東京大学大学院新領域創成科学研究科

情報生命科学専攻

平成 19 年度

修士論文

出芽酵母における新規機能性 RNA の探索

2008 年 3 月提出

指導教員　浅井　潔　教授

66961　古川　貴久

概要

近年の数多くの目覚しい研究成果から、機能性 RNA が生体において基本的な代謝から個体発生や細胞分化までの実に様々な生命現象に関与していることが発見され、これらは以前考えられていたよりもはるかに重要な役割を有すると考えられるようになってきている。
本研究で解析対象とした出芽酵母は生物学の発展に常に重要な役割を担っているモデル生物であり、真核生物の基本的な性質を知る手がかりとなる生物である。この出芽酵母においてはゲノムの網羅的なタイリングアレイ解析や cDNA 解析などから、機能の不明なタンパク質をコードしない転写領域が数多く見つかっており、これらの領域において未知の機能性 RNA が存在する可能性が考えられる。本研究では、情報科学的なアプローチにより新規機能性 RNA の候補となる配列を探索するためのパイプラインを提案し、特に UTR について解析を行った。
まず、出芽酵母の全ゲノムのマルチプルアラインメントデータに対して MXSCARNA を実行することで、２次構造を考慮したアラインメントデータを作成した。そして、それらを入力として RNAz を実行し、進化的に保存されていて熱力学的に安定な２次構造を持つ機能性 RNA 候補を 6211 配列特定した。その予測結果の領域について SGD のアノテーションデータとの重複を調べたところ、既知機能性 RNA の 53%に重複があったうえ、UTR には 591 配列の機能性 RNA 候補があった。次に、RNAz の実行結果で UTR に存在することが予測された機能性 RNA の候補配列を入力として、INFERNAL を実行し、既知の機能性 RNA との相同性において候補となる 44 配列を検出した。また、それらの候補配列を入力として LocARNA を実行することにより候補配列間の構造類似性を調べ、その結果に基づいて WPGMA でクラスタリングを行った。そして、そこで得られた全てのクラスターにおいて GO::TermFinder を実行した結果、各クラスター内の配列を UTR に持つ遺伝子が有意な数で共通の GO タームを持つクラスターを 68 個の GO タームで検出した。これらから得られた結果は情報科学的なアプローチによる予測ではあるが、出芽酵母の UTR において新規機能性 RNA の存在の可能性を示唆している。

目次

1章　導入

1.1　機能性 RNA の重要性

タンパク質が生体において多種多様な機能をもっていることは広く認識されており、その例としては、代謝などの化学反応を起こさせる触媒として働くタンパク質や、コラーゲンやケラチンなど生体構造を形成するタンパク質などが挙げられる。これらタンパク質に関する多くの研究成果から、ゲノム配列においてタンパク質をコードする遺伝子が生体における主要な制御機能を果たしているということが認識されてきた。

しかし、近年の研究成果からは非タンパクコード領域においても生体におけるさまざまな制御に関連している機能性 RNA が多く発見され、非常に注目を集めている分野となっている[Eddy 2001]。これらの領域において従来から機能が特定されていた例として tRNA や rRNA が既に広く知られている。さらに、近年にはさまざまな RNA の部位特異的な修飾を導く snoRNA(small nucleolar RNA)[Kiss 2002]や、転写後の調節装置として働く miRNA(microRNA)[Pasquinelli *et al.* 2002]などがさまざまな生物種において次々と発見されている。これらの例をはじめとして、基本的な代謝から個体発生や細胞分化までの実に様々な生命現象に関与する機能性 RNA が数多く発見されており[Storz 2002]、これらは以前考えられていたよりもはるかに重要な役割を有すると考えられるようになってきている。

１.２　モデル生物としての出芽酵母

出芽酵母は実用面での有用性と、実験面での利便性が両輪となり、生物学の発展に常に重要な役割を担っているモデル生物である。単細胞性の菌類に分類されており、哺乳類などの高等真核生物の共通性を外見から見いだすのは難しいが、生命現象の基本的な分子機構は驚くほど保存されていることがわかっている。よって出芽酵母を解析することにより、真核細胞の基本的な性質について知ることができるといえる。

また、1996 年に真核生物として最初にゲノムの全塩基配列が決定され、そのゲノムサイズは 1200 万塩基程度であり、遺伝子数が約 6000 程度(その後修正されている)であることが特定されている[Goffeau *et al.* 1996]。特にそれ以降は、ゲノム、トランスクリプトーム、プロテオームなど、さまざまな面で多くの目覚しい研究成果が報告されている[Giaever *et al.* 2002, Miura *et al.* 2006, Ghaemmaghami *et al.* 2003]。しかし、比較的シンプルな生物であり、多くの研究が行われているにも関わらず、1000 以上の遺伝子がまだ機能が特定されていないうえ[Pena-castillo *et al.* 2007]、不明な点がまだ多く残されていると考えられている。

