博士論文番号: 1821006

網羅的な切断部位解析による植物 RNA 切断機構の解明

(Comprehensive analysis of cleavage sites in internal cleavage-mediated RNA decay in plant)

上野　大心
奈良先端科学技術大学院大学
先端科学技術研究科　植物代謝制御研究室
（出村　拓　教授）

2021 年 1 月 18 日提出

目次

緒論

遺伝子の DNA 情報は転写、転写後、翻訳、翻訳後調節など様々なステップを経て、タンパク質へと変換される。この一連の過程を遺伝子発現と呼び、細胞が生命を維持する上で重要な役割を担っている。これまでに、遺伝子の発現を理解する目的で、非常に多くの転写（トランスクリプトーム）や翻訳段階（プロテオーム）での研究が行われてきた。加えて、近年では、この転写や翻訳過程に加え、次世代シーケンサーを用いて RNA の安定性を網羅的に評価するなど転写後調節に着目した解析も行われている。この RNA の安定性は、RNA 量、タンパク質量を調節する重要な機構の一つであると考えられ (1)、各生物種はその細胞周期長に対応した固有の RNA 分解速度を持つことが知られている (2)。また、恒常的に細胞内の機能維持に関わるハウスキーピングタンパク質などをコードする RNA の半減期は長い一方で、ストレス応答や細胞増殖など一過的な機能調節に関わる RNA は不安定であるなど、RNA 安定性は様々な生物学的プロセスに関与していることが示唆されている (2, 3)。

これらの RNA 分解を介した遺伝子発現調節が崩れてしまった場合、生体内に危機的な影響を及ぼす場合がある。ヒトにおいて細胞増殖に関与する *c-Myc* RNA は、通常時は Apurinic/apyrimidinic endonuclease 1 (APE1) により積極的に RNA が切断、分解されることで、その蓄積 RNA 量が低く抑えられている (4)。一方で、遺伝子変異などにより *c-Myc* 遺伝子の発現調節が崩れた場合、過剰な血管新生や細胞増殖など、ガンを誘発することが知られている (5)。実際に、*c-Myc* RNA を切断し、分解に導く *APE1* 発現量を抑制することで、*c-Myc* RNA の半減期は 2 倍増加し、RNA 蓄積量に関しては 4.6 倍増加することが報告されている (4)。このように、RNA 安定性は蓄積 RNA 量を調節する重要な機構であり、遺伝子発現調節に大きく寄与していると考えられる。また、これらの RNA 安定性を介した遺伝子発現量の調節は、発生段階や環境ストレス応答など特異的な条件下でも報告されている。例えば、大腸菌では、ストレス応答に関与するタンパク質因子の一つとして RelE が知られている (6)。翻訳過程はエネルギーを大量に消費するステップであり、ストレス条件下で RelE は、翻訳状態が活発な RNA を中心に配列特異的に切断、分解する。このような異なる条件間での RNA 安定性に着目した解析は、個別遺伝子のみに限らず網羅的な解析でも報告されている。植物で C-repeat-binding factor (CBF) 応答遺伝子群は低温順化に関与しており、CCGAC 配列をプロモーター領域内に持っている (7)。これら遺伝子群の RNA は、通常条件下では比較的短い半減期であるが、低温ストレス環境下に暴露されると、その安定性が増すことがシロイヌナズナで報告され

ている (8)。RNA 安定性に着目した解析は、動物細胞でも報告され、胚形成の初期では、母性由来の RNA やタンパク質が多く存在しているが、ある時点を境に父性因子が優勢となり、この転換が初期胚形成において重要なことが知られている (9)。この際、数千種の母方由来の RNA が microRNA などにより選択的、かつ急速に分解される (10)。これらの事例の場合、主要なトランス因子の多くは未だ特定されていないが、大腸菌、植物、動物など様々な生物種の生物学的プロセスにおいて、RNA 安定性を介した遺伝子発現量の緻密な調節、制御が行われている。

これらの RNA 分解機構は、大きく二つに大別することができ、ポリ A 鎖の短縮に依存する分解機構、もしくはエンドヌクレアーゼ等の内部切断に依存する分解機構が存在する (11, 12)。ポリ A 鎖の短縮依存的な分解機構に関しては酵母を中心に解析が行われており、Ccr4/Pop2/Not 複合体によるポリ A 鎖の短縮が起因となり、Dcp1、もしくは Dcp2 によるキャップ除去からの 5' – 3' エキソヌクレアーゼである XRN1 もしくは、3' – 5' エキソヌクレアーゼから構成されるエキソソームにより RNA が分解される (12)。内部切断に依存する分解機構に関しては、エンドヌクレアーゼ等により RNA が切断されることで、ポリ A 鎖の短縮依存的な分解機構と同様に 5' - 3'、もしくは 3' - 5'エキソヌクレアーゼにより RNA が分解されることが酵母で報告されている (12)。

ポリ A 鎖の短縮に依存する分解機構に関わる多くのタンパク質は、酵母において同定されており、植物でも酵母で同定されたホモログ遺伝子を対象に解析が行われている (13)。加えて、この分解機構に関わる配列モチーフも複数報告されており、植物では 5'-3' エキソヌクレアーゼとして知られる AtXRN4 の変異体を用いた実験から、GCUCAG や UUGACU などの配列モチーフを持つ RNA が分解の標的になりやすいことや (14, 15)、ポリ A 鎖の短縮に関わる AtPum2 は UGUAUAUA 配列を認識、結合し RNA 分解を誘導することが報告されている (16)。このように、ポリ A 鎖の短縮に依存する分解機構に関しては、酵母を中心に植物、動物など様々な生物種を対象に分解に関与するタンパク質や配列に着目した詳細な解析が行われている。

内部配列の切断に依存する分解機構に関しても、レアコドンや RNA 高次構造など翻訳伸長反応が阻害される配列を持った RNA を切断、分解する no go decay (NGD) が知られている (17)。NGD に関しては、リボソームが停滞することで、DOM34/Hbs 複合体が形成され、未知のタンパク質因子によって RNA 切断が誘導されることが報告されている (12)。実際に複数の遺伝子由来の RNA が NGD の標的となることが植物でも報告されているが、NGD の阻害剤であるシクロヘキシミドを添加し、網羅的な切断部位解析を行った際も、阻害剤を添

加しない場合と同様に RNA 切断部位が検出されていることから、NGD 非依存的な RNA 切断が多く存在することも示唆されている (18)。また、自身と相補的な塩基配列を持つ RNA に結合し、RNA 切断を引き起こす microRNA に関しても、全遺伝子の数%のみしかターゲット配列を持っていないことから (18, 19)、NGD や microRNA 以外の未知の RNA 切断機構により、多くの RNA が細胞内で切断されていることが推測される。このように、内部配列の切断に依存する分解機構については、いくつか報告されているが、配列的特徴など、これらの切断に関わる要因に関しては未解明である。

そこで本研究では、RNA 内部切断部位を検出する従来手法の問題点を改善した Truncated RNA end sequencing (TREseq) 法をシロイヌナズナにおいて確立し (第一章)、より正確で網羅的な切断部位情報と切断のされやすさ (切断率) に関する情報を基に、RNA 切断に関わる特徴に着目した解析を行った (第二章)。加えて、RNA 切断には複数の要因 (特徴) が複合的に関与すると想定されるため、数理モデルを用いた特徴選択を行うことで、各特徴の RNA 切断への寄与度を評価し (第三章)、未だ不明な点が多い RNA 切断機構に関する理解を深めることを目指した。

第一章

植物における網羅的な RNA 内部切断部位同定法の確立

1-1. 序論

これまで、RNA 切断部位を網羅的に同定する手法として、Parallel analysis of RNA ends (PARE) (20) や genome-wide mapping of uncapped transcripts (GMUCT) (21) が植物で報告されており、Akron-seq (22)、5Pseq (23) など、同様の手法が酵母や動物などを対象とした解析にも使用されている。

これらの手法を用い、microRNA などのターゲットサイトが同定されてきたが、従来方法にはいくつかの実験手法上の問題点が存在した。その一つとしてアダプター付加効率の偏りが挙げられる。最も初期に確立された網羅的な分解産物解析手法は、一本鎖アダプター配列を用いて、RNA にアダプター付加を行なっていた (20, 21, 24)。この方法はアダプター配列と RNA 末端間に強固な二次構造が形成された場合、アダプター付加効率が著しく低下することが報告されている (25, 26)。実際に、従来手法を用いた先行研究では、検出された切断部位周辺で、二次構造が形成されにくい傾向が認められている (22)。また、一本鎖アダプター配列を使用した場合は、RNA の 5' 末端だけではなく、RNA 配列の内部に非特異的にアダプター配列が結合、PCR 増幅によって本来は存在しない RNA の 5' 末端 (シーケンスアーティファクト) が検出されることも報告されている (27)。

このアダプター付加効率を改善した手法として、Cap analysis of gene expression (CAGE) 法が挙げられる (28)。複数の先行研究で報告されているように、アダプター配列を付加する際に、polyethylene glycol (PEG) を添加することでライゲーション効率は劇的に高まることが知られており (25, 26)、CAGE 法でも行われている。また、CAGE 法は二本鎖アダプター配列 (一部分が突出している) を用いることで、アダプターダイマーの形成を抑え、アダプター付加効率を向上させている (28)。加えて、CAGE 法は PCR を用いた増幅を行わないため、従来の一本鎖アダプター配列を使用した際に認められたような、非特異的なアダプター配列の結合とそれに伴う増幅は理論上生じない。この CAGE 法を用いた切断部位の検出は動物細胞で試みられているが、CAGE 法は Cap が付加された (Cap RNA) RNA を濃縮するため、Cap が付加されていない切断部位 (Cap-less RNA) の検出には不向きであった (29)。

加えて、これまでの手法は rRNA を除去するために、ライブラリー作製時にポリ A 鎖付き RNA を濃縮していることも大きな問題点として挙げられる。

RNAseq 法を用いて RNA 蓄積量を測定した解析では、ポリ A 鎖付き RNA の濃縮は、濃縮しなかった場合と比較し、検出されるリードが RNA の 3' 末端側に偏ることが報告されている (30)。この傾向は網羅的な分解産物解析においても認められ (図 1-1)、従来手法では RNA の 3' 末端側で検出される切断部位が過大評価されていた (24)。

これまで、従来手法を用いることで切断部位に関する限られた情報を取得することはできたが、上述したような問題点から正確な切断部位の位置情報や各切断部位での切断のされやすさ (切断率) を数値化することは困難であった。言い換えると、検出される切断部位に偏りがあり、各切断部位での切断率の違い (重み) を考慮した解析が行えず、切断に関わる要因を明らかにすることはできていない。

そこで本研究の第一章では、RNA の切断に関わる配列等の特徴を明らかにすることを目的とし、CAGE 法を改善した Truncated RNA end sequencing (TREseq) 法を確立した。TREseq 法では、Cap-less RNA の濃縮率を高めるために Cap トラップ法を用いて Cap RNA に加え、Cap-less RNA のライブラリーを作製している。加えて、rRNA 除去法、ランダムプライマーを用いることで、ポリ A 鎖付き RNA を濃縮せずにライブラリーを作製した (表 1-1)。第一章では、TREseq 法によって Cap-less RNA の濃縮率がどの程度向上したかを確認するとともに、異なる手法を用いた検証実験を行い、TREseq 法の有効性について考察を行った。

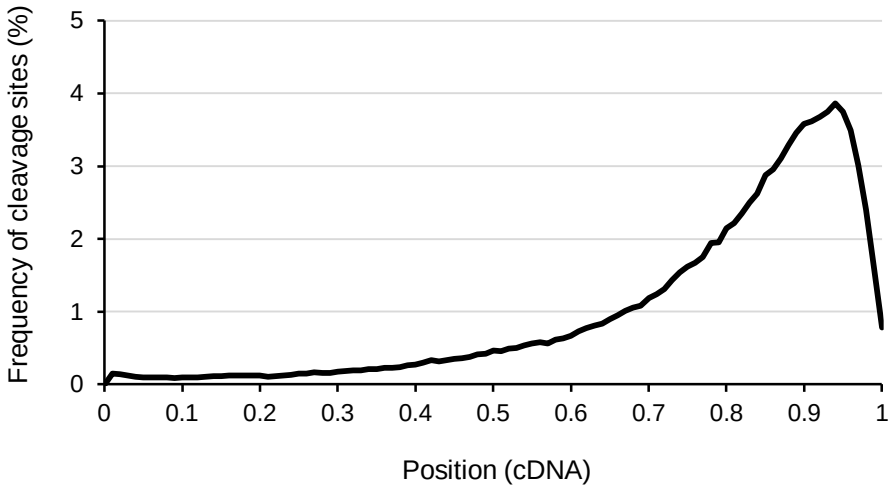


図 1-1. 従来手法における切断部位の偏り

Gregory らが GMUCT 法を用いて取得したシロイヌナズナの網羅的な切断部位情報を Gene Expression Omnibus (GEO) データベースより取得し (21)、マッピング、アノテーションを行った。その後、RNA 長 (cDNA) に対する各切断部位の相対距離を算出し、そのヒストグラムを作成した。X 軸は RNA 上での相対的な切断部位の位置を示し、Y 軸は各位置で検出された切断部位の比率を示す。0、1 はそれぞれ遺伝子の 5' 末端と 3' 末端を示す。

表 1-1. TREseq 法と従来手法との比較

	Adapter ligation	Presence of PEG	how to remove rRNA	PCR cycles
PARE (24)	single strand	No	oligo dT	21 cycles
GMUCT 1.0 (21)	single strand	No	oligo dT	16 cycles
GMUCT 2.0 (31)	single strand	No	oligo dT	11 cycles
TREseq	partially double-stranded	Yes	rRNA depletion	0

従来手法と TREseq 法との比較を示す。比較にはアダプター配列、PEG 添加の有無、rRNA の除去方法、PCR の増幅回数を示した。

1-2. 材料と方法

1-2-1. 植物細胞および培養条件

シロイヌナズナ培養細胞 (*Arabidopsis thaliana* T87) は理化学研究所ジーンバンク室植物開発銀行より分与していただものを使用した。培養は 22ºC、24 時間明期、振とう速度 80 rpm (SLK-3-FS, NK system) の条件で行い、改変 LS 培地を 300 mL 容量の三角フラスコに入れ使用した (32)。一週間ごとに定常期に達した細胞 4 mL を新しい培地 95 mL に移植し継代培養を行った。

1-2-2. Truncated RNA end sequencing (TREseq)

1-2-2-1. 細胞回収、RNA 抽出

回収した細胞を破砕後、TRIzol Reagent (Thermo Fisher Scientific, MA, USA) を用いて Total RNA を抽出した。Total RNA を抽出後、RNeasy kit (Qiagen, Hilden, Germany) を用いてカラム上で DNase I 処理を行い、RNA を精製した。

1-2-2-2. ライブラリー作製

TREseq は non-Amplified non-Tagging Illumina Cap Analysis of Gene Expression (nAnT-iCAGE) を改変した手法である (28)。まず、Ribo-Zero rRNA Removal Kit (Plant Seed/Root) (Illumina, CA, USA) を用いて Total RNA から rRNA を除去した。その後、ランダムプライマー、もしくはオリゴ dT プライマーを用いて逆転写反応を行った。生成された RNA–cDNA hybrids は Cap トラップ法を用いて Cap RNA–cDNA hybrid (Cap RNA) と Cap-less RNA–cDNA hybrid (Cap-less RNA) に分画した。ランダムプライマーを用いた場合の実験手順を図 1-2 に示す。Cap RNA (ランダムプライマー)、Cap less RNA (ランダムプライマー) はそれぞれ transcription start sites (TSS)、および切断部位の情報として使用した。また、Cap RNA (ランダムプライマー) は RNA の蓄積量のデータとしても使用した (各遺伝子単位でリード数が 50 以上の遺伝子を対象)。Cap-less RNA (オリゴ dT プライマー) は、Cap-less RNA (ランダムプライマー) の比較用として使用した。その後、Cap RNA (ランダムプライマー) については nAnT-iCAGE 法に従い cDNA ライブラリーを作製した。Cap-less RNA (ランダム or オリゴ dT) については AMPure XP (Beckman Coulter, Indiana, USA) を用いて精製後、cDNA ライブラリーを作製した。それぞれのライブラリーを Illumina NextSeq 500 (Illumina) に供した。

1-2-2-3. データ解析

次世代シーケンサーを用いて配列情報を取得後、データプロセシングを行った。TREseq のデータ解析は、MOIRAI に追加のプログラムを加え解析を行った (33)。フィルタリングではクオリティが低いリード、もしくは rRNA に由来するリードを解析から除外した。その後、TAIR10 よりゲノム情報を取得し、マッピングを行った。マッピングソフトは HISAT2 を使用した。マッピングの際、ミスマッチがリードの 5' 末端から 3 塩基以内に存在した場合、それらのリードは解析から除外した。マッピングを行った際に、ゲノム上の同じ位置に Cap RNA と Cap-less RNA 由来の read が存在した場合、Cap トラップ時に適切に分画されず、両データに混入した read (ノイズ) の可能性がある。Cap RNA は、末端に Cap 構造に由来する G の塩基が必ず付加されているため、Cap-less RNA の 5' 末端の塩基と、マップされたリードの 5' 末端のゲノム上の塩基が A 塩基、T 塩基、もしくは C 塩基の場合、Cap RNA と Cap-less RNA はそれぞれノイズではないと考えられる (図 1-3A)。その一方で、read の 5' 末端位置のゲノム上の塩基が G であった場合には、[1] Cap RNA、Cap-less RNA の双方が真 (図 1-3B' 1)、[2] Cap RNA が真である (図 1-3B' 2)、[3] Cap-less RNA が真である (図 1-3B' 3)、の 3 パターンの可能性があり、混入したリードかどうかを判断することができない (図 1-3B)。そこで、双方のデータが真であると定義したデータを使用し (図 1-3A; ゲノム上の塩基が A 塩基、T 塩基、もしくは C 塩基の場合)、Cap RNA と Cap-less RNA の Reads Per million Mapped reads (RPM) 値比を、混入した read か否かを決める閾値とした。閾値は Cap RNA の場合は、Cap RNA / Cap-less RNA 比 (図 1-3C)、Cap-less RNA の場合は、Cap-less RNA / Cap RNA 比 (図 1-3D) の 5th percentile とした。閾値を下回るリードは混入したリードと定義し、解析から除外した。

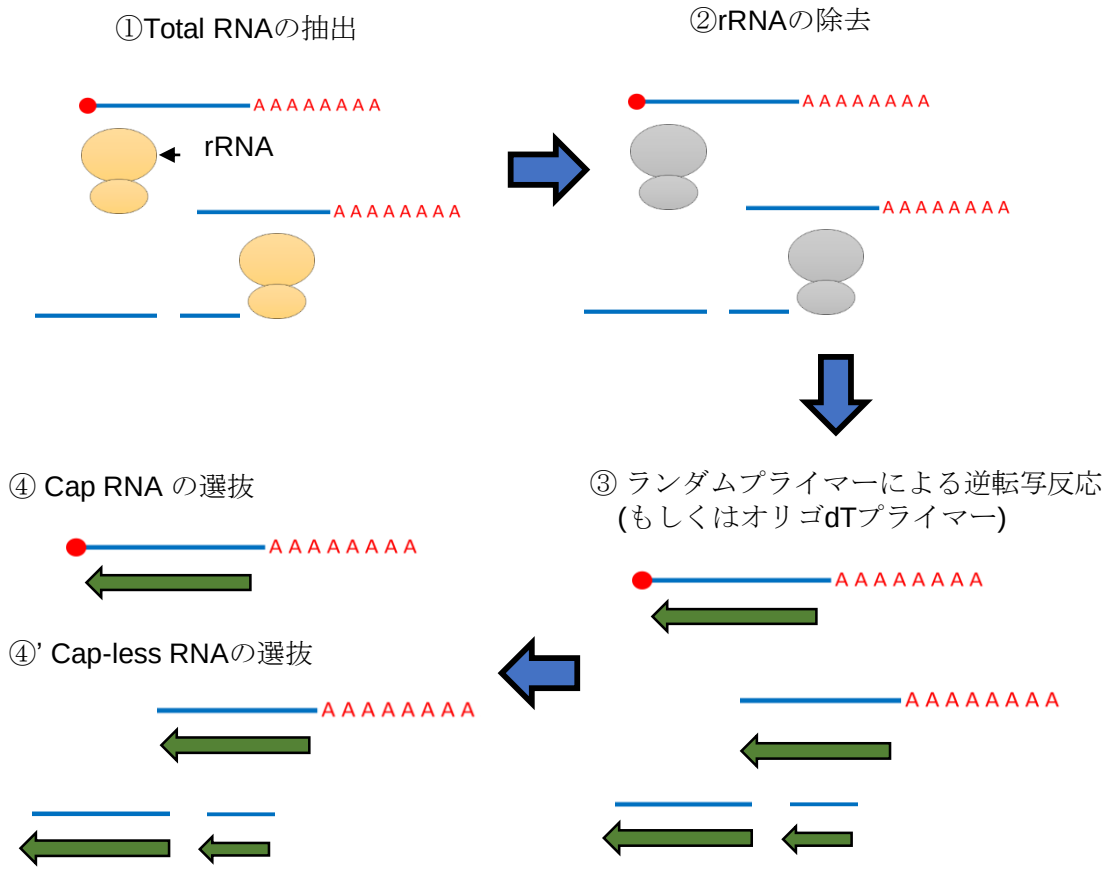


図 1-2. ライブラリー作製の概要図

Total RNA を抽出後、rRNA 除去キットを用い、rRNA を取り除いた。その後、ランダムプライマー、もしくはオリゴ dT プライマーを用い逆転写反応を行った。切断部位データとしては Cap-less RNA を使用するが、Cap RNA が混入する可能性がある。後のステップでそれらのノイズを取り除くために、双方のライブラリーを作製し、次世代シーケンサーを用いてそれぞれ配列情報を取得した。

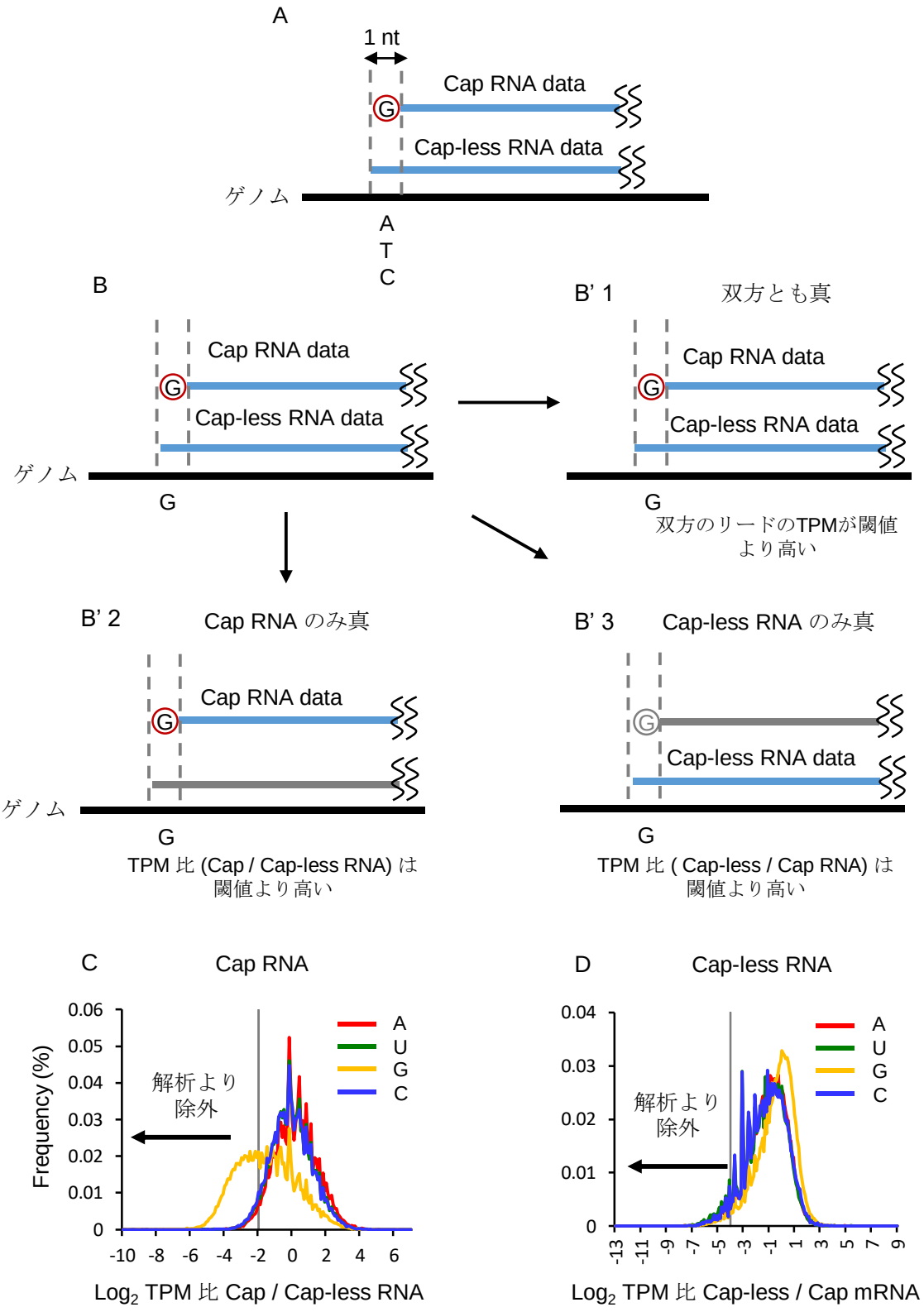

A
1 nt
Cap RNA data
Cap-less RNA data
ゲノム
A
T
C
B
Cap RNA data
Cap-less RNA data
ゲノム
G
B' 1
双方とも真
Cap RNA data
Cap-less RNA data
G
双方のリードのTPMが閾値より高い
B' 2
Cap RNA のみ真
Cap RNA data
ゲノム
G
TPM 比 (Cap / Cap-less RNA) は閾値より高い
B' 3
Cap-less RNA のみ真
Cap-less RNA data
G
TPM 比 (Cap-less / Cap RNA) は閾値より高い
C
Cap RNA
Frequency (%)
0.06
0.05
0.04
0.03
0.02
0.01
0
解析より除外
A
U
G
C
-10 -8 -6 -4 -2 0 2 4 6
Log2 TPM 比 Cap / Cap-less RNA
D
Cap-less RNA
0.04
0.03
0.02
0.01
0
解析より除外
A
U
G
C
-13 -11 -9 -7 -5 -3 -1 1 3 5 7 9
Log2 TPM 比 Cap-less / Cap mRNA

図 1-3. Cap RNA、Cap-less RNA におけるノイズの除去

Cap RNA と Cap-less RNA の 5' 末端が同じ位置にマッピングされた場合 (A)、双方の 5' 末端のゲノム上の塩基が A 塩基、T 塩基、もしくは C 塩基ならば、Cap RNA の 5' 末端の G 塩基は転写後に付加された Cap であると推測できるため、それぞれのリードが Cap RNA、もしくは Cap-less RNA かを区別することができる。Cap RNA と Cap-less RNA の 5' 末端が同じ位置にマッピングされかつ、ゲノム上の塩基が G 塩基だった場合 (B)、[1] 双方のデータが真 (B' 1)、[2] Cap RNA データのみ真 (B' 2)、[3] Cap-less RNA データのみが真 (B' 3) の 3 パターンが考えられ、ゲノム上の塩基では、どのパターンであるかを判断することができない。そこで、(A) のように、Cap RNA と Cap-less RNA データの双方が真であると定義した場合のデータ (ゲノム上の塩基が A 塩基、T 塩基、もしくは C 塩基) を使用し、Cap RNA と Cap-less RNA の RPM 値比を基にデータを除外するかどうかの境界線を設定した。境界線の位置は、Cap RNA の場合は、Cap RNA / Cap-less RNA 比 (C)、Cap-less RNA の場合は、Cap-less RNA / Cap RNA 比 (D) の 5th percentile とした。

1-2-3. 切断率の定義

各部位での切断のされやすさの指標値となる cleavage score (CS_{site} 値) を算出した。CS_{site} 値は、各切断部位でのリード数をその遺伝子の RNA 蓄積量 (本研究で取得した Cap RNA の情報) で除算した値である。また、各 RNA ごとの CS_{site} 値の合計を CS_{gene} 値とした。

CS_{site} = 各切断部位でのリード数 / RNA 蓄積量

CS_{gene} = 各 RNA ごとの CS_{site} 値の合計値

1-2-4. 5'-Bromouridine imunopricipitation cahse (BRIC) 法

1-2-4-1. 細胞回収、RNA 抽出

継代後 2 日目のシロイヌナズナ T87 培養細胞を Bromouridine (BrU) を含む培地で 16 時間培養した。継代 3 日目に BrU を含まない培地に交換し、培地交換後 0、1、3、6 時間後に細胞を回収した。TRIzol Reagent を用いて Total RNA を抽出後、LiCl 沈殿による RNA 精製を行った。Spike-in (MBL, Nagoya, Japan) を加え、5-BrU immunoprecipitation chase (BRIC) を用いて BrU でラベルされた RNA を抽出した (MBL)。

1-2-4-2. ライブラリー作製

SMART-Seq v4 Ultra Low Input RNA キット (Clontech, CA, USA)を用いて cDNA ライブラリーを作製し、Illumina NextSeq 500 (illumina) に供した。

1-2-4-3. データ解析

TAIR10 からゲノム情報を取得し、bwa を用いてマッピングを行い、Cufflinks を用いて fragments per kilobase of exon model per million mapped fragments (FPKM) を算出した。今回の解析では、Tani らの手法を参考に 0 h での FPKM 値が 1 以上の遺伝子を解析対象とした (3)。各タイムポイントにおける FPKM 値は spike-in control を用いて補正した。RNA の蓄積量が 0 h 時と比較し 50% 以下となった場合、それ以降のタイムポイントは解析から除外した (3)。加えて、0、1、3 時間後 の 3 点、もしくは 0、1、3、6 時間後の 4 点でピアソンの積率相関係数が高いタイムポイントのセットを半減期の算出に用いた。半減期の算出には以下の式を使用した：$t1/2 = \ln 2/k_{decay}$ (2)。最終的に、｜半減期 (反復 1) - 半減期 (反復 2)｜≤ 2 の遺伝子を解析対象とした。

1-2-5. Nanopore sequencing 法

1-2-5-1. 細胞回収、RNA 抽出

継代 3 日目のシロイヌナズナ T87 培養細胞を回収した。TRIzol Reagent (Thermo Fisher Scientific) を用いて Total RNA を抽出した。その後、RNeasy kit (Qiagen) を用いてカラム上で DNase I 処理を行い、RNA を精製した。

1-2-5-2. ライブラリー作製

再度、Magnosphere UltraPure RNA Purification Kit (Takara, Shiga, Japan) を用いて RNA を精製後、cDNA-PCR Sequencing Kit (Takara) を用いて逆転写反応、及び switching 反応後、Long Amp Taq を用いて、PCR 増幅を行い、cDNA を得た。ライブラリーを作製後、Nanopore sequencer (Oxford Nanopore Technologies, Oxford, UK) に供した。

1-2-5-3. データ解析

Nanopore sequencer によって検出されたイオン電流シグナルを塩基に置き換え (ベースコール)、pychopper を用いてアダプター配列を除去した。その後、TAIR10 よりゲノム情報を取得し、minimap2 を用いてマッピングを行った。マッピング後、bed ファイルへと変換し、各リードの 5' 末端情報を取得した。

1-2-6. 配列モチーフ等の解析

モチーフ検索ツールである DREME を用いて切断部位の前後 20 塩基から配列モチーフを抽出した (http://meme-suite.org)。microRNA の配列に関しては miRBase (http://www.mirbase.org) より取得し、psRNAtarget (http://plantgrn.noble.org/psRNATarget/) を用いて microRNA のターゲット配列を予測した (18)。Gene ontology (GO) enrichment 解析に関しては、GORILLA (Gene Ontology enRIchment anaLysis and visuaLizAtion tool) (http://cbl-gorilla.cs.technion.ac.il/) を用いて解析を行った。

1-2-7. *in vitro* transcription

in vitro 合成には、Matsuura らが構築したプラスミド pT3-RL-pA を用いた (34)。ポリ A 配列を持つ *in vitro* 転写用プラスミドは *in vitro* 転写反応に先立ち、Ssp I (AATATT) によりポリ A 配列の末端部分を切断し直鎖状にした。Ssp I 処理した DNA 断片は、Gel/PCR Extraction Kit (NIPPON Genetics, Tokyo, Japan) を用いて精製した。精製された DNA 断片を鋳型に、Megascript T3 transcription kit (Ambion, Waltha, USA) を用いて、キ

ャップ構造を持たない mRNA を合成した。合成された RNA はキット付属の DNase I で処理した後、LiCl 沈殿により精製し、付属の RNase-free 水で溶解した。

1-2-8. qRT-PCR

シロイヌナズナから抽出した Total RNA 5 μg に対して、1-2-7 で精製した Cap-less *R-luc* を 1 μg 添加し、1-2-2-2 に示す TREseq 法のライブラリー作製手順を参考に cDNA を合成した。qRT-PCR については、Yamasaki らの研究を参考に行い (35)、Cap-less *R-luc* の遺伝子特異的プライマーセットについては、Matsuura らの研究を参考に 5'-GGATTCTTTTCCAATGCTATTGTT-3'、もしくは 5'-AAGACCTTTTACTTTGACAAATTCAGT-3' を使用した (34)。

1-3. 結果

1-3-1. TREseq 法の解析手順と他の手法を用いた検証実験

1-3-1-1. Cap-less RNA の濃縮比率の算出

これまでの網羅的な分解産物解析では、RNA に対して一本鎖アダプター配列を付加するステップがライブラリー作製に共通しており、また、ポリ A 鎖付き RNA を濃縮しているため検出される RNA 切断部位に偏りが存在していた (24, 31)。CAGE 法では、アダプター付加時に PEG を添加し、二本鎖アダプター配列を使用することで、これらの問題を改善しているが (28)、Cap RNA を濃縮していたため、切断末端である Cap-less RNA は検出できる全 RNA 末端の内 20%ほどであった (29)。TREseq 法では、これらの問題点を改善するために、1-2-2 に示すように Cap トラップ法を行い、Cap RNA ライブラリーに加え、Cap-less RNA ライブラリーを作製している。また、Cap-less RNA の中には rRNA が多量に存在するため、ライブラリー作製時に rRNA 除去法を行っている。

まず、TREseq 法において切断末端 (Cap-less RNA) の濃縮比率が、どの程度向上したかを検証した (図 1-4)。*in vitro* transcription により Cap-less *Renilla luciferase* (*R-luc*) RNA を作製し、シロイヌナズナ由来の Total RNA に添加した。その後、rRNA 除去、Cap トラップ法を用いた分画の有無を組み合わせ、4 つのライブラリーを作製した。これらの 4 つのライブラリー、そして Cap-less *R-luc* RNA を添加した Total RNA を対象に、qRT-PCR を用いて、Cap-less *R-luc* RNA の検出量を比較した。rRNA は、全 RNA の 90%を占めることから (30)、rRNA 除去法を用いることで、少なくとも 10 倍以上は検出量が増加することが想定される。加えて、Cap トラップ法を用いて Cap-less RNA を濃縮するため、数十倍以上 Cap-less *R-luc* RNA の検出量を向上できると考えられた。図 1-4 は、等量の RNA を使用した際に Cap-less *R-luc* RNA をどの程度検出できるかを示し、Total RNA において検出された Cap-less *R-luc* RNA に対する相対比率を示している。図 1-4 に示すように、rRNA を除去した Cap-less RNA ライブラリーでは、Total RNA や rRNA 除去法を行っていない Cap-less RNA ライブラリーと比較して 100 倍以上検出の感度が向上している。rRNA 除去法を行わず Cap-less RNA を濃縮したライブラリーの検出感度が低い理由に関しては、Cap-less RNA である rRNA を濃縮しているためだと考えられた。

以上の結果から、従来手法のアダプター付加効率、Cap-less RNA の濃縮率を改善し、rRNA を除去した TREseq 法を用いることで、より高感度に切断部位 (Cap-less RNA) を検出できることが示された。一方で、Cap が付加された RNA を濃縮している Cap RNA ライブラリーでも、Cap-less *R-luc* RNA が検出されていることを踏まえると、Cap トラップ法を用いて Cap RNA、Cap-less RNA を

完全に分画することは不可能であり、後述するようにデータプロセシングにて双方のデータを比較し、Cap RNA ライブラリーに含まれる Cap-less RNA、もしくは、Cap-less RNA ライブラリーに含まれる Cap RNA をデータから取り除くことが必要であると考えられた。

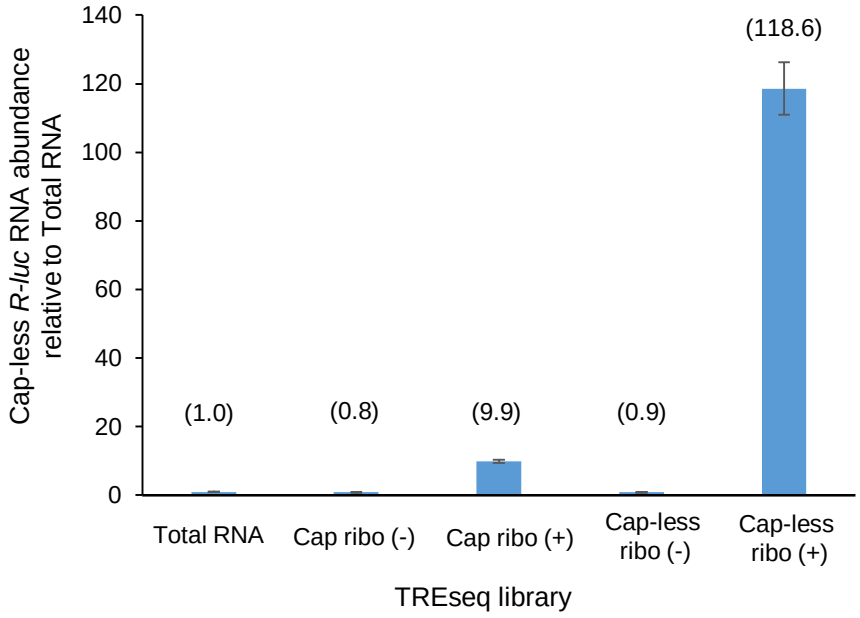


図 1-4. Cap-less RNA 濃縮比率の評価

in vitro transcription にて合成した Cap-less *R-luc* RNA を Total RNA に加え、TREseq 法に従いライブラリー作製を行った。X 軸は、各ライブラリーを示し、Y 軸は Total RNA に含まれる Cap-less *R-luc* RNA 量を基準とした際の相対検出量を示す。Cap ribo (-); rRNA 除去法を行っていない Cap RNA ライブラリー、Cap ribo (+); rRNA 除去法を行った Cap RNA ライブラリー、Cap-less RNA ribo (-); rRNA 除去法を行っていない Cap-less RNA ライブラリー、Cap-less RNA ribo (+); rRNA 除去法を行った Cap-less RNA ライブラリーをそれぞれ示す。括弧内の値は、Total RNA で検出された Cap-less *R-luc* 量に対する検出量の比率を示す。

1-3-1-2. 網羅的な切断部位情報のデータプロセシング

次世代シーケンサーを用いて配列情報を取得した後の手順を図 1-5 に示す。フィルタリング、マッピング後、それぞれのリードは TAIR 10 representative gene models を参考にアノテーションを行った。tRNA、偽遺伝子、もしくは葉緑体、及びミトコンドリアゲノムに存在する遺伝子は解析から除外した (図 1-5)。加えて、より正確に切断部位を同定するために、2 反復で共通の位置に存在する切断部位のみを解析の対象とした。各切断部位でのリード数については、2 反復の平均値を使用した。本研究では、Cap-less RNA (ランダムプライマー) から得られた情報を切断部位として使用するが、このデータには本来の RNA 末端である Cap 構造を持つ Cap RNA のデータも含まれている可能性がある。実際に個別遺伝子について調べてみると、Cap-less RNA には Cap RNA と予想されるリードが存在している (図 1-6)。同様のことは、Cap RNA についても当てはまるため、Cap RNA、Cap-less RNA 相互のデータを比較し、Cap RNA に含まれる Cap-less RNA と予測されるリード、Cap-less RNA に含まれる Cap RNA と予測されるリードを 1-2-2 に従い解析から除外した。最終的に、各切断部位のリード数は RNA の蓄積量で補正するため、Cap-less RNA について RNA の蓄積量情報 (本研究で取得した Cap RNA 情報) がある遺伝子 (遺伝子単位でのリード数が 50 以上) を解析対象とした。

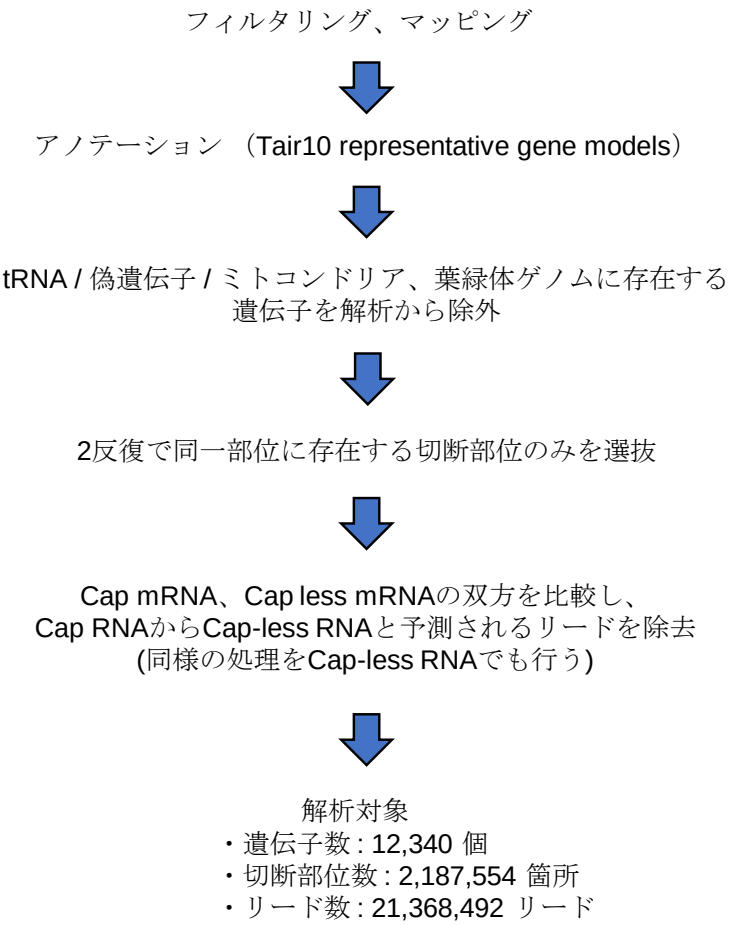


図 1-5. 取得した切断部位情報の解析手順

クオリティが低い、もしくは rRNA に由来するリードを解析から除外し、マッピングを行った。アノテーション後、tRNA、偽遺伝子、もしくは葉緑体、及びミトコンドリアゲノムに存在する遺伝子を解析から除外し、2 反復で共通する切断部位のみを解析対象とした。1-2-2 に示されるように双方のデータからノイズを取り除いた。

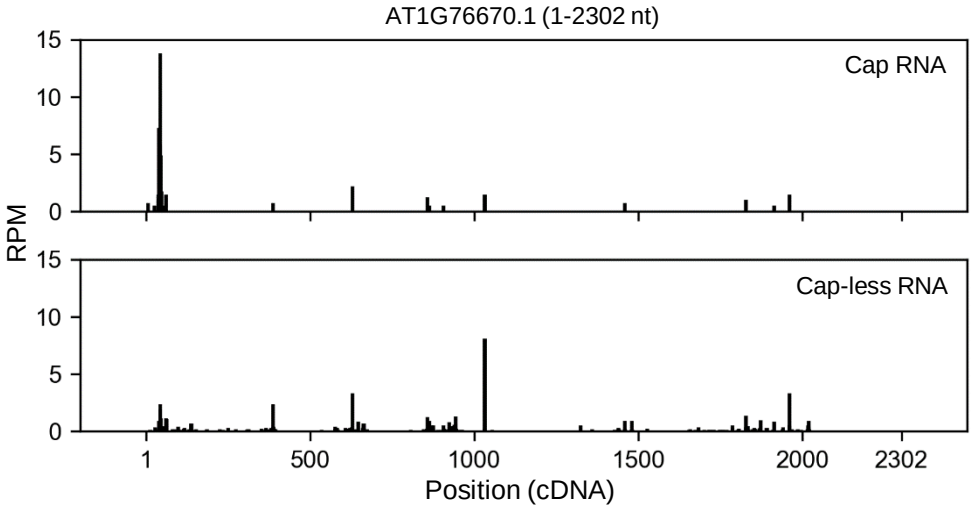


図 1-6. 個別遺伝子を対象とした Cap RNA、Cap-less RNA の比較

上段は Cap RNA (ランダムプライマー)、 下段は Cap-less RNA (ランダムプライマー) を示す。X 軸は RNA 上の位置を示し、Y 軸は各位置でのリード数を示す。1 は TAIR10 に登録されている 5' UTR の 1 塩基目を示している。

1-3-1-3. 検出された切断部位の分布

従来手法の問題点の一つとして、ポリ A 鎖付き RNA の濃縮により検出される切断部位が RNA の 3' 末端側に偏ることが挙げられる。この問題点を改善するために TREseq 法では、rRNA 除去法とランダムプライマーを用いてライブラリーを作製したが、実際に検出される切断部位の偏りが改善するか検証した。個別遺伝子を対象に、ランダムプライマー、もしくはオリゴ dT プライマーを用いて作製した Cap-less RNA ライブラリーでの結果を比較したところ、ランダムプライマーを用いることで、検出される切断部位の 3' 末端側への偏りが大幅に軽減されていた (図 1-7)。同様の結果は、検出された全切断部位を用いて、RNA 内における分布を算出した際でも認められた (図 1-8)。これらの結果は、rRNA 除去法とランダムプライマーを組み合わせライブラリーを作製することで、検出される切断部位の偏りを大幅に軽減したことを示している。

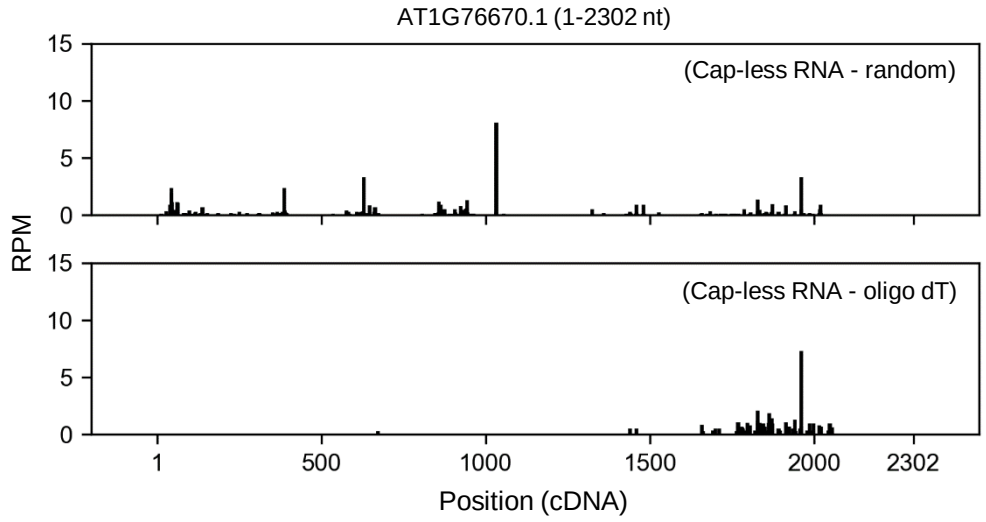


図 1-7. 個別遺伝子を対象とした切断部位の分布

ランダムプライマー (上段)、オリゴ dT プライマー (下段) を用いて作製したライブラリーの切断部位情報 (Cap-less RNA) を示す。X 軸は RNA 上の位置を示し、Y 軸は各位置でのリード数を示す。1 は TAIR10 に登録されている 5' UTR の 1 塩基目を示している。

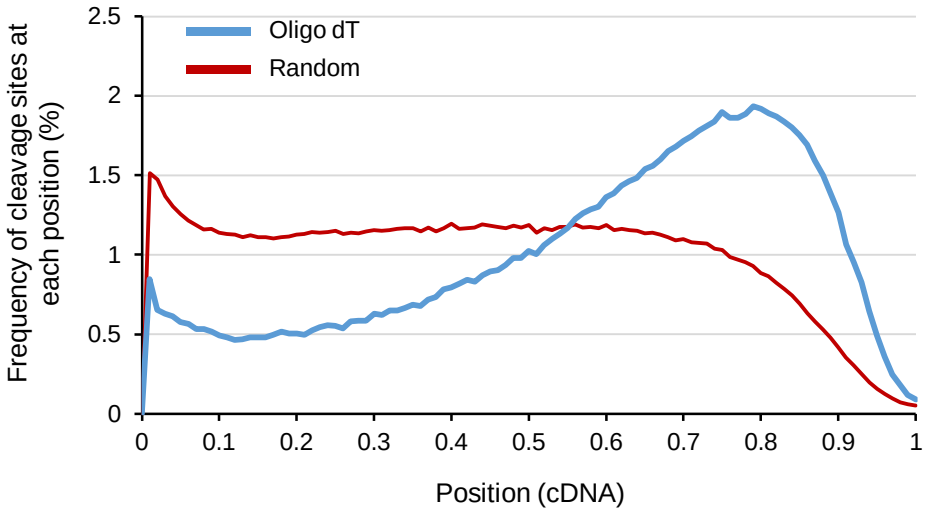


図 1-8. 解析対象となった切断部位の RNA 内での分布

RNA 長 (cDNA) に対する各切断部位の相対距離を算出し、そのヒストグラムを作成した。X 軸は RNA 上での相対的な切断部位の位置を示し、Y 軸は各位置で検出された切断部位の存在比率を示す。0、1 はそれぞれ遺伝子の 5' 末端と 3' 末端を示す。

