://h-invitational.jp/gb2/gbrowse/rnaseq/)。

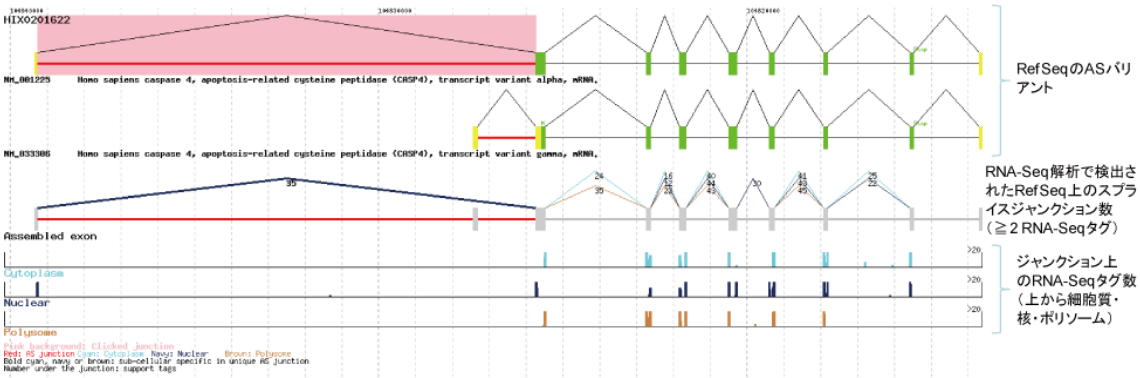


図 4-2　図 4-1 の RNA-Seq タグのマッピング情報をまとめたビューワー。上段には RefSeq の AS バリアントが表示される。赤色の太線は AS ジャンクションであることを示す。中段の灰色の四角は、上段に示した全 AS バリアントのエクソンのゲノム位置をまとめたもの(Assembled exon)を示す。細胞画分ごとに検出されたスプライスジャンクションが表示され、その下にそこにマッピングされた RNA-Seq タグの数が示される。太線は、細胞画分特異的であることを示す。下段には、細胞画分ごとにスプライスジャンクションにマッピングされた RNA-Seq タグの

数がバーの高さで表示される。CDS・UTR の色、細胞画分の色の内訳は、図 4-1 と同じ。データベースの検索結果から選択したスプライスジャンクション（AS ジャンクションかつ各細胞画分のいずれかから検出されたスプライスジャンクション）の背景が、桃色でハイライト表示される。

4.4 考察

4.4.1 RNA-Seq 解析による AS ジャンクションの翻訳検証

次世代シークエンサーであるイルミナ GA を使い、ヒト DLD-1 細胞から分離した細胞質・核・ポリソームの各画分から抽出した mRNA 断片のシークエンシング（RNA-Seq）を行った。細胞質、核、ポリソームから、それぞれ 46,354,139、47,120,831、54,901,628 のシングルエンド 36 bp RNA-Seq タグを生成した。10,923 の RefSeq AS バリアントのうち、ポリソーム由来の RNA-Seq タグがマッピングしたのは 8,440（77%）であった。ポリソームはリボソームが活発にタンパク合成を行っている場所であるため、これらはタンパクに翻訳している AS バリアントだと考えられた。また、タンパクに翻訳してない AS バリアントを同定するために、各細胞画分の RNA-Seq タグをゲノムにマッピングしてスプライスジャンクションを検出した。RefSeq 転写物とのゲノム上の位置による比較から、核特異的かつ AS ジャンクションである 254 本のスプライスジャンクションを同定した。核にはリボソームが存在せずタンパク合成が行われないため、このスプライスジャンクションを有する AS バリアントはタンパクに翻訳されないと考えられる。この AS バリアントを含む遺伝子の例として、カスパーゼ 4 が存在した。本研究で用いた細胞では、カスパーゼ 4 の遺伝子機能において、AS による調節が行われていないかもしれない。あるいは、核特異的に検出されたこのジャンクションは RefSeq 転写物のものではなく、ノンコーディングの可能性を含む未知の転写物のものかもしれない。

本研究では 1 つの細胞のみを用いており、ヒトの AS バリアントの頻度を調べるにはまだカバレッジが足りないと考えている。同じような翻訳検証を複数の細胞で行い、AS バリアントごとにタンパクへの翻訳や頻度情報のアノテーションを付加することによって、それらがヒトの生体内でどのような形で機能しているのか、あるいは生物学的な意味があるのか（内因性の転写のノイズではないのか）についての判断に利用できると考えている。また、ヒトとマウスで同じ機能を有する細胞に対して翻訳検証を行うことにより、種間における細胞内 AS ネットワークの違いを明らかにできるかもしれない。