１.３　非タンパク質コード領域の転写物発現の証拠

近年、出芽酵母ゲノムにおいては網羅的なタイリングアレイ解析[David *et al.* 2006, Juneau *et al.* 2007]や cDNA 解析[Miura *et al.* 2006]が行われ、機能未知の非タンパクコード領域の転写物が多数見つかってきている。David らのタイリングアレイ解析では、出芽酵母のゲノムの少なくとも約 85%はどちらかのストランドにおいて転写している証拠を示し、それら転写領域の 16%はアノテーションのない領域からの転写物であることを示している。また、Miura らの cDNA 解析においても、アノテーションのない領域から 667 個の独立した転写物が特定されている。これらの観察は出芽酵母における非タンパクコード領域の転写物において、未知の機能の可能性を示すと共に、今後の正確な解析の必要性を示唆している。また、これらのことは出芽酵母に限らず、高等真核生物の多くのゲノムにおいても驚くべき多くの非タンパクコード領域の転写物の証拠が提供されてきていることもあり、これらの領域は情報学的、実験的の両方の面において非常に注目されるべき研究対象となっている[Storz 2002]。

１.４　UTR(untranslated region)に存在する機能性 RNA

タンパク質コード領域の転写物である mRNA に存在し、タンパク質として翻訳されない UTR も非タンパク質コード領域の転写物の大きな要素である。この UTR が特定の mRNA において転写後に重要な役割を果たしていることはいくつかの例で既に知られており、近年も mRNA の制御に関わる領域の特定やそれらに関連するタンパク質の特定が広く行われている[Kuersten *et al.* 2003]。代表的なものとして、5'UTR に存在する Gcap 構造や 3'UTR の poly(A)が転写の効率に非常に重要な役割を果たしていることが知られているが、それら以外にも、mRNA の輸送の調節や転写効率の調節などの発現制御[van der Velden *et al.* 1999]、細胞内の局在の調節[Jansen 2001]や安定性の調節[Bashirullah 2001]など、現在までに多くの生物種において、さまざまな種類の機能が特定されている。さらに、UTR の長さ、２次構造の存在、上流の読み枠(uORFs)など、多くの要素が mRNA の制御に関連していることが観察されているうえ、制御タンパク質の結合サイトとして機能している配列が多数含まれていることがわかっている[Kuersten *et al.* 2003]。

出芽酵母のゲノムにおいては約 12-15%が UTR であることが示されており、未だ特定されていない機能性 RNA の存在が十分に考えられる[Hurowitz *et al.* 2003]。また、近年の出芽酵母の UTR に関する研究成果にとして興味深い具体例に次の２つがある。１つ目は配列モチーフに関する成果であり、出芽酵母の 3'UTR には mRNA の安定性に関連している可能性のある 53 種の配列モチーフが存在し、それらの配列モチーフには特定の機能に関連するものがあることが示されている[Shalgi *et al.* 2005]。２つ目は UTR の長さに関する研究成果である。5'UTR と 3'UTR の両側において、UTR の長さが遺伝子の機能や局在などに関連していて、より長い UTR を持つ遺伝子は制御を必要とする遺伝子カテゴリーに分類されていることが報告されている[David *et al.* 2006]。

これらのことから、特に、長い UTR を持つ遺伝子に、生体において特定の機能に関連する配列モチーフが存在している可能性が考えられる。そして、配列モチーフには特定の構造モチーフがあることが推測できるうえ、さらに、配列モチーフではなくとも構造モチーフを持ち、特定の機能に関連しているグループの存在が考えられる。

１.５　情報科学的な機能性 RNA の特定の取り組み

機能性 RNA の重要性は広く認識されてきており、細胞内の仕組みの包括的な理解のためには機能性 RNA の考慮が必要不可欠である。しかし、実験的な取り組みは費用の面や労力の面から制約を受けることが多くなる。それゆえ、情報科学的な手法により、効率的にゲノム配列中から機能性 RNA の候補となる配列を特定することは非常に価値があり、近年、いくつかの手法が提案されている[Rivas *et al.* 2001, Washietl *et al.* 2005, Pedersen *et al.* 2006]。