1-3-1-4. 他の手法を用いた検証実験

RNA の 5' 末端を検出する手法は、転写開始点の同定を目的とした研究で確立され、これまで複数の手法が報告されている (36)。これらの手法は大きく 3 つに大別され、CAGE 法、Template switching 法、Oligo-capping 法 (5' RACE 法) があり (36)、切断部位を同定する網羅的な解析でも応用されている (20–24)。これら 3 つの手法の中でも、CAGE 法は RNA の 5' 末端を検出する精度が高く、ライブラリー作製に由来するバイアスが少ないことが報告され (36)、microRNA のプロセシングサイトや non-coading RNA などの切断部位も検出できることが報告されている (29, 37, 38)。そこで、CAGE 法を改良し本研究で確立した TREseq 法でも、従来手法で報告されている既知の microRNA 切断部位を検出できるか検証した (図 1-9) (39, 40)。加えて、Oligo-capping 法である GMUCT 法 (ポリ A 鎖付き RNA を濃縮) により取得されたデータを GEO データベースより取得、再解析し、TREseq 法との比較に用いた。TREseq 法 (培養細胞) と GMUCT 法 (植物体の花茎) では解析対象とした組織が異なっているが、既知の microRNA 切断部位として解析対象とした AT2G39675.1 と AT1G63130.1 遺伝子については、植物体の花茎で 5' RACE 法による切断部位の検証が Allen らによって行われている (39, 40)。TREseq 法と GMUCT 法で検出された切断部位を比較すると、AT2G39675.1 については、TREseq 法および GMUCT 法の双方で切断部位が検出されていたが、AT1G63130.1 については、TREseq 法でのみ検出された (図 1-9)。GMUCT 法で AT1G63130.1 の切断部位が検出されなかった結果については、切断部位が AT2G39675.1 より RNA の 5' 側に位置することや切断部位へのアダプター付加が効率良く行われなかったことが原因として考えられた。これらの結果は、TREseq 法を用いることで既知の切断部位を検出できることに加え、網羅的に RNA 切断部位を同定する従来の手法では検出が困難な切断部位についても TREseq 法では検出できることを示している。

また、microRNA 以外の他の切断部位についても検証を行った。各遺伝子で最も検出されたリード数が多い切断部位を対象に (1000 遺伝子)、異なる手法を用いて RNA の 5' 末端を検出し、TREseq 法と同様の傾向が認められるか確認した。異なる手法としては、これまでに RNA 末端検出が行われている Nanopore sequencing (Nanopore seq) 法を使用した (41)。その結果、図 1-10 に示すように、TREseq 法で検出された切断部位のピークとわずかにずれてはいたが、Nanopore seq 法でも、TREseq 法とほぼ同様の位置で切断部位が検出された。Nanopore seq 法は 1kb 以上の長い RNA を読むことに適しているロングリードシーケンサーであるため、読まれたリードの正確性 (read accuracy) は TREseq 法のようなシ

ョートリードシーケンサーと比較して低く、挿入や欠失なども生じやすいため、1 塩基から 2 塩基ほど検出されたリードの位置がずれたと考えられる。

検出されたリードの位置は TREseq 法と比較し Nanopore seq 法にて、わずかにずれていたが、これらの結果は、microRNA 以外の切断部位についても、ライブラリー作製に由来するシーケンスアーティファクトではなく、実際に細胞内において存在する切断部位であること示している。

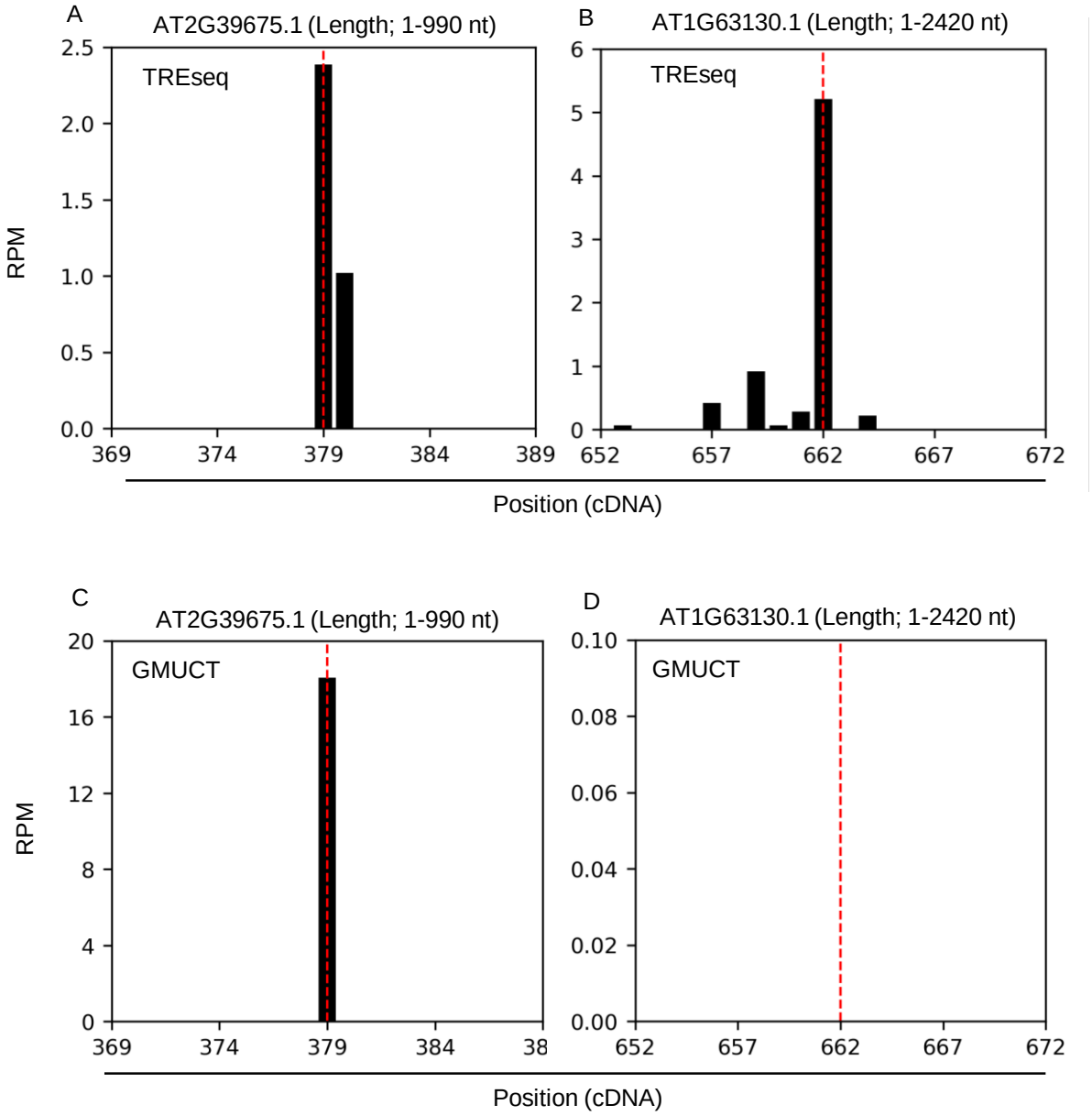


図 1-9. 先行研究で報告されている既知の microRNA 切断部位との比較

5' RACE 法を用いて検証が行われている microRNA の切断部位を対象に、TREseq 法 (A、B)、GMUCT 法 (C、D) との比較を行った。X 軸は RNA 上での位置を示し、Y 軸は各位置で検出されたリード数を示す。赤点線は先行研究において報告されている切断部位を示す。

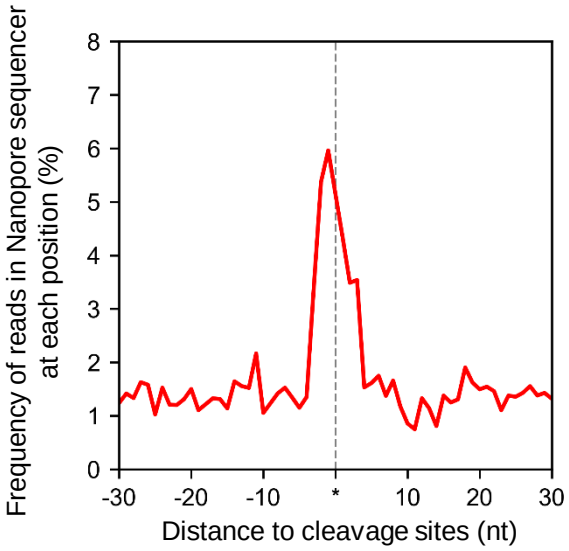


図 1-10. 異なる手法を用いた切断部位の検証

ナノポアシーケンサーを用いて、切断部位情報を取得後、TREseq 法で検出された切断部位周辺のリード数の比率を算出した。X 軸は切断部位からの距離を示し、Y 軸は各位置で検出されたリード数の存在比率を示す。アスタリスクは TREseq 法で検出された切断部位を示す。

1-3-1-5. 切断率の定義

TREseq 法で検出された各切断部位のリード数は RNA の蓄積量に依存する。そこで、各切断部位でのリード数をその遺伝子の RNA 蓄積量 (本研究で取得した Cap RNA データ) で除算した値を cleavage score (CS) とし、各切断部位での切断のされやすさを CS_{site} 値と定義した。また、遺伝子単位での切断のされやすさとして、各遺伝子の CS_{site} 値の合計値、これを CS_{gene} 値と定義した。2 反復の再現性は CS_{site} 値、CS_{gene} 値において r = 0.91、r = 0.99 であった (図 1-11A、1-11B)。また、各部位での CS_{site} 値の分布を見ると正規様に分布しており (図 1-11A)、切断部位のされやすさが片側に偏る傾向は認められなかった。遺伝子単位での CS_{gene} 値についても同様であった(図 1-11B)。

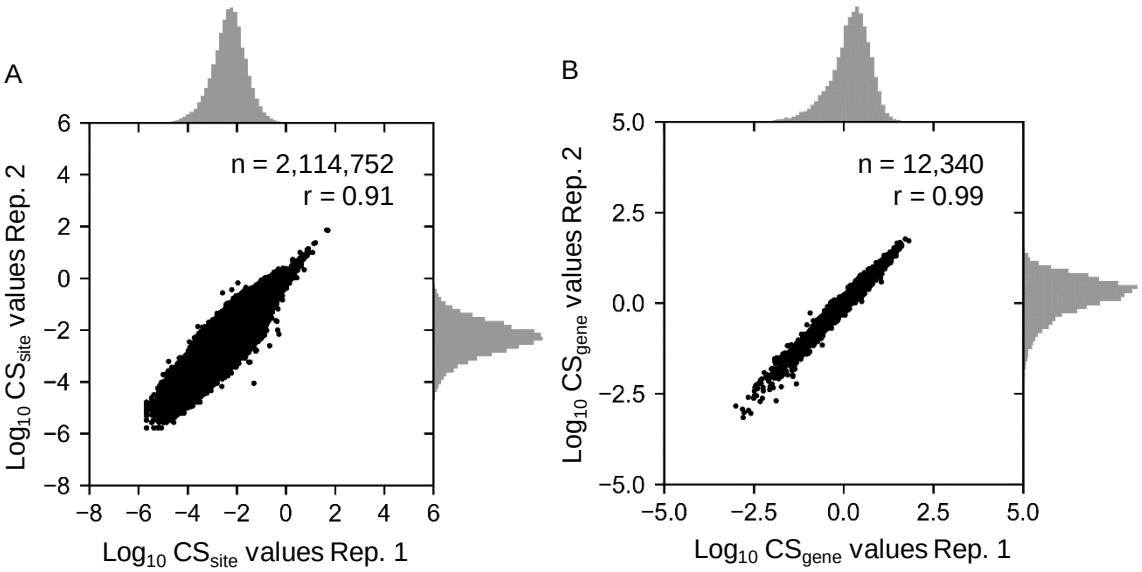


図 1-11. CS_{site} 値、CS_{gene} 値の 2 反復における再現性 (シロイヌナズナ)

CS_{site} 値 (A)、および CS_{gene} 値 (B) の 2 反復の再現性を示す。左辺および右辺は正対する各軸のヒストグラムを示す。

1-3-2. 切断のされやすさと RNA の安定性

1-3-2-1. 5'-Bromouridine imunopricipitation cahse (BRIC) 法を用いた半減期測定

これまで RNA 半減期の測定方法として、転写阻害剤を用いた手法が主に行われてきた (2, 8)。しかし、アクチノマイシン D などの転写阻害剤は細胞内環境に様々な悪影響を与えることが報告されており、生体内での RNA の半減期を正確に測定できていないことが指摘されている (8)。そのため近年では、転写阻害剤を用いない手法として、4-Thiouridine (4SU) や 5'-Bromo-Uridine (BrU) などの合成核酸アナログを用いた半減期の測定が行われている。中でも BrU は他の合成核酸アナログと比較し、細胞に対する悪影響が少ないことが報告されている (3)。そこで本研究では、BRIC 法を用いて RNA 半減期情報を網羅的に取得した。

1-3-2-2. CS_{gene} 値と半減期

これまで、複数の先行研究で RNA の切断のされやすさと RNA 安定性に着目した解析が行われてきたが、切断されやすい RNA ほど半減期が短いなど、両者に関係性は認められなかった (22, 42)。そこで、より正確に切断率を算出できる TREseq 法の切断部位情報を用いて、RNA の切断率と半減期の関係性に着目し解析を行った。遺伝子単位での切断のされやすさとしては、CS_{gene} 値を使用した。CS_{gene} 値が高ければ、切断、分解されやすいと考えられる。1-3-2-1 から半減期情報が存在する遺伝子 (n = 6,825 genes) を用い、CS_{gene} 値と RNA 半減期とのピアソンの積率相関係数を求めた結果、r = -0.18 となり弱い負の相関が認められた (図 1-12A)。

加えて、CS_{gene} 値の TOP 10% (n = 683 genes)、BOTTOM 10% (n = 683 genes) の遺伝子を選抜し、半減期を比較した (図 1-12B)。統計検定を行った結果、CS_{gene} 値の高い遺伝子は半減期が短い傾向が認められた (Welch's t-test, $p < 0.01$)。このことは、CS_{gene} 値が RNA の切断のされやすさ、分解のされやすさを反映していることを示し、TREseq 法の有効性を表していると考えられる。半減期情報が存在する全遺伝子を用いた際に、強い負の相関が認められなかったことに関しては、これらの RNA 半減期情報は、ポリ A 鎖の短縮に依存する分解機構など、RNA 切断に依存する分解機構以外の影響を含んでいることが理由として挙げられる。

また、各 RNA 種の半減期データは手法ごとに異っているため (43)、BRIC 法とは異なる半減期測定法でも同様の傾向が認められるか確認した。転写阻害剤を用いて先行研究にて行われたシロイヌナズナの RNA 半減期情報を Narsai ら

のデータより取得し (2)、切断率と RNA 半減期に着目した解析を行った。半減期情報が存在する全遺伝子を使用した場合 (n = 9,443 genes)、CS_{gene} 値と RNA 半減期とのピアソンの積率相関係数は r = -0.26 となり、切断率を基に TOP 10% (n = 9,44 genes)、BOTTOM 10% (n = 9,44 genes) の遺伝子について半減期を比較したところ、切断率が高いほど RNA 半減期が短い傾向が認められた (図 1-13, Welch's t-test, $p < 0.01$)。異なる半減期測定法で得られたデータを使用した場合でも、切断率と安定性の関係性が認められたことから、これらの切断部位に関する情報は信頼性のあるデータであると考えられる。

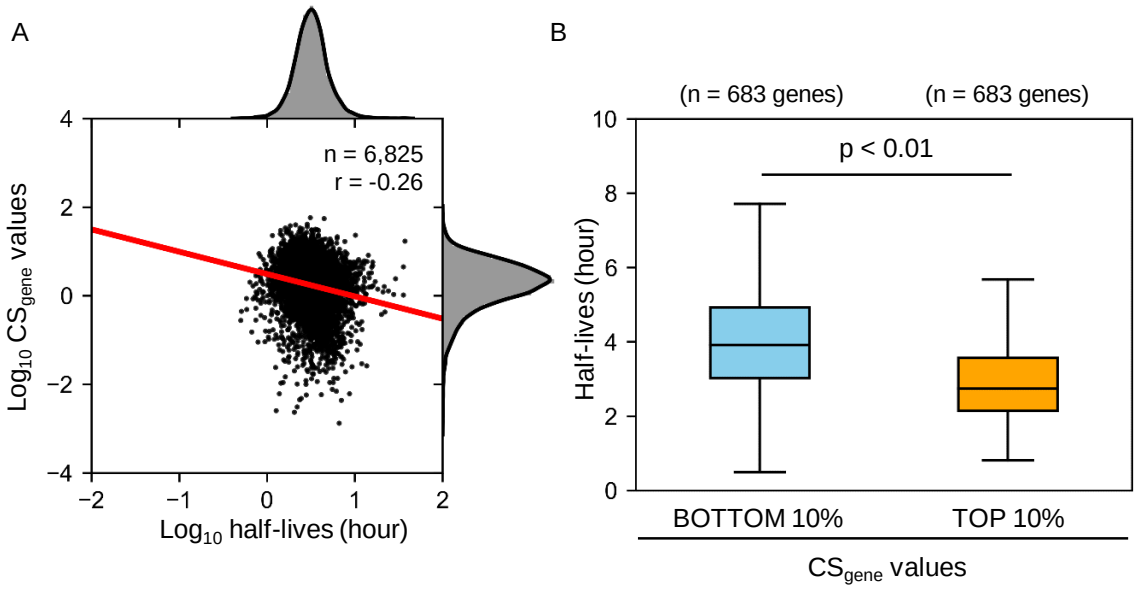


図 1-12. CS_{gene} 値と BRIC 法を用いて取得した半減期情報（シロイヌナズナ）

BRIC 法を用いて半減期情報を取得し、CS_{gene} 値とのピアソンの積率相関係数を求めた (A)。また、CS_{gene} 値が高い、低い順から 10%ずつ遺伝子を選抜し、それらの半減期を比較した (B)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。統計検定には Welch's t-test を使用した。

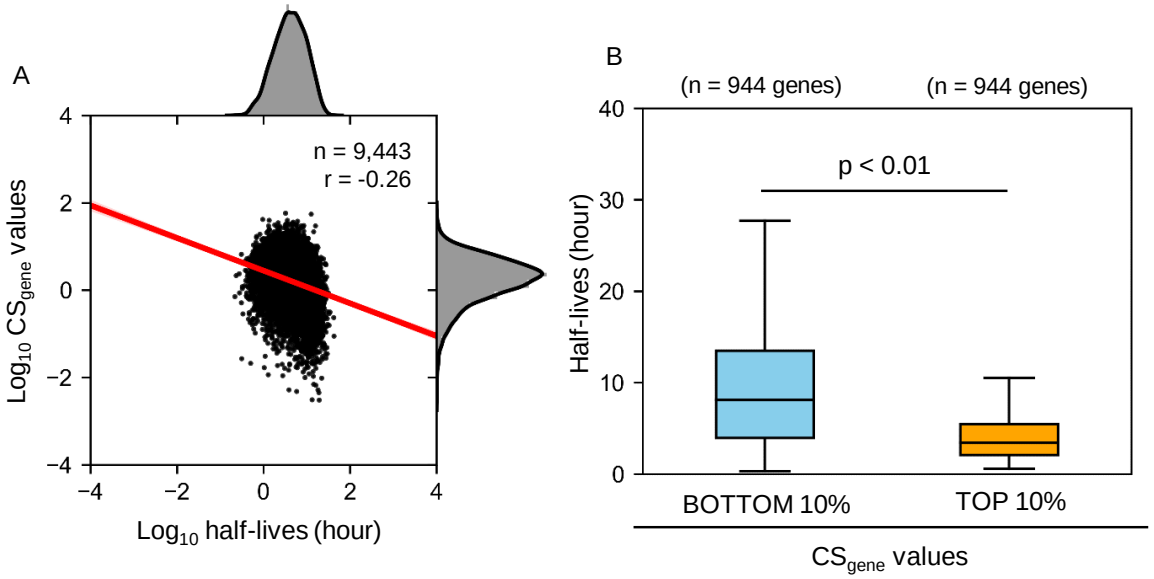


図 1-13. CS_{gene} 値と転写阻害剤を用いて取得した半減期情報（シロイヌナズナ）

先行研究より転写阻害剤を用いた半減期情報を取得し、CS_{gene} 値とのピアソンの積率相関係数を求めた (A)。また、CS_{gene} 値が高い、低い順から 10%ずつ遺伝子を選抜し、それらの半減期を比較した (B)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。統計検定には Welch's t-test を使用した。

1-4. 考察

1-4-1. エキソヌクレアーゼ消化に由来する RNA の 5' 末端

シロイヌナズナやイネにおいて PARE 法や GMUCT 法などの手法により網羅的な分解産物解析が行われている (18, 20, 21, 24)。このように次世代シーケーサーを用いて検出される RNA の 5' 末端情報には、エンドヌクレアーゼにより切断された直後の RNA の 5' 末端に加え、5' – 3' のエキソヌクレアーゼによる分解途中の RNA の 5' 末端が存在すると考えられている。具体的な報告例としては、RNA 上に存在する exon junction complex (EJC) 付近で、エキソヌクレアーゼによる消化が停止することによって検出される RNA 保護断片が挙げられる (44)。EJC は核内でのスプライシングの際に、エキソン-イントロンジャンクションの 20-25 塩基上流に配置され、細胞質でのリボソームの翻訳伸長に伴い RNA から取り除かれる。先行研究において Lee らは、PARE 法を用いて網羅的に切断部位情報を取得し、EJC 領域周辺の RNA 末端の分布を算出している (44)。その結果、検出されたリード数が EJC 領域 (-28 位) で顕著に高かった一方で、エキソヌクレアーゼの変異体では、野生型と比較し、検出されたリード数は減少していた。これらの結果を基に Lee らは、EJC により保護された RNA 断片が次世代シーケンサーを用いた解析により検出できる可能性を主張している。しかしながら、Lee らが用いた 5' – 3' エキソヌクレアーゼの変異体では、結果として転写開始点付近で検出されるリード数が増加するため (44)、EJC 領域での検出リード数が過少評価されている可能性が考えられた。そこで、同じシーケンスデータを GEO データベースより取得し、再解析を行った (図 1-14A)。エキソヌクレアーゼの変異体としては、*xrn4-6*、および *fry1-6* を使用している。再解析では、Lee らの解析のように各ライブラリーで検出された全リード数に対する各部位での検出リード数の存在比率ではなく、EJC 領域付近 (エキソンの 3' 末端から 50 塩基) で検出された全 RNA 末端数に対する各部位での RNA 末端の存在比率を算出した。その結果、*xrn4* の変異体では EJC 領域 (-28 位) で検出された末端数は野生型と比較し、わずかに減少していたが、同じエキソヌクレアーゼ活性を持つ *fry1-6* では、野生型よりも高蓄積している傾向が認められた。この結果は、Lee らの主張とは異なり、EJC 領域での RNA 断片の高蓄積は 5' - 3' のエキソヌクレアーゼ消化に由来しないことを示している。

また、Lee らが用いた PARE 法ではポリ A 鎖付き RNA を濃縮しているが、TREseq 法で取得した切断部位情報を用いた際、図 1-14A のように EJC 領域に RNA 末端が他領域と比較し顕著に検出される傾向は認められなかったことから (図 1-14B)、ポリ A 鎖付き RNA を濃縮した場合でのみ EJC 周辺で RNA 末端が高蓄積する可能性が考えられた。他の先行研究で Nagarajan らは、ポリ A

鎖が付加されていない RNA の分解末端情報を取得するために、Poly (A) RNA セレクション法を用いて、poly A 鎖が付加された RNA (with poly A)とポリ A 鎖が付加されていない RNA (without poly A) に分画し、ライブラリーを作製後、PARE 法を用いて網羅的に RNA 末端情報を取得している (45)。同じシーケンスデータを GEO データベースより取得し、再解析を行った結果、EJC 領域での RNA 末端の高蓄積は、ポリ A 鎖付き RNA を濃縮した場合でのみ検出された (図 1-14B)。ランダムプライマーを使用している TREseq 法では、EJC 領域での RNA 末端の高蓄積は消失していることから、このような傾向は、ポリ A 鎖付き RNA でのみ検出される RNA 断片、もしくは、Poly (A) RNA セレクション法に由来するシーケンスアーティファクトである可能性が考えられた。これらの結果は、網羅的な分解産物解析において検出できる RNA 末端の中で、5' – 3' のエキソヌクレアーゼによる消化途中に由来する RNA の 5' 末端は少ないことを示している。

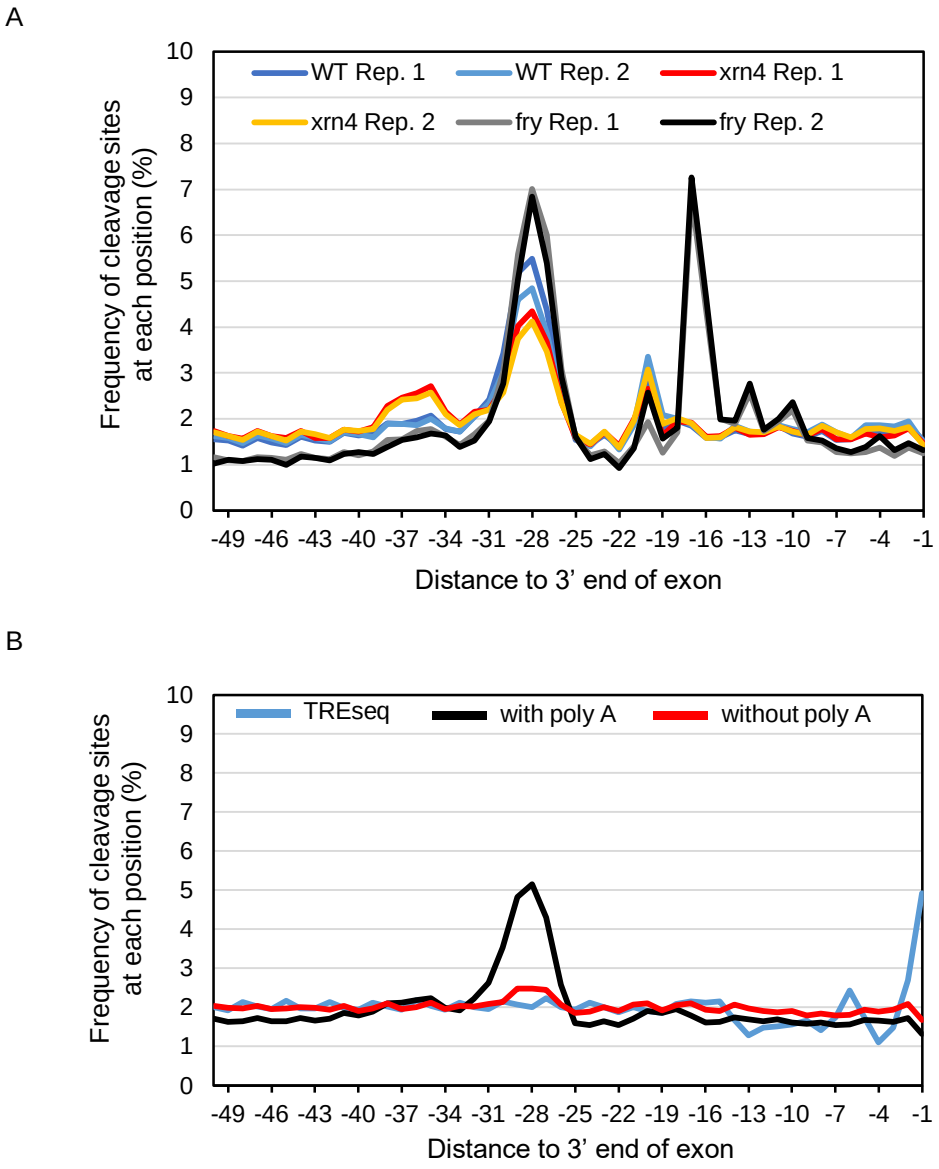


図 1-14. EJC 周辺の切断部位の分布

PARE 法を用いて検出された網羅的な RNA 末端情報を GEO データベースより取得し、EJC 領域周辺の各位置における RNA の 5' 末端の存在比率を新たに算出した (A)。WT Rep. 1; 野生型の反復 1、WT Rep. 2; 野生型の反復 2、*xrn4* Rep. 1; エキソヌクレアーゼの変異体の反復 1、*xrn4* Rep. 2; エキソヌクレアーゼの変異体の反復 2、*fry* Rep. 1; エキソヌクレアーゼの変異体の反復 1、*fry* Rep. 2; エキソヌクレアーゼの変異体の反復 2 を示す。また、Nagarajan らが Poly (A) RNA セレクション法を用いて、RNA を poly A^+ (with poly A)、poly A^- (without poly A) に分画し、作製したライブラリーを用いて検出した網羅的な RNA 末端情報に関しても GEO データベースより取得し、TREseq 法との比較を行った (B)。X 軸はエキソンの 3' 末端からの距離を示し、Y 軸は各位置で検出された RNA 末端の存在比率を示す。

1-4-2. 遺伝子単位の切断率と RNA 半減期の関係性

これまで、動物細胞において RNA 切断のされやすさと半減期に着目した解析が行われてきたが、両者に関係性は認められていなかった (22, 42)。この原因としては、切断部位情報を取得する際に、ポリ A 鎖付き RNA の濃縮を行っていることが理由として挙げられる。Weinberg らが行った RNAseq 法を用いた解析から、ポリ A 鎖付き RNA の濃縮は、RNA の 3' 末端側の切断部位を過大評価するだけではなく、RNA 長に依存して検出効率に偏りを生させていた (30)。そのため、従来の網羅的な切断部位解析では、切断のされやすさを正確に算出できず、半減期との間に関係性が認められなかったと考えられる。一方で、ポリ A 鎖付き RNA を濃縮せず、rRNA 除去法を用いてライブラリーを作製する TREseq 法では、これらの問題点を改善できることから、今回の解析では、切断されやすい RNA ほど半減期が短い傾向が認められたと考えられた。これらの結果は、シロイヌナズナにおいて RNA 切断に依存する分解機構が RNA 安定性を調節する重要な機構の一つであることを示している。

1-4-3. TREseq 法で得られた切断部位情報の有効性

従来の手法と比較し、アダプター付加効率、Cap-less RNA の濃縮率、検出される切断部位の偏りの改善など、より正確な切断部位の検出、切断率の算出が可能になったことにより、本研究において初めて RNA 切断と RNA 安定性との関係性が認められた (図 1-12, 図 1-13)。加えて、1-3-1 に示すように Allen、Yoshikawa らが報告した microRNA の切断部位 (39, 40) と同じ位置で TREseq 法でも切断部位が検出されていることや、異なる手法 (ナノポアシーケンサー) を用いた場合でも同様の位置に切断部位が存在していたことからも、TREseq 法で検出された切断部位はライブラリーに由来するシーケンスアーティファクトではなく、実際に細胞内に存在している切断部位であることを示している。

これまでの網羅的な分解産物解析が植物を対象として行われ、主に microRNA のターゲットサイトに着目し解析が行われてきたが、大部分の RNA 切断は microRNA が関与しないことが示唆されている (18, 19)。加えて、各切断部位の切断のされやすさを数値化し、これらの切断に関わる配列などについて詳細な解析はこれまで行われていない。TREseq 法を用いて算出した、より正確な切断率を用い、切断率が高い配列、低い配列に着目し RNA 切断に関わる特徴を解析することで、これまで不明であった植物 RNA 切断機構の全体像への理解が深まると考えられる。

第二章

植物 RNA 切断に関わる特徴の解析

2-1. 序論

これまで RNA 切断に関わる要因は、酵母を対象に個別遺伝子を中心に解析が行われ、<1> RNA の配列、 <2> RNA 高次構造、<3> 翻訳状態、などが RNA 切断に関与することが示唆されている。

<1> RNA を切断するタンパク質因子は、大腸菌やウィルスを中心に複数同定されており、認識される RNA 配列には異なる特徴があることが明らかとなっている。例えば、ウィルスでエンドヌクレアーゼとして機能する SOX タンパク質は、比較的長い配列モチーフを認識し、C もしくは U 塩基の直前で RNA を切断することが知られている (46)。また、原核生物で広く保存されている RelE は、アミノ酸飢餓状態において G 塩基の直前で RNA を切断する (6)。このように、RNA 切断には RNA 上の配列が大きく関与することが考えられる。

<2> また、このような RNA 切断には配列だけではなく、配列に依存して形成される RNA 高次構造が関与する場合もある。例えば、tRNA のプロセシングに関わる tRNase は、ステムループ構造を認識し pre-tRNA を切断することが動物において報告されている (47)。ショウジョウバエでは、microRNA のプロセシングに関わる Drosha による切断の位置には、primary RNA のステムループ構造の大きさが関与していることが知られており (48)、RNA の高次構造も RNA 切断に関与している。

<3> 加えて、これらの RNA 切断は、配列、高次構造だけではなく、翻訳状態 (RNA 上のリボソームの存在位置や存在量) が関与する可能性もある。RNA 内での切断部位の分布に着目した場合に、終止コドンの周辺で切断部位のピークが検出されることや、CDS 内に 3 塩基単位の切断部位の周期性が認められることがこれまでの解析から明らかとなっている (18, 23)。これらの傾向は、RNA 上でのリボソームの翻訳伸長パターンと類似することから、リボソームの存在位置や存在量が RNA 切断に関与することが示唆されていた。

このように、個別遺伝子を中心にこれまで RNA 切断に関わる特徴が解析されてきたが、植物を対象とした網羅的な切断部位解析においては、microRNA などのターゲット配列が主に着目され、microRNA が関与しない内部切断に関わる配列的な特徴など詳細な知見は少なかった。また、従来手法では、ポリ A 鎖付き RNA を濃縮していたため、検出される切断部位が RNA の 3' 末端に偏るなど、各切断部位での正確な切断率が算出できず、特に切断されやすい (切断

率が高い) 部位に着目した特徴解析など、各特徴が RNA 切断に与える影響を正確に評価することができなかった。加えて、RNA 上での切断部位の分布から翻訳過程が RNA 切断に関与することが示唆されていたが、実際にリボソームの存在位置や存在量情報を取得できるリボソームプロファイリング法を行い、切断部位の位置、切断率との関係性に着目した詳細な解析は植物では行われていない。上記のような理由により、これまで個別遺伝子を中心に RNA の切断に関わる要因が複数挙げられているが、実際にどのような要因が植物での RNA 切断に主体的に関与しているかは不明であった。

そこで本研究の第二章では、第一章で得られたシロイヌナズナにおける各切断部位の切断率に関するデータを基に、特に切断率が高い配列、低い配列などの場合分けを行い、複数の要因が RNA 切断に与える影響に着目し解析を行った。加えて、酵母、ショウジョウバエなどについても同様の解析を行い、シロイヌナズナで得られた結果と比較することで、RNA 切断に関わる特徴の生物種間での保存性について考察を行った。

2-2. 材料と方法

2-2-1. 実験材料および培養条件

2-2-1-1. シロイヌナズナ培養細胞（*Arabidopsis thaliana* T87）

第一章の 1-2-1 と同様に行った。

2-2-1-2. ショウジョウバエ培養細胞 (*Drosophila melanogaster*; S2-R+)

ショウジョウバエ培養細胞 (S2-R+) は、奈良先端科学技術大学院大学岡村勝友教授より分与していただいたものを使用した。培養は 25ºC 、10% FBS、100 U/ml ペニシリン-ストレプトマイシンを含んだシュナイダー培地を用いた。

2-2-1-3. 出芽酵母 (*Saccharomyces cerevisiae*; sigma1278b)

出芽酵母 (sigma1278b 株) は、奈良先端科学技術大学院大学高木博史教授より分与していただいたものを使用した。培養は 30°C 、2% (wt/vol) グルコース、 1% (wt/vol) Difco Bacto yeast extract (Thermo Fisher Scientific)、 2% (wt/vol) Difco Bacto peptone (Thermo Fisher Scientific)を含んだ YPD 培地を用いた。

2-2-2. Truncated RNA end sequencing (TREseq)

2-2-2-1. 細胞回収、RNA 抽出

第一章の 1-2-2-1 と同様に行った。

2-2-2-2. ライブラリー作製、データ解析

Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina) を用い Total RNA から rRNA を除去後、1-2-2-2 と同様にライブラリーを作製した。その後、Illumina NextSeq 500 (Illumina) に供した。マッピングに関しては、各生物種のゲノム情報 (シロイヌナズナ; TAIR10、ショウジョウバエ; FlyBase, 出芽酵母; Saccharomyces Genome Database) を用い、第一章の 1-2-2-3 と同様のデータ解析を行った。

2-2-3. リボソームプロファイリング法

2-2-3-1. 細胞回収

培養 3 日目のシロイヌナズナ培養細胞を回収し、液体窒素で凍結した。凍結したサンプルは乳鉢を用いて破砕し、2 ml 容チューブに分注した。

2-2-3-2. RNase I 処理、ポリソーム分画、RNA 抽出

ポリソーム分画に用いた Extraction Buffer やショ糖密度勾配液は、Yamasaki らの方法に従い作製した (35)。RNase I 処理に関しては、Lei らの手法を参考にして行った (49)。サンプル破砕粉末におおよそ 4 倍量 (w/v) の Extraction Buffer (200 mM Tris-HCl, pH8.5, 50 mM KCl, 25 mM $MgCl_2$, 2 mM EGTA, 100 µg/ml heparin, 100 µg/ml cycloheximide, 2% polyoxyethylene 10-tridecyl ether, 1% sodium deoxycholate) を加え、緩やかに懸濁した。遠心 (14,000 × g, 15 min, 4℃) により細胞残さを除き、さらに遠心 (14,000 × g, 10 min, 4°C) し、その上清を RNA 粗抽出液とした。RNA 粗抽出液に RNase I を 6 µl 加え、室温で 30 min 処理後、RNase I inhibitor (Thermo Fisher Scientific) を 10 µl 加えた。予め作製した 26.25-71.25%ショ糖密度勾配液 (ショ糖, 200 mM Tris-HCl, 200 mM KCl, 200 mM $MgCl_2$) 4.85 ml 上に 300 µl 重層し、超遠心を行った (SW55Ti rotor, 55,000 rpm, 50 min, 4°C, brake-off) (Optima, Beckman Coulter, CA, USA)。ピストン・グラジェント・フラクショネーター (BioComp, Row, Canada) によってショ糖密度勾配の上部より 1.4 ml/min の速さで吸引し、2 分画した場合の後半に回収したポリソーム側の RNA を、リボソームにより保護された断片 (ribosome protected fragment: RPF) として取得した。

2-2-3-3. ライブラリー作製、シーケンス

ライブラリーの作製には TruSeq Ribo Profile kit (Illumina) を使用した。手順としては、ribosomal RNA を除去後、3' 末端へのアダプターライゲーション、逆転写反応、cDNA の環状化を行いライブラリーを作製し、NextSeq 500 (Illumina) を用いてシーケンスを行った。

2-2-3-4. データ解析

次世代シーケンサーを用いてシロイヌナズナの RPF 配列情報を網羅的に取得し、アダプター配列を除去した。ショウジョウバエ、出芽酵母については、GEO データベースより RPF 配列情報を取得し、シロイヌナズナと同様にアダプター配列を除去した。RPF については、両末端の位置情報を使用するため、アダプター配列の全長が読まれているリードのみを解析対象とした。TREseq 法の Cap-less RNA と同様に MOIRAI を使用し、各生物種のゲノム情報 (シロイヌナズナ; TAIR10、ショウジョウバエ; FlyBase, 出芽酵母; Saccharomyces Genome Database) を使用した。マッピングソフトについては bwa を用い、ユニークマップのみを解析に使用した。RNA 上の各位置でのリボソームの存在

量を示す指標として、Ribosome occupancy (RO_{site}) 値を算出した。また、各 RNA ごとの RO_{site} 値の合計を RO_{gene} 値とした。

RO_{site} = RNA 上の各部位での RPF 数 / RNA 蓄積量
RO_{gene} = 各 RNA ごとの RO_{site} 値の合計値

2-2-4. 配列モチーフ等の解析

RNA の高次構造に関しては、RNAfold を用いて、各ポジションにおける塩基対の形成を予測し、解析対象とした配列の塩基対の形成度合い (形成頻度) を算出した。また、切断部位周辺配列の分布に関しては、配列上のモチーフの位置を探索する FIMO を使用し、RNA 上の各位置におけるモチーフの出現頻度を算出した。

2-3. 結果

2-3-1. シロイヌナズナにおける microRNA のターゲット配列と切断部位

網羅的な切断部位解析において、既存の microRNA データベースを基に解析を行った場合、microRNA が関与する切断部位は全体のごくわずかであることが従来手法を用いた解析で示されている (18, 19)。そこで、TREseq 法を用いて検出した切断部位についても、Yu らの解析手法を参考に psRNA target を用いて (18)、microRNA のターゲット配列と重複する (microRNA が関与する) 切断部位数を算出した結果、全切断部位の 1.5%が該当した (表 2-1)。また、Yu らの解析では、microRNA のターゲット配列は全切断部位のごくわずかであるが、microRNA のターゲットとなる遺伝子内での他の切断部位と比較して、その切断率は高いことが報告されている (19)。そこで第一章で解析に使用した AT2G39675.1 と AT1G63130.1 の microRNA が関与する切断部位とその遺伝子内で microRNA が関与しない切断部位を比較したところ、microRNA が関与する切断部位は、他の切断部位と比較し、10 倍から 20 倍程度切断率が高かった (表 2-2)。この結果は、Yu らの解析と同様に、TREseq 法で取得した切断部位情報についても microRNA のターゲットとなる遺伝子内では、microRNA が関与する切断部位の切断率は他の切断部位と比較して高いことを示している。

次に、個別遺伝子を対象とした解析に加えて、TREseq 法で検出された全切断部位を使用し、microRNA が関与する切断部位と他の切断部位との切断率の比較を行った。データベースに登録されている microRNA の全候補配列を用い、microRNA が関与する切断部位と TREseq 法で検出された全切断部位の切断率を比較した結果、microRNA が関与しない切断部位の切断率の方が高い傾向が示され (表 2-3, Welch's t-test, $p < 0.01$)、microRNA が関与する切断部位の切断率に対して 1.06 倍であった (表 2-4)。これらの結果については、シロイヌナズナの組織間で microRNA の発現量が異なる可能性が考えられため、TREseq 法を用いて取得した異なる組織での切断部位情報についても解析を行った。当研究室で取得したシロイヌナズナの幼植物体 (発芽 2 日目)、成熟葉 (展開葉)、未成熟用 (未展開葉) の切断部位情報 (bwa によるマッピング) を使用し (50)、解析した結果、異なる組織および組織の発達状態でも全切断部位に対する microRNA が関与する切断部位の割合は 1.5%ほどであり (表 2-1)、microRNA が関与しない切断部位での切断率は microRNA が関与する切断率より高い傾向が認められた (表 2-3, Welch's t-test, $p < 0.01$, 表 2-4)。

データベースに登録されている microRNA のターゲット配列の中には、実際に機能的であるか検証が行われていない配列が存在するため、それらの配列が解析上のノイズとなっている可能性も考えられるが、これらの結果は、

microRNA が関与せず、かつ、microRNA と同等以上の切断率である切断部位が植物 RNA 内に多く存在することを示している。

表 2-1. シロイヌナズナ培養細胞での microRNA が関与する切断部位の割合

	microRNA が関与する切断部位	他の切断部位
培養細胞	30,502 sites (1.54 %)	1,951,833 sites (98.46 %)
発芽 3 日目	20,454 sites (1.55 %)	1,301,861 sites (98.45 %)
未展開葉	20,064 sites (1.53 %)	1,287,339 sites (98.47 %)
展開葉	19,882 sites (1.51 %)	1,294,377 sites (98.49 %)

表 2-2. microRNA ターゲット遺伝子内での切断率の比較 (培養細胞)

	microRNA が関与する切断部位	同じ遺伝子内での他の切断部位
AT1G63130.1	CS_{site} 値 = 0.073	平均 CS_{site} 値 0.003
AT2G39675.1	CS_{site} 値 = 0.025	平均 CS_{site} 値 0.002

表 2-3. 検出された全切断部位を対象とした際の切断率の比較

	microRNA が関与する切断	他の切断部位
培養細胞	平均 CS_{site} 値 = 0.0148	平均 CS_{site} 値 = 0.0157
発芽 2 日目	平均 CS_{site} 値 = 0.0248	平均 CS_{site} 値 = 0.0256
未展開葉	平均 CS_{site} 値 = 0.0255	平均 CS_{site} 値 = 0.0261
展開葉	平均 CS_{site} 値 = 0.0252	平均 CS_{site} 値 = 0.0264

表 2-4. microRNA が関与する切断部位と関与しない切断部位との切断率の比較

	切断率比 (他の切断部位 / microRNA が関与)
培養細胞	0.0157 / 0.0148 = 1.06
発芽 2 日目	0.0256 / 0.0248 = 1.03
未展開葉	0.0261 / 0.0255 = 1.02
展開葉	0.0264 / 0.0252 = 1.04

2-3-2. シロイヌナズナにおける配列的特徴が RNA 切断に与える影響

次に、第一章で算出した切断率を用いて、切断部位周辺の配列的特徴に着目した解析を行った。エキソン領域に存在する切断部位を対象に、切断部位周辺の配列情報を取得し、塩基比率を算出したところ (n = 1,982,335 sites)、図 2-1 に示されるように切断部位の周辺の G 塩基比率が高かった。また、このような配列が切断率に関与しているならば、CS_{site} 値が高い配列では、よりこの傾向が強まるのではないかと考えた。そこで、CS_{site} 値の TOP 10% (n = 198,234 sites) と BOTTOM 10% (n = 198,234 sites) の切断部位を選抜し、その塩基比率を比較したところ、CS_{site} 値の TOP 10%において、より顕著な塩基の偏りが確認された (図 2-1)。これらの結果は、切断部位の周辺で認められた特異的な配列が RNA 切断の位置、および切断のされやすさに関与していることを示している。

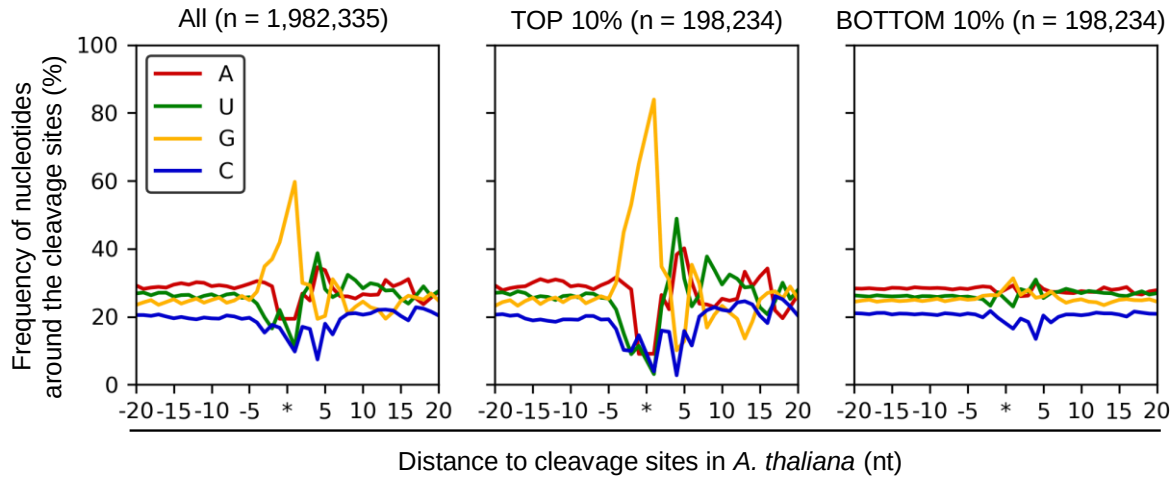


図 2-1. シロイヌナズナにおける切断部位周辺の配列

シロイヌナズナを対象に切断部位周辺の配列情報を取得し、塩基比率を算出した。また、CS_{site} 値を基に場合分けを行った場合の配列に関しても塩基比率を算出した。X 軸は切断部位からの距離を示し、Y 軸は各位置での塩基比率を示す。アスタリスクは切断部位を示す。

2-3-3. シロイヌナズナにおける RNA 高次構造が切断に与える影響

RNA の切断には、その RNA の構造が関与することが tRNA や microRNA のプロセシング過程に着目した解析から明らかとなっている (47, 48)。そこで、TREseq 法で同定した切断部位の前後 30 塩基の配列情報を取得し、RNA 高次構造と切断率に着目した解析を行った。RNA 高次構造の指標値としては、各塩基で塩基対を形成するか、しないかを示す塩基対形成情報 (dot-branket notation) を使用した (図 2-2A)。dot-branket notation では、鍵括弧 (塩基対を形成する)、または点 (塩基対を形成しない) で塩基対の形成の有無を示す。RNAfold を用いて、各塩基での塩基対形成を予測し、CS_{site} 値の BOTTOM 10% (n = 198,234 sites) に対する TOP 10% (n = 198,234 sites) の配列での塩基対の形成頻度 (塩基対の形成度合い) を各位置で算出した。CS_{site} 値の TOP 10%、BOTTOM 10%間の塩基対の形成度合いが同等であった場合、$\log_2$ 対数変換した値は 0 となり、BOTTOM 10%に対して TOP 10%の配列での塩基対の形成度合いが高い場合、$\log_2$ 対数変換した値は正の方向に高い値を示す。図 2-2B に示されるように切断率が高い配列で切断部位周辺の塩基対の形成度合いは高く、下流で低い傾向が認められた。