第五章　総括

本研究の第一章では、完全長 cDNA を用いたヒト選択的スプライシング（AS）バリアント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 AS バリアントの解析を行った。完全長 cDNA を用いた理由は、5'末端が転写開始点であるため、エクソンの数や順番の情報が明確であり、AS イベントの位置やタンパク機能などを正しくアノテーションできるからである。アノテーションに関しては、機械的な自動処理だけでなくマニュアルでも行った。これにより、より精度の高い解析を行うことができた。計 56,419 本の H-Inivitational 2 のヒト完全長 cDNA のうち、18,297 を代表 AS バリアント（RASV）として同定した。以前から報告されていた通り、カセット型 AS パターンが最も多かった。RASV が含まれている 6,877 の AS 遺伝子のうち、4,481（65%）で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）に変化が認められた。新規に同定したタンパク機能モチーフの欠失した RASV は、シグナル伝達に際しモジュレーターとして働くことを示唆した。AS 遺伝子で高頻度に観察される機能は、シグナル伝達や転写制御であることも明らかになった。マニュアルアノテーションにより、典型的な AS パターンに該当しないが、タンパクの多様性に貢献していると考えられる複雑な AS パターンの現象を確認し、ブリッジ型・ネスト型・マルチプル CDS 型と定義して、計 316 遺伝子（5%）同定した。これらは RT-PCR による実験的検証などから、実際に存在することを示した。これらの結果は、完全長 cDNA を用いてゲノムワイドに解析された AS バリアントの初めての例であり、ヒトの細胞内遺伝子機能の多様性を理解するための基盤となる情報である。

第二章では、第一章で行われたヒト AS の解析を拡張し、完全長 cDNA を用いたヒトとマウスにおける AS バリアントの比較ゲノム解析を行った。ヒトとマウスで保存された同一スプライシングバリアント（ESV）は 4,624（23%）、保存 AS 遺伝子は 189（3%）と、種間での保存度は低かった。保存 AS 遺伝子ではカセット型 AS パターンの比率が大きくなり、選択的保持イントロンの比率が小さくなった。選択的保持イントロンについては、cDNA クローニングによって生じたアーティファクトの可能性もあるため、この AS パターンを有すると判定した RASV に対し、全ての保存カテゴリーから CDS がアノテーションされたものをいくつか抽出して RT-PCR で確認した。保存 AS 遺伝子に特徴的に現れるタンパクの機能は、主に細胞の恒常性の維持に関わるものであった。これは、種に関わらず必要不可欠な細胞応答の機能（スイッチのオンオフによる切り替えなど）を保持するため、それをコードする AS 配列も進化を通じて保存されているのだと考えられる。RASV における非保存の AS エクソンは 5'末端で多い一方、CDS に少なく、タンパク機能モチーフにも少なかった。非保存の AS エクソンを持つ遺伝子に特徴的に現れるタンパク機能モチーフは、保存 AS 遺伝子のものと種類が異なり、生殖に関連するものが多かった。さらに、この遺伝子が発現する組織は精巣が多かった。精巣は AS が豊富に見出される組織として知られており、その AS が非保存（すなわちヒト、ひいては種特異的）であることが明らかになったことで、精巣のような生殖組織で生じる AS が種分化の原動力になっているという可能性を示唆した。脳においても数多くの AS が観察されており、ヒト脳特異的な AS を見つけることができれば、ヒトの脳進化における AS 制御の一端を明らかにできるかもしれない。

第三章では、第一章と第二章で行ったヒト AS の解析結果を公開するためのデータベースの構

築を行った。様々な項目からの絞り込み検索や、配列の相同性検索を行う BLAST 検索などの検索システム、およびユーザーがインタラクティブに操作できるビューワーの開発を行った。これらを用いることにより、ユーザーはヒト AS の解析結果を視覚的に観察することが可能となり、ヒト AS の特徴をより分かりやすく理解するのに役立っていると考えている。このデータベースは H-DBAS（Human-transcriptome DataBase for Alternative Splicing）と命名されており、http://www.h-invitational.jp/h-dbas/からフリーでアクセスすることができる。