タンパク質コード遺伝子に開始コドンや終止コドンがあることなどとは異なり、機能性 RNA はアルゴリズムにおいて信頼できる検出の指針となる共通の特徴がゲノム配列には存在していない。しかしながら、tRNA や rRNA が代表するように多くの機能性 RNA は 2 次構造がその機能に強く関連していることがわかっており、2 次構造の特徴を利用した機能性 RNA の検出が非常に重要な手法となっている。加えて、近年多くの生物種においてゲノムが解読されてきたことに伴い、２次構造の進化的な保存という指標が機能性 RNA の探索において強い説得力をもつことが示されている[Washietl *et al.* 2005]。この２点の指標を有効に利用したツールに RNAz[Washietl *et al.* 2005]と Evofold[Pedersen *et al.* 2006]があり、これらのソフトウェアではマルチプルアラインメントデータを入力データとして機能性 RNA の候補を効率よく探すことができることが示されている。

1.6　先行研究と本研究の位置づけ

出芽酵母における新規機能性 RNA をゲノムから網羅的に特定する取り組みとしてはこれまでに２つの研究が行われている[McCutcheon *et al.* 2003, Steigele *et al.* 2007]。特に、Steigele らによる研究において RNAz を出芽酵母のゲノムの全領域に対して実行し、多くの２次構造を持つ RNA 配列の存在を予測している。そして、予測された２次構造を持つ RNA でアノテーションのない領域の一部は UTR に重複し、その UTR を持つ遺伝子群は特定の機能に有意に関連していることが示されている。この成果は、出芽酵母における UTR に新規機能性 RNA の存在の可能性を示唆しているものといえる。

本研究においては、上記してきたいくつかの理由と、この先行研究から、出芽酵母の UTR に注目し、これらの領域に予測される２次構造をもつ RNA において、より具体的な解析を行うパイプラインを提案している。パイプラインの詳細については２章に記述するが、本研究では、既存の研究成果と比べて、(1) 2 次構造を考慮したアラインメントデータの作成、(2)既知機能性 RNA との相同性に基づく探索、(3)構造類似性においてのクラスタリング、

という３点において新しい取り組みとなっている。これらの 3 点について下記に示す。

1.6.(1)　2 次構造を考慮したアラインメントデータの作成

マルチプルアラインメントデータは進化的な情報を得られることから、ゲノム解析において非常に重要なデータであり、RNAz を実行する際に入力として必要となる。出芽酵母におけるデータとしては、7 種の酵母(S.cerevisiae, S.paradoxus, S.mikatae, S.kudriavzevii, S.bayanus, S.castellii, S.kluyberi)のゲノムから、MULTIZ により作成されたマルチプルアラインメントデータが既に作成されている。このとき利用されている MULTIZ は、対象となる複数のゲノムにおいて BLASTZ で作成した局所的なアラインメントのセットをもとにデータを作成しているため、ゲノムの 1 次配列をもとにアラインメントしたデータとなっているという問題点がある。なぜなら、機能性 RNA には 2 次構造が強く関連していることが広く知られており、1 次配列のみに基づくアラインメントでは重要な情報を見落としてしまう可能性があることが示されているからである[Torarinsson *et al.* 2007]。よって、機能性 RNA の探索においては、2 次構造を考慮したマルチプルアラインメントデータを利用することで、より信頼できる予測ができると考えられる。そこで、本研究では、MXSCARNA[Tabei *et al.* 2008]を用いて、既存のマルチプルアラインメントデータをリアラインメントすることにより 2 次構造を考慮したアラインメントデータを得て、そのデータを RNAz の入力として利用している。MXSCARNA は 2 次構造を考慮したアラインメントにおいて、最高精度のツールのひとつであり、かつ、高速に実行できるという利点が示されている[Tabei *et al.* 2008]。

1.6.(2)　既知機能性 RNA との相同性に基づく探索

この探索には Ram[Griffiths-Jones *et al.* 2005]で定義されている ncRNA ファミリーのプロファイルデータ(共分散モデル)と、INFERNAL ソフトウェアパッケージのプログラム cmsearch を利用している[Eddy *et al.* 2002]。

Rfam は既知の ncRNA のファミリーを整理しているデータベースであり、それぞれのファミリーにおける配列のマルチプルアラインメントとプロファイルデータを作成している。そのマルチプルアラインメントは 2 次構造と 1 次配列の両方に基づいて専門家により手作業により管理されており、そのアラインメントがプロファイルデータを作成するために利用されている。このことから、作成されるプロファイルデータは信頼されるデータとして考えられている[Griffiths-Jones *et al.* 2005]。

本研究では INFERNAL のプログラム cmsearch を利用することで、Rfam の ncRNA ファミリーのプロファイルデータを RNAz で得られた UTR の機能性 RNA 候補配列に対して探索し、そのファミリーと相同性を持つ領域を検出している。