一方で、RNA の高次構造は配列に依存しているため、切断部位周辺の各塩基の比率も算出し、塩基対形成情報との比較を行った (図 2-3)。塩基比率に関しては、切断部位の前後 30 塩基の配列情報を使用し、CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基比率を算出した。切断部位周辺の相対的な塩基比率と切断部位周辺の塩基対の形成度合いを比較すると、特に G 塩基の比率と類似する傾向が認められた (図 2-3)。このことから、切断部位周辺において RNA 高次構造が形成されやすいほど切断率が高く、その構造の形成は切断部位周辺の G 塩基の比率に依存していると考えられた。しかし、これらの結果のみでは、RNA 構造が切断に直接的に関わっているか、それとも配列上の G 塩基比率が高い結果として塩基対の形成度合いが高く算出されてしまっているかは判断できないため、形成される RNA 構造が類似しているかなど異なる解析法での評価も今後は必要であると考えられた。

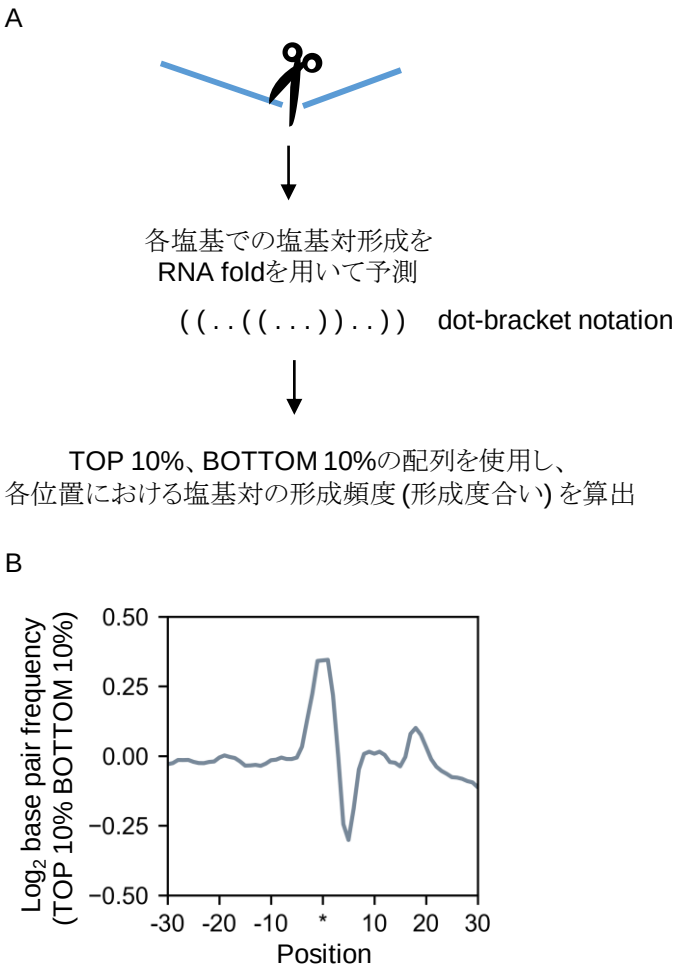


図 2-2. 切断部位周辺の塩基対の形成度合い（シロイヌナズナ）

切断部位周辺の塩基配列を取得し、RNAfold を用いて塩基対が形成される箇所を予測した (A)。CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成頻度 (形成度合い) を算出した (B)。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は解析対象とした CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成度合いを $\log_2$ 対数変換した値を示す。塩基対形成を示す dot-bracket notation において、 鍵括弧は塩基対を形成していることを示し、点は塩基対を形成していないことを示す。

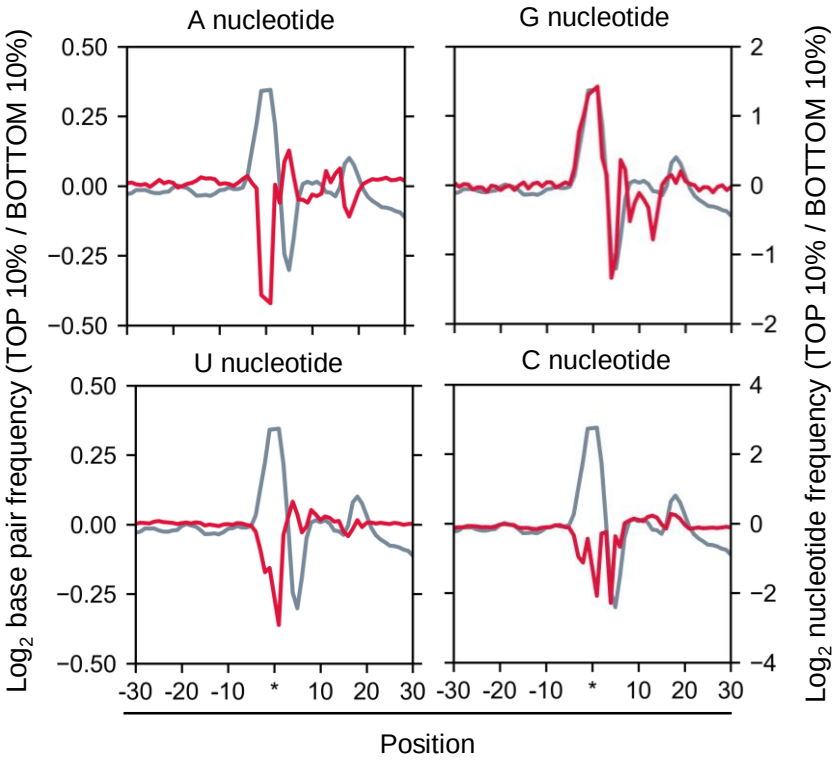


図 2-3. 切断部位周辺の塩基対形成の度合いと塩基比率（シロイヌナズナ）

CS_{site} 値の TOP 10%、BOTTOM10%の配列情報を取得し、A 塩基、G 塩基、U 塩基、C 塩基比率を $\log_2$ 対数変換し、図 2-2 と統合した。灰色は塩基対の形成度合いを示し、赤色は各塩基の比率を示す。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は解析対象とした CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成度合いと塩基比率を $\log_2$ 対数変換した値を示す。

2-3-4. シロイヌナズナにおける翻訳過程と RNA 切断との関係性

2-3-4-1. 開始、終止コドン周辺における切断部位の分布

従来手法を用いた解析では、切断部位の RNA 内での分布に着目すると、終止コドンの周辺に多くの切断部位が存在し、CDS 領域内では 3 塩基単位の周期性が認められている (18, 19)。これらの傾向は、リボソームプロファイリング法を用いて検出されるリボソームの RNA 内での分布と類似しているため、翻訳過程が RNA 切断に関与することが Yu らや、Hou らによる研究で示唆されている (18, 19)。そこで、今回行った TREseq 法でも検出された切断部位に同様の分布が認められるかどうか解析を行った。開始、終止コドンからの距離を算出し、RNA 内での各位置の切断部位の存在比率を調べてみると、Yu らの先行研究と類似するように終止コドンの周辺にて切断部位のピークが認められた (図 2-4)。加えて、従来手法の結果では認められなかった開始コドン付近でも切断部位のピークが検出された。また、CDS 領域の 3 塩基単位の周期性に着目すると、終止コドンに加え、開始コドンの周辺についても周期性が認められた (図 2-5)。TREseq 法では、終止コドンに加え、開始コドン周辺でも切断部位のピーク、CDS 領域の 3 塩基単位の周期性が認められたことについては、検出される切断部位の 3' 末端側への偏りを大幅に改善したことが理由として挙げられる。

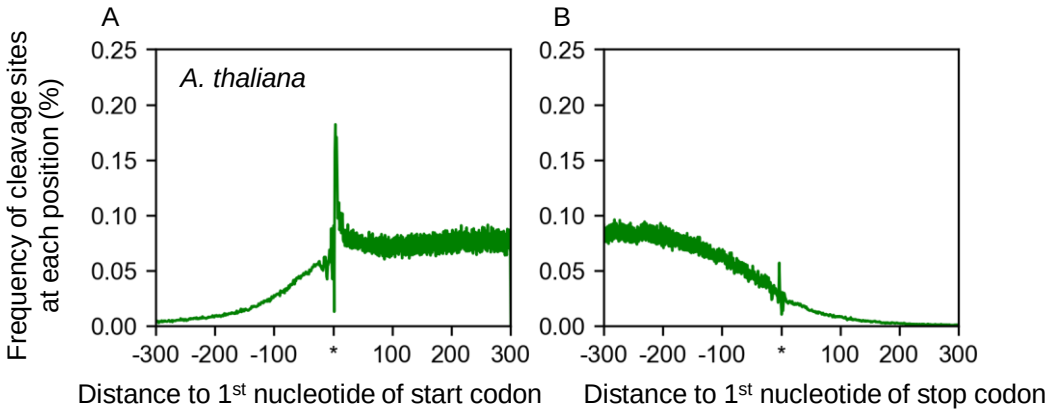


図 2-4. 開始コドン、終止コドン周辺の切断部位の分布（シロイヌナズナ）

開始コドン、終止コドンからの距離を算出し、各位置での切断部位の存在比率を算出した (A, B)。X 軸は、開始コドン、終止コドンからの距離を示す。Y 軸は、各位置での切断部位の存在比率を示す。

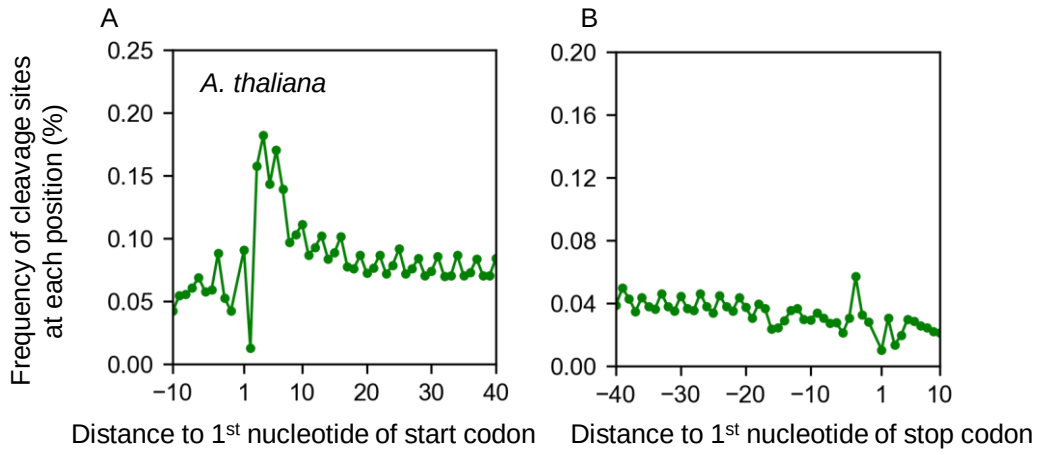


図 2-5. 図 2-4 の拡大図（シロイヌナズナ）

2-3-4-2. リボソームの位置が切断部位の決定に与える影響

これまで、網羅的な切断部位解析から、翻訳過程が RNA 切断に関与することが植物において示唆されていたが (18, 19)、実際に RNA 上のリボソームの位置や存在量に関する情報を取得し、RNA 切断との関係性に着目した研究は植物で報告されていない。そこで本解析では、TREseq 法に用いた同じシロイヌナズナ培養細胞を対象としてリボソームプロファイリング法を行い、リボソームの存在位置、存在量に関する情報を取得し、翻訳過程が切断部位の位置、切断率に与える影響に着目し解析を行った。リボソームプロファイリング法では、Total RNA を抽出後、RNase I で処理をすることで、リボソームが存在する領域は、RNase I による分解から保護される。したがって、これらの保護された断片 (Ribosome Protected Fragment: RPF) に着目し解析を行うことで、リボソームの存在位置、存在量を推定することができる。Weinberg らの解析法を参考に、アダプター配列の除去、マッピングを行い、リボソームに保護された断片の 5' 末端に関する情報を取得し、5' RPF と定義した (30)。その後、RNA 内での 5' RPF の分布を算出した結果、開始コドン、終止コドンの周辺に 5' RPF のピークが検出され、CDS 領域内で 3 塩基単位の周期性が認められた (図 2-6, 図 2-7)。しかし、実際に RPF の 5' 末端と切断部位の分布を比較すると、双方に 3 塩基単位の周期性は認められたが、5' RPF と切断部位の位相が異なるなど、両者に強い位置的な関係性は認めらなかった (図 2-8)。そこで次に、CDS 領域に存在する切断部位の位置を基準とし、周辺の 5' RPF の存在量を算出した。リボソームの存在位置が切断部位の位置決定に重要であるならば、切断部位の周辺に 5' RPF の顕著な偏りが予想される。しかし、実際に 5' RPF の存在比率を調べてみると、切断部位周辺で 5' RPF の存在比率が顕著に高いわけではなく、リボソームの存在位置が切断部位の位置決定に大きく関与する傾向は認められなかった (図 2-9)。一方で、切断部位付近ではわずかに 5' RPF の存在比率が低くなっており、これは切断された RNA の 5' 末端付近にリボソームが存在できないことに起因したと思われる。

また、開始、終止コドン領域に着目し、リボソーム存在量が切断部位の位置に与える影響を解析した。各部位でのリボソーム存在量として、5' RPF 数を RNA 蓄積量で除算した値を RO_{site} 値と定義した。RO_{site} 値が高いほど、RNA 上の任意の部位でのリボソーム存在量が高いことを示す。各遺伝子の開始、終止コドンの前後 50 塩基以内に存在する RO_{site} 値の平均を算出し、リボソーム存在量がが多い遺伝子の TOP 20% (開始コドン n = 2.368 genes, 終止コドン n = 2,185 genes) とリボソーム存在量が少ない遺伝子の BOTTOM 20% (開始コドン n = 2.368 genes, 終止コドン n = 2,185 genes) 間で切断部位の分布を比較したが、図

2-10A、図 2-10B に示されるように両者で大きな違いは認められなかった。加えて、開始コドン周辺の-10 位から+40 位、もしくは、終止コドン周辺の-40 位から+10 位の各位置での切断部位の存在比率について (図 2-10A, 図 2-10B) リボソーム存在量が多い遺伝子の TOP 20%とリボソーム存在量が少ない遺伝子の BOTTOM 20%間のピアソンの積率相関係数を求めたところ、開始コドン側では r = 0.95 (図 2-10C)、終止コドン側では r = 0.91 となり、共に高い正の相関を示し、リボソームの存在量に関わらず切断部位が存在していた (図 2-10D)。これらのことから、リボソーム存在量が切断部位の分布 (位置) に与える影響は小さいと考えられた。

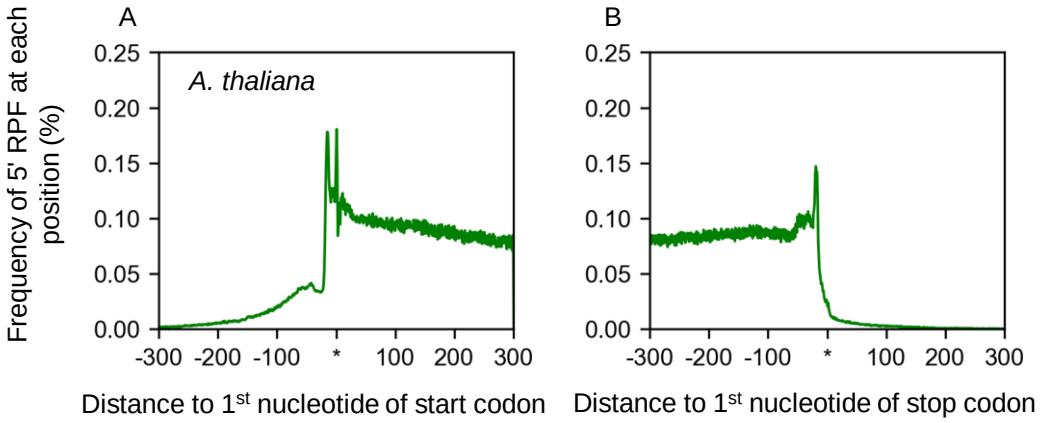


図 2-6. 開始コドン、終止コドン周辺のリボソームの分布（シロイヌナズナ）

開始コドン、終止コドンからの距離を算出し、各位置での 5' RPF の存在比率を算出した (A, B)。X 軸は、開始コドン、終止コドンからの距離を示す。Y 軸は、各位置での 5' RPF の存在比率を示す。

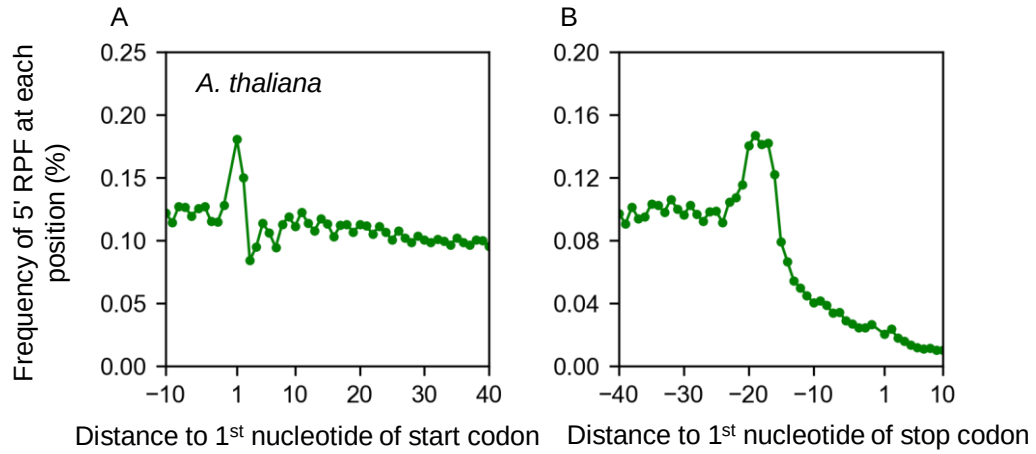


図 2-7. 図 2-6 の拡大図（シロイヌナズナ）

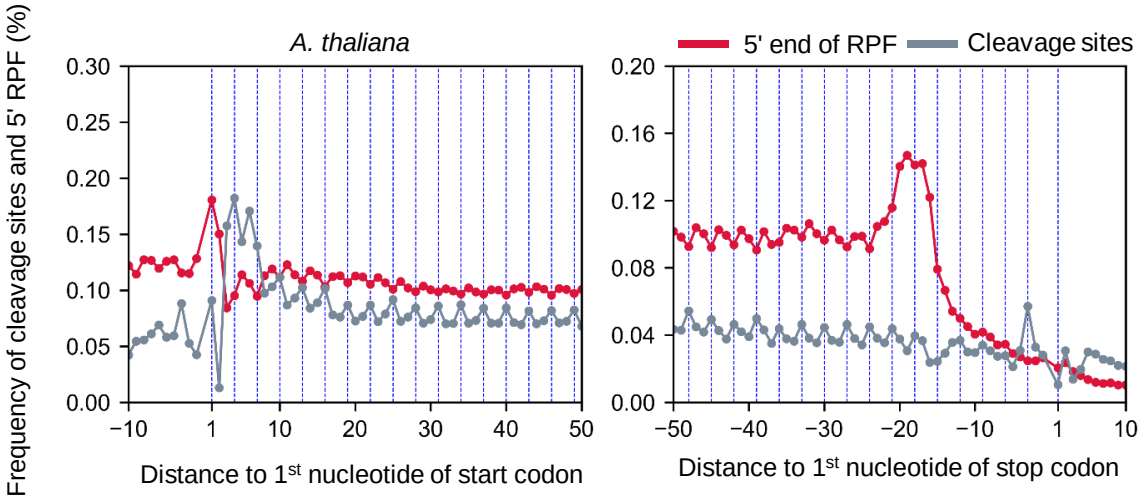


図 2-8. 開始、終止コドン周辺の切断部位とリボソームの分布（シロイヌナズナ）

開始、終止コドンからの距離を算出後、各位置での切断部位、5' RPF の存在比率を算出した。X 軸は、開始コドン、終止コドンからの距離を示す。Y 軸は、各位置での切断部位と 5' RPF 末端の存在比率を示す。

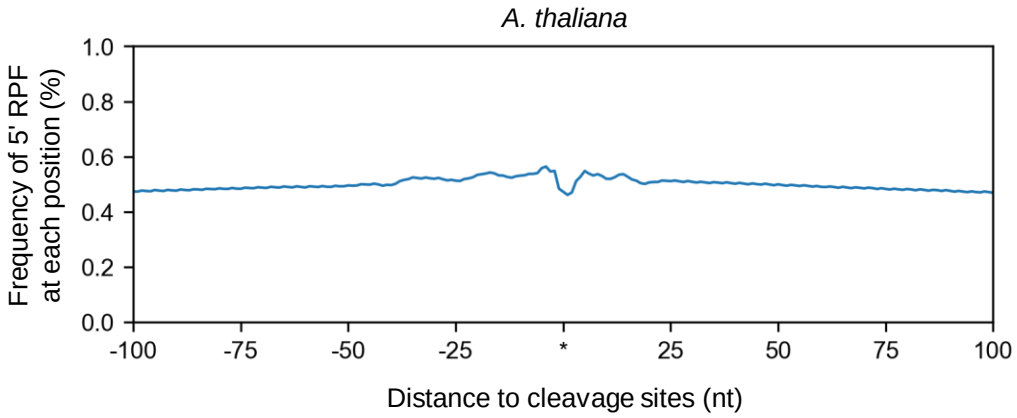


図 2-9. 切断部位周辺の 5' RPF の存在（シロイヌナズナ）

切断部位からの距離を算出後、各位置での 5' RPF の存在比率を算出した。X 軸は、切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は、各位置での 5' RPF 末端の存在比率を示す。

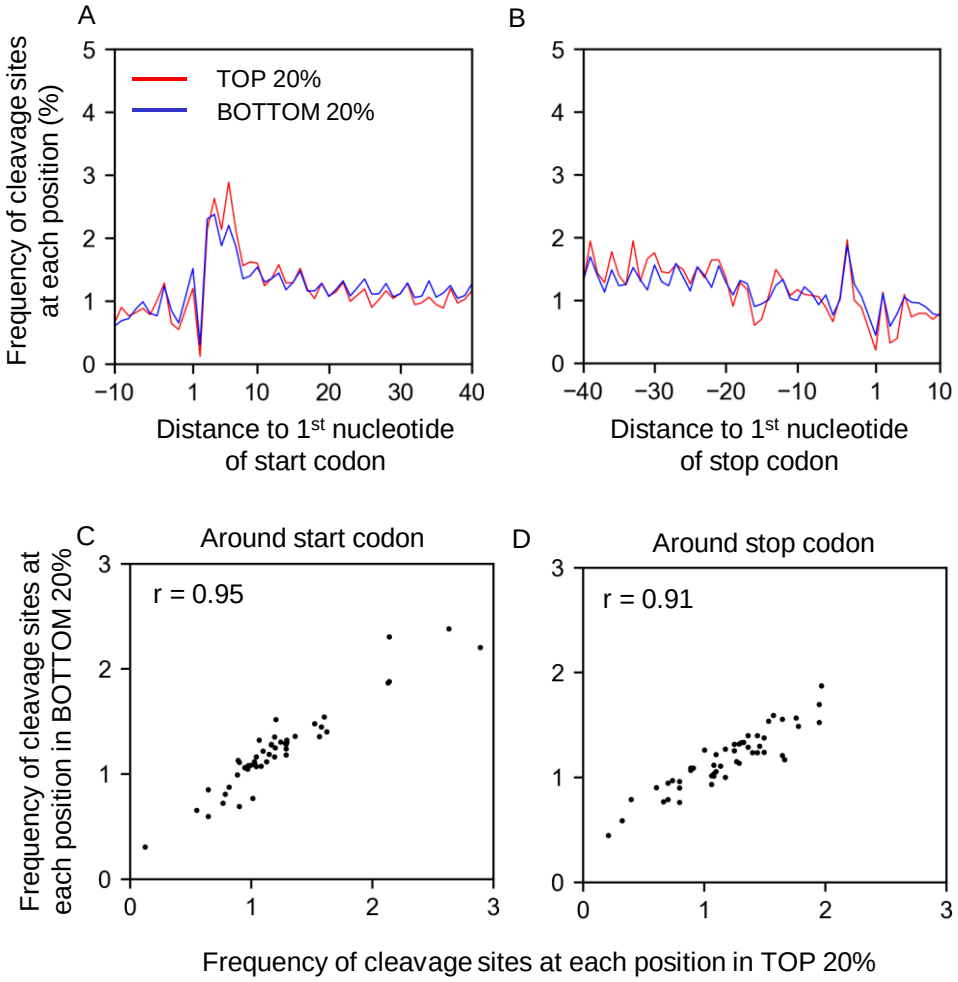


図 2-10. リボソーム存在量が切断部位の分布に与える影響（シロイヌナズナ）

開始、終止コドンの前後 50 塩基以内に存在するリボソーム存在量を基に、TOP 20%、BOTTOM 20%の遺伝子を選抜した。その後、開始、終止コドンからの距離を算出し、各位置での切断部位の存在比率を算出した (A, B)。また、TOP 20%、BOTTOM 20%間での開始コドン (C)、終止コドン (D) 周辺における各位置での切断部位の存在比率のピアソンの積率相関係数を算出した。

2-3-4-3. リボソーム存在量が切断率に与える影響

2-3-4-2 で示したように、翻訳過程が切断部位の位置決定に与える影響は小さいと考えられた。一方で、Pelechano らの研究では、PARE 法と同様の手法である 5Pseq 法を用いて網羅的に RNA 末端を同定し、通常条件下と酸化ストレス条件下で遺伝子単位での切断のされやすさとリボソームの存在量について解析が行われており、通常条件下など単一条件下では両者に関係性は認められなかったが、両条件間での遺伝子単位の切断のされやすさとリボソームの存在量の変動には、正の関係性が酵母で認められている (23)。そこで、リボソーム存在量が切断率に与える影響に着目し、シロイヌナズナを対象に解析を行った。まず、遺伝子単位のリボソーム存在量が遺伝子単位での切断のされやすさの指標値である CS_{gene} 値に与える影響に着目した。遺伝子単位のリボソーム存在量として、RO_{site} 値を遺伝子ごとにまとめた値を RO_{gene} 値と定義した。一般的に、遺伝子単位のリボソーム存在量と翻訳されるタンパク質の量には正の相関が認められることから、RO_{gene} 値が高いほど、翻訳効率は高いと考えられる (51)。図 2-11 に示されるように、CS_{gene} 値と RO_{gene} 値とのピアソンの積率相関係数は $r = 0.67$ となり正の相関が認められ ($n = 12{,}303$ genes)、RNA 上のリボソーム量が多いほど切断されやすい傾向が認められた。Pelechano らの研究で、単一条件下で遺伝子単位での切断のされやすさとリボソーム存在量に正の相関が認められなかった理由については、ポリ A 鎖付き RNA を濃縮することで、RNA 長に依存した RNA 末端の検出効率に偏りが生じたことが考えられる (30)。また、Pelechano らの解析では、酵母を解析対象としていたため、RNA 上のリボソーム存在量が切断率に与える影響は生物種によって異なる可能性も考えられた。

次に、同様の解析を各切断部位単位で行った。CDS 領域に存在する任意の切断部位の前後 50 塩基に存在する RO_{site} 値の平均を算出し、平均 RO_{site} 値の TOP 10% ($n = 174{,}355$ sites)、BOTTOM 10% ($n = 174{,}355$ sites) 間の各切断部位での切断率を比較した結果、切断部位周辺の平均 RO_{site} 値が高い (リボソーム存在量が多い) ほど、CS_{site} 値が高い (切断されやすい) 傾向が認められた (図 2-12A, Welch's t-test, $p < 0.01$)。この切断部位周辺のリボソーム存在量を考えた際に、そもそもの RNA の 5' 末端へのリボソームリクルート (翻訳の開始) 効率が高いため結果として任意の部位でのリボソーム存在量が高くなる、もしくは切断部位周辺でリボソームの翻訳伸長速度が遅くなっている (停止、停滞) 可能性が考えられた。そこで、リボソームの停滞されやすさを概算するために、RO_{gene} 値に対する RO_{site} 値 (RO_{site} / RO_{gene} 値) を算出した。RO_{site} / RO_{gene} 値が高いほど、解析対象とした RNA にリクルートされるリボソーム量に対して、

任意の部位でのリボソーム存在量が高く、その部位での滞在時間も長いと考えられるため、リボソームの翻訳伸長速度が遅くなっていることを示す。図 2-12A と同様に切断部位の前後 50 塩基の平均 RO_{site} / RO_{gene} 値を算出し、平均 RO_{site} / RO_{gene} 値の TOP 10%、BOTTOM 10%間の切断率を比較したが、平均 RO_{site} 値の結果と比べ (図 2-12A)、切断率の差は大きくはならなかった (図 2-12B)。この結果は、各切断部位周辺のリボソーム量が多いほど切断率は高く、そのリボソーム存在量は、停止、停滞に依存せず、そもそもの翻訳の開始効率に依存していることを示唆している。加えて、これらの結果は、翻訳過程 (リボソームの存在位置や存在量) は切断部位の位置決定には大きな影響を与えないが、各切断部位の切断のされやすさには正の影響を与えることを示している。

また、これらの解析に加え、各切断部位周辺に存在するリボソーム量が異なる場合に、切断部位周辺の配列は異なるか否かについても検証を行ったが、平均 RO_{site} 値の TOP 10%、BOTTOM 10%間で切断部位周辺の塩基比率に大きな違いは認められなかった (図 2-12C, 図 2-12D)。この結果は、切断部位周辺のリボソーム存在量は配列に依存せず、切断率に正の影響を及ぼすことを示唆している。

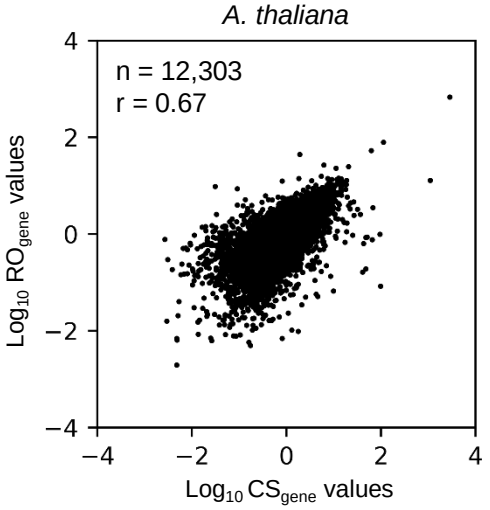


図 2-11. 遺伝子単位でのリボソーム存在量と切断率（シロイヌナズナ）

TREseq 法で解析対象とした遺伝子の内、リボソームプロファイリング情報を持つ遺伝子を対象に RO_{gene} 値と CS_{gene} 値とのピアソンの積率相関係数を算出した。

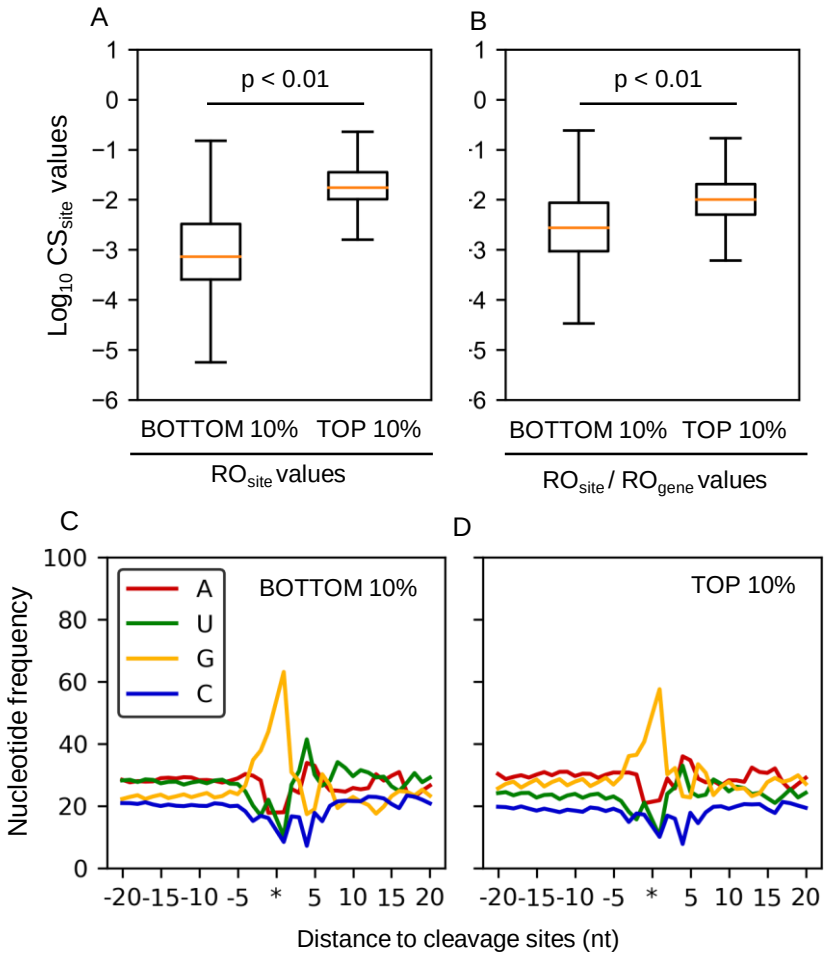


図 2-12. 切断部位周辺のリボソーム密度と切断率（シロイヌナズナ）

切断部位の前後 50 塩基での平均 RO_{site} 値を算出し、平均 RO_{site} 値の TOP 10%、BOTTOM 10%間の切断率を比較した (A)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。また、RO_{site} 値を RO_{gene} 値で除算した場合の RO_{site} / RO_{gene} 値についても算出し、切断部位の前後 50 塩基での平均 RO_{site} / RO_{gene} 値の TOP 10%、BOTTOM 10%間の切断率についても比較を行った (B)。また、平均 RO_{site} 値の TOP 10%、BOTTOM 10%間の切断部位周辺の塩基比率についても比較を行った (C, D)。統計検定には Welch's t-test を使用した。

2-3-5. シロイヌナズナにおける切断部位周辺の配列と RNA 内での切断部位の分布

切断部位の RNA 内での分布に着目すると、開始、終止コドンの周辺にピークが検出され、CDS 領域で 3 塩基単位の周期性が見られた (図 2-8)。一方で、RNA 上の塩基比率の分布についても調べた結果、同様の周期性が認められた (図 2-13, 図 2-14)。この結果を踏まえると、切断部位の分布は RNA 上の特徴的な配列の分布に依存していることが想定された。そこで、切断部位周辺の上流 5 塩基、下流 15 塩基を MEME の motif letter-probability matrix lines 形式 (配列モチーフ) に変換し (図 2-15A, 図 2-15B)、FIMO を用いて RNA 内での配列モチーフの分布を算出した。また、この配列モチーフの 5 塩基と 6 塩基目で切断が生じることから、算出した分布を 5 塩基シフトさせプロットした (図 2-15C, 図 2-15D)。その結果、切断部位周辺の配列モチーフは、開始コドンの周辺で存在比率が高く、CDS 内で 3 塩基単位の周期性を示し、切断部位と同じ位相であった。これらの結果は、切断部位の分布 (位置) はリボソームの存在位置や存在量ではなく、特異的な配列パターンに依存して決定されることを示している。一方で、配列モチーフは終止コドンの周辺でも CDS 内での 3 塩基単位の周期性を同じ位相で示したが、切断部位のピーク (-3 位) の位置で配列モチーフのピークは認められなかった (図 2-15D)。これについては、今回使用した配列モチーフとは異なる配列パターンが関与している可能性も考えられた。

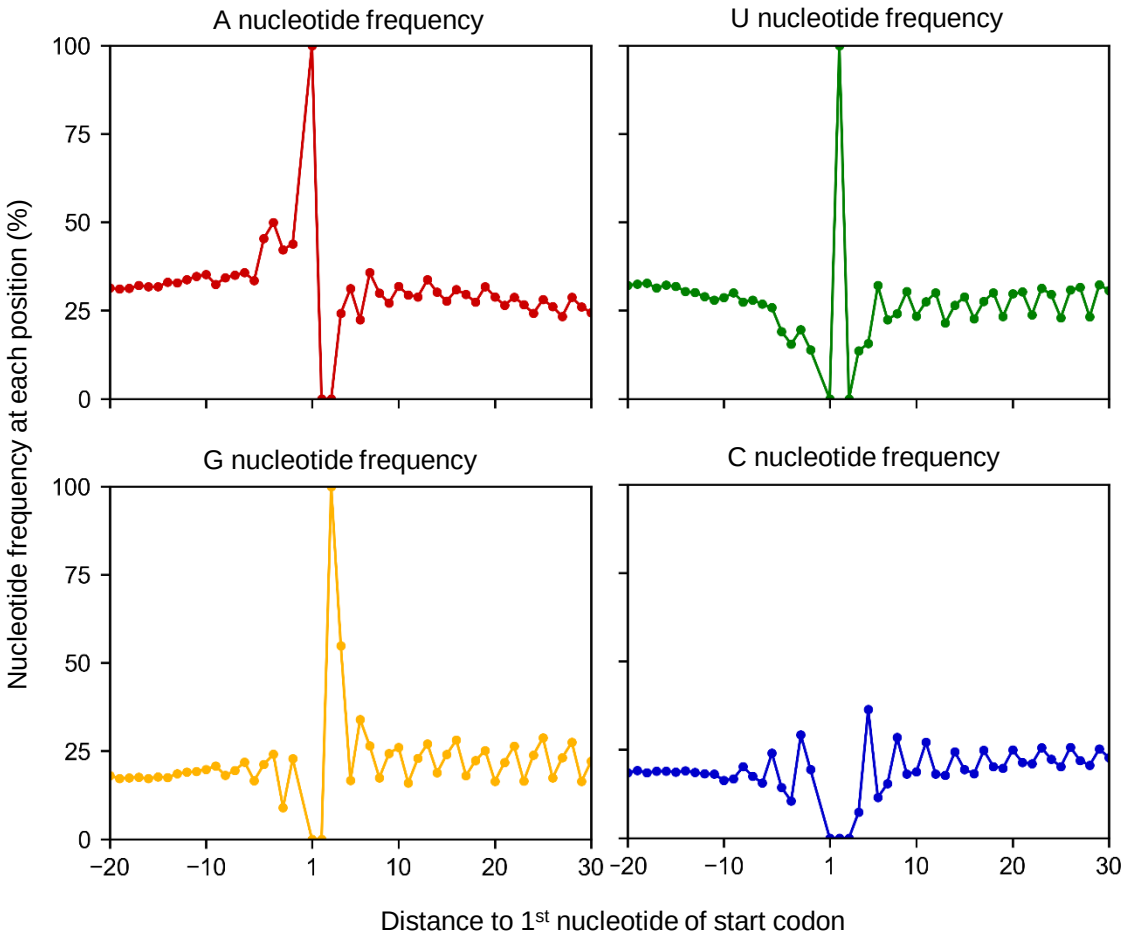


図 2-13. 開始コドン周辺の塩基比率 (シロイヌナズナ)

シロイヌナズナの RNA 配列情報を取得し、開始コドン周辺の塩基比率を算出した。X 軸は開始コドンからの距離を示し、Y 軸は各位置での塩基比率を示す。

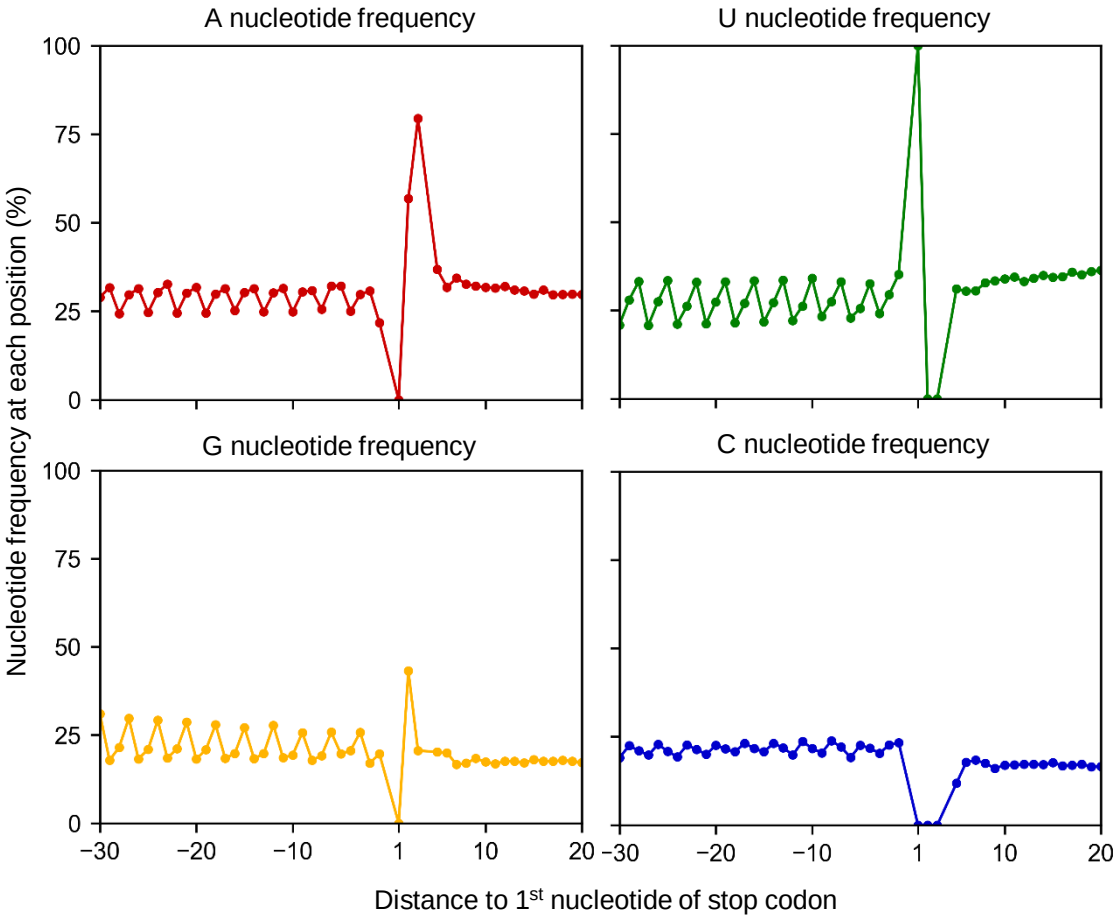


図 2-14. 終止コドン周辺の塩基比率 (シロイヌナズナ)

シロイヌナズナの RNA 配列情報を取得し、終止コドン周辺の塩基比率を算出した。X 軸は終止コドンからの距離を示し、Y 軸は各位置での塩基比率を示す。

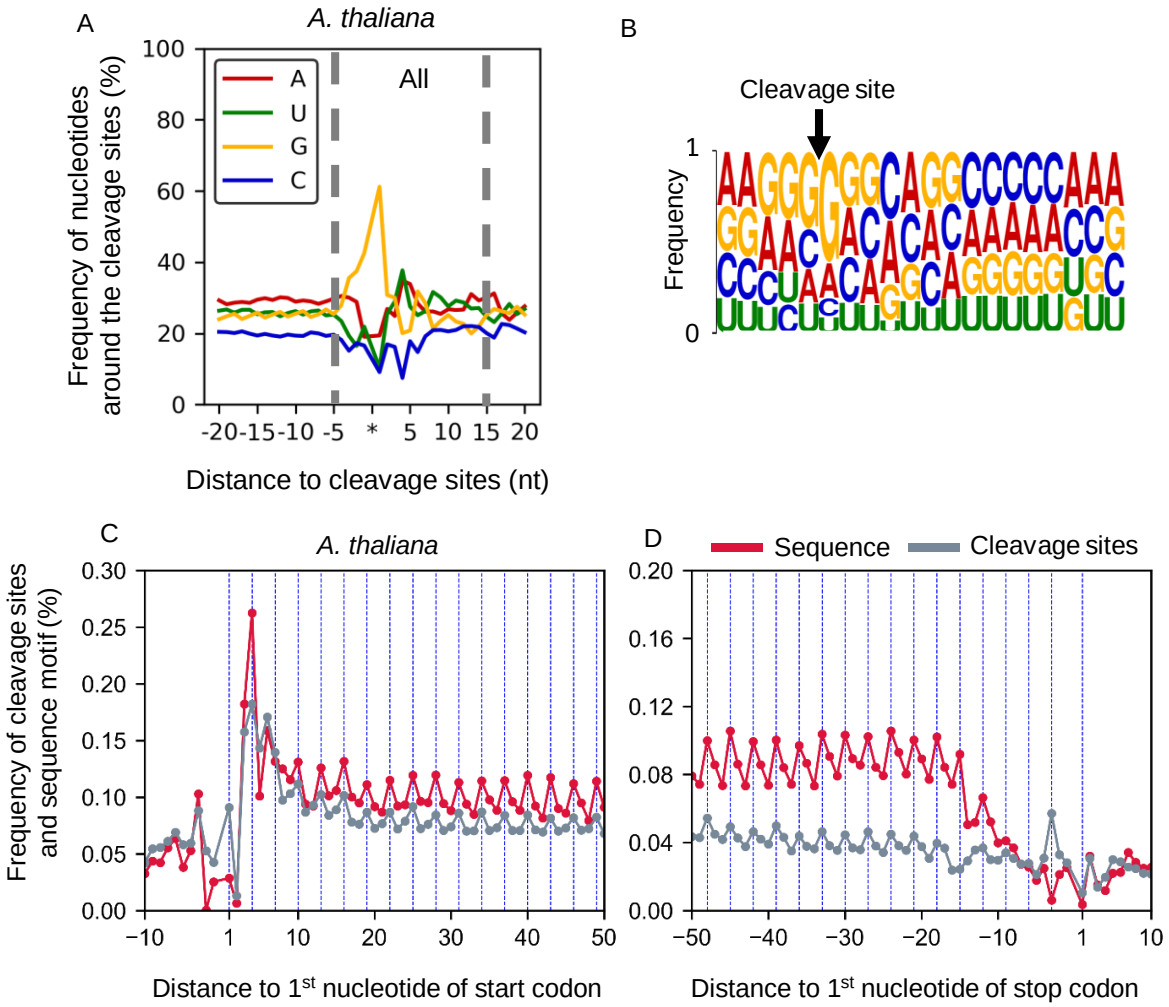


図 2-15. 開始、終止コドン周辺の切断部位、配列モチーフの分布（シロイヌナズナ）

切断部位周辺の配列を用い、MEME の motif letter-probability matrix lines 形式 (配列モチーフ) に変換し、FIMO を用いて RNA 内での分布を算出した (A, B)。開始、終止コドンからの距離を算出後、各位置での配列モチーフ、切断部位の存在比率を算出した (C, D)。X 軸は開始コドン、終止コドンからの距離を示す。Y 軸は各位置での切断部位と配列モチーフの存在比率を示す。

2-3-6. ショウジョウバエ、出芽酵母で検出された切断部位の RNA 内での分布

シロイヌナズナを用いた解析により、RNA 切断には複数の要因が関与することが示された。そこで、異なる生物種においても TREseq 法を行い、RNA 切断に関わる特徴を比較した。解析対象としては、ショウジョウバエ、および、出芽酵母を使用した。シロイヌナズナと同様に RNA を抽出し、TREseq 法のライブラリー作製に従い、網羅的に切断部位情報を取得した。他の生物種を用いた場合でも、従来手法で認められた検出される切断部位の偏りが改善されるか検証したところ、シロイヌナズナと同様に、ほぼ一様に RNA 内に切断部位が分布していた (図 2-16)。この結果は、TREseq 法を用いることで、切断部位の 3' 末端側への偏りを異なる生物種を対象とした解析でも軽減できることを示している。

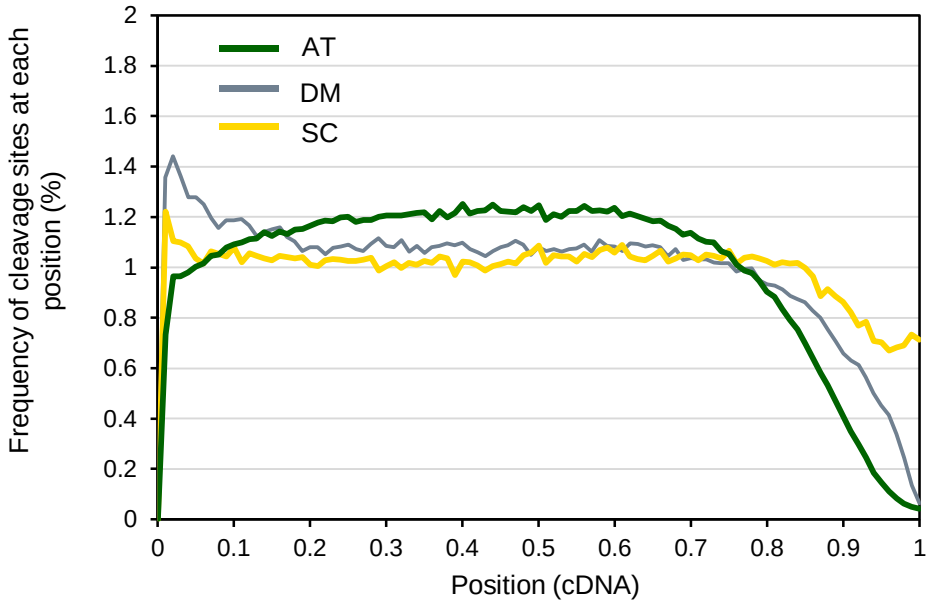


図 2-16. RNA 内での切断部位の分布

網羅的な切断部位情報を取得後、RNA 長 (cDNA) に対する各切断部位の距離を算出し、そのヒストグラムを作成した。X 軸は RNA 長に対する切断部位の距離を示し、Y 軸は各位置での切断部位の存在比率を示す。AT; シロイヌナズナ、 DM; ショウジョウバエ、SC; 出芽酵母をそれぞれ示す。0、1 はそれぞれ遺伝子の 5' 末端と 3' 末端を示す。