第四章では、AS バリアントの翻訳情報を調べるため、次世代シークエンサーのイルミナ GA で配列決定した、ヒト DLD-1 細胞の細胞質・核・ポリソーム由来の mRNA 断片（RNA-Seq タグ）を用いた。各細胞画分の RNA-Seq タグをヒトゲノムにマッピングし、さらにスプライスジャンクションを検出した。ポリソーム由来の RNA-Seq タグがマッピングした既知（RefSeq）の AS バリアントは、8,440（77%）であった。これらは、翻訳する AS バリアントと考えられた。検出されたスプライスジャンクションについては RefSeq 転写物のゲノム位置情報と比較して、核特異的なスプライスジャンクションであり、かつ AS ジャンクションでもあったものを 254 本同定した。このジャンクションを有する AS バリアントは、翻訳しない AS バリアントだと考えられた。これらについては、遺伝子発現の転写後調節が行われているのかもしれない。あるいはこれらがタンパクコード遺伝子中に生じたノンコーディング RNA であれば興味深い。このような様々な細胞における mRNA の翻訳に関する情報を、第一章および第二章で解析した完全長 cDNA から同定した RASV のアノテーションに加えることにより、ヒト AS に関する情報がより充実され、さらに第三章で述べたデータベースやビューワーに取り入れることによって、ヒトの細胞内遺伝子機能の多様性を産み出す AS の役割に対する理解がより進むと考えている。

参考文献

1. Gilbert, W. (1978) Why genes in pieces? *Nature*, **271**, 501.
2. Lander, E.S., Linton, L.M., Birren, B.*, et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
3. Venter, J.C., Adams, M.D., Myers, E.W.*, et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-51.
4. Ewing, B., Green, P. (2000) Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet*, **25**, 232-4.
5. Hattori, D., Millard, S.S., Wojtowicz, W.M., Zipursky, S.L. (2008) Dscam-mediated cell recognition regulates neural circuit formation. *Annu Rev Cell Dev Biol*, **24**, 597-620.
6. Ladd, A.N., Cooper, T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol*, **3**, reviews0008.
7. Modrek, B., Lee, C. (2002) A genomic view of alternative splicing. *Nat Genet*, **30**, 13-9.
8. Larsson, T.P., Murray, C.G., Hill, T., Fredriksson, R., Schioth, H.B. (2005) Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. *FEBS Lett*, **579**, 690-8.
9. Johnson, J.M., Castle, J., Garrett-Engele, P.*, et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141-4.
10. Suzuki, Y., Sugano, S. (2003) Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. *Methods Mol Biol*, **221**, 73-91.
11. Carninci, P., Kvam, C., Kitamura, A.*, et al.* (1996) High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, **37**, 327-36.
12. Cyranoski, D. (2002) Geneticists lay foundations for human transcriptome database. *Nature*, **419**, 3-4.
13. Kikuno, R., Nagase, T., Waki, M., Ohara, O. (2002) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res*, **30**, 166-8.
14. Ota, T., Suzuki, Y., Nishikawa, T.*, et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet*, **36**, 40-5.
15. Wiemann, S., Weil, B., Wellenreuther, R.*, et al.* (2001) Toward a catalog of human

genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res*, **11**, 422-35.

16. Strausberg, R.L., Feingold, E.A., Klausner, R.D., Collins, F.S. (1999) The mammalian gene collection. *Science*, **286**, 455-7.
17. Hu, R.M., Han, Z.G., Song, H.D.*, et al.* (2000) Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. *Proc Natl Acad Sci U S A*, **97**, 9543-8.
18. Imanishi, T., Itoh, T., Suzuki, Y.*, et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*, **2**, e162.
19. Takeda, J., Suzuki, Y., Nakao, M.*, et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res*, **34**, 3917-28.
20. Ewing, B., Hillier, L., Wendl, M.C., Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, **8**, 175-85.
21. Mott, R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci*, **13**, 477-8.
22. Rhead, B., Karolchik, D., Kuhn, R.M.*, et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, **38**, D613-9.
23. The_UniProt_Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, **38**, D142-8.
24. Apweiler, R., Attwood, T.K., Bairoch, A.*, et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, **29**, 37-40.
25. Yamasaki, C., Murakami, K., Takeda, J.*, et al.* (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res*, **38**, D626-32.
26. Kimura, K., Wakamatsu, A., Suzuki, Y.*, et al.* (2006) Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res*, **16**, 55-65.
27. Wang, Z., Burge, C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *Rna*, **14**, 802-13.
28. Ashburner, M., Ball, C.A., Blake, J.A.*, et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-9.
29. Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J*

Mol Biol, **300**, 1005-16.