1.6.(3)　構造類似性においてのクラスタリング

Steigele らの先行研究において予測に利用された RNAz[Washietl *et al.* 2005]では、進化的な配列保存性と熱力学的安定性という 2 点に基づいて、SVM によって機能性 RNA の候補となる配列を検出している。しかし、機能性 RNA 候補特定のためには、これら以外にもさまざまな特徴が考えられ、具体的には、配列モチーフ、構造モチーフ、塩基組成などがある。本研究では、RNAz によって検出された機能性 RNA 候補配列を構造類似性に基づいてクラスタリングすることで、構造モチーフに注目した解析を提案している。

機能性 RNA 候補配列の構造類似性に基づいたクラスタリングは先行研究で行われている[Will *et al.* 2007]。この研究では局所的な RNA 配列のアラインメントを行う LocARNA を利用することで、Rfam に存在する既知の RNA の配列に対してクラスタリングを行い、それらにおいてよい分類ができることが示されている。さらに、Ciona intestinalis ゲノムにおいて RNAz で予測されていた 3332 個の 2 次構造を持つ RNA に対して実行したところ、tRNA など既知の RNA を取り戻すことができたことに加えて、構造類似性に基づいたいくつかの新しい RNA ファミリーを特定している。

本研究では、このクラスタリング手法を、出芽酵母の UTR において RNAz によって予測された 2 次構造をもつ RNA 配列において適用し、さらに、それぞれのクラスターにおいて GO::TermFinder を実行することで、有意に共通の機能を持つ遺伝子が含まれるクラスターを探索している。このことにより、UTR に類似した 2 次構造を持つ配列群と、特定の機能(GO ターム)との関連を調べることができる。

2 章　方法と構成要素

初めに、本章の概要を記す。
機能性 RNA を探索するための強力なツールとして INFERNAL や LocARNA があるが、これらは計算コストがかかるという問題点がある。この理由から、本研究では、まず、網羅的なゲノム解析に適した機能性 RNA の予測ツールである RNAz を利用して、候補領域の絞込みを行った。このとき、RNAz は入力としてマルチプルアラインメントデータを必要とするが、機能性 RNA の探索により適したデータとするために、MXSCARNA を用いてリアラインメントを行った。
おおまかにパイプラインを述べると、MXSCARNA を実行したマルチプルアラインメントデータを入力として RNAz で絞込みを行い、RNAz で UTR から得られた候補配列において INFERNAL、LocARNA、GO::TermFinder を利用して新規機能性 RNA 候補の特定を行った。このプロセスは図 1 に示すようなパイプラインで示すことができる。以下に、このパイプライン中のそれぞれの過程の詳細を述べる。

2.1　出芽酵母のマルチプルアラインメントデータの取得

本研究ではパイプラインの過程で RNAz というソフトウェアを用いる(詳細は 2.4)。このソフトウェアを利用して、出芽酵母のゲノムから網羅的に機能性 RNA の候補配列を得るためには入力としてマルチプルアラインメントのデータが必要となる。
そこで、既存する 7 種の酵母(S.cerevisiae, S.paradoxus, S.mikatae, S.kudriavzevii, S.bayanus, S.castellii, S.kluyberi)の MULTIZ により計算されたマルチプルアラインメント(Multiple alignments of 7 Yeast species)を UCSC ゲノムブラウザよりダウンロードした。

2.2　2 次構造を考慮したアラインメントデータの作成（図 1－①）

導入にも示しているように、2.1 で得た MULTIZ により作成されているデータは、対象となるゲノムにおいて BLASTZ で作成した局所的なアラインメントのセットをもとに全ゲノムのマルチプルアラインメントデータを作成しているため、ゲノムの 1 次配列をもとにアラインメントしたデータとなっているという問題点がある。なぜなら、機能性 RNA には 2 次構造が強く関連していることが広く知られており、これらを探索する際には 1 次配列のみに基づくアラインメントでは重要な情報を見落としてしまう可能性があることが示されているからである[Torarinsson et al. 2007]。
この理由から、MXSCARNA[Tabei *et al.* 2008]を利用し 2.1 で取得したデータをリアラインメントして 2 次構造を考慮に入れたマルチプルアラインメントデータを作成した。MXSCARNA は 2 次構造を考慮したアラインメントにおいて、最高精度のツールのひとつ

であり、かつ、高速に実行できるという利点が示されている。

2.3　マルチプルアラインメントデータの前処理（図 1－②）

RNAz では 400 カラムを超えるアラインメントデータでは評価ができないため、マルチプルアラインメントデータのうち長いアラインメントブロックにおいては前処理が必要となる。