2-3-7. ショウジョウバエ、出芽酵母における切断率の算出

シロイヌナズナと同様のデータプロセシングを行い、ショウジョウバエ、出芽酵母について CS_{site} 値、CS_{gene} 値の 2 回の反復実験の再現性を確認したところ、少なくともピアソンの積率相関係数は、r = 0.89 以上を示した (図 2-17)。

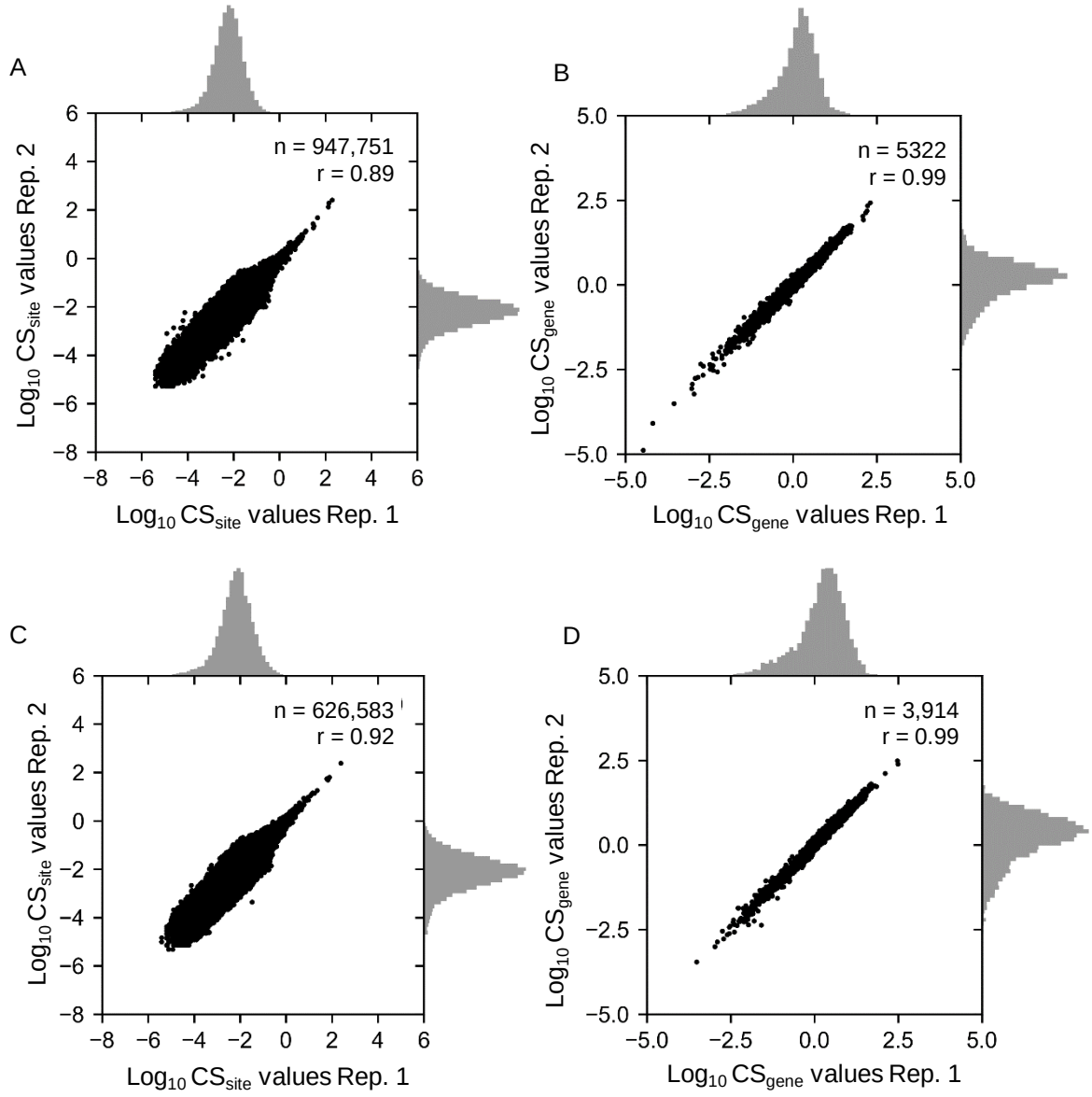


図 2-17. CS_{site} 値、CS_{gene} 値の 2 反復実験での再現性

ショウジョウバエ (A, B)、出芽酵母 (C, D) での CS_{site} 値 (A, C)、および CS_{gene} 値 (B, D) の 2 反復の再現性を示す。左辺および右辺は正対する各軸のヒストグラムを示す。

2-3-8. ショウジョウバエ、出芽酵母における切断率と RNA 安定性との関係

2-3-8-1. 遺伝子単位の切断率が半減期に与える影響

ショウジョウエ、出芽酵母の RNA 半減期情報を公開されているデータより取得し (52, 53)、シロイヌナズナと同様の解析を行った。解析対象となった全遺伝子の半減期情報と遺伝子単位の切断率とのピアソンの積率相関係数を算出したところ、ショウジョウバエでは r = -0.36 (n = 1,235 genes)、出芽酵母では、r = -0.61 (n = 3,914 genes) となった (図 2-18, 図 2-19)。また、算出した遺伝子単位の切断率を基に TOP 10% (ショウジョウバエ, n = 124 genes; 出芽酵母, n = 391 genes)、BOTTOM 10% (ショウジョウバエ, n = 124 genes; 出芽酵母, n = 391 genes) の遺伝子について比較を行ったところ、切断率が高い RNA ほど半減期が短い傾向がショウジョウバエ、出芽酵母ともに認められた (図 2-18, 図 2-19, Welch's t-test, $p < 0.01$)。これらの結果は、RNA 切断に依存する分解は、シロイヌナズナのみならずショウジョウバエや出芽酵母など、異なる生物種においても RNA 安定性に関与することを示している。

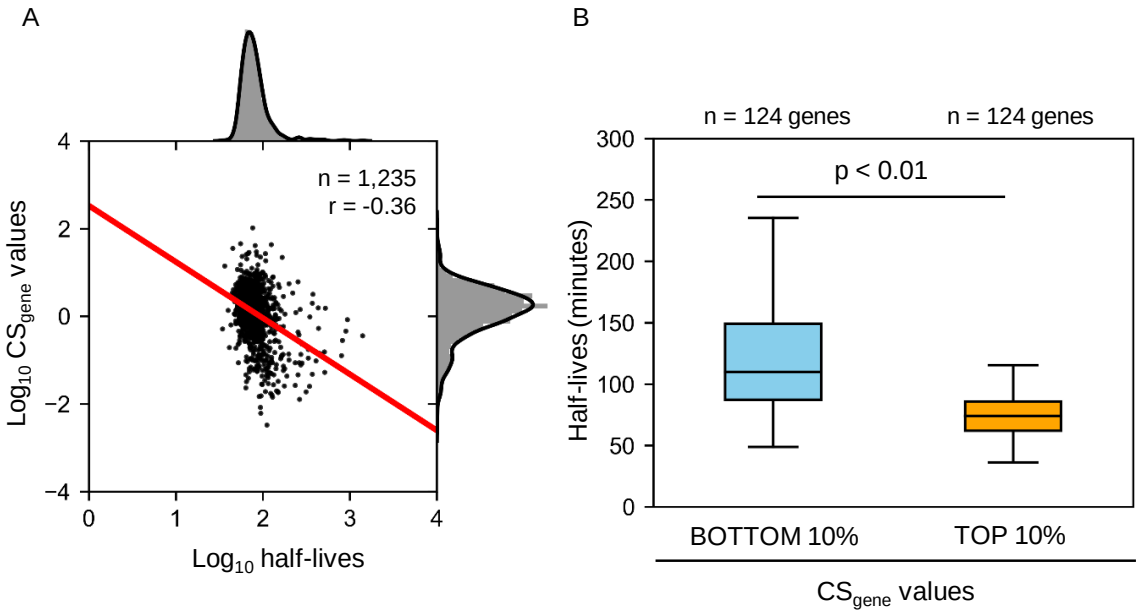


図 2-18. CS_{gene} 値と半減期 (ショウジョウバエ)

公開されているデータより半減期情報を取得し、CS_{gene} 値とのピアソンの積率相関係数を求めた (A)。また、CS_{gene} 値が高い、低い順から 10%ずつ遺伝子を選抜し、それらの半減期を比較した (B)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。統計検定には Welch's t-test を使用した。

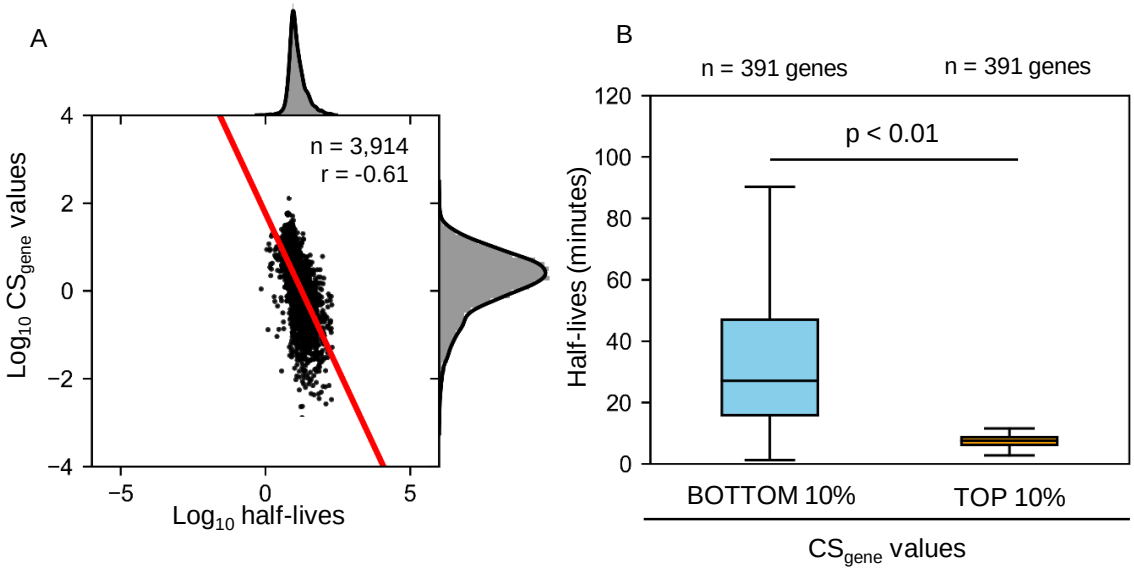


図 2-19. CS_{gene} 値と半減期 (出芽酵母)

公開されているデータより半減期情報を取得し、CS_{gene} 値とのピアソンの積率相関係数を求めた (A)。また、CS_{gene} 値が高い、低い順から 10%ずつ遺伝子を選抜し、それらの半減期を比較した (B)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。統計検定には Welch's t-test を使用した。

2-3-8-2. 遺伝子単位の切断率と GO ターム

mRNA の安定性とコードするタンパク質の機能には関連性が存在することが動植物において報告されている (2, 3, 8)。例えば、半減期が短い mRNA には転写や刺激応答に関わる RNA 種が多く存在する一方で、半減期が長い mRNA はリボソームに関連する RNA 種が多いことが知られている (2, 3, 8)。そこで、シロイヌナズナ、ショウジョウバエ、出芽酵母の CS_{gene} 値の TOP 10%、および BOTTOM 10%の遺伝子を用い GO enrichment 解析を行った。第二章の 2-3-8-1 で示されるように半減期と CS_{gene} 値は逆相関を示すことから、CS_{gene} 値が高い RNA 種には半減期が短い RNA 種の GO term が、CS_{gene} 値が低い RNA 種には半減期が長い RNA 種の GO term が生物種間で共通して存在するのではないかと考えた。予想されたように、CS_{gene} 値が高い集団においては、signal transduction や regulation of transcription など、刺激応答や転写に関わる GO term が確認され (表 2-5)、その反対に CS_{gene} 値が低い RNA 種には、翻訳過程に関連する GO term が 3 種間で共通して存在していた (表 2-6)。これらの結果は、これまでの RNA 半減期解析から提唱されているように (2, 3)、環境応答など迅速な制御に関わる RNA 種は不安定であるが、恒常的な機能に関わる RNA 種は安定であり、RNA 切断機構は、このような遺伝子群の安定性に関わっていることを示している。このことから、RNA 切断機構は細胞が生命を維持する上で重要な生物学的プロセスに関与し、多くの生物種で細胞の機能調節に関与している可能性が考えられた。

表 2-5. GO enrichment analysis (CS_{gene} TOP 10%)

GO Term	GO term CS_{gene} TOP 10%	AT	DM	SC
GO:0006355	regulation of transcription, DNA-templated	3.12E-05	1.23E-08	5.96E-12
GO:0006464	cellular protein modification process	5.75E-14	1.45E-05	3.95E-05
GO:0006468	protein phosphorylation	5.11E-14	8.71E-10	6.55E-06
GO:0006996	organelle organization	3.27E-10	7.35E-12	2.50E-06
GO:0007165	signal transduction	2.81E-07	5.04E-16	5.65E-07
GO:0009889	regulation of biosynthetic process	0.000593	5.09E-11	1.09E-14
GO:0009892	negative regulation of metabolic process	0.000295	1.46E-05	1.17E-08
GO:0009893	positive regulation of metabolic process	0.00024	9.04E-13	4.86E-11
GO:0009987	cellular process	6.47E-11	6.43E-05	1.75E-04
GO:0010468	regulation of gene expression	3.68E-08	2.39E-13	3.86E-14

CS_{gene} 値の TOP 10%の遺伝子を選抜後、GOrilla を用いて GO 解析を行った。$p < 0.01$ かつ、3 種間で共通の GO term を示す。AT; シロイヌナズナ、 DM; ショウジョウバエ、SC; 出芽酵母をそれぞれ示す。

表 2-6. GO enrichment analysis (CS_{gene} BOTTOM 10%)

GO Term	GO term CS_{gene} BOTTOM 10%	AT	DM	SC
GO:0006412	translation	8.41E-14	5.46E-21	6.69E-19
GO:0006518	peptide metabolic process	5.13E-14	2.53E-20	2.04E-18
GO:0009059	macromolecule biosynthetic process	5.30E-08	2.38E-14	1.02E-05
GO:0034645	cellular macromolecule biosynthetic process	7.15E-08	3.26E-13	9.43E-10
GO:0043043	peptide biosynthetic process	8.41E-14	5.46E-21	6.69E-19
GO:0043603	cellular amide metabolic process	3.68E-13	1.80E-20	4.13E-17
GO:0043604	amide biosynthetic process	8.41E-14	1.31E-20	1.79E-17
GO:0044249	cellular biosynthetic process	0.000564	1.69E-12	7.01E-06
GO:0044271	cellular nitrogen compound biosynthetic process	6.47E-08	2.75E-16	1.28E-07
GO:1901566	organonitrogen compound biosynthetic process	3.82E-08	2.83E-17	1.46E-15

CS_{gene} 値の BOTTOM 10%の遺伝子を選抜後、GOrilla を用いて GO 解析を行った。$p < 0.01$ かつ、3 種間で共通の GO term を示す。AT; シロイヌナズナ、 DM; ショウジョウバエ、SC; 出芽酵母をそれぞれ示す。

2-3-9. ショウジョウバエにおける microRNA のターゲット配列と切断部位

TREseq 法で検出した切断部位解析について、microRNA のターゲット配列に着目した解析を行った結果、ターゲットサイトは全切断部位の中で、ごくわずかであることが第二章の 2-3-1 で示された。そこで、他の生物種についても同様の傾向が認められるかどうか解析を行った。microRNA のターゲット配列のデータベースには出芽酵母の情報が登録されていないため、ショウジョウバエを対象に解析を行った。2-3-1 と同様に、psRNA target を用いて microRNA のターゲット配列と重複する (microRNA が関与する) 切断部位数を算出した結果、全切断部位の 1.3%が microRNA が関与する切断部位に該当した (表 2-7)。このことは、ショウジョウバエでも microRNA が関与しない RNA 切断が多く存在することを示している。また、TREseq 法で検出された全切断部位の切断率に着目した場合、microRNA が関与しない切断部位での切断率は、microRNA が関与する切断部位の切断率より統計的に高く (表 2-8, Welch's t-test, $p < 0.01$)、1.03 倍程度であった(表 2-9)。これらの結果は、第二章の 2-3-1 の結果と同様に microRNA が関与せず、かつ、microRNA と同等の切断率である切断部位がショウジョウバエの RNA にも多く存在することを示している。

表 2-7. ショウジョウバエでの microRNA が関与する切断部位の割合

	microRNA が関与する切断	他の切断部位
培養細胞	11,463 sites (1.33%)	848,109 sites (98.67%)

表 2-8. 検出された全切断部位を対象とした際の切断率の比較

	microRNA が関与する切断	他の切断部位
培養細胞	平均 CS_{site} 値 = 0.0151	平均 CS_{site} 値 0.0156

表 2-9. microRNA が関与する切断部位と関与しない切断部位との切断率の比較

	切断率比 (他の切断部位 / microRNA が関与)
培養細胞	0.0156 / 0.0151= 1.03

2-3-10. ショウジョウバエ、出芽酵母における配列的特徴が切断に与える影響

これまでの解析から、RNA 切断には特異的な配列パターンが関与することがシロイヌナズナで示された。そこで、ショウジョウバエ、出芽酵母の切断部位周辺の配列に着目し解析を行ったところ、一部の領域ではシロイヌナズナと異なる傾向が認められたが (+5 塩基目の A、U の比率など)、切断部位周辺の G 塩基比率が高いなど、シロイヌナズナと類似する傾向が認められた (図 2-20)。また、CS_{site} 値の TOP 10%と BOTTOM 10%の切断部位を選抜し、その塩基比率を比較したところ、CS_{site} 値の TOP 10%において、より顕著な塩基の偏りが確認された (図 2-20)。一方で、各生物種の CDS 領域、UTR 領域の平均的な塩基比率は異なることから (表 2-10, 表 2-11)、それぞれの領域ごとに CS_{site} 値の BOTTOM 10%に対する TOP 10%の切断部位周辺の配列での相対的な塩基比率を算出した。その結果でも切断部位周辺の配列パターンは生物種間を通じて、同様の傾向が認められた (図 2-21)。

次に、切断部位周辺の配列を中心領域 (-10～+10)、上流領域 (-50～-10)、下流領域 (+10～+50) に分け、DREME を用いてモチーフ検索を行い、そのモチーフの類似性を STAMP で評価した。また、当研究室では、イネ、バラ、レタスなど異なる植物種についても TREseq 法を行い、網羅的な切断部位に関する情報を取得している (54)。そこで、STAMP を用いたモチーフ配列を比較する際は、シロイヌナズナ、ショウジョウバエ、出芽酵母のデータに加え、イネ、バラ、レタスの配列モチーフ情報も使用した。解析には最も有意に検出された配列モチーフを使用した。また、いくつかの生物種については UTR 領域に関するゲノム情報が不足しているため、各生物種の CDS 領域に存在する切断部位のみを対象とした。加えて、配列モチーフを抽出する際は、各生物種の CDS 領域の塩基比率が異なることを考慮し、CS_{site} 値の BOTTOM 10%の配列をコントロール配列として使用した。図 2-22 に示すように、切断部位の中心領域に関しては (-10～+10) 、いずれの生物種でも G リッチな配列モチーフが検出された。一方で、上流、下流領域については、植物や出芽酵母で AU リッチな配列モチーフが検出されたが、ショウジョウバエのみ他の生物種と異なる配列モチーフが検出された (図 2-23, 図 2-24)。これらの結果は、切断部位の中心領域の配列は生物種を問わず保存されているが、その上流、下流領域の配列は動物（ショウジョウバエ）では異なる配列が切断に関与している可能性を示唆している。

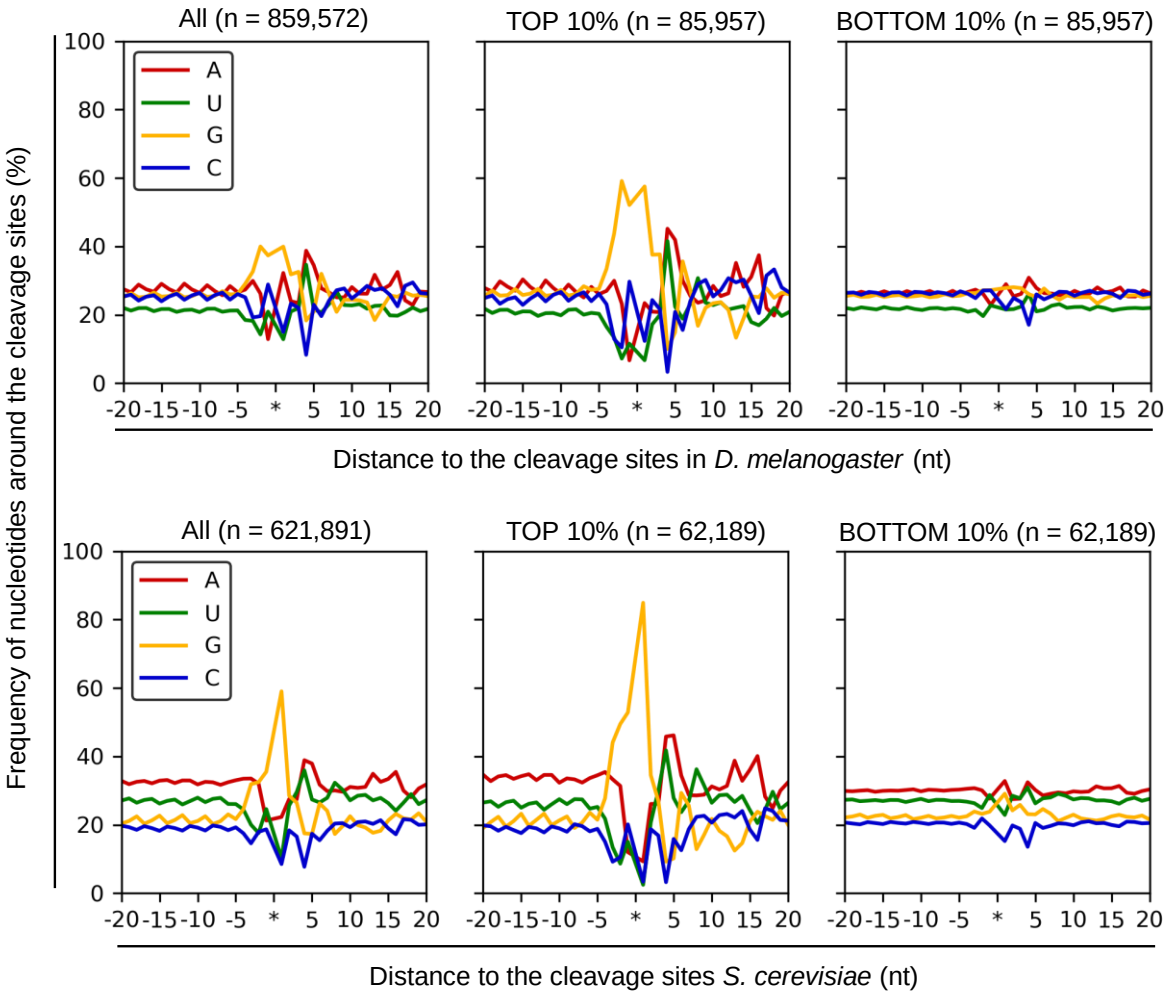


図 2-20. ショウジョウバエ、出芽酵母における切断部位周辺の塩基比率

ショウジョウバエ、出芽酵母を対象に切断部位周辺の配列情報を取得し、塩基比率を算出した。また、切断率を基に場合分けを行った場合の配列についても塩基比率を算出した。X 軸は切断部位からの距離を示し、Y 軸は各位置での塩基比率を示す。アスタリスクは切断部位を示す。

表 2-10. 各生物種における CDS 領域の塩基比率

	A (%)	U (%)	G (%)	C (%)
A. thaliana	28.69	27.18	23.82	20.31
D. melanogaster	25.63	20.44	26.78	27.15
S. cerevisiae	32.63	27.75	20.48	19.14

表 2-11. 各生物種における UTR 領域の塩基比率

	A (%)	U (%)	G (%)	C (%)
A. thaliana	29.40	17.56	17.62	35.42
D. melanogaster	33.35	28.16	18.69	19.80
S. cerevisiae	32.87	27.90	20.31	18.92

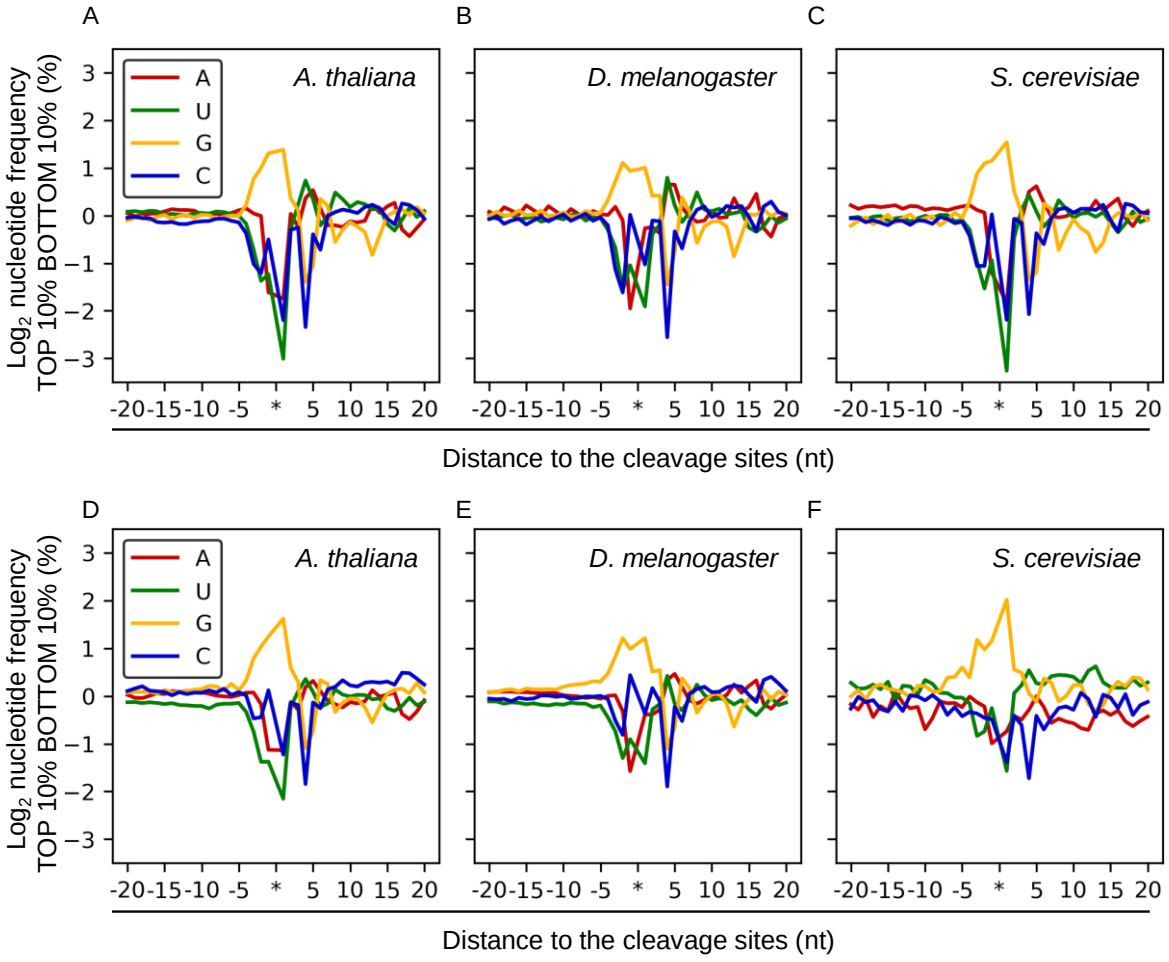


図 2-21. 各生物種における CDS、UTR 領域での切断部位周辺の塩基比率

各生物種ごとに CDS 領域の塩基比率は異なることから、CS_{site} 値の TOP 10%、BOTTOM 10%の配列を用いて相対的な塩基比率を算出した (A-C)。また、UTR 領域に関しても、同様に解析を行った (D-F)。X 軸は切断部位からの距離を示し、Y 軸は BOTTOM 10%に対する TOP 10%の切断部位周辺の配列での相対的な塩基比率を示す。アスタリスクは切断部位を示す。

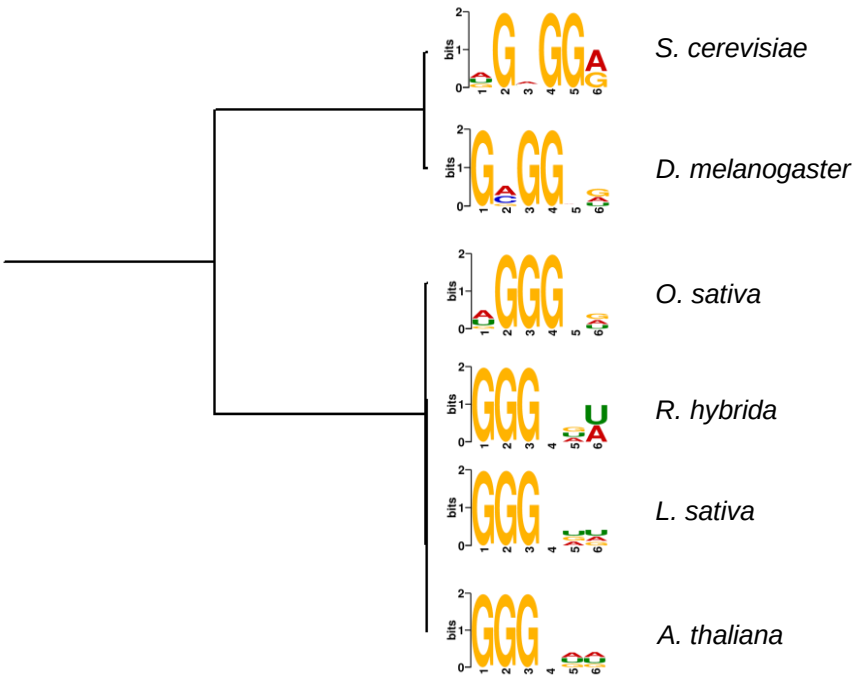


図 2-22. 切断部位の中心領域（-10 から+10）に着目したモチーフ解析

切断部位の-10 から+10 の領域に着目し、MEME のソフトである DREME を用いてモチーフ配列を抽出した。その後、STAMP を用いてモチーフ配列の類似度を算出した。

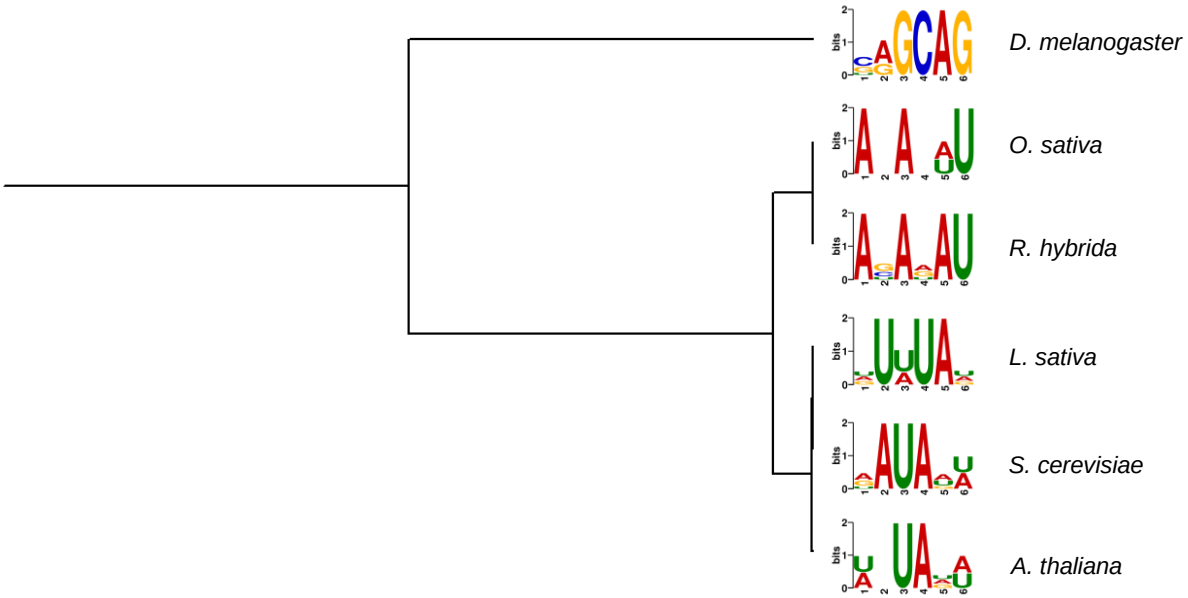


図 2-23. 切断部位の上流領域（-50 から-10）に着目したモチーフ解析

切断部位の-50 から-10 の領域に着目し、MEME のソフトである DREME を用いてモチーフ配列を抽出した。その後、STAMP を用いてモチーフ配列の類似度を算出した。

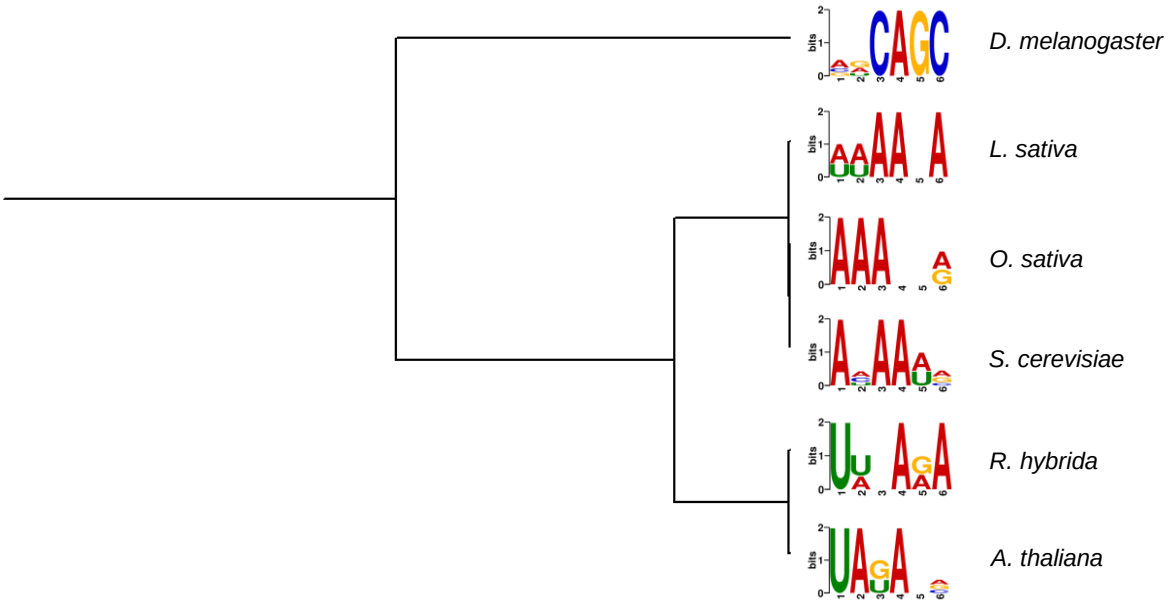


図 2-24. 切断部位の下流領域（+10 から+50）に着目したモチーフ解析

切断部位の+10 から+50 の領域に着目し、MEME のソフトである DREME を用いてモチーフ配列を抽出した。その後、STAMP を用いてモチーフ配列の類似度を算出した。

2-3-11. ショウジョウバエ、出芽酵母で RNA 高次構造が切断に与える影響

第二章の 2-3-3 の結果から、切断部位周辺の塩基対の形成度合いは高く、その度合いは切断部位周辺に頻出する G 塩基に依存すると考えられた。そこで、ショウジョウバエ、出芽酵母についても同様の解析を行い、異なる生物種で RNA 高次構造が RNA 切断に与える影響を評価した。図 2-2 と同様に、切断部位周辺の前後 30 塩基の配列を使用し、各位置での塩基対の形成の有無を RNAfold を用いて予測し、CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成頻度(塩基対の形成度合い) を算出した。その結果、ショウジョウバエ、および出芽酵母で切断率が高い部位で切断部位周辺の塩基対の形成度合いはシロイヌナズナと同様に高い傾向にあった (図 2-25)。また、切断部位周辺の各塩基の比率も算出し、塩基対の形成度合いとの比較を行ったところ、シロイヌナズナと同様に切断部位周辺の塩基対形成の度合いは、特に G 塩基の比率と類似する傾向が認められた (図 2-26, 図 2-27)。これらの結果は、異なる生物種でも切断部位の周辺で RNA 高次構造が形成しやすく、その構造の形成は切断部位周辺の G 塩基の比率に依存していることを示唆している。

RNAfold は、配列を基に簡易的に RNA 構造を予想するソフトウェアであるため、異なる観点からの解析も行った。切断部位周辺で認められた塩基対の形成度合いが RNA 切断に大きく関与しているならば、類似する RNA 構造が各生物種で検出されると考えられる。図 2-22 で用いた切断部位の前後 10 塩基の配列を CS_{site} 値が高い順から 50 個選抜し、RNAfold の Structure Conservation Analysis web server を使用し、RNA 高次構造の類似性を算出したところ、どの生物種からも共通する RNA 高次構造は抽出されなかった。RNAfold などのソフトウェアは、配列情報を基に RNA 構造を予想しているため、細胞内での実際の RNA 構造とは異なる可能性も考えられるが、Structure Conservation Analysis の結果を踏まえると、切断部位周辺の配列が重要であり、RNAfold の結果は単に G 塩基の比率を反映したものと考えられた。

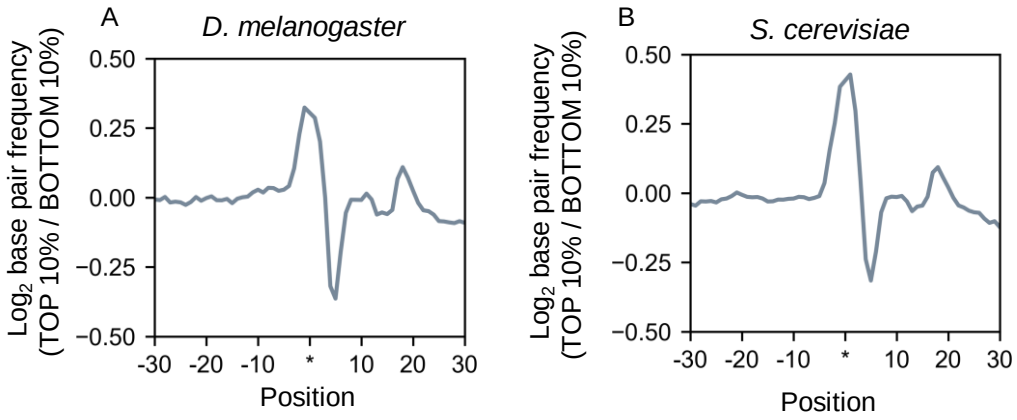


図 2-25. 切断部位周辺の塩基対形成の度合い (ショウジョウバエ、出芽酵母)

切断部位周辺の塩基配列を取得し、RNAfold を用いて塩基対が形成される箇所を予測した。その後、CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成頻度 (形成度合い) をショウジョウバエ (A)、出芽酵母 (B) について算出した。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成度合いを log_2 対数変換した値を示す。

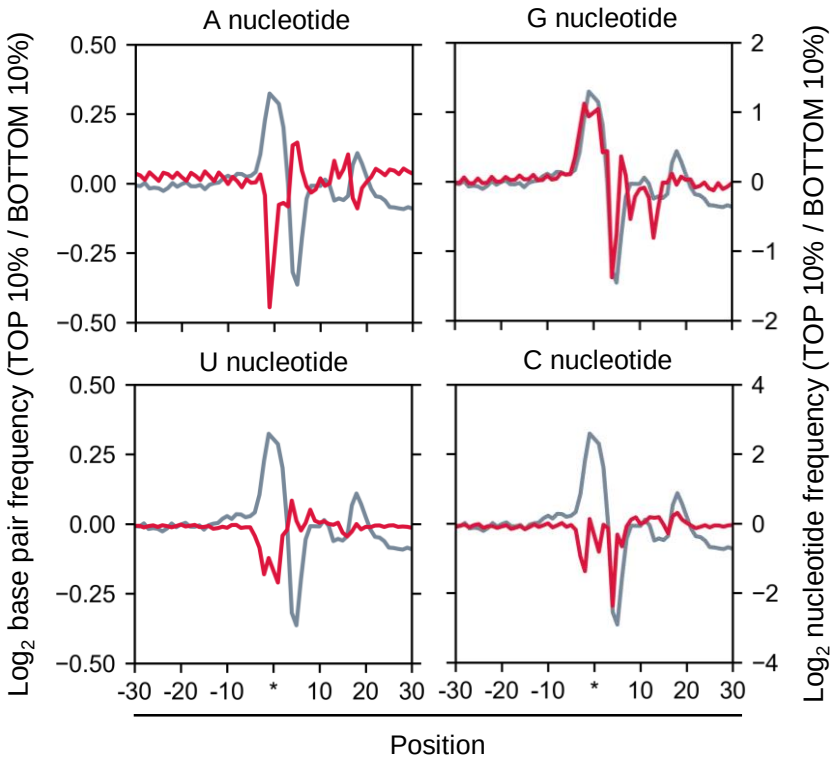


図 2-26. 切断部位周辺の塩基対形成の度合いと塩基比率（ショウジョウバエ）

CS_{site} 値の TOP 10%、BOTTOM 10%の配列情報を取得後、A 塩基、G 塩基、U 塩基、C 塩基比率を $\log_2$ 対数変換し、図 2-25 と統合した。灰色は塩基対の形成頻度を示し、赤色は各塩基の比率を示す。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は解析対象とした CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成度合いと塩基比率を $\log_2$ 対数変換した値を示す。

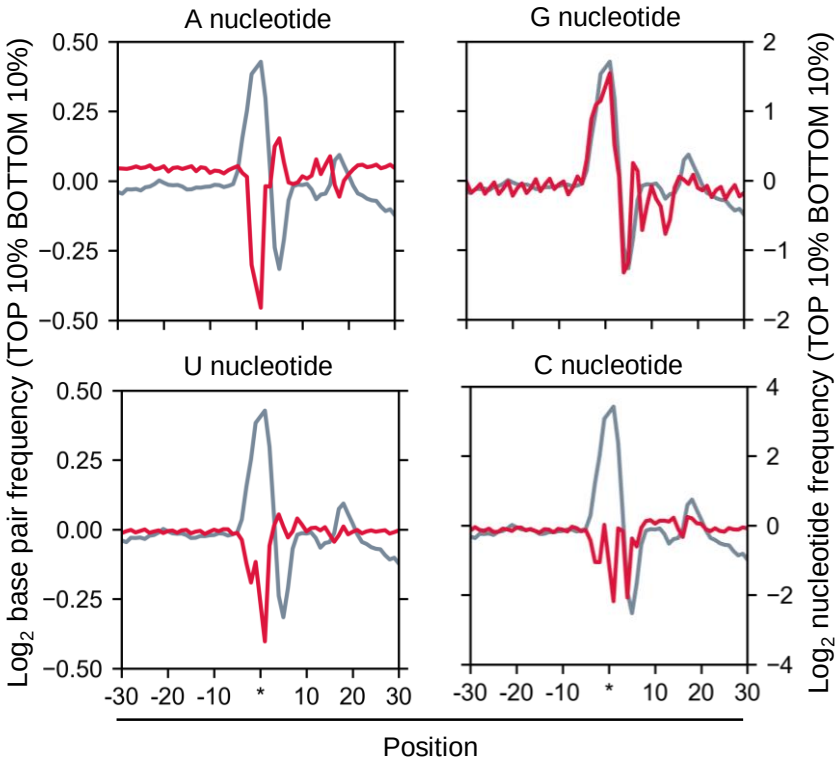


図 2-27. 切断部位周辺の塩基対形成の度合いと塩基比率（出芽酵母）

CS_{site} 値の TOP 10%、BOTTOM 10%の配列情報を取得後、A 塩基、G 塩基、U 塩基、C 塩基比率を $\log_2$ 対数変換し、図 2-25 と統合した。灰色は塩基対の形成頻度を示し、赤色は各塩基の比率を示す。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は解析対象とした CS_{site} 値の BOTTOM 10%に対する TOP 10%の配列での相対的な塩基対の形成度合いと塩基比率を $\log_2$ 対数変換した値を示す。

2-3-12. ショウジョウバエ、出芽酵母における翻訳過程と RNA 切断との関係性

2-3-12-1. 開始、終止コドン周辺における切断部位の分布

Ibrahim らや、Pelechano らの研究から、酵母や動物などで、翻訳過程が RNA 切断に関与することが示唆されている (22, 23)。しかし、2-3-4 で示したように、リボソームの存在位置や存在量は、切断部位の位置決定には関与せず、切断率のみに影響を与える可能性が考えられた。そこで、先行研究からリボソームプロファイリング情報を取得し (55, 56)、シロイヌナズナと同様の解析をショウジョウバエ、出芽酵母についても行った。まず、取得した切断部位の RNA 内での分布に着目し解析を行ったところ、ショウジョウバエ、出芽酵母ともに開始、終止コドンの周辺で検出された切断部位のピークが認められ、CDS 内で 3 塩基単位の周期性が認められた (図 2-28, 図 2-29)。

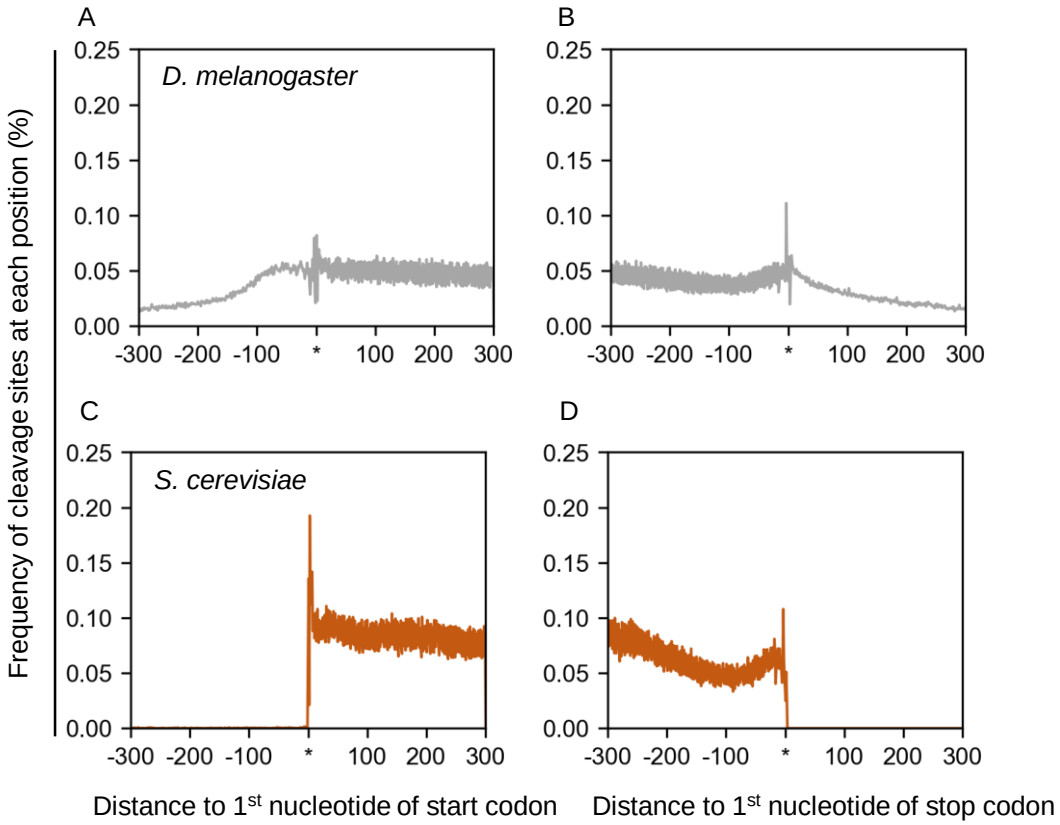


図 2-28. 開始コドン、終止コドン周辺の切断部位の分布 (ショウジョウバエ、出芽酵母)

開始コドン、終止コドンからの距離を算出し、各位置での切断部位の存在比率をショウジョウバエ (A, B)、出芽酵母について算出した (C, D)。X 軸は、開始コドン、終止コドンからの距離を示す。Y 軸は、各位置での切断部位の存在比率を示す。