30. Nakao, M., Nakai, K. (2002) Improvement of PSORT II Protein Sorting Prediction for Mammalian Proteins. *Genome Informatics*, **13**, 441-442.
31. Hirokawa, T., Boon-Chieng, S., Mitaku, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378-9.
32. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-80.
33. Landry, J.R., Mager, D.L., Wilhelm, B.T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet*, **19**, 640-8.
34. Cordaux, R., Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat Rev Genet*, **10**, 691-703.
35. Lev-Maor, G., Sorek, R., Shomron, N., Ast, G. (2003) The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science*, **300**, 1288-91.
36. Hide, W.A., Babenko, V.N., van Heusden, P.A., Seoighe, C., Kelso, J.F. (2001) The contribution of exon-skipping events on chromosome 22 to protein coding diversity. *Genome Res*, **11**, 1848-53.
37. Castle, J.C., Zhang, C., Shah, J.K.*, et al.* (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nat Genet*, **40**, 1416-25.
38. Lejeune, F., Maquat, L.E. (2005) Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr Opin Cell Biol*, **17**, 309-15.
39. Karin, M. (1999) How NF-kappaB is activated: the role of the IkappaB kinase (IKK) complex. *Oncogene*, **18**, 6867-74.
40. Peters, R.T., Liao, S.M., Maniatis, T. (2000) IKKepsilon is part of a novel PMA-inducible IkappaB kinase complex. *Mol Cell*, **5**, 513-22.
41. Xing, Y., Resch, A., Lee, C. (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res*, **14**, 426-41.
42. Nakao, M., Barrero, R.A., Mukai, Y., Motono, C., Suwa, M., Nakai, K. (2005) Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res*, **33**, 2355-63.
43. Scharf, J.M., Endrizzi, M.G., Wetter, A.*, et al.* (1998) Identification of a candidate modifying gene for spinal muscular atrophy by comparative genomics. *Nat Genet*, **20**, 83-6.

44. Faber, P.W., Barnes, G.T., Srinidhi, J., Chen, J., Gusella, J.F., MacDonald, M.E. (1998) Huntingtin interacts with a family of WW domain proteins. *Hum Mol Genet*, **7**, 1463-74.

45. Wiemann, S., Kokocinski, A.K., Poustka, A. (2005) Alternative pre-mRNA processing regulates cell-type specific expression of the IL4l1 and NUP62 genes. *BMC Biology*, **3**, 16.

46. Oyama, M., Itagaki, C., Hata, H.*, et al.* (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res*, **14**, 2048-52.

47. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., Lander, E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*, **10**, 950-8.

48. Modrek, B., Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*, **34**, 177-80.

49. Pritsker, M., Doniger, T.T., Kramer, L.C., Westcot, S.E., Lemischka, I.R. (2005) Diversification of stem cell molecular repertoire by alternative splicing. *Proc Natl Acad Sci U S A*, **102**, 14290-5.

50. Takeda, J., Suzuki, Y., Sakate, R.*, et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res*, **36**, 6386-95.

51. Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., Sachidanandam, R. (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res*, **34**, 3955-67.

52. Wahl, M.C., Will, C.L., Luhrmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701-18.

53. Fairbrother, W.G., Yeo, G.W., Yeh, R.*, et al.* (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res*, **32**, W187-90.

54. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., Burge, C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831-45.

55. Kuroyanagi, H. (2009) Fox-1 family of RNA-binding proteins. *Cell Mol Life Sci*, **66**, 3895-907.

56. Licatalosi, D.D., Mele, A., Fak, J.J.*, et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464-9.

57. Kawahara, Y., Sakate, R., Matsuya, A.*, et al.* (2009) G-compass: a web-based

comparative genome browser between human and other vertebrate genomes. *Bioinformatics*, **25**, 3321-2.