まず、200 カラム以上の長さを持つマルチプルアラインメントブロックにおいて、サイズ 120 カラム、スライド幅 40 カラムの重複する窓幅(window)でアラインメントをスライスした。このサイズとスライド幅は、長い ncRNA に含まれる局所的な 2 次構造を検出するのに十分な長さであると同時に、短い領域で構成される 2 次構造を検出することにおいても長すぎない長さであることが製作者らにより報告されている[Washietl *et al.* 2005]。

次に、先行研究を参考に予測に不適当な領域を予め除去した[Washietl *et al.* 2005]。出芽酵母の配列に対して、それぞれのアラインメントされている配列をペアワイズアラインメントし、ギャップが 25%以上の配列を除去した。また、アラインメントされているそれぞれの配列において長さが 50 塩基以下、GC 含量が 75%以上になる配列を除去した。これらの過程で出芽酵母の配列が除去された場合と、出芽酵母の配列のみがのこった場合はそのアラインメント領域を全て除去した。

また、RNAz は 7 配列以上のマルチプルアラインメントを計算することができないという制約があるため、7 以上の配列があ

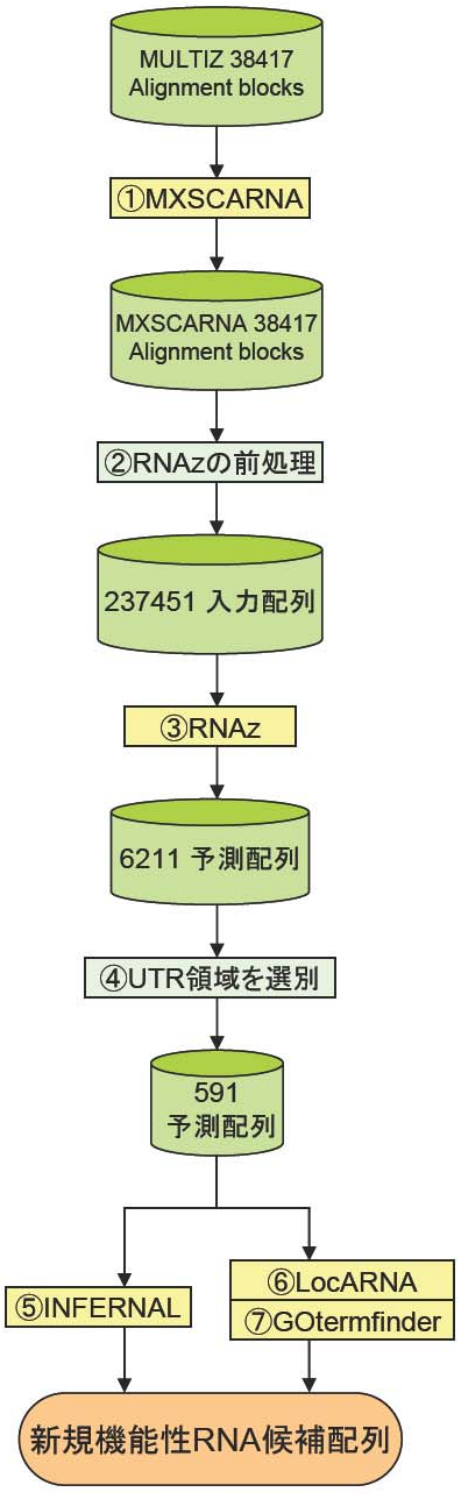


図1
パイプラインの流れ図を示している。
緑の円柱:配列データ、黄色の四角:ツール名、青色の四角:処理の過程、をそれぞれ示す。

る場合、それらの配列間の平均のペアワイズアラインメントの類似度が 80%により近づくように 6 配列を選択した。これは、RNAz の予測において適度な配列の相違が信頼性を高めることが知られているからである[Washietl *et al.* 2005]。さらに、マルチプルアラインメントのそれぞれの配列間で完全一致している配列は除去している。ここで、出芽酵母の配列が除かれることはなく、全ての配列が一致している場合はペアワイズアラインメントデータのみを保持した。この作業も、上記したように RNAz の予測において適度な配列の相違が信頼性を高めることが理由である。
以上のような過程により、RNAz の入力配列とするマルチプルアラインメントデータを作成した。