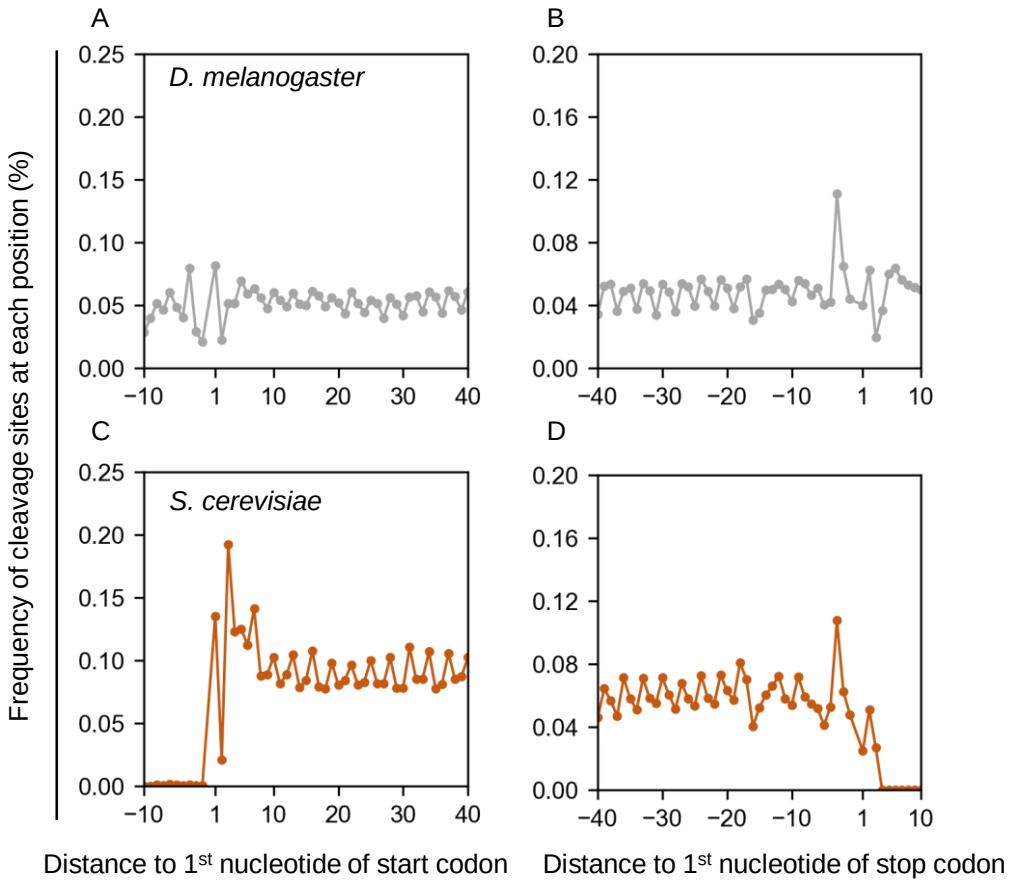


図 2-29. 図 2-28 の拡大図

2-3-12-2. リボソームの位置が切断部位の決定に与える影響

次に、先行研究より取得したリボソームプロファイリング情報を使用し、2-3-4-2 と同様にリボソームに保護された断片の 5' 末端を 5' RPF と定義した。また、RNA 上の各部位での 5' RPF 数を RNA 蓄積量で除算した値を RO_{site} 値 (リボソーム存在量)、RO_{site} 値を遺伝子ごとにまとめた値を RO_{gene} 値と定義し、切断部位との位置関係に着目した解析を行った。まず、5' RPF の分布を単独で見ると、切断部位と同様に開始、終止コドンの周辺にピークが検出され、CDS 内で 3 塩基単位の周期性が認められた (図 2-30, 図 2-31)。しかしながら、切断部位の分布と比較してみると、3 塩基単位の周期性は共通して認められたが、シロイヌナズナと同様に位相は一致していなかった (図 2-32)。また、CDS 領域に存在する切断部位の位置を基準とし、周辺の 5' RPF の存在量を算出し、リボソームの存在位置が RNA 切断の位置決定に与える影響を評価したが、5' RPF の切断部位周辺での顕著な偏りはシロイヌナズナと同様に認められなかった(図 2-33)。

また、第二章の 2-3-4-2 に示すように、開始、終止コドン周辺領域に着目し、リボソームの存在量が切断部位の位置に与える影響を解析した。まず、ショウジョウバエについて、開始、終止コドンの前後 50 塩基以内に存在する RO_{site} 値の平均を算出し、リボソームプロファイリング情報のある遺伝子をリボソーム存在量の多い TOP 20% (開始コドン n = 1,035 genes; 終止コドン n = 995 genes) とリボソーム存在量の少ない BOTTOM 20% (開始コドン n = 1,035 genes; 終止コドン n = 995 genes) に分け、開始、終止コドン周辺での切断部位の分布を比較したが、図 2-34A、図 2-34B に示すように、リボソームの存在量の違いで切断部位の分布に顕著な違いは認められなかった。加えて、開始コドン周辺の-10 位から+40 位、終止コドン周辺の-40 位から+10 位の各位置での切断部位の存在比率について (図 2-34A, 図 2-34B)、リボソーム存在量が多い遺伝子の TOP 20%とリボソーム存在量が少ない遺伝子の BOTTOM 20%間のピアソンの積率相関係数を算出したところ、開始コドン側で r = 0.81 (図 2-34C)、終止コドン側で r = 0.89 を示した (図 2-34D)。出芽酵母については、UTR 領域のゲノム情報が不足しているため、開始コドンの下流 50 塩基、終止コドンの上流 50 塩基以内での RO_{site} 値の平均を算出し、リボソームプロファイリング情報のある遺伝子をリボソーム存在量が多い遺伝子の TOP 20% (開始コドン n = 767 genes; 終止コドン n = 758 genes) とリボソーム存在量が少ない遺伝子の BOTTOM 20% (開始コドン n = 767 genes; 終止コドン n = 758 genes) に分け、同様の解析を行った。その結果、シロイヌナズナやショウジョウバエと類似する傾向が認められたことから (図 2-35)、異なる生物種でも翻訳過程 (リ

ボソームの存在位置や存在量) は切断部位の位置決定には大きく関与しないと考えられた。

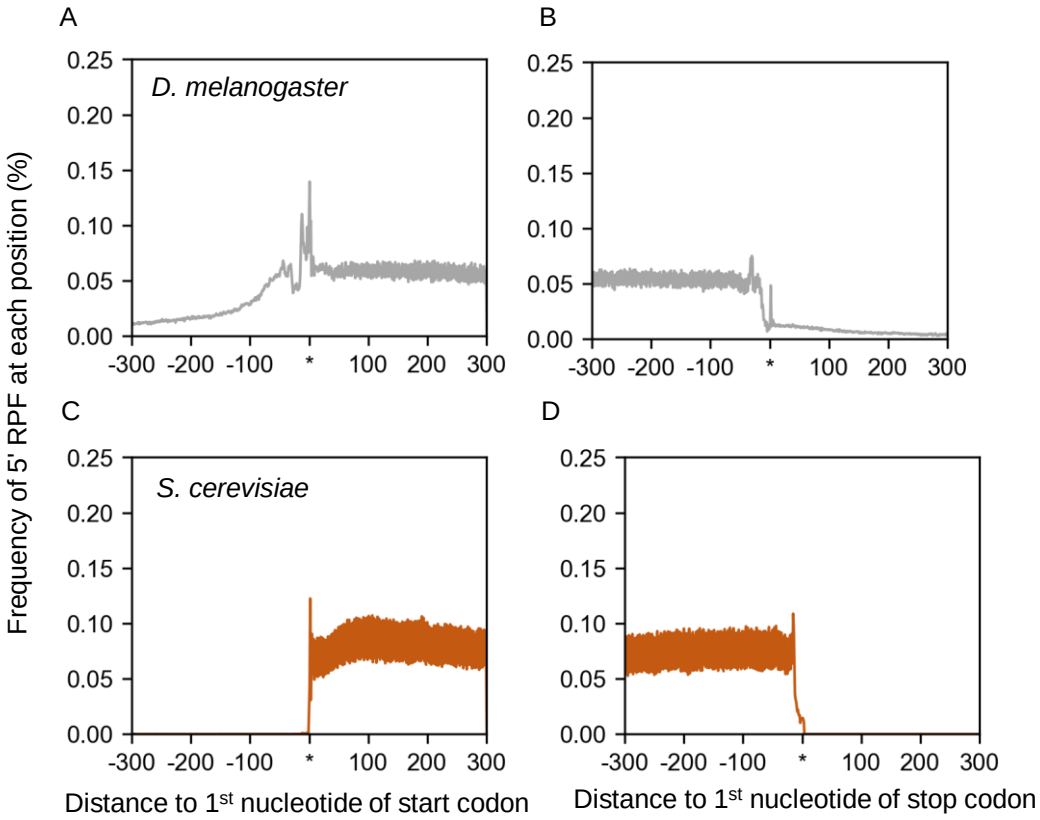


図 2-30. 開始コドン、終止コドン周辺のリボソームの分布（ショウジョウバエ、出芽酵母）

開始コドン、終止コドンからの距離を算出し、各位置での 5' RPF の存在比率をショウジョウバエ (A, B)、出芽酵母について算出した (C, D)。X 軸は開始コドン、終止コドンからの距離を示す。Y 軸は各位置での 5' RPF の存在比率を示す。

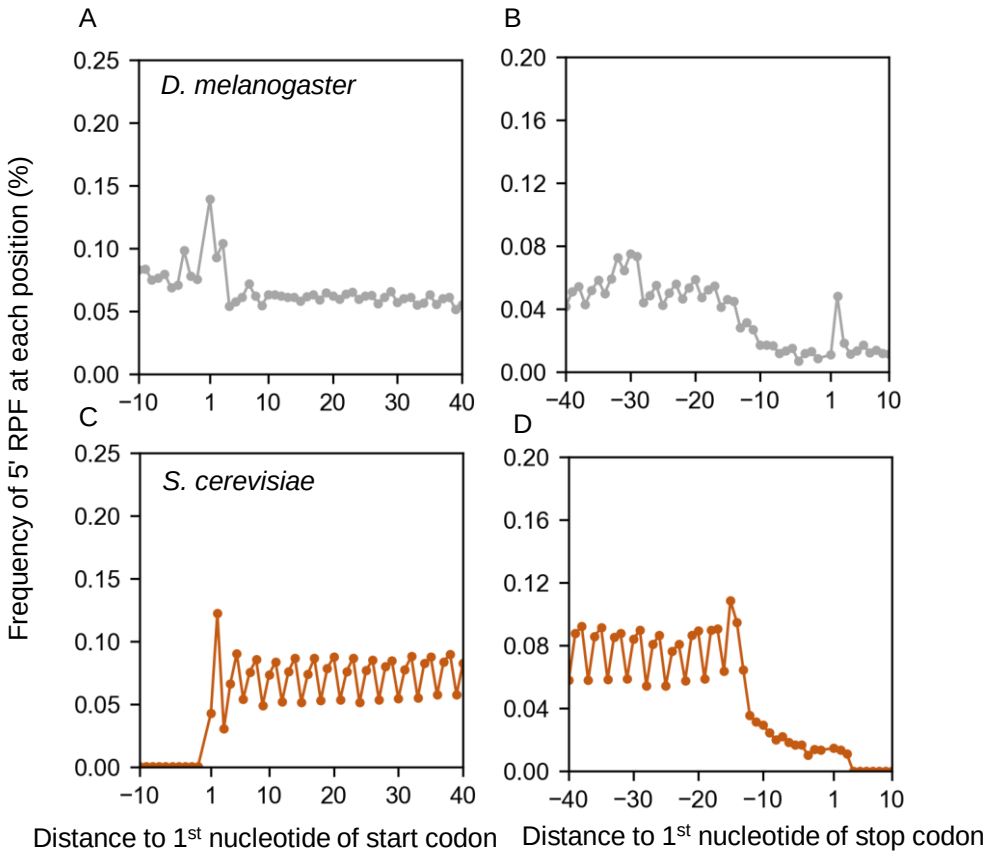


図 2-31. 図 2-30 の拡大図（ショウジョウバエ、出芽酵母）

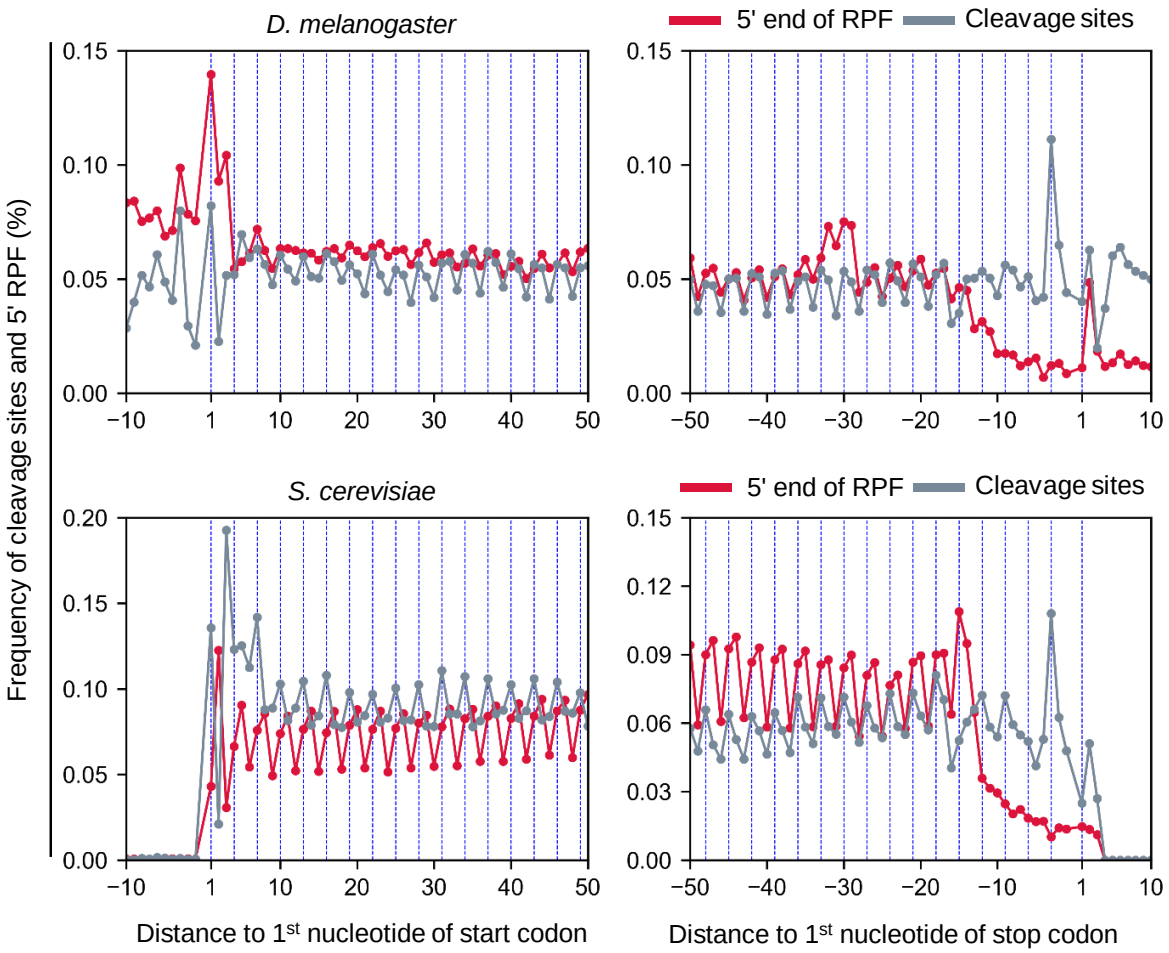


図 2-32. 開始、終止コドン周辺の切断部位とリボソームの分布（ショウジョウバエ、出芽酵母）

開始、終止コドンからの距離を算出後、各位置での切断部位と 5' RPF の出現頻度を算出した。X 軸は開始コドン、終止コドンからの距離を示す。Y 軸は各位置での切断部位と 5' RPF の出現頻度を示す。

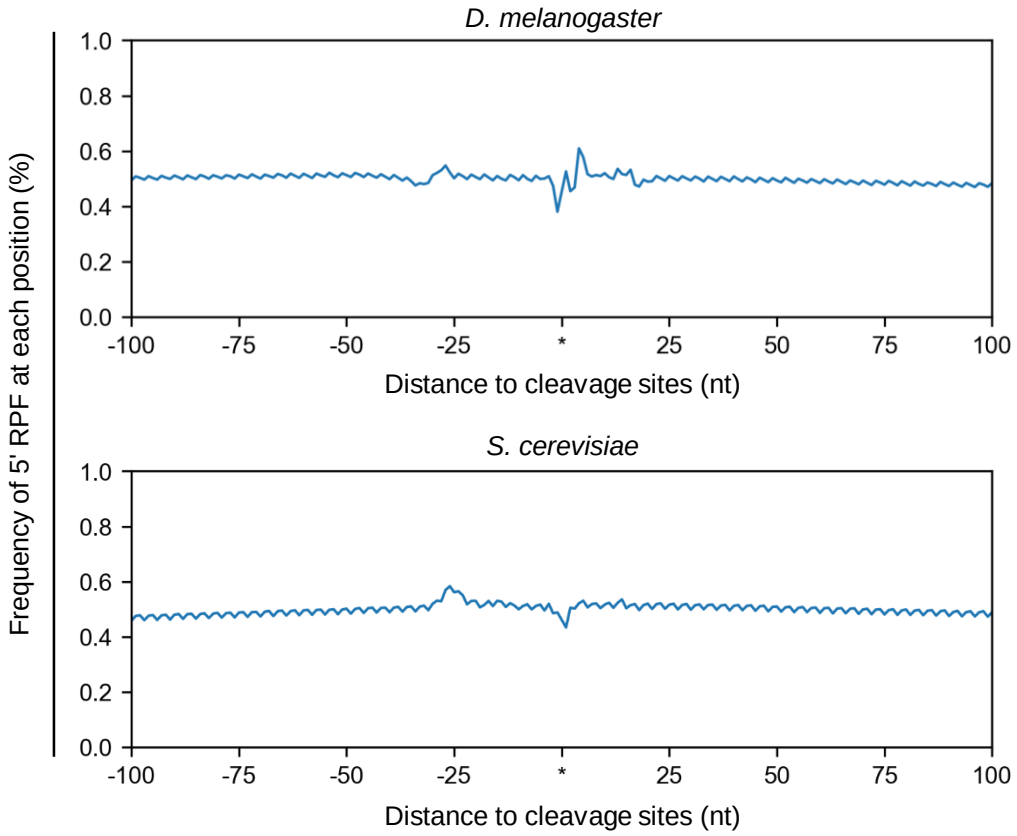


図 2-33. 切断部位周辺の 5' RPF の存在比率（ショウジョウバエ、出芽酵母）

切断部位からの距離を算出後、各位置における 5' RPF の存在比率を算出した。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は各位置での 5' RPF 末端の存在比率を示す。

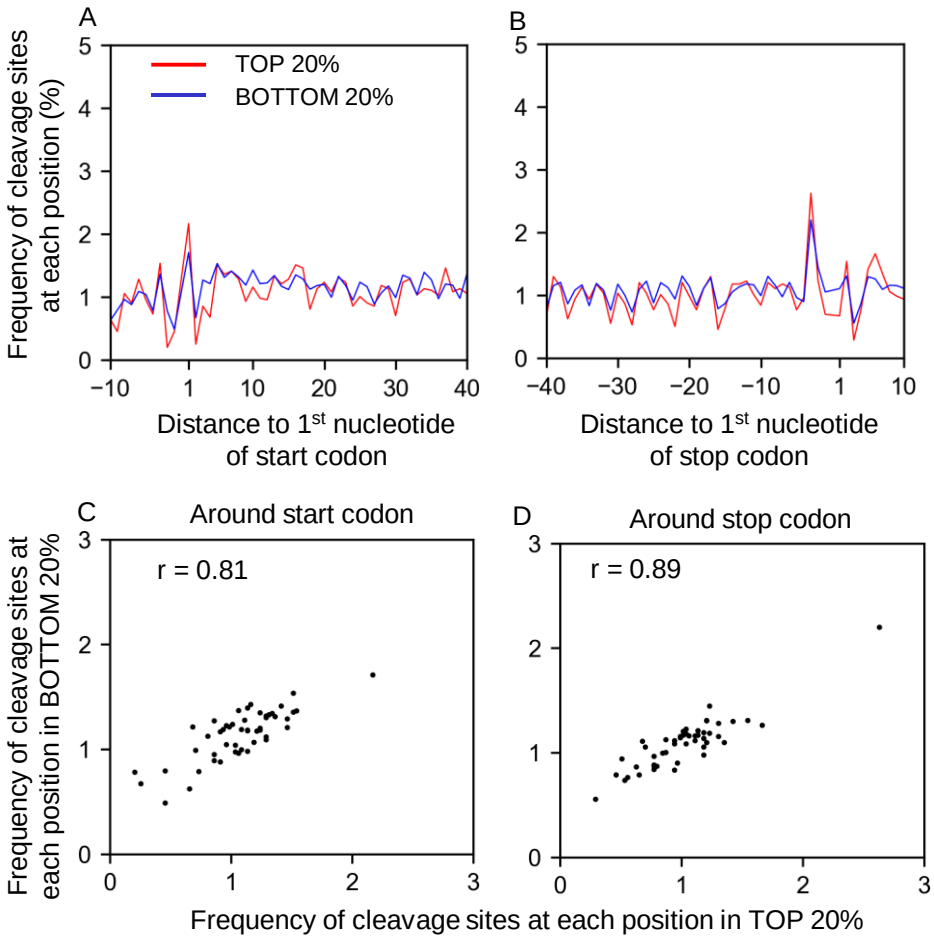


図 2-34. リボソーム存在量が切断部位の分布に与える影響（ショウジョウバエ）

開始、終止コドンの前後 50 塩基以内に存在するリボソーム存在量を基に、TOP 20%、BOTTOM 20%の遺伝子を選抜した。その後、開始、終止コドンからの距離を算出し、各位置の切断部位の存在比率を算出した (A, B)。また、TOP 20%、BOTTOM 20%間での、開始コドン (C)、終止コドン (D) 周辺における各位置の切断部位の存在比率のピアソンの積率相関係数を算出した。

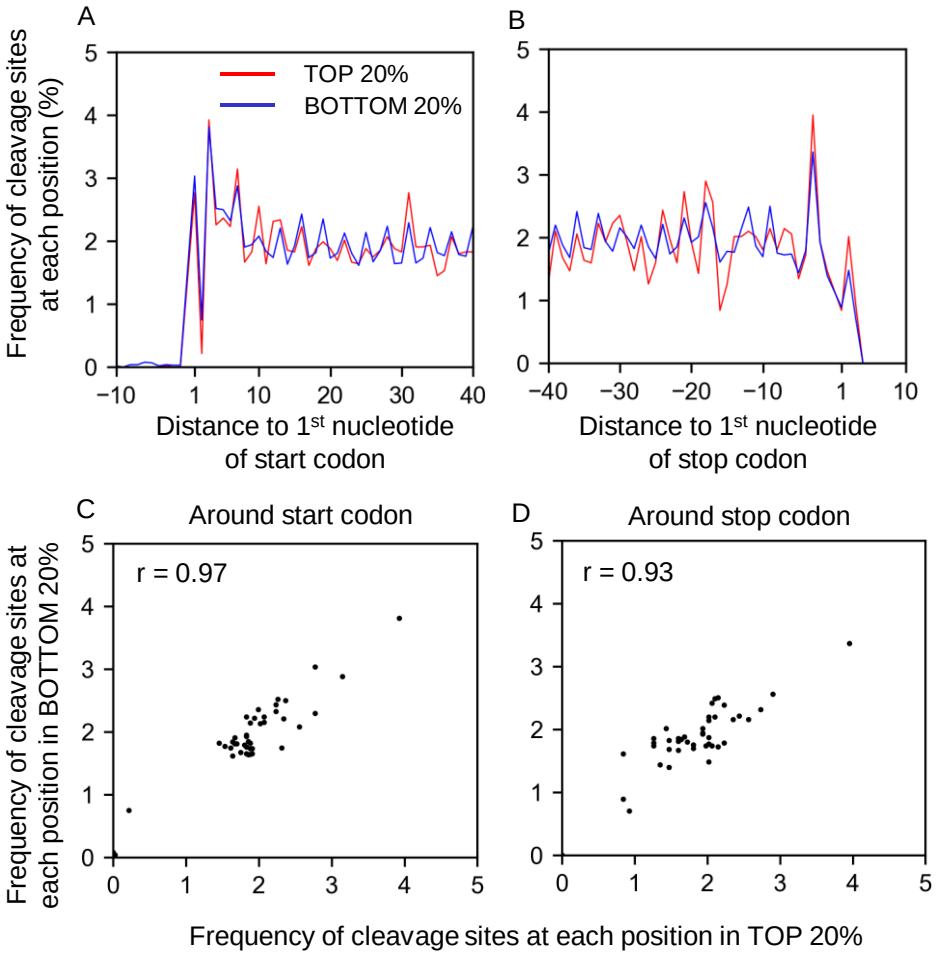


図 2-35. リボソーム存在量が切断部位の分布に与える影響（出芽酵母）

開始コドンの下流 50 塩基、終止コドンの上流 50 塩基以内に存在するリボソーム存在量を基に、TOP 20%、BOTTOM 20%の遺伝子を選抜した。その後、開始、終止コドンからの距離を算出し、各位置の切断部位の存在比率を算出した (A, B)。また、TOP 20%、BOTTOM 20%間での、開始コドン (C)、終止コドン (D) 周辺における各位置の切断部位の存在比率のピアソンの積率相関係数を算出した。

2-3-12-3. リボソーム存在量が切断率に与える影響

第二章の 2-3-4-3 の解析により、遺伝子単位でのリボソーム存在量は切断率に正の影響を及ぼすことが示されている。そこで、ショウジョウバエ、出芽酵母でも同様の傾向を示すかどうか解析を行ったところ、ピアソンの積率相関係数はショウジョウバエで r = 0.46 (n = 5,299 genes)、出芽酵母で r = 0.39 (n = 3,892 genes) となり、正の相関が認められた (図 2-36)。この結果は、異なる生物種でもシロイヌナズナと同様に、RNA 上のリボソーム存在量が多いほど切断されやすいことを示している。Pelechano らの酵母を用いた解析では、遺伝子単位での切断のされやすさとリボソーム存在量に正の相関が認められなかったが (23)、図 2-36 に示すように、TREseq 法を用いた解析では、出芽酵母でも通常条件下で両者に正の関係性が認められた。これらの結果から、Pelechano らの解析では、PARE 法と類似する手法である 5Pseq 法を用い網羅的に RNA 末端を検出しているため、ライブラリー作製時にポリ A 鎖付き RNA を濃縮していた点など、実験手法上の問題により両者の関係性が認められなかったと考えられた。

次に、リボソーム存在量が切断率に与える影響に着目し解析を行った。ショウジョウバエについて、CDS 領域に存在する切断部位の前後 50 塩基での平均 RO_{site} 値を算出し、平均 RO_{site} 値の TOP 10% (n = 59,554 sites)、BOTTOM10% (n = 59,554 sites) の各切断部位での切断率を比較した場合、切断部位周辺の平均 RO_{site} 値が高い (リボソーム存在量が多い) ほど、切断されやすい (CS_{site} 値が高い) 傾向が認められた (図 2-37A, Welch's t-test, $p < 0.01$)。また、リボソームの停滞率を概算するために、RO_{gene} 値に対する RO_{site} 値 (RO_{site} / RO_{gene} 値) を算出し、図 2-37A と同様に、切断部位の前後 50 塩基での平均 RO_{site} / RO_{gene} 値の TOP 10%、BOTTOM 10%の切断率を比較したが、平均 RO_{site} 値の結果と比べ (図 2-37A)、切断率の差は大きくならなかった (図 2-37B)。加えて、切断部位の前後 50 塩基での平均 RO_{site} 値の TOP 10%、BOTTOM 10%間の切断部位周辺の塩基比率を比較したが、両者に大きな違いは認められなかった (図 2-37C, 図 2-37D)。出芽酵母についても、CDS 領域に存在する切断部位の前後 50 塩基での平均 RO_{site} 値や RO_{site} / RO_{gene} 値を算出し、TOP 10% (n = 61,513 sites)、BOTTOM 10% (n = 61,513 sites) 間の切断率を比較した場合も、シロイヌナズナ、ショウジョウバエと同様の傾向が認められたことから (図 2-38)、異なる生物種においても切断部位周辺のリボソーム存在量は配列に依存せず、切断率に正の影響を及ぼすことが考えられた。

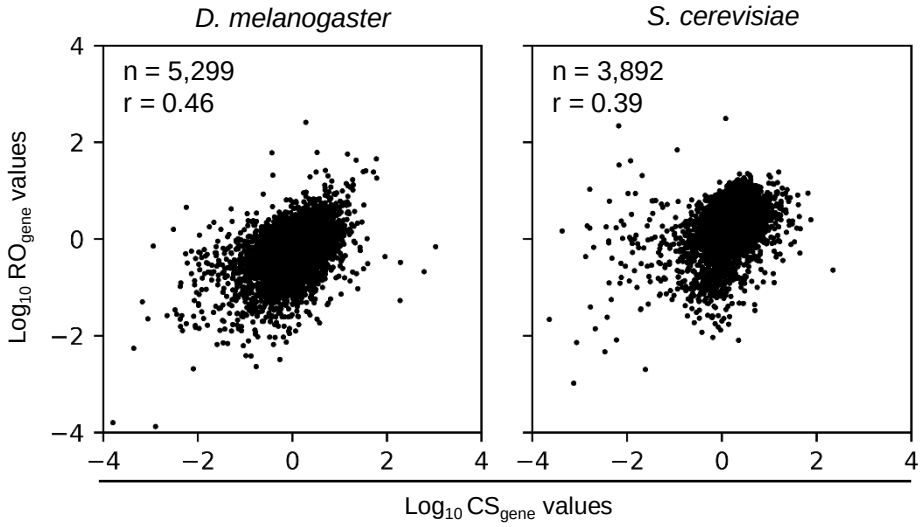


図 2-36. 遺伝子単位でのリボソーム存在量と切断率（ショウジョウバエ、出芽酵母）

TREseq 法で解析対象とした遺伝子の内、リボソームプロファイリング情報を持つ遺伝子を対象に RO_{gene} 値と CS_{gene} 値のピアソンの積率相関係数を算出した。

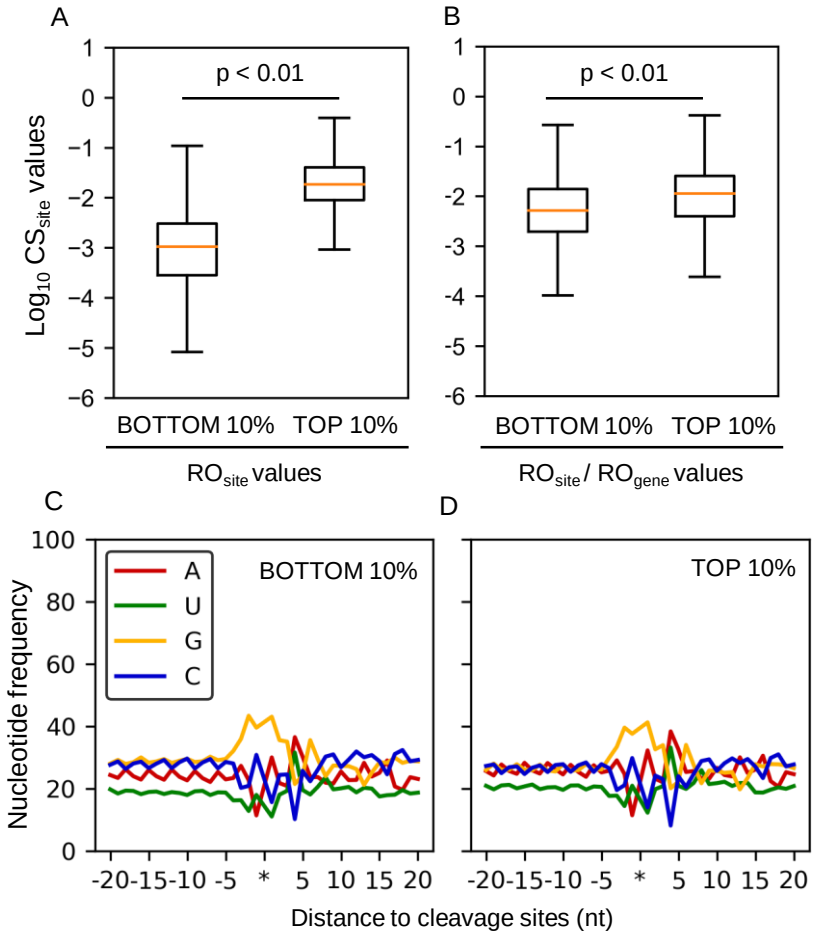


図 2-37. 切断部位周辺のリボソーム密度と切断率（ショウジョウバエ）

切断部位の前後 50 塩基での平均 RO_{site} 値を算出し、平均 RO_{site} 値の TOP 10%、BOTTOM 10%の切断率を比較した (A)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。また、RO_{site} 値を RO_{gene} 値で除算した場合の RO_{site} / RO_{gene} 値についても算出し、切断部位の前後 50 塩基での平均 RO_{site} / RO_{gene} 値の TOP 10%、BOTTOM 10%の切断率についても比較を行った (B)。また、平均 RO_{site} 値の TOP 10%、BOTTOM 10%間の切断部位周辺の塩基比率についても比較を行った (C, D)。統計検定には Welch's t-test を使用した。

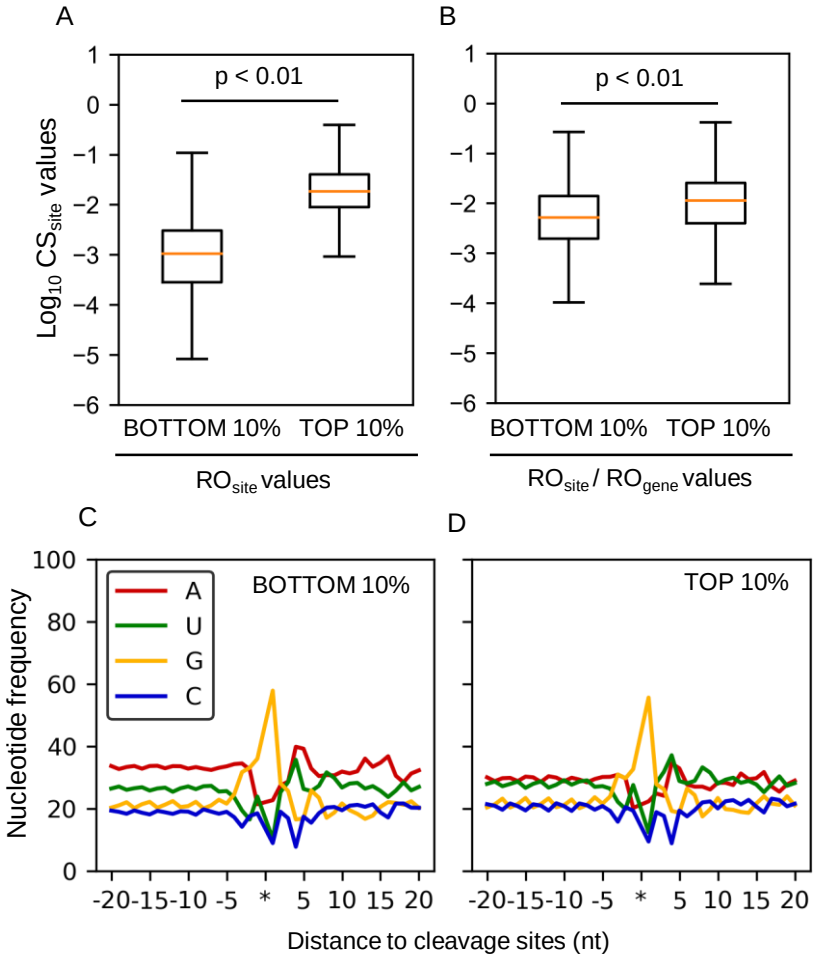


図 2-38. 切断部位周辺のリボソーム密度と切断率（出芽酵母）

切断部位の前後 50 塩基での平均 RO_{site} 値を算出し、平均 RO_{site} 値の TOP 10%、BOTTOM 10%の切断率を比較した (A)。箱ひげ図は上から、最大値、第 1 四分位数 (75%)、中央値、第 3 四分位数 (25%)、最小値を示す。外れ値は省略した。また、RO_{site} 値を RO_{gene} 値で除算した場合の RO_{site} / RO_{gene} 値についても算出し、切断部位の前後 50 塩基での平均 RO_{site} / RO_{gene} 値の TOP 10%、BOTTOM 10%の切断率についても比較を行った (B)。また、平均 RO_{site} 値の TOP 10%、BOTTOM 10%間の切断部位周辺の塩基比率についても比較を行った (C, D)。統計検定には Welch's t-test を使用した。

2-3-13. ショウジョウバエ、出芽酵母における切断部位周辺の配列と RNA 内での切断部位の分布

第二章の 2-3-5 で示されたように、シロイヌナズナを用いた解析で、切断部位周辺の配列モチーフが CDS 内で 3 塩基単位の周期性を持ち、切断部位の RNA 内での分布と同様の位相が示された。そこで、これらの傾向が、ショウジョウバエ、出芽酵母についても認められるか解析を行った。まず、開始、終止コドン周辺の塩基比率を算出したところ、シロイヌナズナと同様に RNA 内での切断部位の分布と類似する傾向が各塩基比率について認められた (図 2-39, 図 2-40, 図 2-41, 図 2-42)。次に、ショウジョウバエ、出芽酵母の切断部位周辺の配列情報を取得し、切断部位の上流 5 塩基、下流 15 塩基を MEME の motif letter-probability matrix lines 形式 (配列モチーフ) に変換し、FIMO を用いて RNA 内での配列モチーフの分布を算出した (図 2-43, 図 2-44)。その結果、配列モチーフは開始コドンの周辺で存在比率のピークが認められ、CDS 内で 3 塩基単位の周期性が存在し、位相に関しても切断部位の分布と同じであった (図 2-43, 図 2-44)。一方で、終止コドン-3 位付近の切断部位の存在比率については、シロイヌナズナ、ショウジョウバエ、出芽酵母などの 3 種間で、配列モチーフや 5' RPF の分布では説明できないことから、今回の解析で使用した特徴とは異なる配列や他の要因が関与している可能性が考えられた。

今回の解析から、シロイヌナズナ、ショウジョウバエ、出芽酵母など異なる生物種で共通して配列モチーフが RNA の切断部位の決定に重要であることが示された。

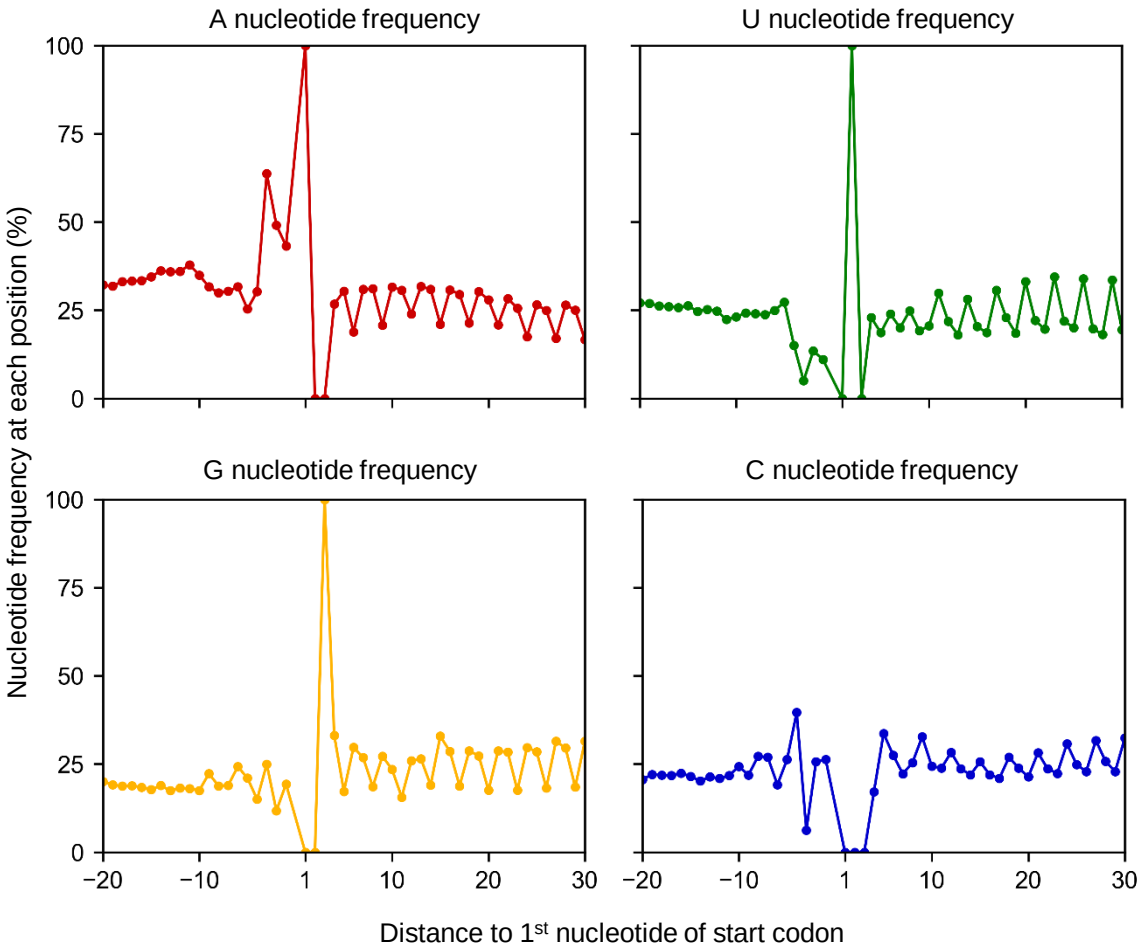


図 2-39. 開始コドン周辺の塩基比率 (ショウジョウバエ)

ショウジョウバエの RNA 配列情報を取得し、開始コドン周辺の塩基比率を算出した。X 軸は開始コドンからの距離を示し、Y 軸は各位置での塩基比率を示す。

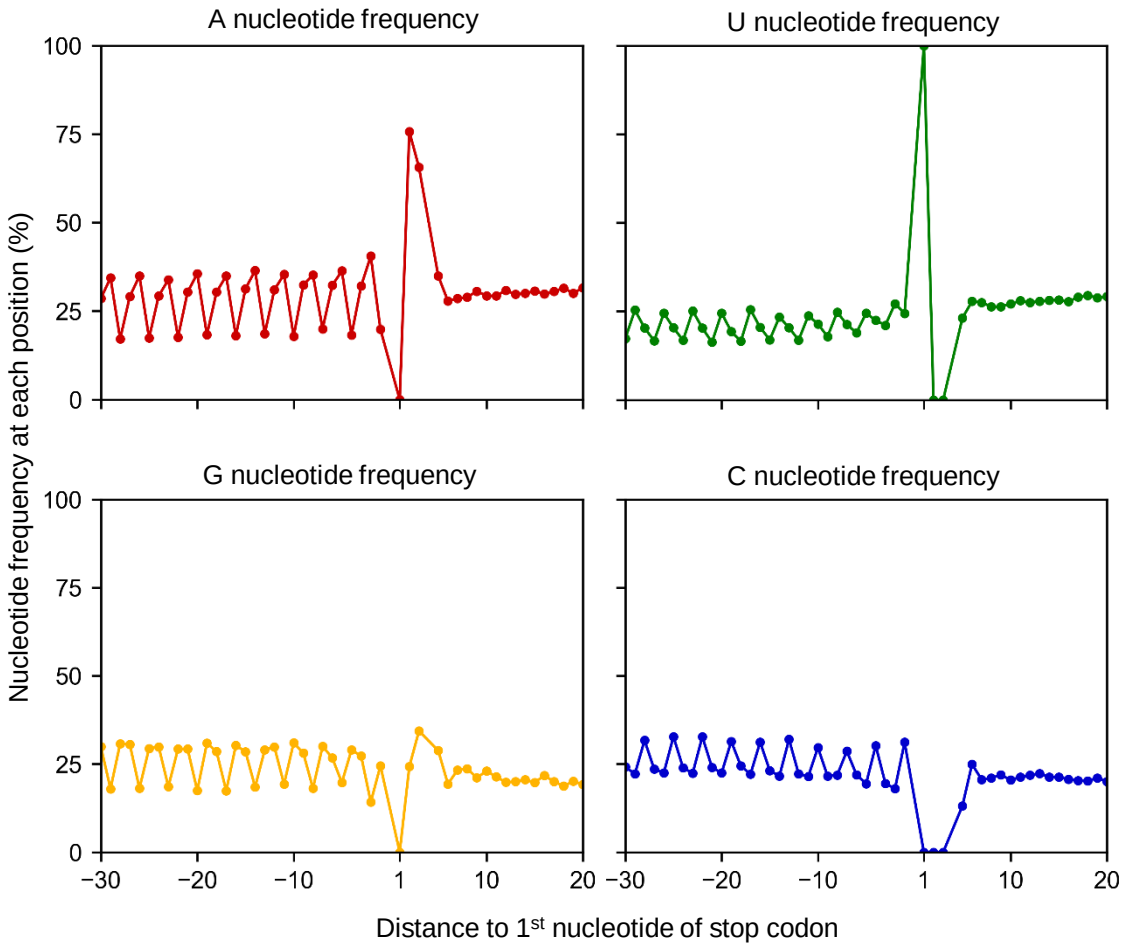


図 2-40. 終止コドン周辺の塩基比率 (ショウジョウバエ)

ショウジョウバエの RNA 配列情報を取得し、終止コドン周辺の塩基比率を算出した。X 軸は終止コドンからの距離を示し、Y 軸は各位置での塩基比率を示す。

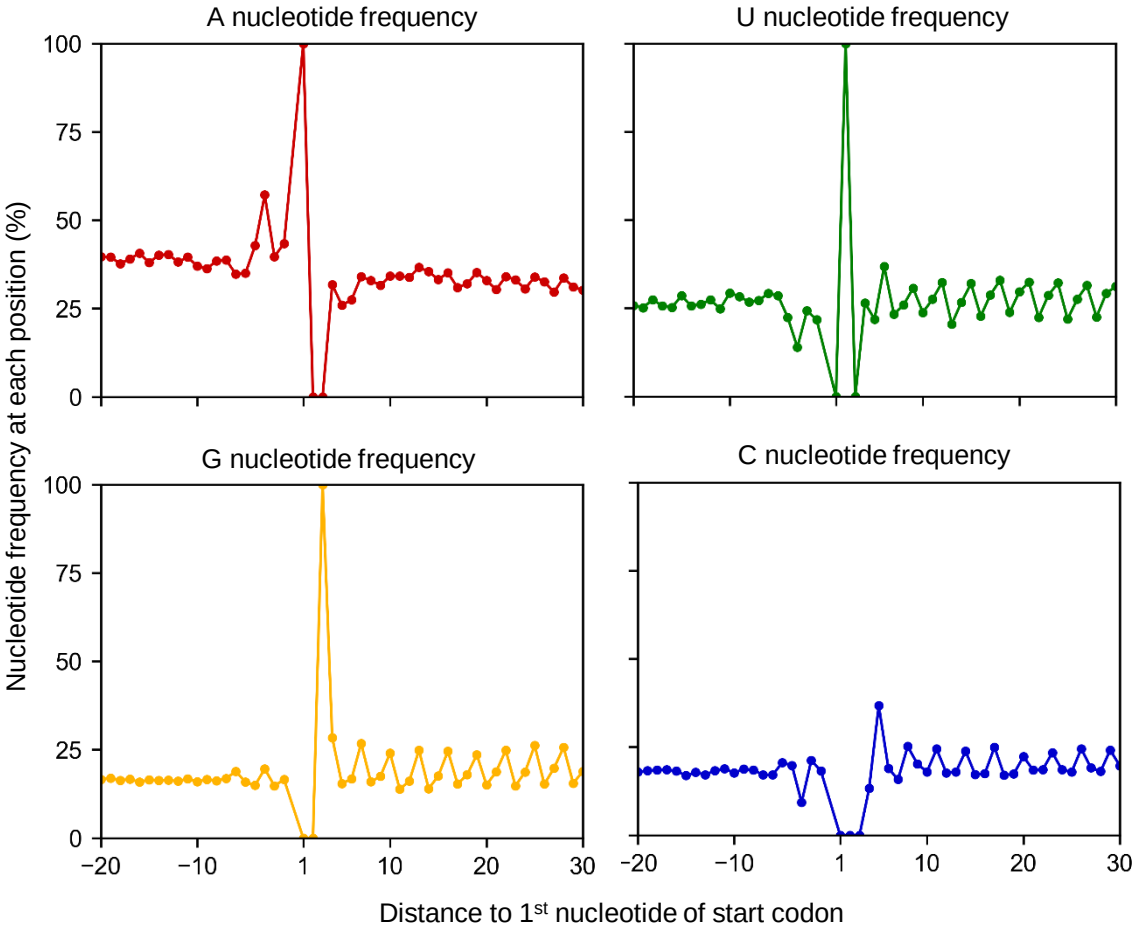


図 2-41. 開始コドン周辺の塩基比率 (出芽酵母)