58. Matsuya, A., Sakate, R., Kawahara, Y.*, et al.* (2008) Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res*, **36**, D787-92.
59. Schwartz, S., Kent, W.J., Smit, A.*, et al.* (2003) Human-mouse alignments with BLASTZ. *Genome Res*, **13**, 103-7.
60. Carninci, P., Kasukawa, T., Katayama, S.*, et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559-63.
61. Horton, P., Park, K.J., Obayashi, T.*, et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res*, **35**, W585-7.
62. Takeuchi, N., Ueda, T. (2003) Down-regulation of the mitochondrial translation system during terminal differentiation of HL-60 cells by 12-O-tetradecanoyl-1-phorbol-13-acetate: comparison with the cytoplasmic translation system. *J Biol Chem*, **278**, 45318-24.
63. Jin, Y., Suzuki, H., Maegawa, S.*, et al.* (2003) A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J*, **22**, 905-12.
64. Kaminuma, E., Mashima, J., Kodama, Y.*, et al.* (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res*, **38**, D33-8.
65. Pruitt, K.D., Tatusova, T., Klimke, W., Maglott, D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37**, D32-6.
66. Flicek, P., Aken, B.L., Ballester, B.*, et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res*, **38**, D557-62.
67. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-8.
68. Li, H., Handsaker, B., Wysoker, A.*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-9.
69. Stein, L.D., Mungall, C., Shu, S.*, et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*, **12**, 1599-610.
70. Inukai, K., Funaki, M., Ogihara, T.*, et al.* (1997) p85alpha gene generates three isoforms of regulatory subunit for phosphatidylinositol 3-kinase (PI 3-Kinase), p50alpha, p55alpha, and p85alpha, with different PI 3-kinase activity elevating responses to insulin. *J Biol Chem*, **272**, 7873-82.
71. Ueki, K., Algenstaedt, P., Mauvais-Jarvis, F., Kahn, C.R. (2000) Positive and

negative regulation of phosphoinositide 3-kinase-dependent signaling pathways by three different gene products of the p85alpha regulatory subunit. *Mol Cell Biol*, **20**, 8035-46.

72. Simader, H., Hothorn, M., Kohler, C., Basquin, J., Simos, G., Suck, D. (2006) Structural basis of yeast aminoacyl-tRNA synthetase complex formation revealed by crystal structures of two binary sub-complexes. *Nucleic Acids Res*, **34**, 3968-79.
73. Kim, J.E., Kim, K.H., Lee, S.W., Seol, W., Shiba, K., Kim, S. (2000) An elongation factor-associating domain is inserted into human cysteinyl-tRNA synthetase by alternative splicing. *Nucleic Acids Res*, **28**, 2866-72.
74. Kawamata, H., Fujimori, T., Imai, Y. (2004) TSC-22 (TGF-beta stimulated clone-22): a novel molecular target for differentiation-inducing therapy in salivary gland cancer. *Curr Cancer Drug Targets*, **4**, 521-9.
75. Deppmann, C.D., Alvania, R.S., Taparowsky, E.J. (2006) Cross-species annotation of basic leucine zipper factor interactions: Insight into the evolution of closed interaction networks. *Mol Biol Evol*, **23**, 1480-92.
76. Tsuritani, K., Irie, T., Yamashita, R.*, et al.* (2007) Distinct class of putative "non-conserved" promoters in humans: comparative studies of alternative promoters of human and mouse genes. *Genome Res*, **17**, 1005-14.
77. Urrutia, R. (2003) KRAB-containing zinc-finger repressor proteins. *Genome Biol*, **4**, 231.
78. Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., Lee, C. (2004) Assessing the impact of alternative splicing on domain interactions in the human proteome. *J Proteome Res*, **3**, 76-83.
79. Oh, H.J., Lau, Y.F. (2006) KRAB: a partner for SRY action on chromatin. *Mol Cell Endocrinol*, **247**, 47-52.
80. Johnson, M.E., Viggiano, L., Bailey, J.A.*, et al.* (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature*, **413**, 514-9.
81. Zendman, A.J., Van Kraats, A.A., Weidle, U.H., Ruiter, D.J., Van Muijen, G.N. (2002) The XAGE family of cancer/testis-associated genes: alignment and expression profile in normal tissues, melanoma lesions and Ewing's sarcoma. *Int J Cancer*, **99**, 361-9.
82. Tanino, M., Debily, M.A., Tamura, T.*, et al.* (2005) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res*, **33**, D567-72.
83. Ma, Y., Zhang, S., Xia, Q.*, et al.* (2002) Molecular characterization of the TCP11 gene which is the human homologue of the mouse gene encoding the receptor of

fertilization promoting peptide. *Mol Hum Reprod*, **8**, 24-31.