2.4　2 次構造を持つ RNA の検出（図 1－③）
2 次構造を持つ RNA の検出には RNAz というソフトウェアを用いる。このソフトウェアは進化的な情報が含まれるマルチプルアラインメントデータを利用して判別を行ううえ、網羅的なゲノム解析に適した速度で実行できることが報告されている[Washietl *et al.* 2005]。また、既存する同様の目的のソフトウェアである QRNA[Rivas *et al.* 2001]、DDBRNA[Bernardo *et al.* 2003]、MSARI[Coventry *et al.* 2004]と比べて性能がよいことが示されているうえ、高い検出感度と特異性があることも示されている[Washietl *et al.* 2005]。
RNAzを 2.3 で得られた全ての入力配列に対して実行し、出芽酵母のゲノム中から進化的に保存されていて熱力学的に安定な 2 次構造を持つRNA配列を検出した。RNAzはSVMに基づいて 2 次構造を持つRNAを予測する分類器であり、マルチプルアラインメントデータを入力として、それから予測される構造の熱力学的安定性と、共通構造に一致する配列の共変動に基づいて、進化的に保存された 2 次構造を持つ確率（P_{SVM}）の値を計算する[Washietl *et al.* 2005]。このとき、P_{SVM}のカットオフ値を 0.5 とするときに、多くの機能性RNA候補の分類の特異性と感受性においてよい性能を示すことが報告されていることから、本研究におけるカットオフ値としてP_{SVM} =0.5 を設定した。

2.5　2 次構造を持つ RNA のアノテーションデータとの比較（図 1－④）
RNAz の出力結果において 2 次構造を持つ RNA がゲノムにおいてどのように分布しているか調べるために、SGD(saccharomyces genome database)において作成されている出芽酵母ゲノムのアノテーションの情報(saccharomyces_cerevisiae.gff)を SGD からダウンロードした。
そして、アノテーションデータとの比較を行うため、RNAz により 2 次構造を持つ RNA として検出された予測配列のそれぞれをアノテーションの対象となっているゲノムデータに対して BLAST を実行し、そのゲノムにおける位置を特定した。このとき、多くの場合にスコアの高い領域が複数検出されたので、もっとも適当な箇所を選択できるように次のよう

な処理をした。まず、スコアの高い領域から順に、MULTIZ によるアラインメントデータの開始点の座標と BLAST でヒットした領域の開始点の座標を比較して、同一染色体の10000bp 以内に存在しているかどうかを調べ、この条件を満たす場合は、その BLAST のヒットを適切な位置として採用した。この条件を満たすものが上位 3 位ヒットまでに含まれない場合、その RNA の配列は除去した。この作業を行うことで、2 次構造を持つ RNA の配列をアノテーションデータとの比較に適切な位置を決定した。

これらの過程により決定された座標を利用し、アノテーションデータの CDS、ncRNA、intron、UTR 候補に含まれる 2 次構造を持つ RNA の配列の数を調べた。このとき、「UTR 候補」をイントロンに接しない CDS の両側 300bp 以内と定義してそれらの領域に含まれる数を調べ、アノテーションと重なる領域は除いている。この定義の主な理由としては、多くの遺伝子は転写物の構造が明らかではなく、5'UTR と 3'UTR の領域がはっきりとわからないことがある。また、先行研究において出芽酵母の UTR の平均の長さが約 260bp、遺伝子間領域の平均の長さが約 530bp であることなどを考慮に入れて決めた[Hurowitz *et al.* 2003]。

そして、先行研究において、RNAz による予測で UTR に存在する 2 次構造を持つ RNA は特定の機能に有意な関連が示されていることや、多くの UTR がその mRNA の制御に関連していることが知られていることから、UTR 候補に含まれていた 2 次構造をもつ RNA に注目した[Steigele *et al.* 2005、Kuersten *et al.* 2003]。

2.6　既知機能性 RNA との相同性の検索（図 1－⑤）

RNAz で UTR 候補に存在すると予測された 2 次構造を持つ RNA において、INFERNAL を用いて、既知機能性 RNA との相同性を調べることにより、新規機能性 RNA 候補配列を探索した。

この探索に必要なデータとして、まず Rfam から既知の 607 ファミリー(Version8.1)のプロファイルデータ(共分散モデル)をダウンロードした。Rfam は、ncRNA のファミリーを包括的に収集しているデータベースである。Rfam のそれぞれのファミリーのアラインメントデータ(seed alignment)は専門家により人手で管理されており、それらは信頼できる ncRNA の 2 次構造と 1 次配列の両方に基づいて作成されている。また、そのアラインメントデータ(seed alignment)から INFERNAL のプログラム cmbuild により、そのファミリーのプロファイルデータ(共分散モデル)が作成されている。これらのことから、Rfam と INFERNAL は ncRNA のファミリーの構成を定義するのに最も正確で一般的なツールと考えられている[Griffiths *et al.* 2005, Wang *et al.* 2007]。

ここでは Rfam で定義される 607 ファミリーのプロファイルデータを、RNAz で UTR 候補に存在すると予測された 2 次構造を持つ RNA に対して、INFERNAL のプログラム cmsearch を実行し、それらの配列における既知 ncRNA ファミリーとの相同性を調べた。この時、相同性のある候補の判別の閾値には、Rfam で定義される TC(Trusted cutoff)と