出芽酵母の RNA 配列情報を取得し、開始コドン周辺の塩基比率を算出した。X 軸は開始コドンからの距離を示し、Y 軸は各位置での塩基比率を示す。

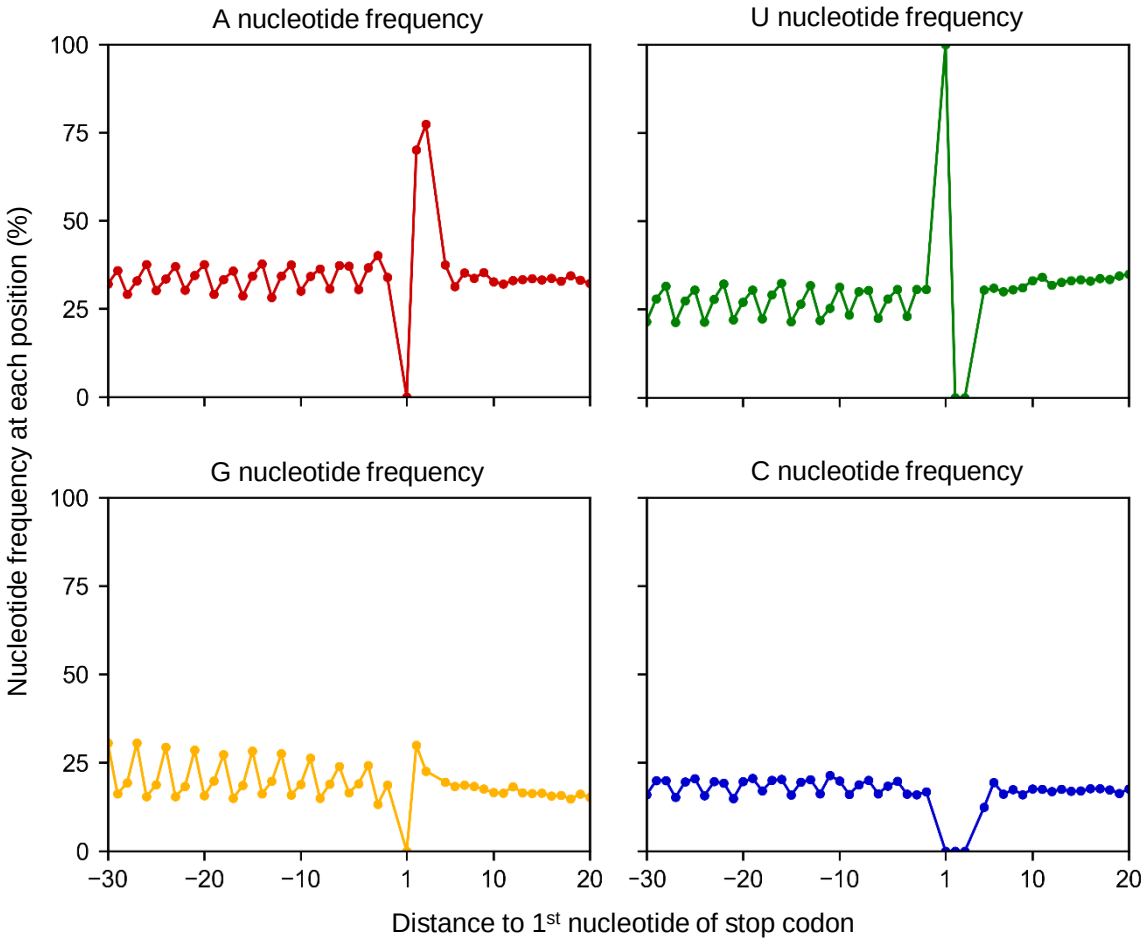


図 2-42. 終止コドン周辺の塩基比率 (出芽酵母)

出芽酵母の RNA 配列情報を取得し、終止コドン周辺の塩基比率を算出した。X 軸は終止コドンからの距離を示し、Y 軸は各位置での塩基比率を示す。

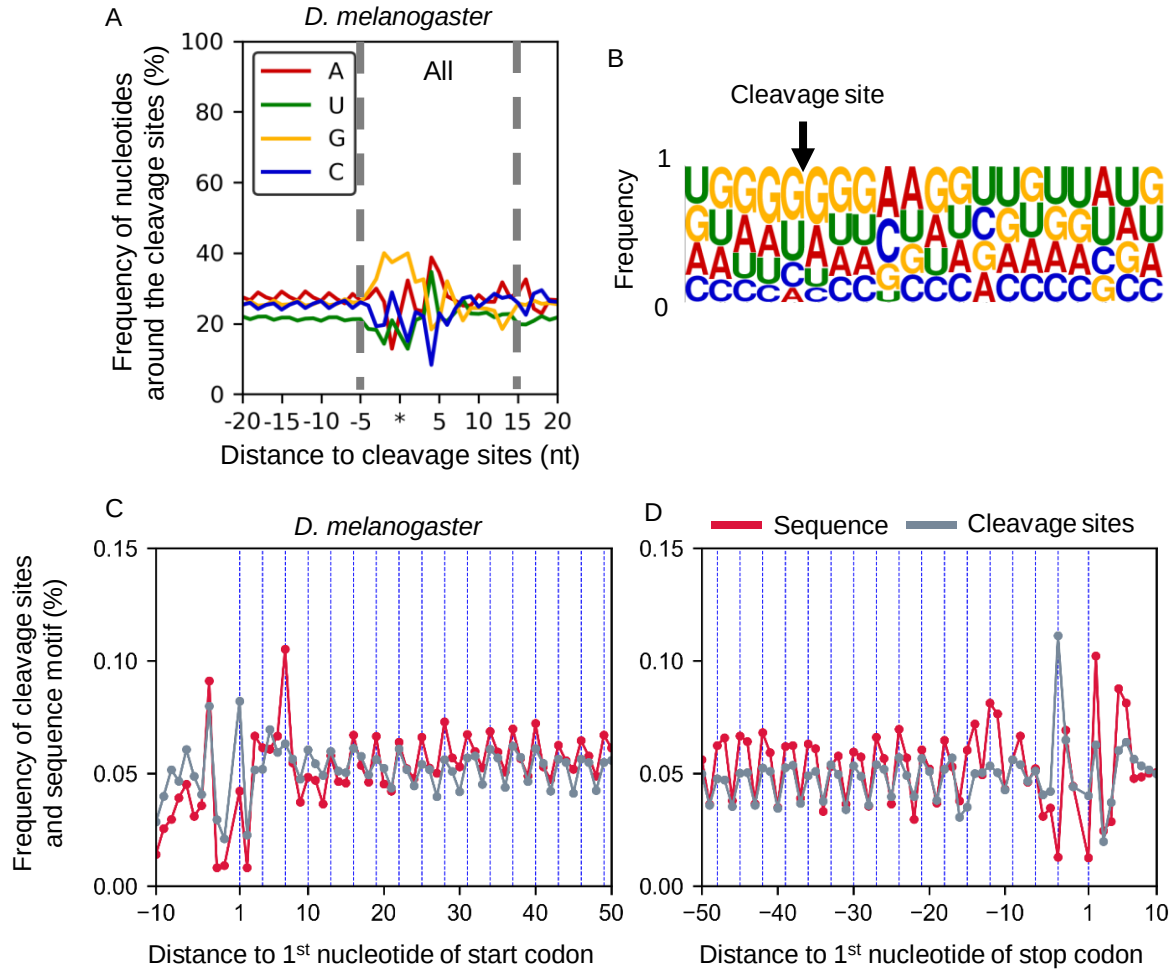


図 2-43. 開始、終止コドン周辺の切断部位、配列モチーフの分布（ショウジョウバエ）

切断部位周辺の配列を用い、MEME の motif letter-probability matrix lines 形式 (配列モチーフ) に変換し、FIMO を用いて RNA 内での分布を算出した (A, B)。開始、終止コドンからの距離を算出後、各位置での配列モチーフ、切断部位の存在比率を算出した (C, D)。X 軸は開始コドン、終止コドンからの距離を示す。Y 軸は各位置での切断部位と配列モチーフの存在比率を示す。

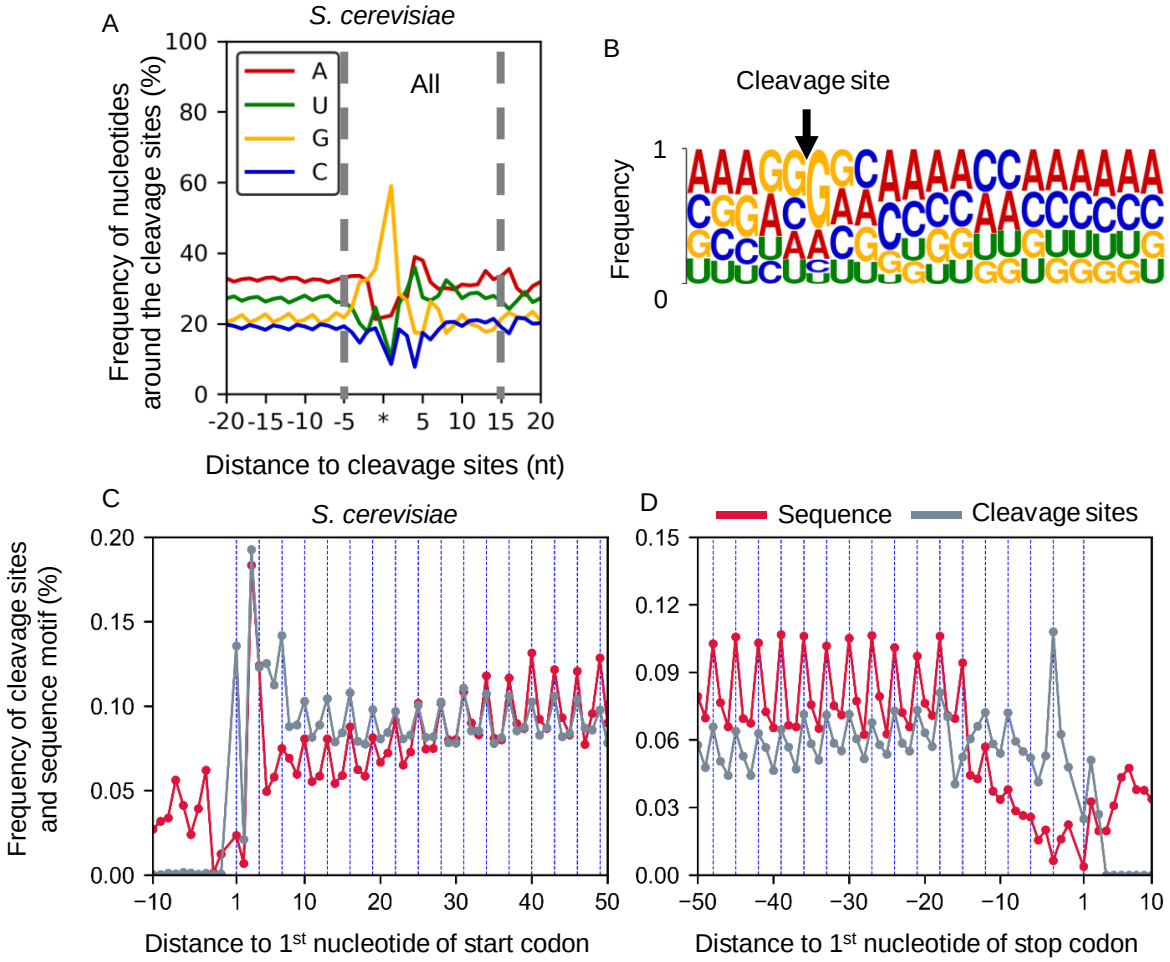


図 2-44. 開始、終止コドン周辺の切断部位、配列モチーフの分布（出芽酵母）

切断部位周辺の配列を用い、MEME の motif letter-probability matrix lines 形式（配列モチーフ）に変換し、FIMO を用いて RNA 内での分布を算出した（A, B)。開始、終止コドンからの距離を算出後、各位置での配列モチーフ、切断部位の存在比率を算出した（C, D)。X 軸は開始コドン、終止コドンからの距離を示す。Y 軸は各位置での切断部位と配列モチーフの存在比率を示す。

2-4. 考察

2-4-1. 生物種間における切断率と RNA 安定性

RNA 半減期情報は、ポリ A 鎖の短縮に依存する分解、そして、エンドヌクレアーゼ等による RNA 切断に依存する分解の影響が混在した結果である。酵母では、一般的にポリ A 鎖の短縮に依存する分解が主要な RNA 分解機構であると考えられているが (12)、今回の解析からは、他の生物種と比べ、遺伝子単位の切断率と半減期との間により強い負の相関が認められた (図 1-12、図 1-13、図 2-18、図 2-19)。一般的に RNA 半減期を再現性良く定量的に測定することは難しく、各手法間で全体的な傾向は類似するが、算出される各 RNA 種の半減期は手法ごとに異なることが知られている (43)。そのため、CS_{gene} 値と RNA 半減期との関係性は、対象とした生物種や半減期測定法によって異なることが予想されるため、一概に今回の解析から他の生物種と比較して、出芽酵母で RNA 切断が半減期により強い影響を与えているとは断言できない。しかし、少なくとも、各生物種間で共通して CS_{gene} 値が高いほど半減期が短い傾向が認められたことから、RNA 切断機構は生物種を問わず、RNA の安定性に大きく寄与していると考えられる。この知見は、従来の切断部位解析では認められなかったものである。

2-4-2. RNA 構造が切断部位に与える影響

第二章の 2-3-3、2-3-11 で示されたように、RNA 切断には RNA 高次構造が関与し、その構造は G 塩基の比率と密接に関わっていることが考えられた。G 塩基と RNA 構造を考えた際に、グアニン四重鎖が知られている。グアニン四重鎖は非古典的な RNA の立体構造として知られており、G リッチな配列が連続することで、強固な立体構造が形成される。このような RNA の立体構造は、リボソームが RNA 上を伸長する際の障壁となるため、グアニン四重鎖は翻訳抑制に関与することが哺乳類で知られている (57, 58)。しかし、今回の解析では、図 2-9、図 2-33 に示されるように、切断部位周辺でリボソーム存在比率が顕著に高い傾向は認められず、また、リボソームの停滞率を概算した際も、停滞率を算出する前のリボソーム存在量を基にした結果と大きな違いは認められなかった (図 2-12, 図 2-37, 図 2-38)。加えて、グアニン四重鎖は G リッチな配列が複数回くり返されることから、G 比率が高い領域は、10 塩基から 20 塩基ほどになるが、今回の解析では、切断部位周辺の G 塩基比率が高い領域は長くても 5、6 塩基ほどであった (図 2-21)。これらの結果を踏まえると、切断部位周辺の RNA 構造は G 塩基に依存しているが、グアニン四重鎖ではないと考えられる。

このような、RNA 構造に着目した解析は、配列情報のみを使用し、RNA 高次構造を予測するソフトウェアが主に用いられてきたが、実際に細胞内で形成される RNA 構造とは異なる可能性も指摘されている。これまでの方法を改良した手法として、近年、 DMS-MaPseq 法が哺乳類で確立されている (59)。この手法では DMS 処理をすることで塩基対を形成していない塩基がメチル化修飾されることを利用し、次世代シーケンサーを用いて実際に RNA のメチル化修飾を解析することで、RNA の構造を推定する。そのため、従来の手法と比較し、より正確に細胞内での RNA の高次構造を把握することが可能になっている。将来的には、これらの手法を用いて取得した情報を組み合わせることで、RNA 構造が切断に与える影響をより詳細に解析できると考えられる。

しかし、少なくとも今回の解析からは、第二章の 2-3-9 で示したように、切断率が高い配列から共通の RNA 構造が検出されなかった結果を踏まえると、RNA 構造は切断には大きく関与せず、切断部位の周辺に存在する配列が重要であると考えられた。

2-4-3. RNA 内で認められた切断部位の 3 塩基単位の周期性

第二章の 2-3-4 や 2-3-12 の結果から、遺伝子単位、切断部位単位の双方で、リボソームの存在量は切断率に正に関与することが示されている。これまでに、翻訳過程が関わる RNA 切断機構として no-go decay (NGD) が報告されている。NGD では二次構造やレアコドン、コードするアミノ酸配列などによって、リボソームが停止、停滞し、RNA が切断される (17, 60)。この NGD はリボソームが複数個連なって停滞することで生じると考えられており、リボソームの翻訳伸長を止めるシクロヘキシミドによって NGD は阻害される (60)。RNA 内での切断部位の分布に着目すると、CDS 内で 3 塩基単位の周期性が認められ、この傾向は TREseq 法、そして従来手法を用いた場合も同様に検出されている。CDS 領域で認められた 3 塩基単位の周期性が NGD による RNA 切断であるならば、NGD を阻害するシクロヘキシミド処理によって、これらの周期性は消失すると考えられる。しかし、Yu らや、Ibrahim らの研究で示されているように、シクロヘキシミド処理後の切断部位の分布を解析した場合でも、開始コドン側でより明確な 3 塩基単位の周期性が検出される、もしくは処理の前後で変化は認められていない (18, 22, 23)。これらの結果は、酵母、植物、動物など、生物種を問わず報告されていることから (18, 22, 23)、CDS 領域での切断部位の 3 塩基単位の周期性は主として NGD に由来する RNA 切断ではないと考えられる。

一方で、CDS 領域での 3 塩基単位の周期性は、RNA 内部の切断ではなく、5' - 3' のエキソヌクレアーゼによる分解途中の RNA 末端である可能性が酵母で

提唱されている (23)。Pelechano らは、従来手法を用いて網羅的に RNA の 5' 末端情報を取得後、リボソームの RNA 内での分布との比較を行った。その結果、両者は CDS 領域で 3 塩基単位の周期性を持つこと、そして、終止コドン周辺の CDS 領域に着目するとエキソヌクレアーゼの変異体では、RNA 末端の 3 塩基単位の周期性が消失することから、これらの周期性は、5' - 3' エキソヌクレアーゼによる消化がリボソームにより保護されることによって生じると Pelechano らは主張している。それに対して Ibrahim らは、従来手法を一部改変した手法を用いて動物細胞を対象に解析を行ったところ、5' - 3' エキソヌクレアーゼの変異体でも、RNA 末端の明確な 3 塩基単位の周期性が CDS 領域で認められることを報告している。Ibrahim らは、この結果から 3 塩基単位の周期性は RNA 切断によって引き起こされたものであると主張している (22)。Pelechano らが行った解析では、検出されたリード数を基に RNA 内での切断部位の分布が算出されているため、第一章の 1-4-1 に述べたように、エキソヌクレアーゼ変異体では CDS 領域で検出されるリード数が過少評価されていることや、RNA 蓄積量が高い一部の遺伝子の結果を反映している可能性が考えられた。そこで、Pelechano らが行った解析データを GEO データベースより取得し (23)、再解析を行った。再解析では、開始、終止コドン付近で検出された全 RNA 末端数に対する各部位での存在比率を算出した。また、Pelechano らは終止コドン側のみに着目し解析を行っていたため、再解析では開始コドン側についても RNA 末端の分布を算出した。その結果、終止コドン側では野生型と比べエキソヌクレアーゼの変異体でわずかに 3 塩基単位の周期性は弱まっていたが、開始コドン側では野生型と変異体間でほぼ同等の周期性が認められた (図 2-45)。この結果は、開始コドン側から 5' - 3' エキソヌクレアーゼによる消化により、CDS 内で 3 塩基単位の周期性が形成されるという Pelechano らの主張とは異なることから、これらの RNA 末端はエキソヌクレアーゼによる RNA 保護断片ではなく、NGD 以外の切断機構により切断された RNA 切断部位であると考えられた。

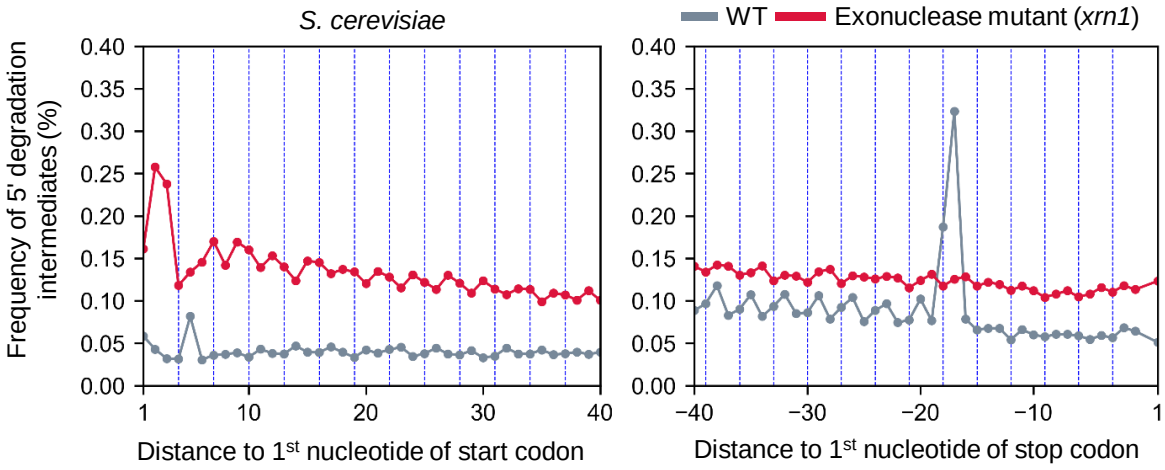


図 2-45. エキソヌクレアーゼ変異体における RNA の 5' 末端の分布

Pelechano らが行った網羅的な分解産物データを GEO データベースより取得し、開始、終止コドンからの距離を算出した。その後、各位置での RNA の 5' 末端の存在比率を求めた。灰色は野生型、赤色は 5' - 3' エキソヌクレアーゼの変異体を示す。X 軸は開始コドン、終止コドンからの距離を示す。Y 軸は各位置での RNA の 5' 末端の存在比率を示す。

2-4-4. 切断部位周辺のコドン、コードするアミノ酸配列と RNA 切断

翻訳の際、RNA 上をリボソームが進み、コドンに対応するアミノ酸が運搬され、リボソームが翻訳伸長する。この際、各コドンに対応する tRNA 量が豊富に存在するほど、リボソームが進む速度は速く、対応する tRNA 量が少ないコドンほど、リボソームが進む速度も遅いと考えられている (43)。また、非最適コドン配列をレポーター遺伝子に挿入すると、リボソームの停止、停滞が誘導され、RNA が切断されることが報告されている (17, 60)。リボソームの停止、停滞に依存する RNA 切断は、コードするアミノ酸配列によっても引き起こされることが知られている (17, 60)。新規に合成されたペプチド鎖は、負の電荷を帯びたリボソームタンパク質が存在するリボソームの狭窄部位を通過するため、新生ペプチド鎖にアルギニンやリシンなど、正の電荷を帯びたアミノ酸残基が存在する場合、リボソームは停止、停滞し、RNA が切断されやすい (30)。これらの個別遺伝子を対象にした解析から、RNA 切断にはコドンやコードするアミノ酸配列が関与することが報告されていた。

一方で、網羅的な切断部位解析では、切断部位周辺のコドンやコードするアミノ酸配列が RNA 切断に与える影響についてこれまで考察は行われてこなかった。TREseq 法で検出した切断部位に着目した場合、切断部位の周辺配列には G 塩基比率が高い傾向が認められていた。この配列的特徴は、コドンやコードするアミノ酸配列に由来する可能性も考えられたが、第二章の 2-3-10 の結果から、CDS 領域と UTR 領域の切断部位周辺の配列を比較した際に、両者の領域で同様の傾向が認められている (図 2-21)。通常、翻訳の伸長反応が生じない UTR 領域においても、CDS 領域と同様の塩基比率の傾向が認められたことを考慮すると、切断部位の位置、切断率に関わる配列的特徴は、コドンやコードするアミノ酸配列ではなく塩基配列が直接関与していると考えられる。これらの結果は、個別遺伝子の解析から、コドンやコードするアミノ酸配列が切断に関与する場合があるが (17, 60)、全体としては塩基の配列パターンが RNA 切断に主として関与することを示している。

2-4-5. 翻訳過程が RNA 切断の位置に与える影響

これまで、翻訳過程が RNA 切断に関与することが示唆されていたが、植物では、リボソームの存在位置、および存在量に関する情報を取得し、切断部位情報との比較を行なった解析は存在しなかった。本研究では、シロイヌナズナを対象に、リボソームの存在位置、存在量に関する情報を取得し、切断部位の位置に与える影響に着目し解析を行なった結果、第二章の 2-3-4 に示すように、リボソームの存在位置や存在量は RNA 切断の位置に大きな影響を与えないこ

とが示唆された。加えて、切断部位周辺に頻出する配列モチーフを用いた解析を行った結果 (図 2-15)、配列モチーフは RNA 内の切断部位の分布と同様であったことから、切断部位の位置決定には、配列パターンが大きく関与していると考えられた。加えて、第二章の 2-3-12、2-3-13 で示されたように、出芽酵母、ショウジョウバエでも同様の傾向が認められたことから、異なる生物種においても RNA 切断部位の決定には配列パターンが重要であることが示された。

2-4-6. RNA 上に存在するリボソーム量と切断率との関係性

これまで様々な生物種で翻訳過程と RNA の安定性に着目した解析が行われており、リボソームの翻訳伸長を抑制するシクロヘキシミド処理を行うことで、RNA が安定的になることが植物、酵母、動物などで報告されている (61–63)。

RNA の分解機構に着目した場合、ポリ A 鎖の短縮に依存する分解機構、RNA 切断に依存する分解機構に大別することができるが、翻訳過程が両者に与える影響は異なっている。ポリ A 鎖の短縮に依存する分解機構に着目した場合、脱キャッピング酵素である Dhha1p やポリ A 鎖の短縮に関わる CAF は、RNA 上のリボソーム存在量が少ない、翻訳効率が低い RNA に積極的に作用することが酵母で報告されている (64, 65)。その一方で、RNA の切断に着目した場合、RelE のようにリボソームが多く存在する RNA を積極的に分解するなど、分解機構ごとに翻訳過程が与える影響は異なると考えられる (6)。これまで、植物で実際にリボソームの存在量に関する情報を取得し、RNA の切断とリボソームの存在量を比較した解析は存在しなかった。本解析において、遺伝子単位、そして各切断部位での切断率とリボソームの存在量を比較した結果、両者には正の関係性が認められ (図 2-11、図 2-12)、植物でも、リボソームの存在量が RNA 切断に関与することを初めて明らかにした。加えて、ショウジョウバエ、出芽酵母でも同様の傾向が認められたことから (図 2-36, 図 2-37, 図 2-38)、広い生物種で RNA 上のリボソーム存在量が切断率に正に関与し、RNA 上のリボソーム量が多いほど、RNA が切断、分解されやすいことが示された。

これまで、RNA 切断の位置にリボソームの位置が大きく関与すると考えられてきたが、今回の解析で、リボソームの存在量は切断率に正の影響を与えるが、切断の位置決定には関わらず、切断部位の周辺の配列モチーフに依存して RNA が切断されることが示された。

2-4-7. 想定される RNA 切断に関与するトランス因子

第二章の図 2-1 の結果から、RNA 切断には特異的な配列が重要であり、G 塩基の直前で RNA が切断されると考えられた。しかし、このような切断に関わ

るタンパク質因子は植物ではこれまで報告されていない。他の生物種に着目すると、リボソームと相互作用し、G 塩基の直前で RNA を切断する RelE (6) が原核生物で同定されている。TREseq 法で検出された切断部位の解析から、RNA 上のリボソーム存在量が多いほど切断が生じやすいことや、切断部位周辺の塩基比率を調べると G 塩基比率が高いことから、RelE のホモログ遺伝子がシロイヌナズナでの RNA 切断に関与することが想定された。しかし、RelE のアミノ酸配列を基にブラスト検索を行った結果、RelE と類似するアミノ酸配列を持つ遺伝子はシロイヌナズナで存在しなかった (data not shown)。一方で、今回の解析から、植物、酵母、ショウジョウバエで RNA 切断に関わる特徴が類似していたことから、切断に関与するタンパク質因子も保存されている可能性が考えられた。真核生物で幅広く保存されており、細胞質でのエンドヌクレアーゼ活性を持つタンパク質因子は複数知られているため (rRNA のプロセシングに関与する Nob1 など) (66)、このような異なる生物種間で高度に保存されているタンパク質因子の変異体を作出し、TREseq 法を用いた切断部位の検出を行うことで、図 2-21 で示したような切断部位周辺の G 塩基比率が高い領域を切断するタンパク質因子の同定が期待できる。

第三章

数理モデルを用いた植物 RNA 切断に関わる要因の特徴選択

3-1. 序論

第一章、第二章では、網羅的に RNA の切断部位を同定し、それらの切断に関わる要因について個別に解析を行った。切断部位の位置に着目した場合、第二章の 2-3-5、2-3-13 に示すように、切断部位の位置決定には RNA 上のリボソーム位置ではなく配列モチーフが重要であること示された。それに対して、切断率については、塩基配列に加えて、翻訳状態など複数の要因が関与することが想定された。このような複合的な要因が関連する現象に対して、重要な特徴を選択、評価する方法として、分類、もしくは回帰モデルを用いた特徴選択が挙げられる。これらのモデル構築により、説明変数と目的変数から統計的手法を用いることで式を推定した後、各説明変数に割り当てられる係数を基に、その特徴の重要性を評価することができる。例えば、切断・非切断のような、データが属するクラスを予測 (分類) する手法として、K-近傍法やサポートベクターマシーン (SVM)、ランダムフォレストなどが知られており、ゲノム情報の配列から遺伝子領域を予測することや、遺伝子発現量を基にした疾病の診断予測などに利用されている (67)。中でも、アンサムブル学習として知られているランダムフォレストは、高い識別性能を持つことが報告されている (67)。アンサンブル学習とは、複数の識別器 (弱識別器) を統合させて 1 つの学習モデルを生成させる手法であり、ランダムフォレストの場合は、決定木と呼ばれる弱識別器を使用している。決定木とは、任意の基準を設け、その基準を基にデータを分類していく手法である。ランダムフォレストでは、重複を許したデータ抽出を行い決定木を生成後 (ブーストトラップ)、各弱学習機の結果を統合し、データの識別、分類を行う (バギング) (図 3-1)。

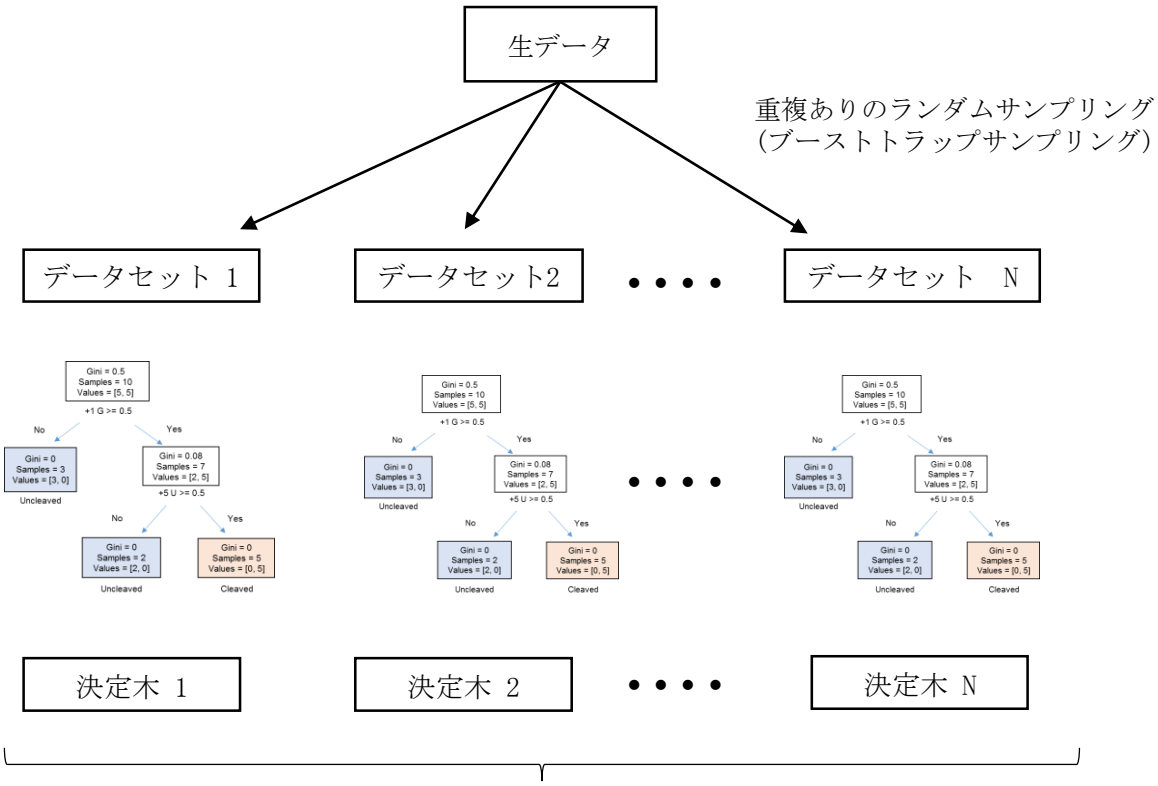


図 3-1. ランダムフォレストの概念図

ランダムフォレストは、生データより重複ありのランダムサンプリングを行い、決定木を構築していく。最終的には、決定木の結果を統合し、データの分類を行う (バギング)。

また、各切断部位の切断率のようなデータ値を予測する手法として、回帰モデルが挙げられる。回帰モデルとしては、予測値と実測値の二乗誤差を最小化(最小二乗法) するように係数を変化させ式を算出する重回帰分析が古くから用いられている。この際、説明変数の係数を基に各特徴が目的変数（切断率）に与える影響を評価でき、正の方向に数値が高いほど目的変数に正の影響を与え (切断率が高い)、負の方向に数値が高いほど、目的変数に負の影響 (切断率が低い) を与える要因であると推定できる。特に、最近では重回帰分析を改良した、Least Absolute Shrinkage and Selection Operator (ラッソ回帰) やリッジ回帰などが遺伝子発現における転写や翻訳過程に着目した解析に用いられている (68, 69)。これらの回帰モデルは、訓練データでの過剰な学習 (過学習) を防ぎ未知データに対する汎化能力を高めるために、予測値と実測値の残差に加えて、回帰係数の総和、もしくは回帰係数の二乗和を最小化する (70)。特に、回帰モデルの解釈性や次元削減を重要視した場合、スパースモデリングであるラッソ回帰を用いた特徴選択が行われている (68, 69)。スパースは「密度が低い」ことを意味し、スパースモデリングは現象を構成する本質的な情報はごくわずかであるという仮説 (スパース性) に基づき、入力された情報から重要な情報を抽出するモデルである。実際に、先行研究にて転写過程や翻訳過程など、複合的な要因の関与が想定されるデータを対象に、ラッソ回帰を用いた特徴選択が行われている。例えば、Qabaja らは、RNAseq 法によって得られた網羅的な RNA 蓄積量データや microRNA 発現量データを用いて、疾患に関与する約 20 種の microRNA を選抜している (68)。また、Hu らの研究では、シロイヌナズナの各 mRNA の翻訳状態に着目し、約 60 の特徴から RNA の翻訳状態に関わる十数の配列情報などの特徴を抽出している (69)。

このように、転写、翻訳過程に関する情報を対象にランダムフォレスト分類、もしくはラッソ回帰を用いた複数要因の評価が行われているが、RNA の切断については、先行研究を含め、単一の相関解析に留まり、真に重要な特徴の選抜や各特徴が切断率に与える影響の大小など複合的な要因を考慮した総合的な知見はない。上述したような数理モデルを用い、切断・非切断部位の決定、切断率に関わる複数の要因とその重要度を明らかにすることで、RNA の安定性という観点から、遺伝子発現機構を理解するための重要な知見が得られると考えられる。

そこで、本研究の第三章では、第一章、第二章で得られた知見を基に、RNA 切断、非切断部位の決定に多くの要因が想定される中で、何が重要であるかをランダムフォレスト分類を用いて検証するとともに、ラッソ回帰を用いて各切断部位の切断率関わる重要な特徴の選択、評価を行った。

3-2. 方法

3-2-1. ランダムフォレスト分類モデルの構築

3-2-1-1. 使用するモデルと解析対象とした RNA 切断部位

第二章の結果より、切断・非切断部位の決定には、RNA 上の配列が重要であることが認められたため、数理モデルを用いた検証を行った。数理モデルとしては、任意の RNA 上の部位が与えられた際に、その部位が切断、非切断かを予測するモデルを構築した。各遺伝子ごとに最も CS_{site} 値が高い部位を切断部位と定義し、前後 30 塩基以内に切断部位が検出されなかった部位を非切断部位と定義し解析を行った。加えて、各切断部位に関しては、第一章で使用した psRNAtarget を用いて microRNA の切断と予測される切断部位を解析から除外した。解析対象となった遺伝子を 9:1 の割合で、トレーニングデータ (遺伝子数 = 900; 切断部位数 = 900; 非切断部位数 = 3,289) とテストデータ (遺伝子数 = 100; 切断部位数 = 100; 非切断部位数 = 369) に分割後、トレーニングデータを用いてモデルを構築し、テストデータを用いてモデルの精度を検証した。分類モデルとしては、python の skit-learn ライブラリーのランダムフォレスト分類を用いた。

3-2-1-2. モデルに使用した説明変数

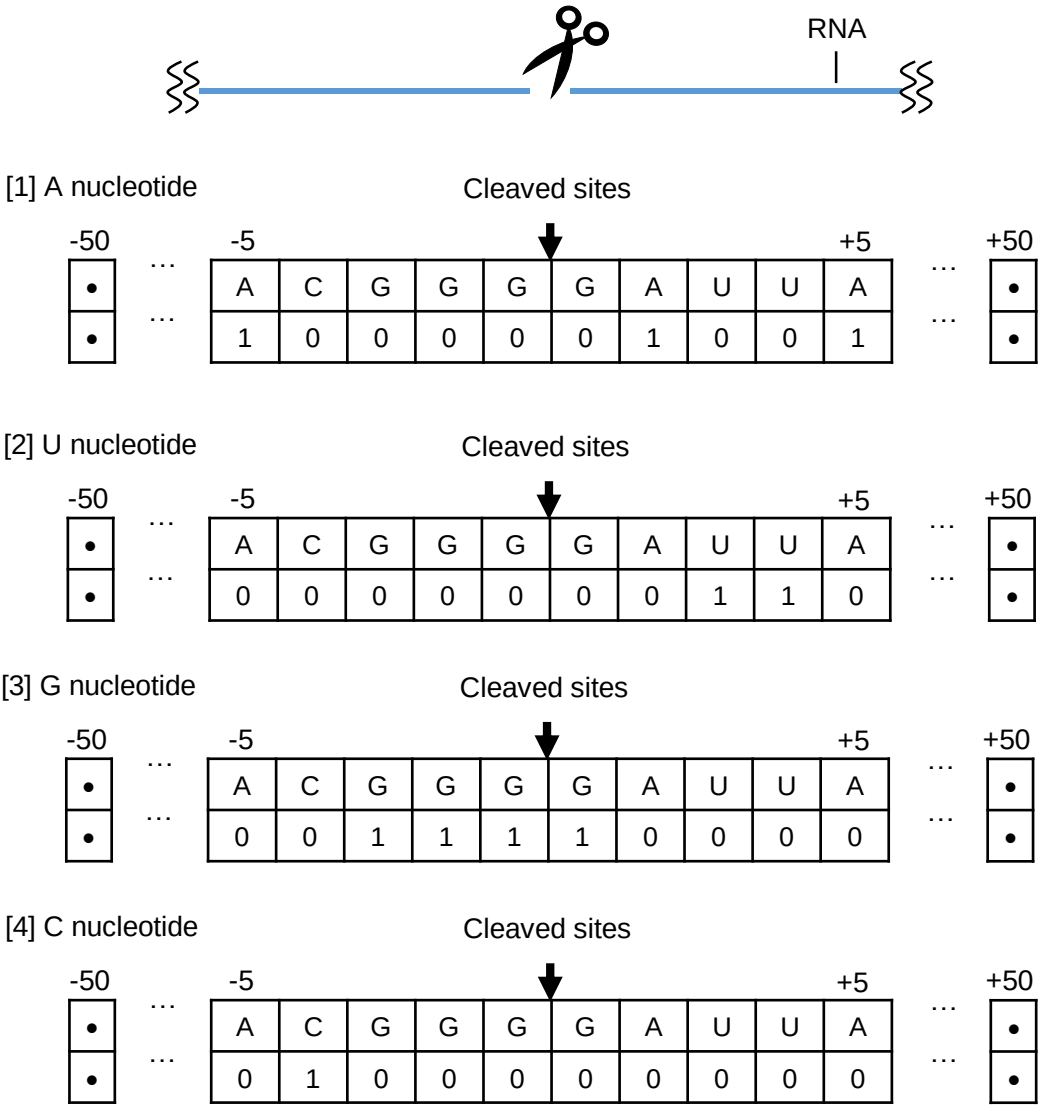
ランダムフォレストを用いた特徴選択では、切断部位周辺の塩基、RNA の高次構造、リボソームの位置および存在量情報を図 3-2 のようなデータ形式に整え解析を行った。RNA の高次構造情報に関しては、第二章の図 2-2 で示した各塩基での塩基対形成の有無を RNAfold を用いて予測し、解析に使用した。

3-2-1-3. ハイパーパラメーターの調整

python の skit-learn ライブラリーのランダムフォレストの決定木では、ジニ不純度 (Gini impurity) という指標値を用いて任意の基準を設定し、その基準を基にデータを分類していく(図 3-3)。ジニ不純度とは、分類したデータの不純度を示し、ジニ不純度が低いほどデータをきれいに分類できていることを意味している。例えば、対象となる部位が 10 個ある場合、切断部位が 5 つ、非切断部位が 5 つ存在すると、ジニ不純度は 0.5 となる (図 3-3 [A])。また、+1 位が G 塩基であるかという基準を設け、データを分割すると、+1 位が G 塩基である群 (7 つ) の中で非切断部位が 2 つ、切断部位が 5 つの場合は、ジニ不純度は 0.408… となる (図 3-3 [B])。加えて、分割前後のそれぞれのデータのグループをノードと呼び、それぞれ図 3-4 に示す名称である。各決定木のハイパーパラメーターについては、トレーニングデータを用い、表 3-1 に示す全通りの組み合わせを

行い、python の skit-learn ライブラリーのランダムフォレスト分類で算出できる Out of bag (OOB) score が最も高くなる値をテストデータ用のハイパーパラメーターとして使用した。図 3-1 で述べたように、ランダムフォレストは重複ありの選抜を行うため (ブーストトラップサンプリング)、一部のデータについては決定木の構築には使われないデータが存在する (OOB)。OOB score とは、これらの決定木に使用されなかったデータを用いて、各決定木の予測精度の平均値を算出した値である。

また、各特徴のジニ重要度 (Gini importance) を調べることで、どの特徴がデータの分類 (切断、非切断の決定) にとって重要であったかを知ることができる。決定木を構築する際に、各特徴ごとにジニ不純度が算出されるが、ジニ重要度は親ノードと子ノードのジニ不純度の減少率を基に算出される。ジニ重要度については、python のランダムフォレスト分類における feature_importances_ を用いて取得した。



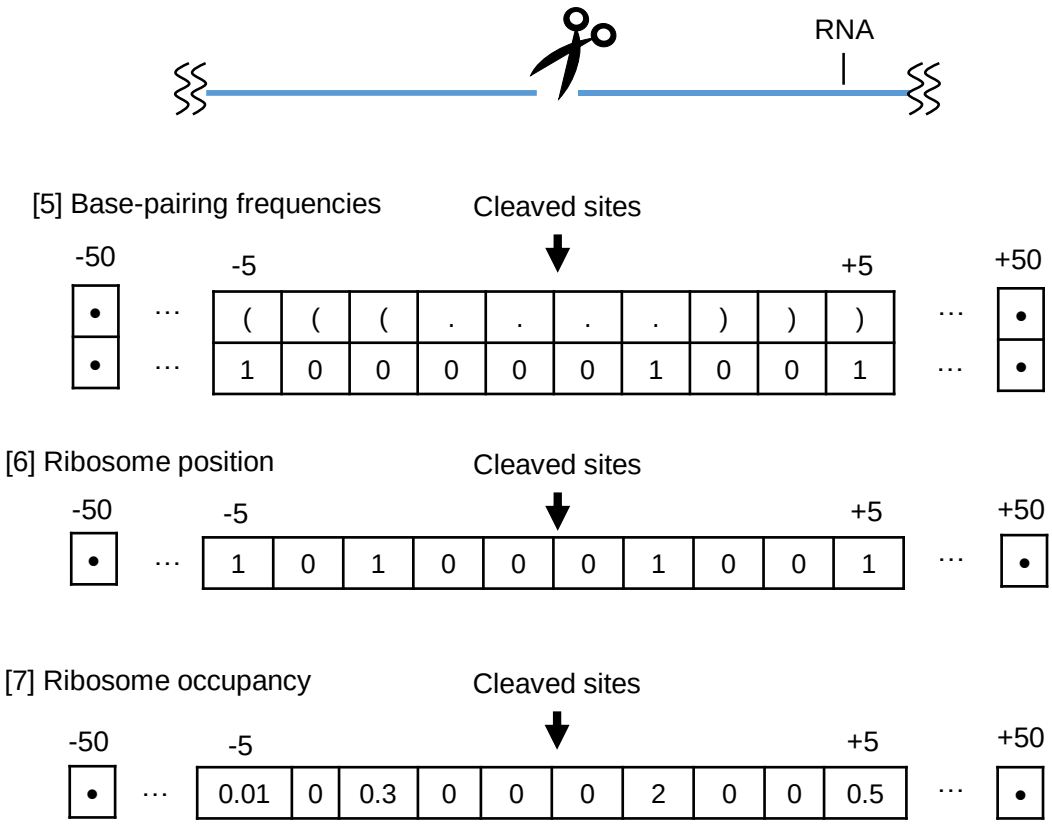


図 3-2. ランダムフォレストにおける特徴の抽出

切断部位の前後 50 塩基を使用し、配列情報を数値に変換した。各位置で対象とする塩基が存在した場合は 1、存在しない場合は 0 とした。また、塩基対形成情報については、塩基対の形成を示す鍵括弧を 1、塩基対をなさない点を 0 とした。加えて、リボソームの位置については、リボソームが存在する場合 1 を、存在しない場合は 0 とした。リボソーム存在量については、リボソームの位置情報に加え、その位置での RO_{site} 値を情報に加えた。

$$G(t) = \sum_{i=1}^{K} P(x_i|t)\,(1 - P(x_i|t)) = 1 - \sum_{i=1}^{K} P(x_i|t)^2$$

Gはジニ不純度、Kはクラス数、$P(x_i|t)P(x_i|t)$はあるノードtにおいて
クラスx_iが選ばれた確率を示す

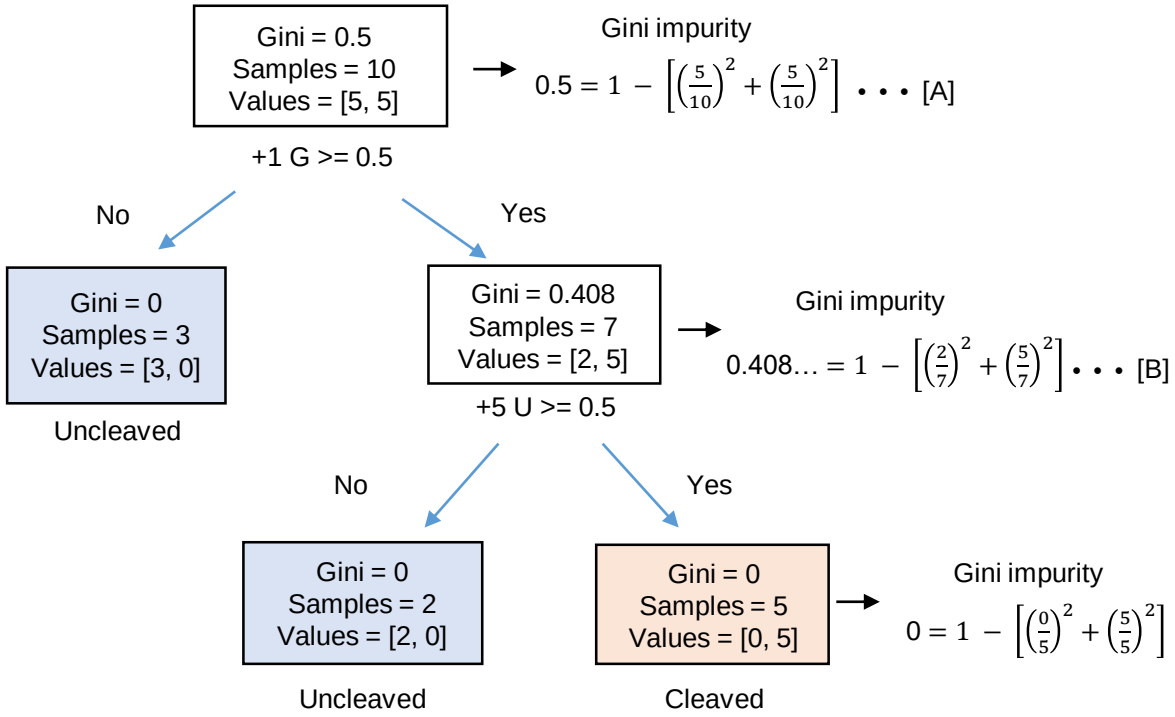


図 3-3. 決定木について

python の skit-learn ランダムダムフォレストの決定木において、データ分割に使用する特徴はジニ不純度 (Gini impurity) を基に決定される。ジニ不純度はデータを分割した際の不純度を示し、完全に分類を行えた場合は、ジニ不純度は 0 となる。Samples はサンプル数を示し、values は各ノードでの切断部位 (cleaved)、非切断部位 (uncleaved) の数を示す。決定木は、python の tree.plot_tree を参考に作成した。

表 3-1. ハイパーパラメーターの調整 (ランダムフォレスト分類)

parameters	values
n_estimators	1, 10, [100]
max_features	[0.1], 0.3, 0.6
max_depth	5, 10, [15], 20
min_samples_split	[5], 10, 15, 20
min_samples_leaf	5, 10, [15], 20