84. Sorek, R. (2007) The birth of new exons: mechanisms and evolutionary consequences. *Rna*, **13**, 1603-8.
85. Yeo, G., Holste, D., Kreiman, G., Burge, C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol*, **5**, R74.
86. Stamm, S., Riethoven, J.J., Le Texier, V.*, et al.* (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*, **34**, D46-55.
87. Lee, C., Atanelov, L., Modrek, B., Xing, Y. (2003) ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res*, **31**, 101-5.
88. Koscielny, G., Le Texier, V., Gopalakrishnan, C.*, et al.* (2009) ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, **93**, 213-20.
89. Kim, N., Alekseyenko, A.V., Roy, M., Lee, C. (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res*, **35**, D93-8.
90. Takeda, J., Suzuki, Y., Nakao, M.*, et al.* (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res*, **35**, D104-9.
91. Sayers, E.W., Barrett, T., Benson, D.A.*, et al.* (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **38**, D5-16.
92. Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res*, **36**, W5-9.
93. Hiller, M., Huse, K., Szafranski, K.*, et al.* (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet*, **36**, 1255-7.
94. Lareau, L.F., Inada, M., Green, R.E., Wengrod, J.C., Brenner, S.E. (2007) Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, **446**, 926-9.
95. Shimada, M.K., Hayakawa, Y., Takeda, J., Gojobori, T., Imanishi, T. (2010) A comprehensive survey of human polymorphisms at conserved splice dinucleotides and its evolutionary relationship with alternative splicing. *BMC Evol Biol*, **10**, 122.
96. Li, R., Yu, C., Li, Y.*, et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966-7.
97. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
98. Wang, E.T., Sandberg, R., Luo, S.*, et al.* (2008) Alternative isoform regulation in

human tissue transcriptomes. *Nature*, **456**, 470-6.

99. Takeda, J., Suzuki, Y., Sakate, R.*, et al.* (2010) H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Res*, **38**, D86-90.

100. Trapnell, C., Pachter, L., Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105-11.

101. Lamkanfi, M., Festjens, N., Declercq, W., Vanden Berghe, T., Vandenabeele, P. (2007) Caspases in cell survival, proliferation and differentiation. *Cell Death Differ*, **14**, 44-55.

102. Lakshmanan, U., Porter, A.G. (2007) Caspase-4 interacts with TNF receptor-associated factor 6 and mediates lipopolysaccharide-induced NF-kappaB-dependent production of IL-8 and CC chemokine ligand 4 (macrophage-inflammatory protein-1). *J Immunol*, **179**, 8480-90.

謝辞

本研究の遂行にあたり、終始ご指導・ご鞭撻賜りました鈴木穣准教授に心から感謝いたします。研究の方向性の指導からまとめ方の指導まで、本当に素晴らしいご指導をいただきました。先生のご指導がなければここにあるほとんどの成果はありませんでした。

また、研究に関する様々の有用なご指導・ご教授を賜りました国立遺伝学研究所の五條堀孝教授、産業技術総合研究所バイオメディシナル情報研究センター分子システム情報統合チームの今西規研究チーム長に深く心から感謝いたします。先生方には、分子進化学や生物学データベースに関する様々な知識を授けていただいただけでなく、学会での発表や様々な人達とのつながりの機会を与えてくださいました。

研究の機会を与えて下さり、ご指導・ご助言を頂きました本研究室の長である菅野純夫教授にも同じく心から感謝申し上げます。研究室セミナーでのご助言やご意見などがとても参考になり、勉強になりました。

HL60 細胞を用いた選択的保持イントロンおよび DLD-1 細胞における RNA の翻訳検証の実験（主に細胞分画）については、富田野乃准教授、上田卓也教授に行っていただきました。RT-PCR による RNA 発現の実験的検証については、関真秀様、入江拓磨様に行っていただきました。また、H-DBAS の構築には、株式会社メイズの湯野川春信様、黒田毅様、伊利夫様、新田美智子様、小関洋平様に行っていただきました。非常に感謝しております。

産業技術総合研究所バイオメディシナル情報研究センター分子システム情報統合チームのポスドク・テクニカルスタッフ・事務の方々にも、様々なご指導とご協力をいただきました。特に、坂手龍一様、佐藤慶治様には、比較ゲノム解析や自動アノテーションの効率化に多大なご協力をいただきました。また、ライフサイエンス統合データベースセンターの中尾光輝様には、バイオインフォマティクス解析におけるアドバイスをいただきました。ここに感謝いたします。

その他、メディカルゲノム専攻の研究室の方々にも多くの有用なサポートをいただきました。特に、若栗浩幸様、関森悦子様には、次世代シークエンサーのデータ処理などに多大のご協力をいただきました。ここに感謝いたします。

様々な方の支えがあったおかげで、本研究をまとめあげることができました。本当にありがとうございました。

2011 年　3 月
武田　淳一