NC(noise cutoff)を用いた。TC は、Rfam のデータベースで管理される配列データにおいて、そのファミリーに含まれている配列データの中で最も低いスコアを表しており、NC はそのファミリーには含まれない配列において最も高いスコアを表している。

2.7　構造類似性の評価（図 1－⑥）

RNAz で UTR 候補に存在すると予測された 2 次構造を持つ RNA において、LocARNA [Will *et al.* 2007]を利用して配列間の構造類似性を評価した。

導入でも示したように、先行研究で、Rfam に存在する既知の RNA の配列群において LocARNA のスコアに基づいてクラスタリングを行った結果、それらにおいてよい分類ができることが示されている[Will *et al.* 2007]。さらに、Ciona intestinalis ゲノムにおいて RNAz で予測されていた 3332 個の 2 次構造を持つ RNA に対して実行したところ、tRNA など既知の RNA を取り戻すことができたことに加えて、構造類似性に基づいたいくつかの新しい RNA ファミリーを特定している。

このクラスタリング手法を、出芽酵母の UTR 候補において RNAz によって予測された 2 次構造をもつ RNA 配列において適用した。LocARNA は、2 次構造を考慮したうえで RNA 配列の局所的なアラインメントを行い、その構造類似性のスコアを算出できる。ここで、まず、構造類似性を評価する全ての配列において、McCaskill のアルゴリズムを利用してその塩基対確率行列を得た。そして、その結果を入力として、予測配列の全ての 2 配列間の組み合わせにおいて LocARNA を実行することで、2 次構造を考慮したうえで RNA 配列の局所的なアラインメントを行い、それらの配列間の構造類似性のスコアを算出した。

2.8　構造類似性に基づく樹形図の作成

RNAz で UTR 候補に存在すると予測された 2 次構造を持つ RNA において、2.7 で得た LocARNA のスコアから算出する距離の行列に基づいて、WPGMA(weighted pair-group method algorithm)によりクラスタリングを行い、樹形図を作成した。この樹形図は入力となる配列における構造類似性に基づいた樹形図となる。このとき、例外的に大きなスコアの値が距離に与える影響を避けるために次のような式で配列 i と配列 j の距離 d(i,j)を決めた。

$$d(i,j) = \max(0, q\text{-}score(i,j))$$

ここで、score(i,j)は配列 i と配列 j の LocARNA によるアラインメントのスコアであり、q は全ての 2 組の配列のスコアにおいて上位から 1%の位置にあたるスコアの値である。

2.9　有意に共通の機能を持つ遺伝子が含まれるクラスターの検出（図 1－⑦）

得られた樹形図における全てのクラスターにおいて GO::TermFinder を利用することにより、有意に共通の GeneOntology(GO)タームを持つ遺伝子のクラスターを検出した。この

とき、UTR 候補に予測された 2 次構造を持つ RNA の配列は mRNA に含まれることを仮定しているため、それぞれの配列に隣接している遺伝子を割り当てている。

GO::TermFinder はあるリストに含まれる遺伝子において、有意(P-value<0.05)に共通の GO タームを持っているかを調べ、そのリストに存在する全ての有意な GO タームを出力する。ある遺伝子のリストにおける、ある GO タームの P-value は、超幾何分布を利用し、ゲノムの全ての遺伝子におけるその GO タームの遺伝子数の割合をバックグラウンドとして、そのリストにおいて GO アノテーションのある遺伝子のうち、その GO タームを含んでいる遺伝子の確率を示している。つまり、そのリストの中のその GO タームの割合が、ゲノム全体においての割合と比較されている。

ここでは 2.8 で得られた樹形図における全てのクラスター(全てのノード以下の葉の集まり)で出芽酵母の全ての遺伝子数(6605 個：SGD からダウンロードしたアノテーションの情報(saccharomyces_cerevisiae.gff)に基づく)をバックグラウンドとして、GO::TermFinder を実行することで、有意に共通の GO タームを持つ遺伝子が含まれるクラスターを探索した。このことにより、出芽酵母の UTR に類似した構造を持つ RNA のクラスターとその UTR を mRNA に含む遺伝子の機能の関連を調べた。