上記のハイパーパラメーターを全通り組み合わせ、OOB score が最も高いハイパーパラメーターを選択した。[] はテストデータに使用したハイパーパラメーターを示す。n_estimators; 決定木の数、 max_features; 各決定木に使用する最大の特徴数、max_depth; 各決定木の深さ、min_samples_split; 内部ノードのサンプル数の最小値、min_samples_leaf; 葉ノードに属するサンプル数の最小値をそれぞれ示す。

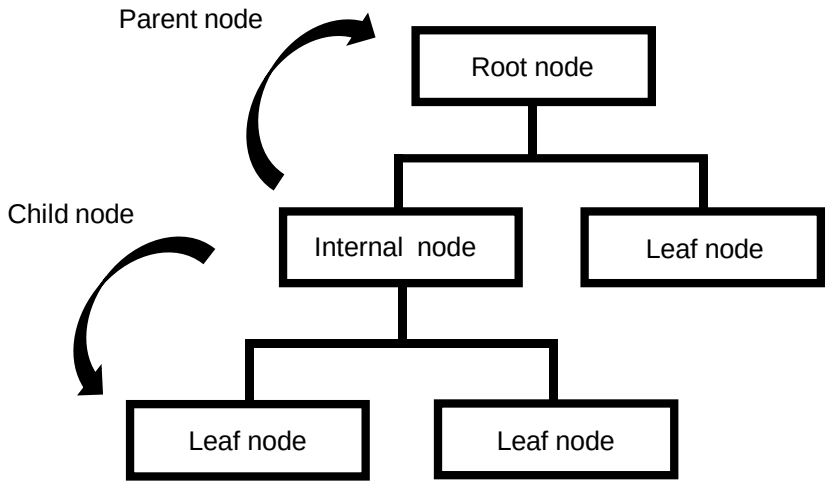


図 3-4. 決定木の名称について

決定木には、最上層の根ノード (root node)、最下層の葉ノード (leaf node)、そして、中間ノードである (internal node) が存在する。任意のノードに対して、上層を親ノード (parent node)、下層を子ノード (child node) と呼ぶ。

3-2-2. ラッソ回帰、リッジ回帰モデルの構築

3-2-2-1. 使用するモデルと解析対象とした RNA 切断部位

TREseq 法により、約 15,000 個の遺伝子と約 200 万カ所の切断部位が同定されている。これらのデータの中には、切断部位が数カ所しか存在しない遺伝子が複数存在する。より切断部位情報としての信頼性を高めるために、遺伝子長に対して 20%以上の切断部位が検出されている遺伝子を解析対象とした。また、TREseq 法で検出された遺伝子の CS_{gene} 値、および TAIR10 ゲノムデータベースに登録されている RNA 長が上位 95%以上、下位 5%以下の遺伝子を除き、解析を行った。加えて、各切断部位については、第一章で使用した psRNAtarget を用いて microRNA の切断と予測される切断部位を解析から除外した。解析対象となった遺伝子を 9:1 の割合で、トレーニングデータ (遺伝子数 = 996; 切断部位数 = 395,375) とテストデータ (遺伝子数 = 111; 切断部位数 = 43,742) に分割した。トレーニングデータを用いてモデルを構築し、テストデータを用いてモデルの精度を検証した。回帰モデルとしては、python の skit-learn ライブラリーのラッソ回帰、もしくは、リッジ回帰を用いた。

3-2-2-2. モデルに使用した説明変数

切断率に関する特徴選択では、切断部位周辺の特徴と RNA 全体の特徴に分け数理モデルに使用した (図 3-5)。切断部位周辺の特徴については、切断部位周辺の塩基、コドン、コードするアミノ酸配列、RNA 高次構造の形成度合い、リボソーム存在量に関する情報を使用した。RNA 構造の形成度合いについては、自由エネルギー (ΔG) を使用した。これらの特徴を対象に、塩基配列を用いた例に示すように切断部位周辺の前後 30 塩基の領域から網羅的に探索した (図 3-6)。RNA 高次構造の形成度合いについては、配列が短い場合に自由エネルギーを算出できないため、最低塩基長を 5 塩基とした。

また、RNA 全体の特徴については、5' UTR、CDS、3' UTR 領域、各領域の 5' 末端、3' 末端領域での、塩基、コドン、コードするアミノ酸配列、RNA 高次構造の形成度合い、リボソーム存在量に関する情報を算出した。各領域の 5’ 末端、3' 末端については、末端から 50 塩基の範囲とした。開始コドン、終止コドン周辺のコドン、コードするアミノ酸配列が翻訳の開始効率に関与することが Volkova らの研究で報告されていることから (71)、開始、終止コドンから 10 コドンと、対応するアミノ酸配列についても情報を取得した。これらの特徴を使用し、ラッソ回帰、リッジ回帰を用いた特徴選択を行った。

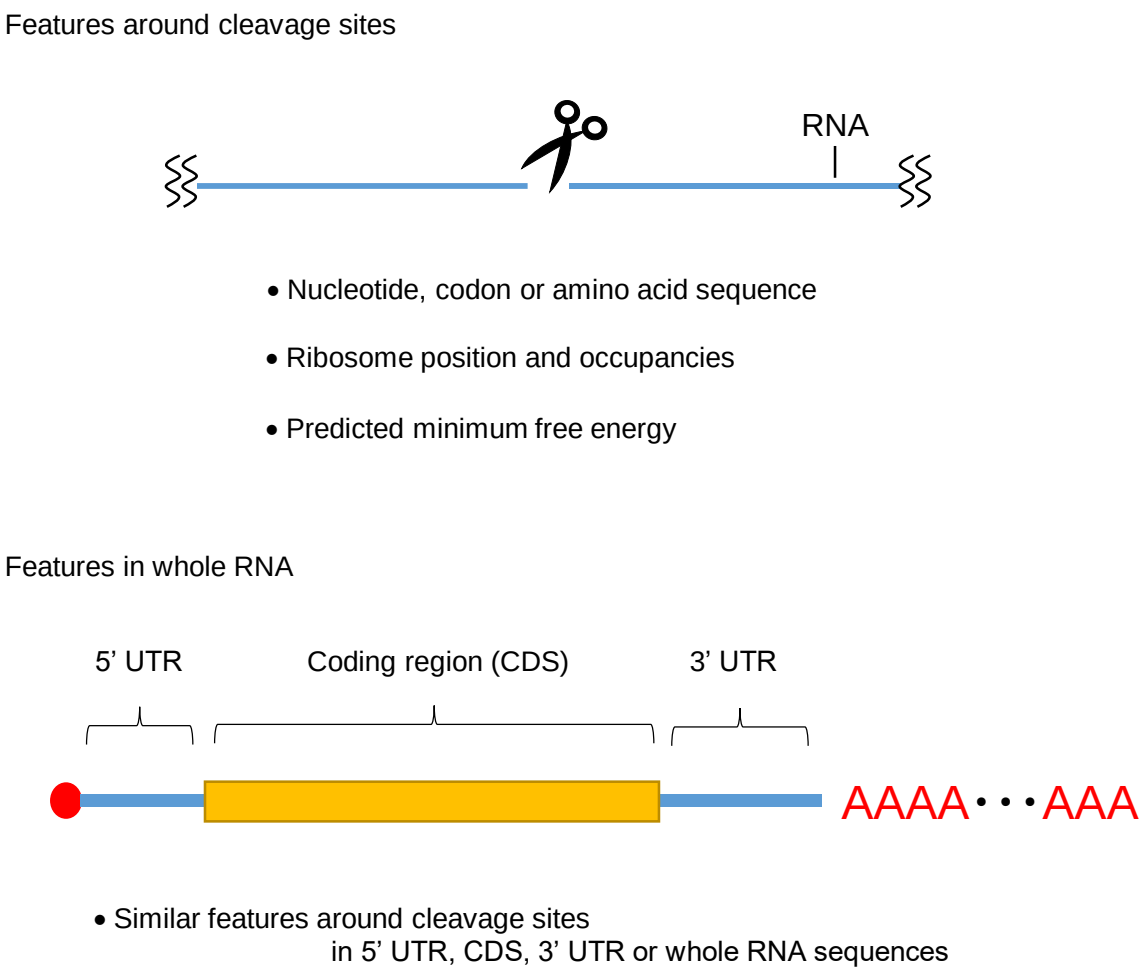
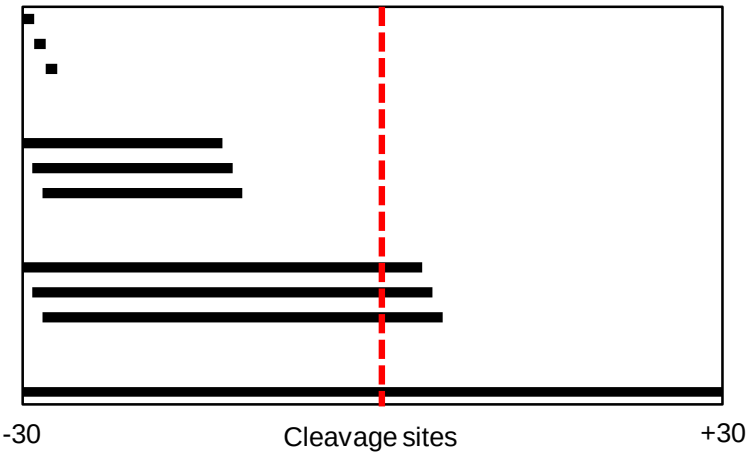


図 3-5. ラッソ回帰、リッジ回帰モデルに使用した特徴

回帰モデルには、切断部位周辺と RNA 全体の特徴を使用した。切断部位周辺の特徴については、切断部位周辺の塩基、コドン、コードするアミノ酸配、リボソーム存在量、RNA fold を用いて算出される予測自由エネルギーを使用した。RNA 全体の特徴としては、塩基、コドン、コードするアミノ酸配、リボソーム存在量、RNA の高次構造については、5' UTR、CDS、3' UTR 領域からも情報を取得した。

A



B

Calculating nucleotide frequency in each region

Region	A	U	G	C
-30	--	--	--	--
-29	--	--	--	--
-27	--	--	--	--
•				
-30 ~ -10				
-29 ~ -9				
-28 ~ -8				
•				
-30 ~ +10				
-30 ~ +11				
-30 ~ +12				
•				
-30 ~ +30				

図 3-6. 切断部位周辺の特徴抽出 (塩基配列の例)

切断部位の前後 30 塩基を対象に、可変長領域 (最小 1 塩基、最大 60 塩基) を 1 塩基ずつシフトさせ、網羅的に領域を探索する (A)。各切断部位ごとに指定した領域の塩基比率を算出し、モデルに情報を加えた (B)。

3-2-2-3. 共線性の除去

説明変数どうしの相関が高い場合、モデルの予測精度が安定しないなどの問題が生じる。そこで、スピアマンの順位相関係数を用いて、説明変数間の相関係数が高い場合 (r > 0.6)、切断率との相関係数が低い説明変数を解析から除外した (69)。

3-2-2-4. ハイパーパラメーターの設定、モデルの評価方法

一般的な線形回帰は、以下に示す公式から成り立ち、$x_{i1}, x_{i2}, \ldots, x_{ip}$ は i 番目の切断部位の特徴を示し、y_i は i 番目の切断部位の切断率を示す。

$$\hat{y}_i = \alpha + \boldsymbol{\beta} \cdot \mathbf{x}_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} \quad ,$$

$\boldsymbol{\beta} \cdot \mathbf{x}_i$ は $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)$ と $\boldsymbol{x}_{ij}$ ベクターの内積を表し、β_j は j 番目の特徴の係数を示し、αは切片、p は特徴の数を示す。

ラッソ回帰は、残差平方和に加えて、回帰係数の絶対合計値を最小化する。従って、

$$\boldsymbol{\beta}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

となる。$\lambda \sum_{j=1}^{p} |\beta_j|$ はβ_j に対する L1 正則化項を示し、$\lambda \geq 0$ となる。

リッジ回帰では、回帰係数の二乗和を最小化するため、

$$\boldsymbol{\beta}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

となる。$\lambda \sum_{j=1}^{p} \beta_j^2$ はβ_j に対する L2 正則化項を示し、$\lambda \geq 0$ となる。ラッソ回帰のλ値は、数理モデルの解釈性を増すために、トレーニングデータを対象に 10 分割交差検証を行い、予測値と実測値との平均二乗誤差平方根と係数が 0 以外の特徴数を基にλ値を 0.01 とした (図 3-7)。また、配列情報のみを用いたラッソ回帰に関しても、同様に構築した (図 3-8)。リッジ回帰については、予測値と実測値との平均二乗誤差平方根が最小であるλ値 ($\lambda = 10^5$) を使用した (図 3-9)。ハイパーパラメーターを決定後、テストデータを用いてモデルの精度を検証した (図 3-7、図 3-8、図 3-9)。

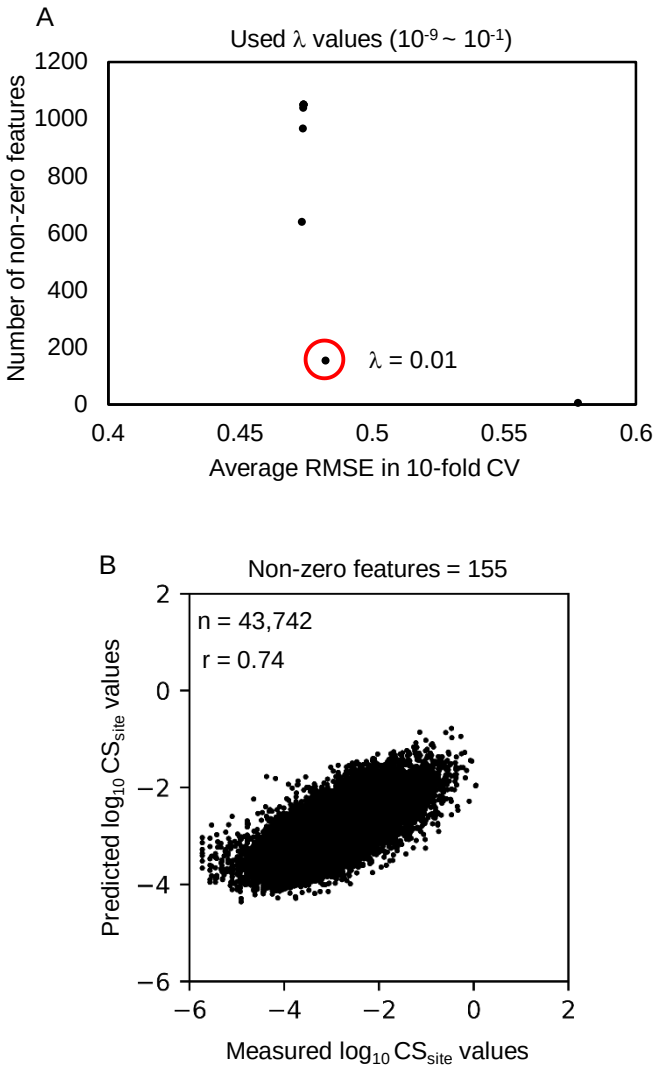


図 3-7. トレーニングデータを用いたハイパーパラメーター調整とテストデータでのモデルの性能評価 (ラッソ回帰)

トレーニングデータを用いて 10 分割交差検証 (10-fold cross validation; 10-fold CV) を行い、平均二乗誤差平方根 (root mean squared error; RMSE) と係数が 0 ではない特徴の数を基にλ値を決定した (A)。X 軸は、RMSE の平均値を示し、Y 軸は係数が 0 ではない特徴の数を示す。λ = 0.01 (赤丸) とし、テストデータを用いて CS_{site} 値の予測を行った (B)。X 軸は TREseq 法での実測値を示し、Y 軸は数理モデルを用いて予測した CS_{site} 値を示す。

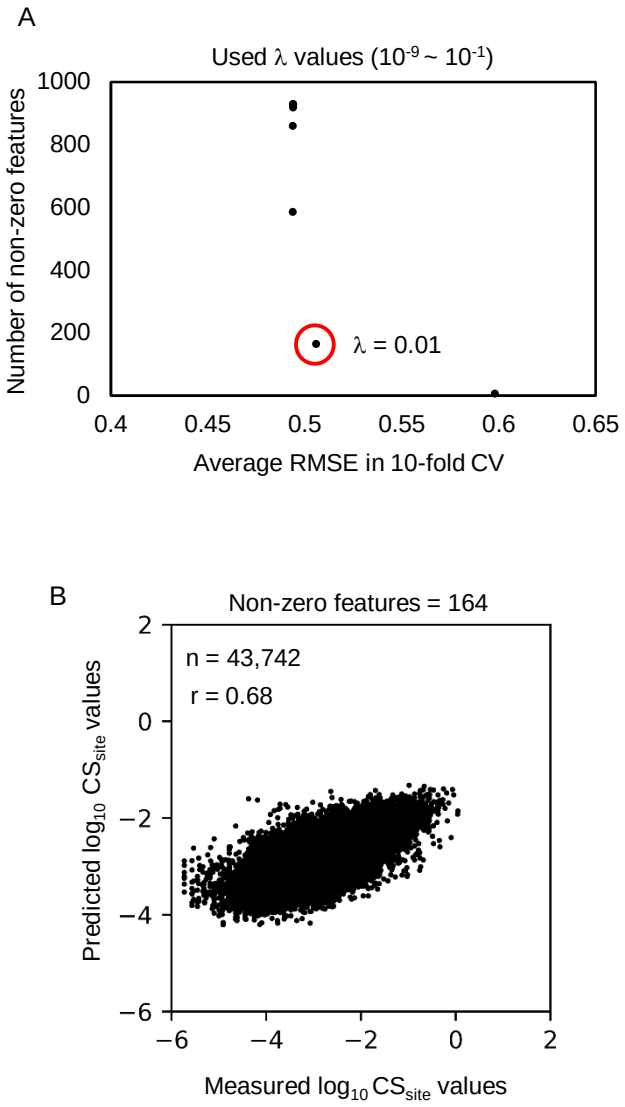


図 3-8. トレーニングデータを用いたハイパーパラメーター調整とテストデータでのモデルの性能評価 (配列情報のみのラッソ回帰)

トレーニングデータを用いて 10 分割交差検証 (10-fold cross validation; 10-fold CV) を行い、平均二乗誤差平方根 (root mean squared error; RMSE) と係数が 0 ではない特徴の数を基にλ値を決定した (A)。X 軸は、RMSE の平均値を示し、Y 軸は係数が 0 ではない特徴の数を示す。λ = 0.01 (赤丸) とし、テストデータを用いて CS_{site} 値の予測を行った (B)。X 軸は TREseq 法での実測値を示し、Y 軸は数理モデルを用いて予測した CS_{site} 値を示す。

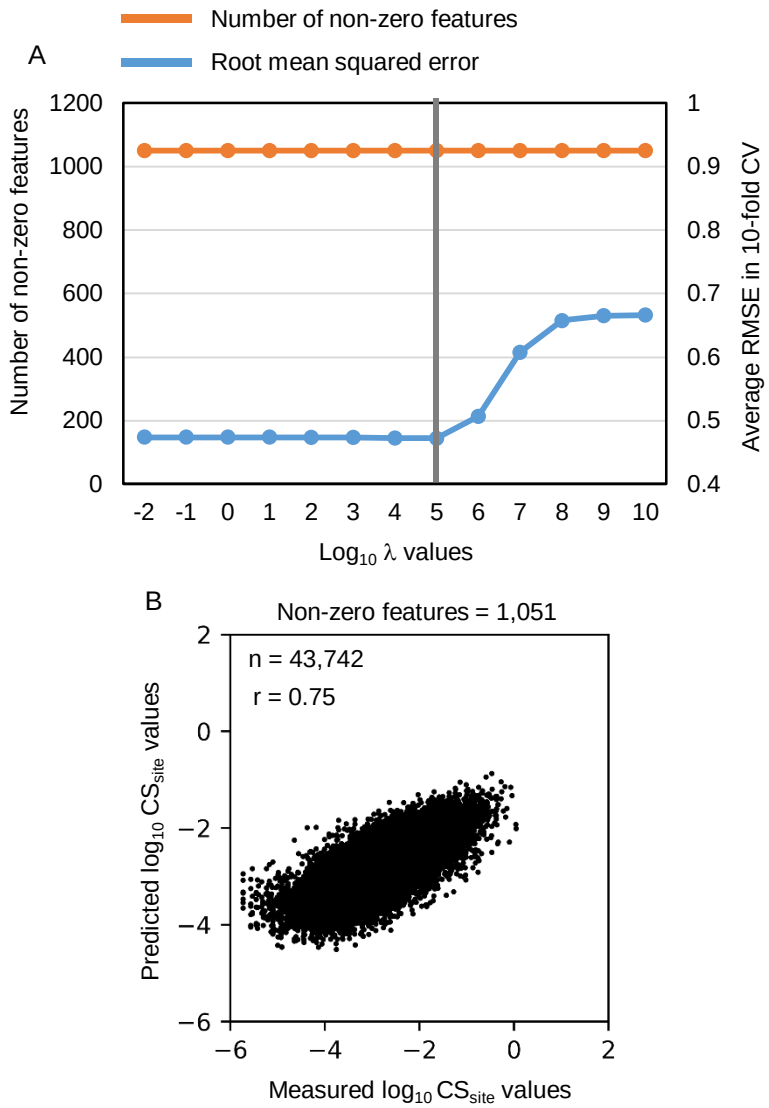


図 3-9. トレーニングデータを用いたハイパーパラメーター調整とテストデータでのモデルの性能評価 (リッジ回帰)

ラッソ回帰とは異なり、リッジ回帰では大幅な特徴数の削減はできないため、最も平均二乗誤差平方根 (root mean squared error; RMSE) が小さくなるλ値を探索した (A)。X 軸は、トレーニングデータを用いて、10 分割交差検証 (10-fold cross validation; 10-fold CV) で使用したλ値を示し、Y 軸は使用したλ値における、係数が 0 ではない特徴の数と RMSE の平均値を示す。灰色線は、テストデータに使用するλ値 ($\lambda = 10^5$) を示す。λ値を 10^5 とし、テストデータにて、CS_{site} 値の予測を行った (B)。X 軸は TREseq 法での実測値を示し、Y 軸は数理モデルを用いて予測した CS_{site} 値を示す。

3-3. 結果

3-3-1. ランダムフォレスト分類モデルを用いた RNA 切断、非切断に関わる要因の特徴選択

3-3-1-1. データプロセシング

第二章の結果より、切断部位の位置 (切断、非切断部位) の決定には、リボソームの存在位置や存在量ではなく、RNA 上の配列が重要であることが示された。そこで、第三章では RNA 上の配列が切断、非切断の決定に重要であるかの検証をランダムフォレスト分類を用いて行った (図 3-10)。全切断部位を解析対象とした場合、各遺伝子の RNA 蓄積量によって切断部位の検出率が異なることが想定される。より結果の解釈性を単純化するために、切断部位としては各遺伝子の CS_{site} 値が最も高い部位を切断部位と定義し、同じ遺伝子内の前後 30 塩基以内に CS_{site} 値が存在しない配列を非切断部位として使用した。これらの切断、非切断部位を遺伝子単位で 9 : 1 の割合でトレーニングデータ (遺伝子数 = 900; 切断部位数 = 900; 非切断部位数 = 3,289)とテストデータ (遺伝子数 = 100; 切断部位数 = 100; 非切断部位数 = 369) に分割し、ランダムフォレストを用いた分類を行った (図 3-11)。

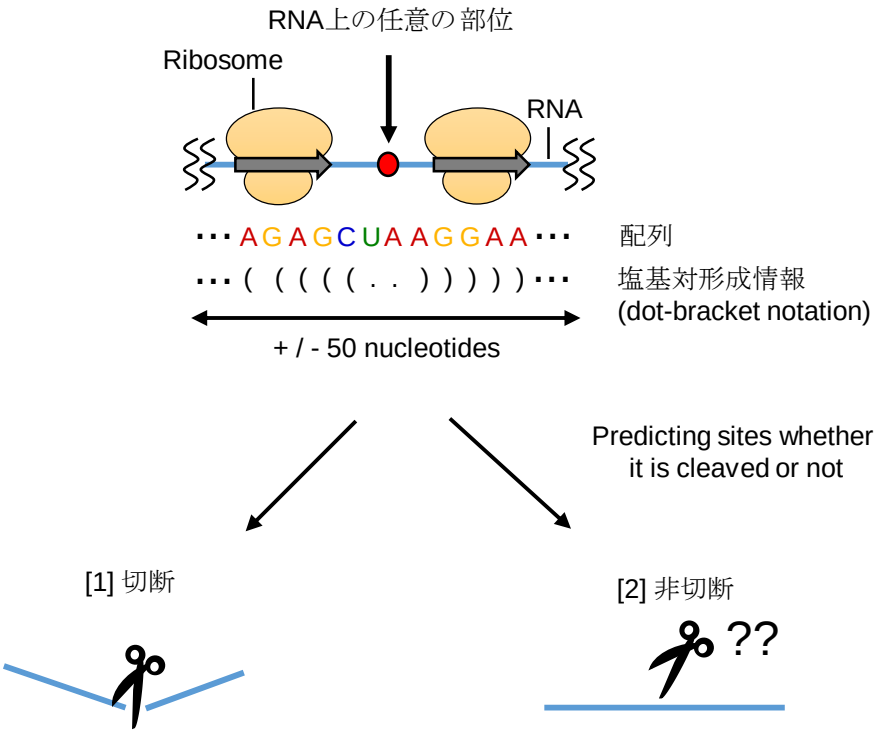


図 3-10. 切断・非切断部位の決定に関わる要因の特徴選択

RNA 上の任意の部位とその周辺の配列、リボソームの存在位置、存在量、塩基対形成情報が与えられた際に、対象となる部位が切断、非切断部位かを分類するモデルを構築した。

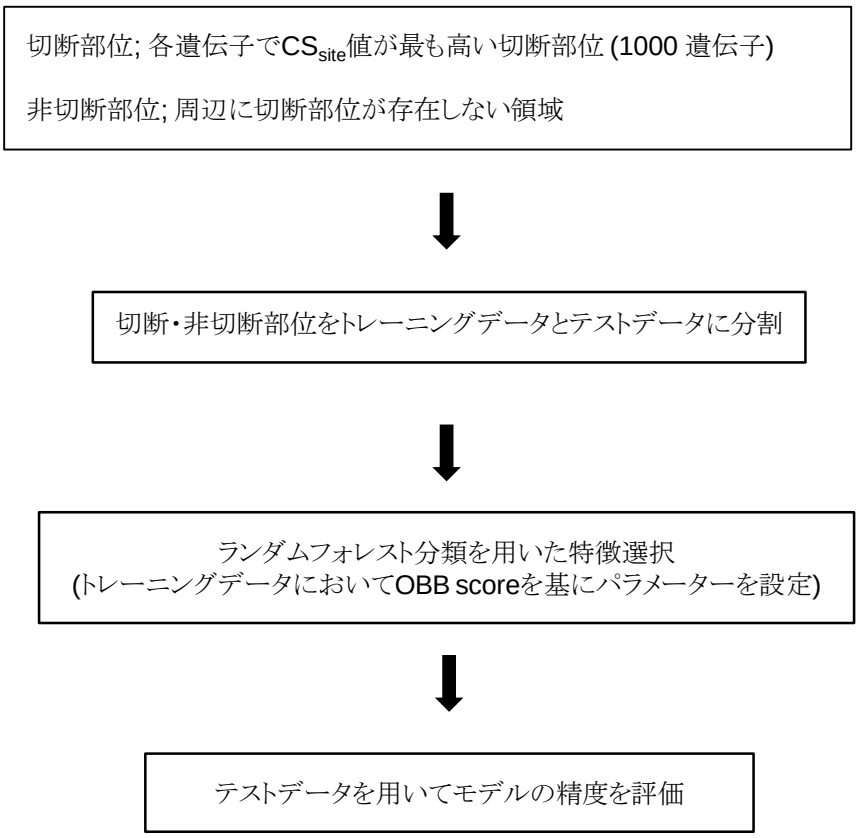


図 3-11. データプロセシング (ランダムフォレスト分類)

切断および非切断部位を選抜後、トレーニングデータとテストデータに分割し、トレーニングデータを用いて、ハイパーパラメーターの調整を行った。その後、テストデータを用いてモデルの精度を評価した。

3-3-1-2. テストデータを用いたランダムフォレスト分類モデルの精度の評価

トレーニングデータにより決定したハイパーパラメーターを用いて、テストデータでのモデルの精度を評価した。分類モデルの評価としては、Receiver Operating Characteristic (ROC) 曲線と Are Under Curve (AUC) 値を使用した。ランダムフォレスト分類を用いることで、各配列が切断部位、もしくは非切断部位である確率が算出される。ROC 曲線は、ランダムフォレスト分類から抽出した切断部位である確率を基に、複数の閾値 (カットオフ) で切断、非切断部位を定義し、真陽性率 (True positive rate; TP) と偽陽性率 (False positive rate; FP) を算出し (表 3-2)、プロットした図である。真陽性率は、切断部位と定義した部位のうち、ランダムフォレスト分類で正しく切断と判別できた部位の割合を示す。一方で、偽陽性率は、非切断部位と定義した部位のうち、ランダムフォレスト分類で誤って切断と分類された部位を示す。例えば、図 3-12 のように、1.0 以上を切断部位と定義した場合、真陽性率および偽陽性率の割合は、それぞれ 0 となる (図 3-12A)。また、閾値を 0.6 以上とした場合、真陽性率は 1、偽陽率は 0.33…となる (図 3-12B)。理想的なモデルとしては、カットオフの値を段階的に下げた場合でも、偽陽性率を低く抑えながらも高い真陽性率を維持するため、グラフは左上に凸の形になる (図 3-12C)。その一方で、分類精度が低い場合は、カットオフの値を下げると真陽性率、および偽陽性率が共に上昇するため、各点は対角線上を通る (図 3-12C)。実際に、テストデータを用いて予測した結果を基に ROC 曲線を描写したところ、左上に凸のグラフになった (図 3-13A)。AUC 値はグラフ曲線の下部面積を示し、値が高いほど数理モデルの分類精度が高いことを示している。今回構築したモデルの AUC 値を算出したところ、0.99 となったことからも十分に分類できていると考えられる (図 3-13A)。加えて、構築した数理モデルから各特徴の係数に関する情報 (ジニ重要度) を抽出したところ、配列情報が RNA 切断に大きく関与することが示された (図 3-13B)。一方で、RNA の塩基対形成情報やリボソームの位置、存在量のジニ重要度は配列情報と比べ低く、切断および非切断部位の決定に及ぼす影響は小さいという結果となった (図 3-13B)。また、各位置でのジニ重要度を抽出したところ、第二章での結果と類似するように解析対象となった部位周辺の G 塩基が重要であることが示された (図 3-14)。これらの結果は、第二章で示された結果と一致するように、RNA 切断の位置 (切断、非切断部位) の決定には切断部位周辺の配列が重要であることを示している。

表 3-2. 真陽性と偽陽性について

		Measured	
		Cleaved	Non-cleaved
Predicted	Cleaved	a	b
	Non-cleaved	c	d

True positive rate = a / a+c

False positive rate = b / b+d

TREseq 法で切断部位として検出された配列であり、かつランダムフォレスト分類でも切断部位と判別された配列を真陽性 (a)、TREseq 法で切断部位として検出されていない配列であったが、ランダムフォレスト分類で切断部位と判断された配列を偽陽性 (b)、TREseq 法で切断部位として検出された配列であり、かつランダムフォレスト分類において非切断部位と判別された配列を偽陰性 (c)、TREseq 法で切断部位として検出されていない配列であり、かつランダムフォレスト分類でも非切断部位と判別された配列を真陰性 (d) と定義した。真陽性率は a / a+c となり、偽陽性率は b / b+d となる。

A

Probabilityが1.0以上を切断と定義

Sample	Measured	Cleaved probability	Definition
Seq. 1	Cleaved	0.8	Non-cleaved
Seq. 2	Non-cleaved	0.7	Non-cleaved
Seq. 3	Cleaved	0.6	Non-cleaved
Seq. 4	Non-cleaved	0.3	Non-cleaved
Seq. 5	Non-cleaved	0.1	Non-cleaved

TP = 0, FP = 0

B

Probabilityが0.6以上を切断と定義

Sample	Measured	Cleaved probability	Definition
Seq. 1	Cleaved	0.8	Cleaved
Seq. 2	Non-cleaved	0.7	Cleaved
Seq. 3	Cleaved	0.6	Cleaved
Seq. 4	Non-cleaved	0.3	Non-cleaved
Seq. 5	Non-cleaved	0.1	Non-cleaved

TP = 1, FP = 0.333…

C

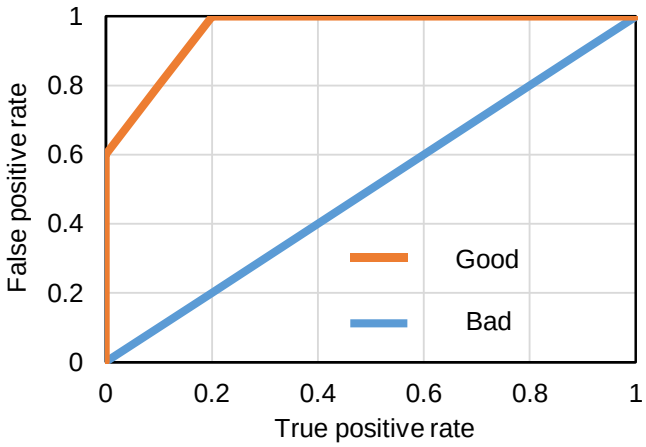


図 3-12. ROC 曲線の概要

ランダムフォレスト分類を用いることで、配列ごとに切断部位である確率が算出される。例えば、切断部位である確率が 1.0 以上のものを切断部位と定義した場合、解析対象に使用した Seq. 1 ~ Seq. 5 の配列は非切断部位として定義され、真陽性率 (TP)、偽陽性率 (FP) は 0 となる (A)。また、切断部位である確率が 0.6 以上のものを切断と定義した場合、解析対象に使用した Seq. 1 ~ Seq. 3 は切断部位と定義され、Seq. 4 ~ Seq. 5 は非切断部位と定義される。この時の TP は 1、FP は 0.333…となる (B)。このように、切断部位である閾値の定義を変更し、各点を描写した図が ROC 曲線と呼ばれている。構築したモデルの精度が高い場合は、左上に凸の形となり、予測精度が低い場合、対角線上を通る (C)。また、分類予測が実測値と反対の判別をすると右下に凸の形となる。

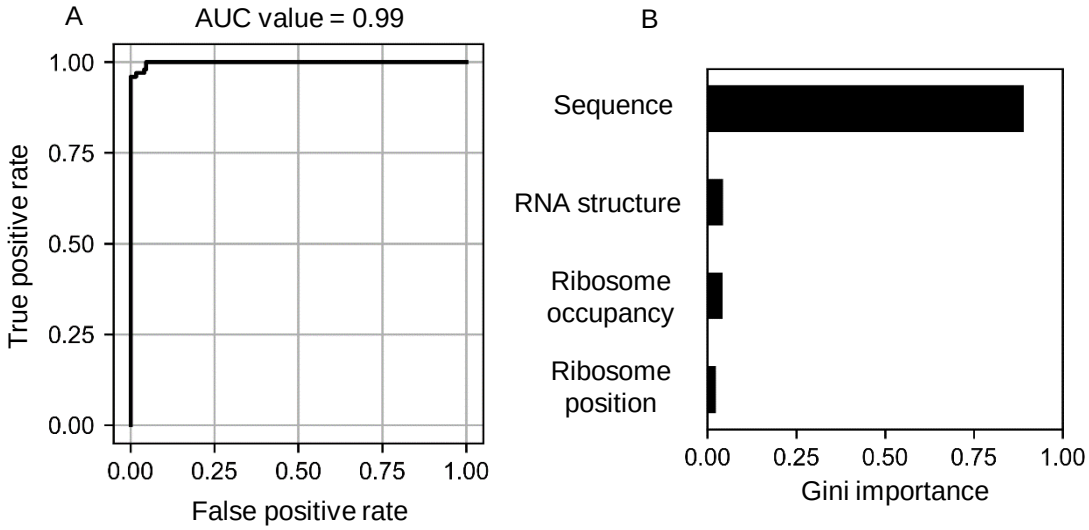


図 3-13. ランダムフォレスト分類の精度、及び各特徴の評価

モデルの評価に関しては、ROC 曲線を用い、その AUC 値を算出した (A)。また、Gini importance より、各特徴の重要度を評価した (B)。

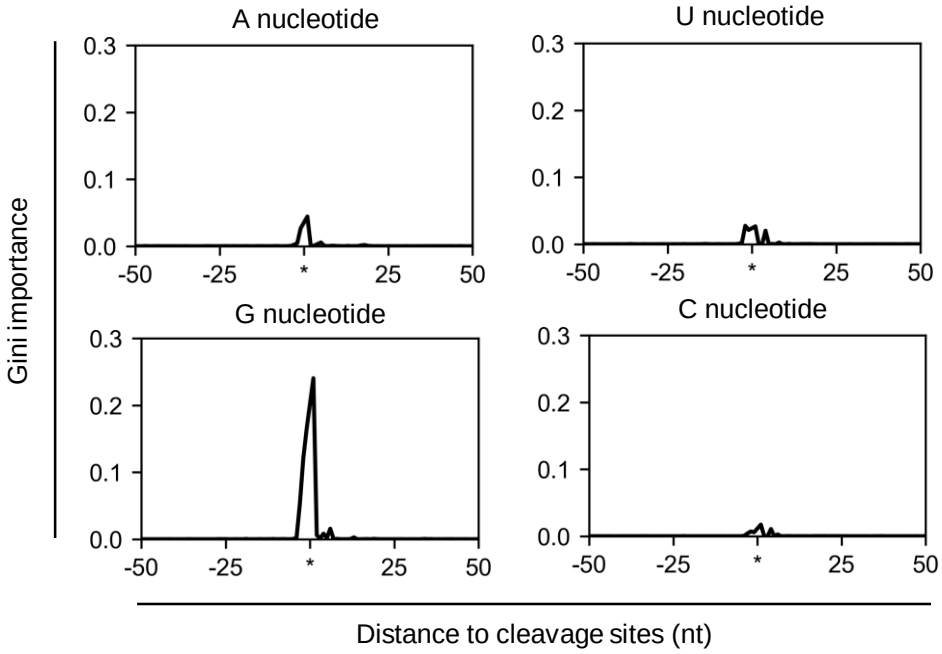


図 3-14. 配列情報のジニ重要度 (Gini importance)

各塩基ごとに Gini importance を算出した。X 軸は切断部位からの距離を示し、アスタリスクは切断部位を示す。Y 軸は各位置における Gini importance を示す。

3-3-2. ラッソ回帰モデルを用いた RNA 切断に関わる要因の特徴選択

3-3-2-1. データプロセシング

切断および非切断部位の決定とは異なり、切断率には翻訳過程や切断部位周辺の配列など複数の要因が関与することが想定されている。そこで第三章では、切断率に着目し、スパースモデリングであるラッソ回帰を用いた特徴選択を行った。

切断率を説明できるモデルでは、CS_{site} 値の全体的な傾向を把握することを目的としているため、RNA 長、CS_{gene} 値が極端に高い、低い遺伝子を解析から除外した。また、各遺伝子で検出される切断部位数が少ない場合、検出限界に達していた可能性があるため、RNA 長に対して 20%以上の領域で切断部位が検出されている遺伝子を解析対象とした。各切断部位については、第一章で使用した psRNAtarget を用いて microRNA の切断と予測される切断部位を解析から除外した。切断部位情報を遺伝子単位で 9：1 の割合でトレーニングデータ (遺伝子数 = 996; 切断部位数 = 395,375) とテストデータ (遺伝子数 = 111; 切断部位数 = 43,742) に分割し、CS_{site} 値に関するモデルを構築した。

ラッソ回帰に使用する特徴については、第二章の結果を基に切断部位周辺の塩基配列、リボソーム存在量、RNA の高次構造の形成度合いに関する情報を用いた。また、Presnyak らの研究で、RNA の安定性には RNA 全体の配列が関与することが報告されていることから (43)、RNA 全体の塩基比率や CDS 内のコドン、コードするアミノ酸配列に関する情報、RNA 上のリボソーム存在量情報も加えた。翻訳状態については、開始、終止コドン周辺の配列が翻訳状態に影響を及ぼすことが報告されているため、各遺伝子の開始コドン、終止コドン周辺の塩基、コドン、コードするアミノ酸配列に関する情報も加えた。

切断部位周辺の特徴を抽出する際、図 3-6 に示すように、網羅的に特徴を探索しているため、説明変数間の相関が高くなる (多重共線性)。多重共線性はモデルの予測精度が安定しないなどの問題を生じさせるため、説明変数間の相関係数が高い場合、切断率との相関が弱い特徴を解析から除外した。特徴の選択圧を調整するハイパーパラメーターであるλ値 (ハイパーパラメーター) については、10 分割交差検証を用いて、実測値と予測値の平均二乗誤差平方根と係数が 0 以外の特徴を基にλ値を決定した (図 3-15)。

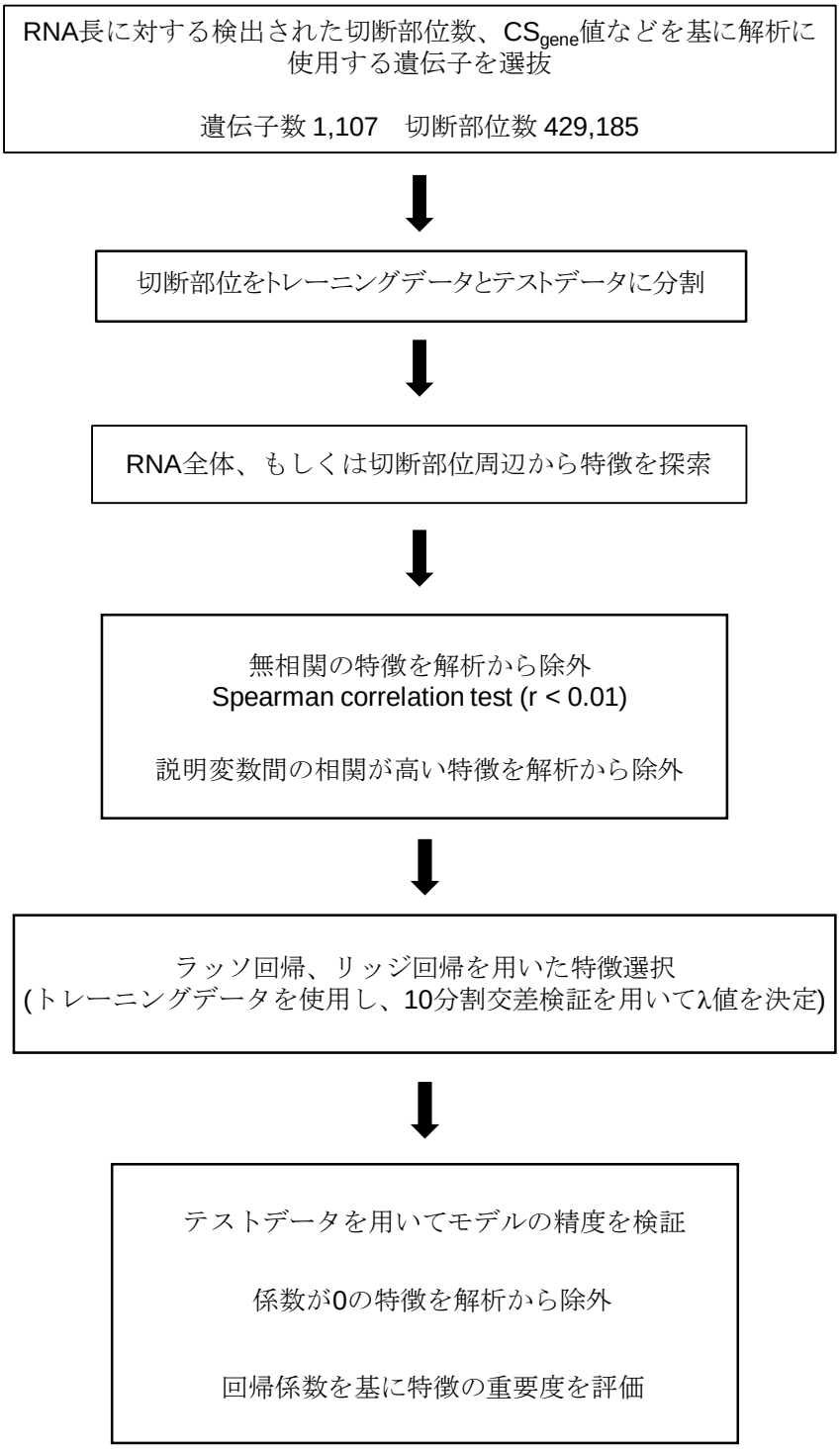


図 3-15. ラッソ回帰およびリッジ回帰におけるデータプロセシング

解析対象とするデータをトレーニングデータとテストデータへと分割後、切断に関わる特徴を探索し、共線性を持つ特徴を除去した。その後、トレーニングデータを用いてハイパーパラメーターを調整し、テストデータを用いてモデルの精度を検証した。

3-3-2-2. テストデータを用いたラッソ回帰モデルの精度の検証

構築したモデルの精度をテストデータを用いて検証した結果、ピアソンの積率相関係数は r = 0.73 となり、全体的な CS_{site} 値の傾向を予測できていることが示された (図 3-7)。これらの特徴の中で、係数が 0 の特徴を除くと 155 個の特徴が残った。ラッソ回帰より選ばれた特徴には、切断率に正 (切断されやすくなる)、負 (切断されにくくなる) の係数を持つ特徴が存在する。そこで、各特徴の係数を正、負の係数に分けた後、各特徴をグループ分けし、どのような特徴が重要であるかの評価を行った。正の係数を持つ特徴に着目し、切断部位周辺の特徴を調べてみると、塩基配列に関わる特徴が最も大きな割合を占めていた (図 3-16A)。一方で、RNA 全体の特徴に着目すると、塩基配列に加え、コドン、コードするアミノ酸配列、リボソーム存在量など翻訳過程に関わる特徴が一定の割合を占めていた (図 3-16B)。また、負の特徴に着目し解析を行ったところ、切断部位周辺の特徴については、正の係数と同様に塩基配列に関わる特徴が大きな割合を占めており (図 3-17A)、RNA 全体の特徴についても、塩基配列に関わる特徴が大きな割合を占めていたが、切断部位周辺の特徴と比べると全体的に係数は小さいものであった (図 3-17B)。

次に、各特徴の係数を基に、正の係数値が高い順から 5 つ (positive 5)、負の係数値が高い順から 5 つずつ (negative 5) 特徴を選抜した (表 3-3, 表 3-4)。Positive 5 の特徴をみると、RNA 全体に存在するリボソーム存在量が最も切断率に正に関与する傾向が認められた。また、切断部位の-4～+3 位周辺の G 塩基比率など、第二章の図 2-21 に示されたように、切断部位周辺での G 塩基比率が高い領域が抽出された (図 2-21、表 3-3)。Negative 5 の特徴に着目すると、+4~+5 位、+8～+12 位での G 塩基比率など G 塩基比率が低い領域が抽出されていた。加えて、他の塩基に着目すると、-2 位の U 塩基や+4 位の C 塩基などについても同様の傾向が認められた (図 2-21, 表 3-4)。

これらの結果をまとめると、各切断部位の切断率には切断部位周辺の塩基配列が切断率に大きく関与することに加え、RNA 全体のリボソーム存在量についても切断率の調節に関わることが示され、第二章での結果と一致していた。

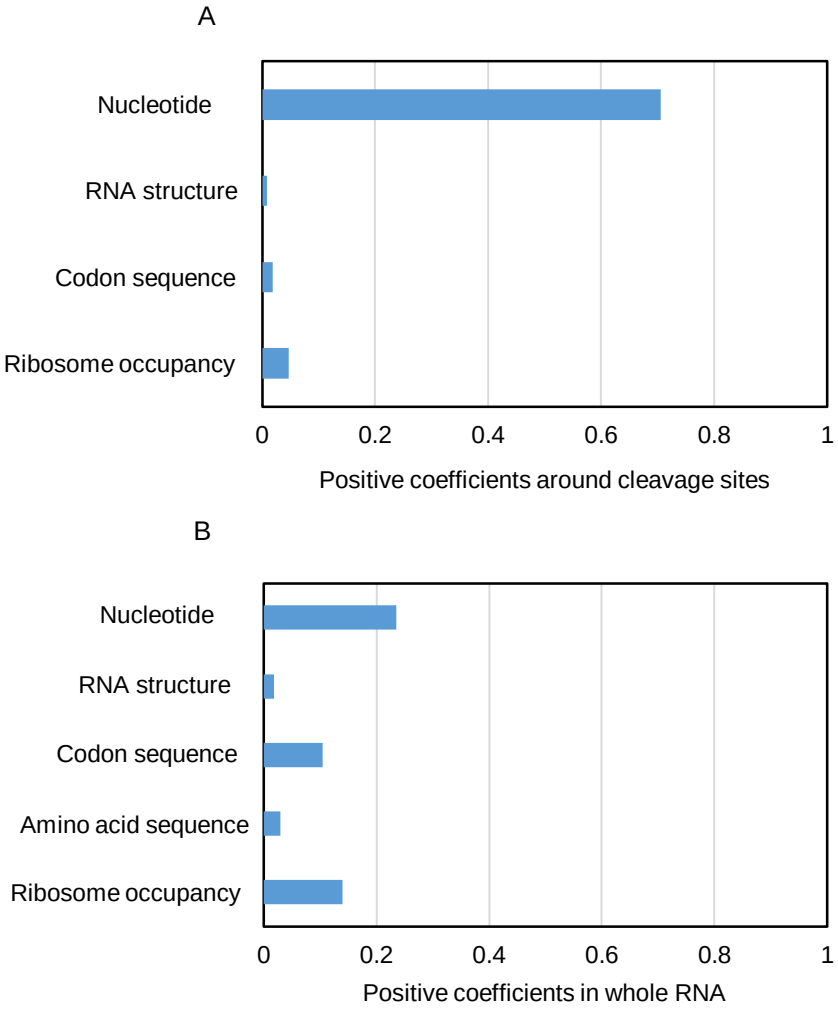


図 3-16. ラッソ回帰から抽出した各特徴の係数 (正の係数)

ラッソ回帰より各特徴の係数に関する情報を取得し、切断部位周辺の特徴 (A)、もしくは、RNA 全体の特徴 (B) にカテゴリー分けを行った。

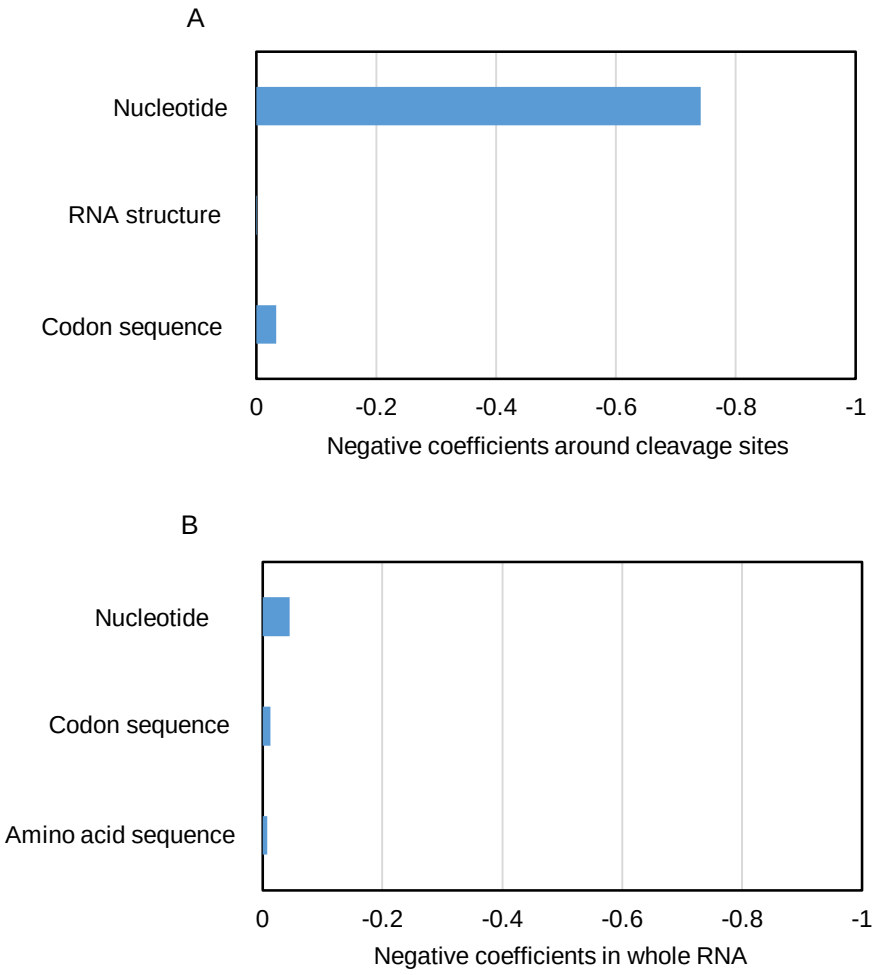


図 3-17. ラッソ回帰から抽出した各特徴の係数 (負の係数)

ラッソ回帰より各特徴の係数に関する情報を取得し、切断部位周辺の特徴 (A)、もしくは、RNA 全体の特徴 (B) にカテゴリー分けを行った。

表 3-3. ラッソ回帰における正の係数を持つ特徴 (positive 5)

Features (positive)	Coefficient
Ribosome occupancy in RNA	0.135
GG frequency around cleavage sites -4 to +2	0.108
G frequency around cleavage sites at +1	0.090
G frequency around cleavage sites -2 to +1	0.067
G frequency around cleavage sites -1 to +3	0.045

表 3-4. ラッソ回帰における負の係数を持つ特徴 (negative 5)

Features (negative)	Coefficient
G frequency around cleavage sites +4 to +5	-0.061
G frequency around cleavage sites +8 to +14	-0.052
U frequency around cleavage sites at -2	-0.051
C frequency around cleavage sites at +4	-0.050
A frequency around cleavage sites +17 to +19	-0.042

3-3-2-3. 別の回帰モデルを用いた再現性の確認

ラッソ回帰により選ばれた特徴が別の回帰モデルを用いても、同様の傾向が認められるか検証を行った。比較用の回帰モデルとしては、リッジ回帰モデルを使用した。ラッソ回帰と同様の手順でモデルを作成し (特徴数 1,051 個)、テストデータを用いてモデルの検証を行ったところ、ラッソ回帰での CS_{site} 値の予測結果と同程度の相関係数が得られた (図 3-9)。双方のモデルで共通の 155 個の特徴を用いて、ラッソ回帰、リッジ回帰で選ばれた特徴の回帰係数のピアソンの積率相関係数を求めたところ、r = 0.84 という正の相関関係が認められた (図 3-18)。また、各特徴の正の係数値が高い順から 5 つ (positive 5)、負の係数値が高い順から 5 つずつ (negative 5) 特徴を選抜した際も、切断部位周辺の配列 (切断部位周辺の G 塩基比率など) に加え、翻訳過程 (RNA 上のリボソーム量) に関わる特徴が抽出されるなど、別の回帰モデルを使用した場合でもラッソ回帰と同様の傾向が認められた (表 3-5、表 3-6)。

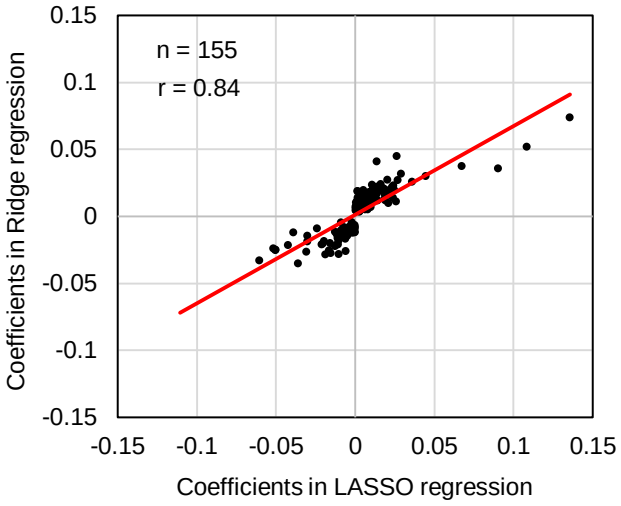


図 3-18. ラッソ回帰より取得した特徴の再現性の確認

ラッソ回帰、リッジ回帰で共通して得られた特徴を抽出し、共通した特徴のピアソンの積率相関係数を求めた。

表 3-5. リッジ回帰における正の係数を持つ特徴 (positive 5)

Features (positive)	Coefficient
Ribosome occupancy in RNA	0.074
GG frequency around cleavage sites -4 to +2	0.052
GG frequency around cleavage sites -1 to +1	0.045
GG frequency around cleavage sites -3 to -1	0.041
G frequency around cleavage sites -2 to +1	0.037

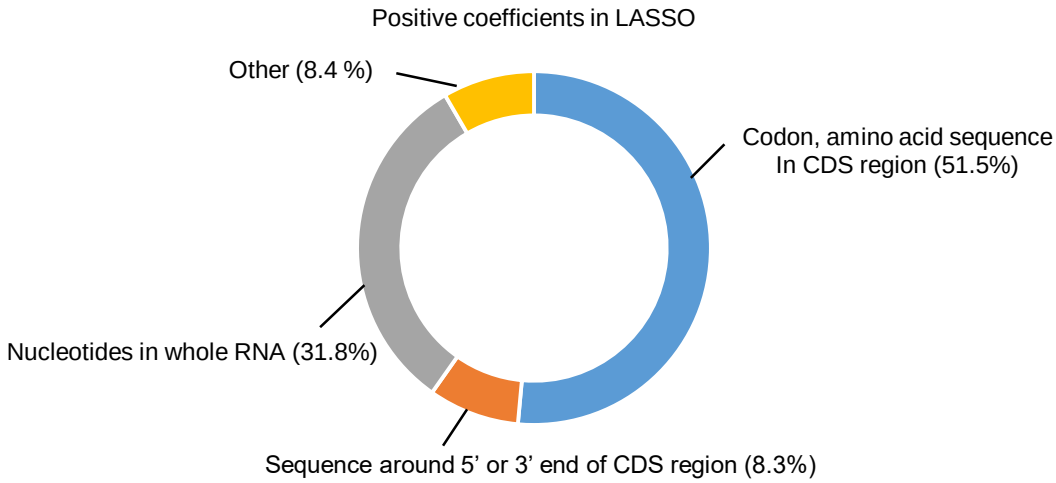
表 3-6. リッジ回帰における負の係数を持つ特徴 (negative 5)

Features (negative)	Coefficient
AA frequency around cleavage sites -1 to +2	-0.035
G frequency around cleavage sites +4 to +5	-0.033
GU frequency around cleavage sites +4 to +6	-0.028
A frequency around cleavage sites -1 to +1	-0.028
UG frequency around cleavage sites +1 to +2	-0.027

3-3-3. 配列情報のみを用いた切断率の予測

第三章の 3-3-2-2 の結果より、切断率には塩基配列と翻訳過程が大きく関与する結果となった。Hu らの研究で、RNA 上のリボソーム量は配列情報から説明できることから (69)、配列情報のみを用いた場合でも切断率を予測できると考えた。実際に配列情報のみ (塩基、コドン、コードするアミノ酸配列) を用いた場合でも、図 3-8 に示されるように、切断率を高い精度で予測できることが示された (相関係数 r = 0.68)。また、配列情報のみを用いて構築したモデルで抽出された特徴をみると、予想されたようにコドン、アミノ酸配列、RNA 領域、CDS 領域の末端など、RNA 上のリボソーム量 (翻訳状態) に関与する配列特徴が新たに抽出された (図 3-19) (69)。

加えて、今回構築した数理モデルについて、外来遺伝子を対象とした検証を行った。配列情報のみを用いて構築したラッソ回帰モデルは、内在遺伝子の配列情報と切断率を用いている。このモデルが、実際に植物細胞内で生じている RNA 切断の切断率を説明できるならば、シロイヌナズナ植物体に導入したレポーター (外来) 遺伝子の切断率についても同様に予測できると考えられる。当研究室では、*firefly luciferase* (*F-luc*) 遺伝子を導入したシロイヌナズナ植物体 (芽生え 2 日目) を対象に TREseq 法を行っている。*F-luc* RNA についての切断部位、および切断率に関する情報取得し、構築したラッソ回帰モデルで予測した CS_{site} 値とのピアソンの積率相関係数を求めたところ相関係数は r = 0.71 となった (図 3-20)。これらの結果は、配列情報のみで RNA の切断率を説明できることを示すとともに、実際に植物細胞内で生じている RNA 切断に関わる特徴を数理モデルで抽出しているものと考えられる。



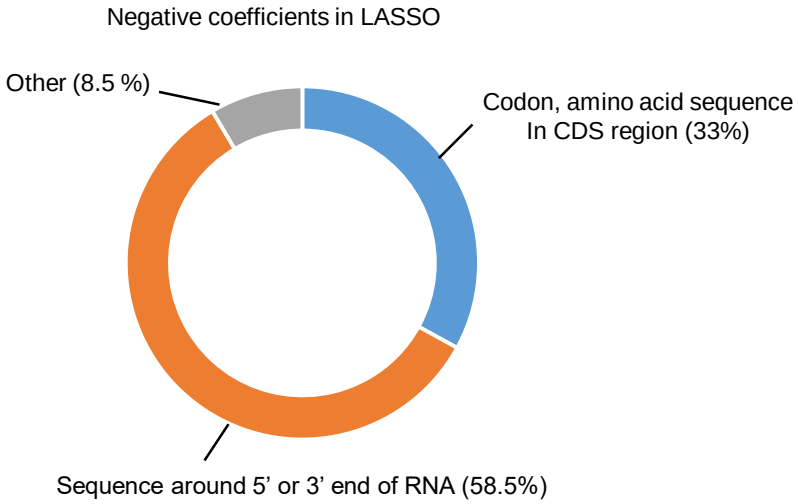


図 3-19. 配列情報のみを用いて構築したラッソ回帰モデルで新たに抽出された特徴

図 3-7 で構築したモデルと比較し、配列情報のみを用いて構築したラッソ回帰モデルで新たに抽出した特徴を正、負の係数ごとにまとめた。

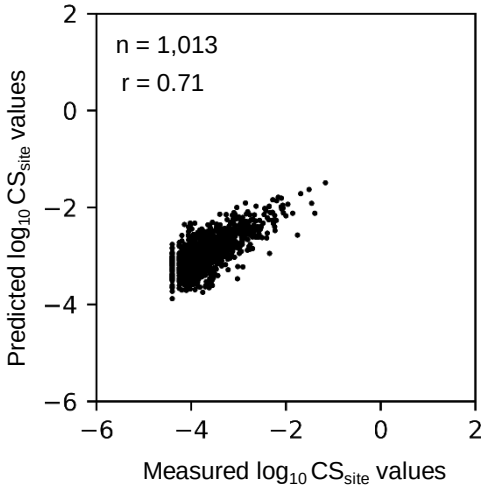


図 3-20. 外来遺伝子の予測 (*F-luc* RNA)

配列情報のみを使い構築した数理モデルを用いて、*F-luc* RNA 内の CS_{site} 値を予測した。X 軸は TREseq 法での実測値を示し、Y 軸は数理モデルを用いて予測した CS_{site} 値を示す。

3-4. 考察

3-4-1. 切断、非切断部位の決定に配列が及ぼす影響

第二章の結果より、RNA 切断部位の位置 (切断・非切断部位) の決定には、配列が重要であることが示された。そこで、第三章では切断、非切断部位の決定に関わる要因について、ランダムフォレスト分類モデルを用いた特徴選択を行ったところ、切断部位の位置決定には、RNA 高次構造やリボソーム存在量といった特徴と比較し、切断部位周辺の配列特徴が重要であることが示された。Ibrahim らの研究で、RNA 切断にはリボソームの位置が重要であることが報告されていたように (22)、リボソームの存在位置も切断部位の決定に関与する傾向はわずかながら認められたが、その影響は非常に弱いものであった (図 3-13)。これらの結果は、RNA の切断、非切断部位の決定に関わる複数要因の寄与度を分類モデルから明らかにし、特に切断部位周辺の配列特徴が大きく関与していることを新たに明らかにした。

3-4-2. 切断部位周辺の特徴が切断率に及ぼす影響

第一章、第二章の結果を基に、ラッソ回帰による特徴選択を行ったところ、約 160 種の特徴が選抜された。切断部位周辺の特徴に着目し、各特徴の係数を見ると、塩基配列に関わる特徴が大きな係数である一方で、コドン、およびコードするアミノ酸配列に関する特徴は大きな係数ではなかった (図 3-16A)。同様の結果は第二章でも示されており、RNA 切断には切断部位周辺のコドンやコードするアミノ酸配列ではなく、塩基配列自身が重要であることが示唆されていた (図 2-21)。これらの結果は、ラッソ回帰を用いた特徴選択からも、切断部位周辺のコドンやコードするアミノ酸配列が切断率に与える影響は小さく、3-3-2 で示されたように切断率には切断部位周辺の配列特徴が大きく関与することが明らかとなった。これらの結果は、第三章の 3-3-1 で示すように、切断部位周辺の塩基配列が切断部位の位置決定に重要であることに加え、各切断部位での切断のされやすさにも大きく関与することを示している。

3-4-3. RNA 全体の特徴が各切断部位の切断率に及ぼす影響

ラッソ回帰により抽出された各特徴の回帰係数を見ると、RNA 全体の特徴として RNA 上のリボソーム量や塩基配列などの特徴が一定の割合を占めていた (図 3-16B)。RNA 全体の特徴が RNA 分解に寄与することは、ポリ A 鎖の短縮依存的な分解機構でも報告されている。例えば、脱キャップに重要な Dhh1p や CAF のターゲット RNA の認識には、RNA の末端配列だけではなく、RNA 内のコドン配列、翻訳状態が重要であることが酵母、動物において報告されている

(43, 64, 65)。また、核内での pre-RNA の切断でも、pre-RNA の切断される領域に加え、上流、下流の 100 塩基ほどの領域が重要であるなど、幅広い領域の配列的特徴が切断という現象にとって重要であることが知られている (72)。ポリ A 鎖の短縮依存的な RNA の分解機構、pre-RNA の切断機構で見られるように、植物 RNA の切断機構でも、切断部位周辺だけではなく、RNA 全体の特徴が重要であると考えられる。特に今回の解析では、RNA の翻訳状態を反映するリボソーム量に関する特徴がラッソ回帰での positive 5 の特徴に含まれていたことからも、RNA 全体としての翻訳状態が各切断部位の切断率に関与することが示されている (表 3-3)。これまで RNA 切断機構については、切断部位周辺の特徴 (配列など) が主に解析されてきたが、今回得られた結果から RNA 全体の特徴、特に RNA 上のリボソーム存在量が多いほど、各部位での切断が生じやすいことを新たに明らかとした。

3-4-4. 構築したモデルより得られた知見の実証

これまでの解析から、切断部位周辺の配列的特徴が切断、非切断部位の決定に重要であること、そして、切断率に関与することが示されている。加えて、RNA 上のリボソーム存在量が各切断部位の切断率に正の影響を及ぼすが、切断、非切断部位の決定には大きく関与しないことが示されている。このような、切断部位周辺の配列、RNA 上のリボソーム量が RNA 切断に与える影響については、当研究室において renilla luciferase (*R-luc*) を用いて行った DNA 一過性発現実験で検証されている。この実験では、[1] 通常の *R-luc* 配列に加え、[2] *R-luc* RNA の 5' UTR 配列を置換し、翻訳効率 (RNA 上でのリボソーム存在量) を向上させたコンストラクト、および [3] *R-luc* RNA 内の切断率が最も高い部位の G 塩基をアミノ酸置換が生じないように A 塩基に置換した 3 種類の発現カセットを使用している (Kaneko, unpublished)。これらの発現カセットをシロイヌナズナ培養細胞のプロトプラストに導入し、その後精製した RNA を用いて TREseq 法を行っている。まず、*R-luc* RNA については、リボソームプロファイリング情報が存在しないため、配列情報のみを用いて構築したラッソ回帰で最も切断部位の CS_{site} 値が高い部位を予測したところ (図 3-21)、TREseq 法で検出した実測 CS_{site} 値が最も高い切断部位と同じであった。また、この切断部位の +1 位の G 塩基をアミノ酸置換が生じないように A 塩基に置換した場合の CS_{site} 値についてもラッソ回帰を用いて予測したところ、改変後の切断率は改変前と比べ大幅に減少することが予想された (図 3-21)。実際に、シロイヌナズナ培養細胞のプロトプラストに *R-luc* を導入し、精製した RNA から TREseq 法を用いて実測 CS_{site} 値を算出したところ、改変前、改変後の塩基置換した部位での切

断率は、ラッソ回帰を用いて切断率を予想したように 50 分の 1 程度まで大幅に減少した (Kaneko, unpublished)。また、5' UTR 配列を置換し翻訳効率を上昇させた場合、切断部位の位置に大きな変動は認められなかったが、各切断部位での切断率は上昇していた (Kaneko, unpublished)。これらの結果は、本研究で提唱した切断部位決定の配列依存性と翻訳過程が切断率に正に関与することを実証したものである。内在遺伝子の配列情報のみを用いて構築したモデルから、植物体における *F-luc* RNA 内の切断率を予測可能なことからも、今回構築したモデルから抽出された特徴は信頼性がある情報であると考えられる。

これらの結果は、植物において RNA 切断に関与する複数要因の寄与度を明らかにし、RNA 切断における配列依存性、翻訳過程の重要性を明らかにしたものである。

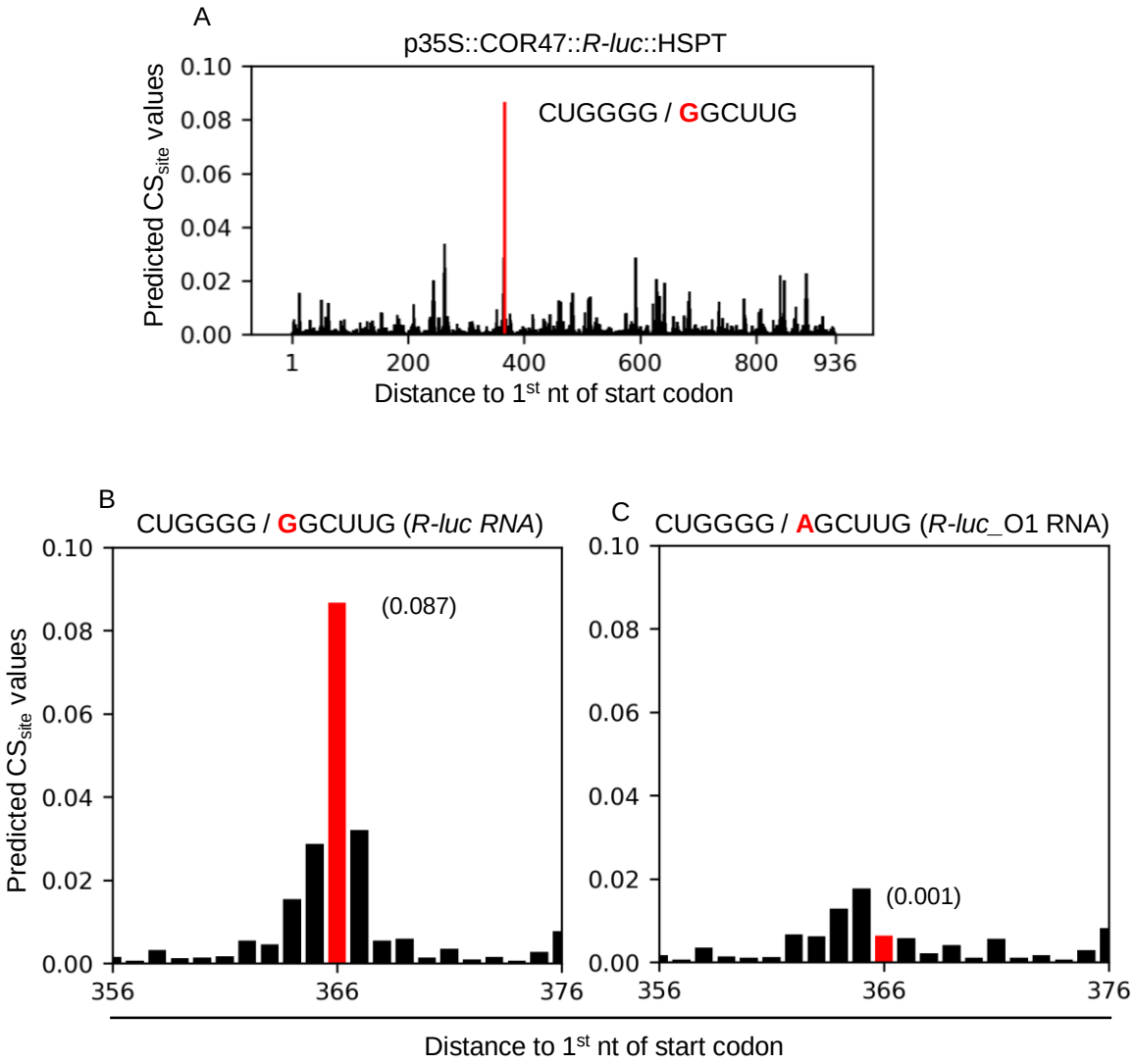


図 3-21. 外来遺伝子の切断部位および切断率の予測

p35S::COR47::*R-luc*::HSPT の配列情報を取得し、ラッソ回帰を用いて各部位での CS_{site} 値を予測した (A)。予測 CS_{site} 値が最も高い部位の G 塩基を A 塩基に置換した配列についてもラッソ回帰を用いて CS_{site} 値を予測した (B, C)。赤棒は改変前の *R-luc* (p35S::COR47::*R-luc*::HSPT) にて、TREseq 法での実測 CS_{site} 値が最も高い部位を示す。括弧内は塩基置換した部位 (赤棒) での予測 CS_{site} 値を示す。

総括

本研究は、遺伝子発現において重要な調節機構として知られている RNA 分解機構の中でも、特に切断に依存する分解機構について、切断に関わる配列特徴など複数の視点に着目し解析を行った。本研究の第一章にて、植物 RNA 切断機構の全体像を把握するするために、従来行われていた網羅的な切断部位解析法を改良した Truncated RNA end sequencing (TREseq) 法をシロイヌナズナにおいて確立した。この手法を用いることで、検出される切断部位の偏りを大幅に軽減し、より正確な切断部位の同定、切断率の算出を可能にした (図 1-7, 図 1-8)。この TREseq 法を用いて取得したシロイヌナズナでの切断部位、切断率に関する情報を使用することで、切断されやすい配列に着目するなど、RNA 切断に関わる特徴について、より詳細な解析が可能となった。また、シロイヌナズナにおいて切断されやすい RNA ほど半減期が短いことからも (図 1-12, 図 1-13)、RNA 切断に依存する分解機構が RNA 安定性を調節する重要な機構の一つであることが明らかとなった。

続く第二章では、第一章で取得した網羅的な切断部位、切断率情報を使用し、配列、翻訳状態など様々な視点から解析を行った。切断率を基に各切断部位の配列に着目し解析を行ったところ、切断部位の周辺で G 塩基比率が高い傾向が認められた (図 2-1)。加えて、出芽酵母やショウジョウバエを用いて TREseq 法を行ったところ、種間ごとにわずかに異なるがシロイヌナズナと同様の配列パターンが認められ、RNA 切断に関わる配列は異なる種間で保存されていることが明らかとなった (図 2-21)。シロイヌナズナ、ショウジョウバエ、出芽酵母で共通して濃縮されていた GO term に着目した際も、切断されやすい遺伝子群では、環境応答など迅速な制御に関わる GO term が濃縮されていた一方で (表 2-5)、切断されにくい遺伝子群では翻訳過程など恒常的な機能に関わる GO term が濃縮されているなど (表 2-6)、RNA 切断機構は細胞が生命を維持する上で重要な生物学的プロセスに関与し、多くの生物種で細胞の機能調節に関与している可能性が考えられた。また、RNA 切断機構として microRNA が関与する RNA 切断が報告されているが、検出された全切断部位に対して microRNA のターゲット配列と重複する切断部位はごくわずかであったことから (表 2-1, 表 2-7)、網羅的な切断部位解析によって検出された多くの切断部位はこれらとは異なる機構によるものと考えられた。この RNA 切断機構については、これまで翻訳過程が切断に関与することが示唆されていたが、実際に リボソーム存在位置や存在量と切断率との関係性に着目した解析はこれまで植物で行われてこなか

った。本研究において、TREseq 法と同じ培養条件のシロイヌナズナ培養細胞から RNA 上のリボソームの存在位置、存在量に関する情報を取得し、切断率に与える影響に着目し解析を行ったところ、遺伝子単位、切断部位単位の解析で RNA 上のリボソーム存在量が多いほど切断が生じやすいことを植物で初めて明らかにした (図 2-11, 図 2-12)。一方で、切断部位の周辺のリボソーム存在比率に顕著な偏りが認められなかったことや (図 2-9)、リボソーム存在量が多い遺伝子と少ない遺伝子で、切断部位の分布に顕著な違いは認められなかったことから (図 2-10)、翻訳過程 (リボソーム存在位置と存在量) は切断の位置決定には主として関与しないことがシロイヌナズナで示された。同様の傾向は、第二章の 2-3-12 で示したようにショウジョウバエ、出芽酵母でも認められたことから、多くの真核生物では、翻訳過程は切断率に正の影響を与えるが、切断の位置決定には大きく関与しないと考えられた。これらの結果は、RNA 切断に関わる配列依存性、翻訳過程の重要性が真核生物で保存されていることを示している。

第二章では、RNA 切断に関わる要因について個別に解析を行ったが、第三章では、想定される多くの要因について、数理モデルを用いて各特徴の RNA 切断への寄与度を評価した。第二章の結果を基に、RNA 切断部位の位置 (切断・非切断部位) の決定には配列が重要であるかをランダムフォレスト分類を用いて検証した。これまで、リボソームの存在量や存在位置が切断部位の決定に重要であると考えられていたが (18, 22, 23)、ランダムフォレスト分類の結果からは、翻訳過程が切断部位の位置決定に与える影響は弱く、配列特徴が最も大きく関与することが明らかとなった (図 3-13)。この結果は、第二章の結果と一致しており、リボソーム存在位置や存在量ではなく、切断部位周辺の配列が切断部位の位置決定にとって重要であることを示している。また、各切断部位の切断率に関わる特徴について、ラッソ回帰を用いた特徴選択を行った結果、切断部位周辺の G 塩基比率や RNA 上のリボソーム存在量に関わる特徴が高い係数を示し (図 3-16, 表 3-3)、切断率には切断部位周辺の配列だけではなく、RNA 上のリボソーム存在量など RNA 全体の特徴が切断率に正に関与することが示された。加えて、T87 培養細胞にて *R-luc* を一過的に発現させた検証実験で、切断部位周辺の G 塩基を A 塩基に置換することで、切断率は 50 分の 1 程度まで減少することや (Kaneko, unpublished)、翻訳効率を高める (RNA 上のリボソーム存在量が多い) と各切断部位の切断率は増加したことから (Kaneko, unpublished)、実際に植物細胞内でも切断部位周辺の配列と RNA 上のリボソーム存在量が各切断部位の切断率に正に関与することが明らかとなった。

本研究の結果は、植物 RNA 切断機構について、切断部位の位置や切断率を決定する複数の要因を評価し、それぞれの要因が切断率に関与する寄与度を明らかにした。これらの知見は、RNA 切断機構への理解を大きく進歩させ、将来の植物 RNA 分解機構の解明にとって非常に重要な情報となる。また、これらの知見は、第三章にて植物体で発現した外来遺伝子 (*F-luc* RNA および *R-luc* RNA) の切断率を予測できたように (図 3-20, 図 3-21)、外来遺伝子の発現を調節する上でも非常に有効であると考えられる。1989 年に Hiatt らによって組換え抗体の生産が植物体を用いて初めて報告されて以降、植物細胞での外来遺伝子の発現、特に医療用タンパク質を生産させる試みは、さまざまな外来タンパク質を対象に盛んに行われてきた (73, 74)。近年では、エボラ出血熱やコロナウィルスに対するワクチンが植物を用いて生産されるなど、これらのバイオ医薬品の市場規模は著しい速度で拡大している (75, 76)。本研究で得られた情報を基に、対象とする外来遺伝子から切断率が高い配列をあらかじめ除去することが可能であり、導入遺伝子のより効率的な発現が期待できることから、植物細胞を用いた有用物質生産に貢献する可能性も秘めている。

謝辞

本研究を遂行するにあたり、御指導、御鞭撻を賜りました出村拓教授に厚く御礼申し上げます。加藤晃教授には大変お忙しい中、直接の懇切なる御指導ならびに格別なる御高配を賜り、深く御礼申し上げます。久保稔特任准教授 (現熊本大学　特任講師)、大谷美沙都助教 (現東京大学　准教授)、國枝正助教、中田未友希助教、津川暁特任助教、ならびに Yoichiro Watanabe 特任助教には貴重な御助言と多大なる御配慮を賜り、感謝申し上げます。

また、植物代謝制御研究室の皆様には本当にお世話になりました。原田麻記氏には、事務的な面でお世話になりました。金城聖子氏には試薬類の作製に関し大変お世話になりました。さらに、川邊陽文博士、山﨑将太朗博士にはデータ解析や研究における御指導のみならず、多岐に渡ってお世話になりました。厚く御礼申し上げます。鈴木淳展氏、西村侑美氏には、私の至らないところもあり、迷惑をかけることもありましたが、研究の遂行にあたって様々な面でお世話になりました。ここに全ての方のお名前を挙げることはできませんが、植物代謝制御研究室の皆様の御指導、御助言、御協力等に対し、心から御礼申し上げます。皆様からの励ましによって修士、博士課程を含め有意義な 5 年間を過ごすことができました。

また、本学の友人達にも大変お世話になりました。研究や就職活動などで行き詰った時には支えて頂き、また時には他愛のない話で笑いあえる最高の仲間でした。

最後に、家族にはいつも自分の意思を尊重し、暖かく見守っていただきました。この場を借りて深く御礼申し上げます。

参考文献

1. Keene, J. D. (2010) Minireview: Global regulation and dynamics of ribonucleic acid. *Endocrinology*. **151**, 1391–1397
2. Narsai, R., Howell, K. A., Millar, A. H., O'Toole, N., Small, I., and Whelan, J. (2007) Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *Plant Cell*. **19**, 3418–3436
3. Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012) Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res.* **22**, 947–956
4. Barnes, T., Kim, W. C., Mantha, A. K., Kim, S. E., Izumi, T., Mitra, S., and Lee, C. H. (2009) Identification of Apurinic/apyrimidinic endonuclease 1 (APE1) as the endoribonuclease that cleaves c-myc mRNA. *Nucleic Acids Res.* **37**, 3946–3958
5. Melnik, S., Werth, N., Boeuf, S., Hahn, E. M., Gotterbarm, T., Anton, M., and Richter, W. (2019) Impact of c-MYC expression on proliferation, differentiation, and risk of neoplastic transformation of human mesenchymal stromal cells. *Stem Cell Res. Ther.* **10**, 1–18
6. Hwang, J. Y., and Buskirk, A. R. (2017) A ribosome profiling study of mRNA cleavage by the endonuclease RelE. *Nucleic Acids Res.* **45**, D327–D336
7. Jaglo, K. R., Kleff, S., Amundsen, K. L., Zhang, X., Haake, V., Zhang, J. Z., Deits, T., and Thomashow, M. F. (2001) Components of the Arabidopsis C-repeat/dehydration-responsive element binding factor cold-response pathway are conserved in Brassica napus and other plant species. *Plant Physiol.* **127**, 910–917
8. Chiba, Y., Mineta, K., Hirai, M. Y., Suzuki, Y., Kanaya, S., Takahashi, H., Onouchi, H., Yamaguchi, J., and Naito, S. (2013) Changes in mRNA stability associated with cold stress in arabidopsis cells. *Plant Cell Physiol.* **54**, 181–194
9. Bashirullah, A., Halsell, S. R., Cooperstock, R. L., Kloc, M., Karaiskakis, A., Fisher, W. W., Weili, F., Hamilton, J. K., Etkin, L. D., and Lipshitz, H. D. (1999) Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster. *EMBO J.* **18**, 2610–2620
10. Mishima, Y., and Tomari, Y. (2016) Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. *Mol. Cell*. **61**, 874–885
11. Parker, R., and Song, H. (2004) The enzymes and control of eukaryotic mRNA turnover. *Nat. Struct. Mol. Biol.* **11**, 121–127
12. Parker, R. (2012) RNA degradation in Saccharomyces cerevisae. *Genetics*. **191**, 671–702

13. Chiba, Y., and Green, P. J. (2009) mRNA degradation machinery in plants. *J. Plant Biol.* **52**, 114–124
14. Rymarquis, L. A., Souret, F. F., and Green, P. J. (2011) Evidence that XRN4, an Arabidopsis homolog of exoribonuclease XRN1, preferentially impacts transcripts with certain sequences or in particular functional categories. *Rna*. **17**, 501–511
15. Basbouss-Serhal, I., Pateyron, S., Cochet, F., Leymarie, J., and Bailly, C. (2017) 5' to 3' mRNA decay contributes to the regulation of arabidopsis seed germination by dormancy. *Plant Physiol.* **173**, 1709–1723
16. Tam, P. P. C., Barrette-Ng, I. H., Simon, D. M., Tam, M. W. C., Ang, A. L., and Muench, D. G. (2010) The Puf family of RNA-binding proteins in plants: Phylogeny, structural modeling, activity and subcellular localization. *BMC Plant Biol.* 10.1186/1471-2229-10-44
17. Doma, M. K., and Parker, R. (2006) Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature*. **440**, 561–564
18. Yu, X., Willmann, M. R., Anderson, S. J., and Gregory, B. D. (2016) Genome-wide mapping of uncapped and cleaved transcripts reveals a role for the nuclear mrna cap-binding complex in cotranslational rna decay in arabidopsis. *Plant Cell*. **28**, 2385–2397
19. Hou, C. Y., Lee, W. C., Chou, H. C., Chen, A. P., Chou, S. J., and Chen, H. M. (2016) Global analysis of truncated RNA ends reveals new insights into Ribosome Stalling in plants. *Plant Cell*. **28**, 2398–2416
20. German, M. A., Pillay, M., Jeong, D. H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008) Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat. Biotechnol.* **26**, 941–946
21. Gregory, B. D., O'Malley, R. C., Lister, R., Urich, M. A., Tonti-Filippini, J., Chen, H., Millar, A. H., and Ecker, J. R. (2008) A Link between RNA Metabolism and Silencing Affecting Arabidopsis Development. *Dev. Cell*. **14**, 854–866
22. Ibrahim, F., Maragkakis, M., Alexiou, P., and Mourelatos, Z. (2018) Ribothrypsis, a novel process of canonical mRNA decay, mediates ribosome-phased mRNA endonucleolysis. *Nat. Struct. Mol. Biol.* **25**, 302–310
23. Pelechano, V., Wei, W., and Steinmetz, L. M. (2015) Widespread co-translational RNA decay reveals ribosome dynamics. *Cell*. **161**, 1400–1412
24. Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008) Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. *Curr. Biol.* **18**, 758–762
25. Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y., and Robb, G. B. (2012) Structural bias in T4 RNA ligase-mediated 3′-adapter ligation. *Nucleic Acids Res.* **40**, e54

26. Song, Y., Liu, K. J., and Wang, T. H. (2014) Elimination of ligation dependent artifacts in T4 RNA ligase to achieve high efficiency and low bias microRNA capture. *PLoS One*. **9**, e94619
27. Hou, C. Y., Wu, M. T., Lu, S. H., Hsing, Y. I., and Chen, H. M. (2014) Beyond cleaved small RNA targets: Unraveling the complexity of plant RNA degradome data. *BMC Genomics*. **15**, 15
28. Murata, M., Nishiyori-Sueki, H., Kojima-Ishiyama, M., Carninci, P., Hayashizaki, Y., and Itoh, M. (2014) Detecting expressed genes using CAGE. *Methods Mol. Biol*. **1164**, 67–85
29. Mercer, T. R., Dinger, M. E., Bracken, C. P., Kolle, G., Szubert, J. M., Korbie, D. J., Askarian-Amiri, M. E., Gardiner, B. B., Goodall, G. J., Grimmond, S. M., and Mattick, J. S. (2010) Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.* **20**, 1639–1650
30. Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P. (2016) Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.* **14**, 1787–1799
31. Willmann, M. R., Berkowitz, N. D., and Gregory, B. D. (2014) Improved genome-wide mapping of uncapped and cleaved transcripts in eukaryotes-GMUCT 2.0. *Methods*. **67**, 64–73
32. Matsui, T., Takita, E., Sato, T., Kinjo, S., Aizawa, M., Sugiura, Y., Hamabata, T., Sawada, K., and Kato, K. (2011) N-glycosylation at noncanonical Asn-X-Cys sequences in plant cells. *Glycobiology*. **21**, 994–999
33. Hasegawa, A., Daub, C., Carninci, P., Hayashizaki, Y., and Lassmann, T. (2014) MOIRAI: A compact workflow system for CAGE analysis. *BMC Bioinformatics*. **15**, 144
34. Matsuura, H., Shinmyo, A., and Kato, K. (2008) Preferential translation mediated by Hsp81-3 5′-UTR during heat shock involves ribosome entry at the 5′-end rather than an internal site in Arabidopsis suspension cells. *J. Biosci. Bioeng.* **105**, 39–47
35. Yamasaki, S., Matsuura, H., Demura, T., and Kato, K. (2015) Changes in Polysome Association of mRNA Throughout Growth and Development in Arabidopsis thaliana. *Plant Cell Physiol.* **56**, 2169–2180
36. Adiconis, X., Haber, A. L., Simmons, S. K., Levy Moonshine, A., Ji, Z., Busby, M. A., Shi, X., Jacques, J., Lancaster, M. A., Pan, J. Q., Regev, A., and Levin, J. Z. (2018) Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nat. Methods*. **15**, 505–511
37. Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A. M. M., Sheng, Y., Abdelhamid, R. F., Anand, S., Gehrig, J., Akalin, A., Kockx, C. E. M., Van Der Sloot, A. A. J., Van IJcken, W. F. J., Armant, O., Rastegar, S., Watson, C., Strahle, U., Stupka, E., Carninci, P., Lenhard, B., and Muller, F. (2013) Dynamic

regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res.* **23**, 1938–1950

38. De Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., Åström, G., Babina, M., Bertin, N., Burroughs, A. M., Carlisle, A. J., Daub, C. O., Detmar, M., Deviatiiarov, R., Fort, A., Gebhard, C., Goldowitz, D., Guhl, S., Ha, T. J., Harshbarger, J., Hasegawa, A., Hashimoto, K., Herlyn, M., Heutink, P., Hitchens, K. J., Hon, C. C., Huang, E., Ishizu, Y., Kai, C., Kasukawa, T., Klinken, P., Lassmann, T., Lecellier, C. H., Lee, W., Lizio, M., Makeev, V., Mathelier, A., Medvedeva, Y. A., Mejhert, N., Mungall, C. J., Noma, S., Ohshima, M., Okada-Hatakeyama, M., Persson, H., Rizzu, P., Roudnicky, F., Sætrom, P., Sato, H., Severin, J., Shin, J. W., Swoboda, R. K., Tarui, H., Toyoda, H., Vitting-Seerup, K., Winteringham, L., Yamaguchi, Y., Yasuzawa, K., Yoneda, M., Yumoto, N., Zabierowski, S., Zhang, P. G., Wells, C. A., Summers, K. M., Kawaji, H., Sandelin, A., Rehli, M., Hayashizaki, Y., Carninci, P., Forrest, A. R. R., and De Hoon, M. J. L. (2017) An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872–878
39. Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*. **121**, 207–221
40. Yoshikawa, M., Peragine, A., Mee, Y. P., and Poethig, R. S. (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. *Genes Dev.* **19**, 2164–2175
41. Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J. W., Barton, G. J., and Simpson, G. G. (2020) Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification. *Elife*. **9**, e49658
42. Bracken, C. P., Szubert, J. M., Mercer, T. R., Dinger, M. E., Thomson, D. W., Mattick, J. S., Michael, M. Z., and Goodall, G. J. (2011) Global analysis of the mammalian RNA degradome reveals widespread miRNA-dependent and miRNA-independent endonucleolytic cleavage. *Nucleic Acids Res.* **39**, 5658–5668
43. Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., and Coller, J. (2015) Codon optimality is a major determinant of mRNA stability. *Cell*. **160**, 1111–1124
44. Lee, W. C., Hou, B. H., Hou, C. Y., Tsao, S. M., Kao, P., and Chen, H. M. (2020) Widespread exon junction complex footprints in the RNA degradome mark mRNA degradation before steady state translation. *Plant Cell*. **32**, 904–922
45. Nagarajan, V. K., Kukulich, P. M., Von Hagel, B., and Green, P. J. (2019) RNA degradomes reveal substrates and importance for dark and nitrogen stress responses of Arabidopsis XRN4. *Nucleic Acids Res.* **47**, 9216–9230
46. Gaglia, M. M., Rycroft, C. H., and Glaunsinger, B. A. (2015) Transcriptome-Wide Cleavage Site Mapping on Cellular mRNAs Reveals Features Underlying Sequence-Specific Cleavage by the Viral Ribonuclease SOX. *PLoS Pathog.* **11**, 1–25

47. Nashimoto, M., Geary, S., Tamura, M., and Kaspar, R. (1998) RNA heptamers that direct RNA cleavage by mammalian tRNA 3' processing endoribonuclease. *Nucleic Acids Res.* **26**, 2565–2571
48. Zeng, Y., and Cullen, B. R. (2005) Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J. Biol. Chem.* **280**, 27595–27603
49. Lei, L., Shi, J., Chen, J., Zhang, M., Sun, S., Xie, S., Li, X., Zeng, B., Peng, L., Hauck, A., Zhao, H., Song, W., Fan, Z., and Lai, J. (2015) Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J.* **84**, 1206–1218
50. Ueno, D., Mikami, M., Yamasaki, S., Kaneko, M., Mukuta, T., Demura, T., and Kato, K. (2020) Changes in mRNA Degradation Efficiencies under Varying Conditions Are Regulated by Multiple Determinants in Arabidopsis thaliana. *Plant Cell Physiol.* 10.1093/pcp/pcaa147
51. Liu, T. Y., Huang, H. H., Wheeler, D., Xu, Y., Wells, J. A., Song, Y. S., and Wiita, A. P. (2017) Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. *Cell Syst.* **4**, 636-644.e9
52. Neymotin, B., Athanasiadou, R., and Gresham, D. (2014) Determination of in vivo RNA kinetics using RATE-seq. *RNA*. **20**, 1645–1652
53. Burow, D. A., Martin, S., Quail, J. F., Alhusaini, N., Coller, J., and Cleary, M. D. (2018) Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. *Cell Rep.* **24**, 1704–1712
54. Ueno, D., Mukuta, T., Yamasaki, S., Mikami, M., Demura, T., Matsui, T., Sawada, K., Katsumoto, Y., Okitsu, N., and Kato, K. (2020) Different plant species have common sequence features related to mRNA degradation intermediates. *Plant Cell Physiol.* **61**, 53–63
55. Luo, S., He, F., Luo, J., Dou, S., Wang, Y., Guo, A., and Lu, J. (2018) Drosophila tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. *Nucleic Acids Res.* **46**, 5250–5268
56. Gerashchenko, M. V., and Gladyshev, V. N. (2017) Ribonuclease selection for ribosome profiling. *Nucleic Acids Res.* **45**, e6
57. Khateb, S., Weisman-Shomer, P., Hershco-Shani, I., Ludwig, A. L., and Fry, M. (2007) The tetraplex (CGG)n destabilizing proteins hnRNP A2 and CBF-A enhance the in vivo translation of fragile X premutation mRNA. *Nucleic Acids Res.* **35**, 5775–5788
58. Beaudoin, J. D., and Perreault, J. P. (2010) 5′-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res.* **38**, 7022–7036
59. Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., and Rouskin, S. (2016) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods*. **14**, 75–82

60. Simms, C. L., Yan, L. L., and Zaher, H. S. (2017) Ribosome Collision Is Critical for Quality Control during No-Go Decay. *Mol. Cell*. **68**, 361-373.
61. Green, P. J. (1993) Control of mRNA stability in higher plants. *Plant Physiol.* **102**, 1065–1070
62. Beelman, C. A., and Parker, R. (1994) Differential effects of translational inhibition in cis and in trans on the decay of the unstable yeast MFA2 mRNA. *J. Biol. Chem.* **269**, 9687–9692
63. Ross, J. (1995) mRNA Stability in Mammalian Cells. *Microbiol Rev*. **59**, 423–450
64. Radhakrishnan, A., Chen, Y. H., Martin, S., Alhusaini, N., Green, R., and Coller, J. (2016) The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell*. **167**, 122-132.e9
65. Webster, M. W., Chen, Y. H., Stowell, J. A. W., Alhusaini, N., Sweet, T., Graveley, B. R., Coller, J., and Passmore, L. A. (2018) mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases. *Mol. Cell*. **70**, 1089-1100.e8
66. Lamanna, A. C., and Karbsteina, K. (2009) Nob1 binds the single-stranded cleavage site D at the 3′-end of 18S rRNA with its PIN domain. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14259–14264
67. Qi, Y. (2012) Random forest for bioinformatics. *Ensemble Mach. Learn. Methods Appl.* 10.1007/9781441993267_10
68. Qabaja, A., Alshalalfa, M., Bismar, T. A., and Alhajj, R. (2013) Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions Computational methods for biomarker discovery and systems biology research. *Eurasip J. Bioinforma. Syst. Biol.* **2013**, 1–11
69. Hu, Q., Merchante, C., Stepanova, A. N., Alonso, J. M., and Heber, S. (2015) Mining transcript features related to translation in Arabidopsis using LASSO and random forest. *2015 IEEE 5th Int. Conf. Comput. Adv. Bio Med. Sci. ICCABS 2015*. 10.1109/ICCABS.2015.7344713
70. Kyung, M., Gilly, J., Ghoshz, M., and Casellax, G. (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5**, 369–412
71. Volkova, O. A., and Kochetov, A. V. (2012) Interrelations between the Nucleotide Context of Human Start AUG Codon, N-end Amino Acids of the Encoded Protein and Initiation of Translation. *J. Biomol. Struct. Dyn.* **27**, 611–618
72. Tian, B., and Graber, J. H. (2012) Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*. **3**, 385–396
73. Hiatt, A., Cafferkey, R., and Bowdish, K. (1989) Production of antibodies in transgenic plants. *Nature*. **342**, 76–78

74. Shanmugaraj, B., Bulaon, C. J. I., and Phoolcharoen, W. (2020) Plant molecular farming: A viable platform for recombinant biopharmaceutical production. *Plants*. **9**, 1–19
75. Rosales-Mendoza, S., Nieto-Gómez, R., and Angulo, C. (2017) A perspective on the development of plant-made vaccines in the fight against ebola virus. *Front. Immunol.* **8**, 1–13
76. Rosales-Mendoza, S., Márquez-Escobar, V. A., González-Ortega, O., Nieto-Gómez, R., and Arévalo-Villalobos, J. I. (2020) What does plant-based vaccine technology offer to the fight against COVID-19? *Vaccines*. **8**, 1–19