3章　結果

この章では結果を次の 4 つのパラグラフに分けて述べる。

3.1　MXSCARNA によるリアラインメントと RNAz によって２次構造を持つ RNA を網羅的に検出した結果(パイプライン 2.1-2.4)

3.2　RNAz の予測配列をアノテーションデータと比較し、既知機能性 RNA との重複を調べた結果と UTR 候補に検出した数を調べた結果(パイプライン 2.5)

3.3　UTR 候補に検出された RNAz の予測配列において、INFERNAL による探索の結果(パイプライン 2.6)

3.4　UTR 候補に検出された RNAz の予測配列において、LocARNA と GO::TermFinder により有意なクラスターを検出した結果(パイプライン 2.7-2.9)

以下に詳細を述べていく。

３.１　リアラインメントと２次構造を持つ RNA の探索

パイプライン 2.4 で網羅的なゲノム解析に適した機能性 RNA の予測ツールである RNAz を利用して候補領域の絞込みを行うが、この際に入力としてマルチプルアラインメントデータが必要となる。そこで、まず、7 種の酵母(S.cerevisiae, S.paradoxus, S.mikatae, S.kudriavzevii, S.bayanus, S.castellii, S.kluyberi)の MLUTIZ により作成されている既存のマルチプルアラインメントデータを UCSC ゲノムブラウザから得た。そして、このアラインメントデータに含まれる 38417 アラインメントブロックの全てに対して、MXSCARNA を実行してリアラインメントを行い、２次構造を考慮に入れたアラインメントデータを作成した。

次に、作成されたリアラインメントデータにおいて RNAz を実行するために、適当な前処理を行い(詳細は方法に記載)、入力配列となるアラインメントデータを 237451 個得た。また、リアラインメントを行わないアラインメントデータにおける入力配列数は 240767 個であった(表 1)。

表 1. 前処理後のアラインメントデータの入力配列の数

	MXSCARNA※1	MULTIZ※2
入力配列	237451	240767

それぞれのマルチプルアラインメントデータに対して前処理を行うことで得られる入力配列の数を示している。

(※1)MXSCARNA を実行したマルチプルアラインメントデータ

(※2)MULTIZ によって作成されたマルチプルアラインメントデータ(既存)

(※1、※2 は以下の表においても同様である)

前処理によって得られた全ての入力配列に対して、RNAz を実行することにより、進化的に保存されていて熱力学的に安定な 2 次構造を持つ RNA を検出した(表 2)(以下これらを RNAz の予測配列とする)。

このとき、RNAzにおいてP_{SVM}のカットオフ値を 0.5 とするときに、多くの機能性RNA候補の分類の特異性と感受性においてよい性能を示すことが報告されていることから[Washietl *et al.* 2005]、本研究におけるカットオフ値としてP_{SVM} =0.5 を設定した。

そして、RNAz の実行した結果として 6211 個の機能性 RNA 候補配列を検出し、領域が重なる予測配列を結合することにより、4645 個の独立した機能性 RNA 候補領域を得て、これを RNAz の予測領域とした。また、MULTIZ によるアラインメントにおいては、それぞれ 5883 個、4343 個であり、MXSCARNA を実行したアラインメントデータでは検出される予測配列の数が多くなっていた。

表 2. RNAz による予測配列と予測領域の数

	MXSCARNA	MULTIZ
予測配列	6211	5883
予測領域	4645	4343

それぞれのアラインメントデータにおいてRNAzの実行から得られる予測配列の数を示している。予測領域は予測配列が重なる場合にそれらを結合して得た独立した領域の数を示している。

ここで、RNAz の予測領域において、それぞれのアラインメントデータ間での重複している領域の数を調べたところ、2983 領域は重複した配列となっていたことがわかった。そして、MXSCARNA を実行することにより、新たに得られる予測配列は 1662 領域であることがわかった(図 2)。

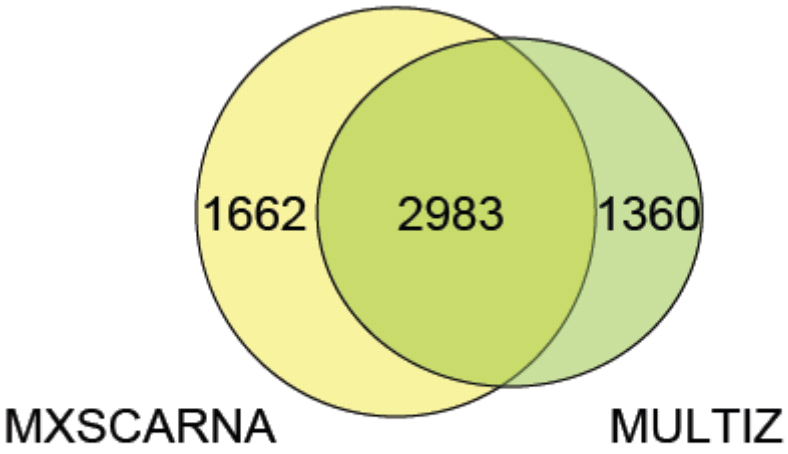


図2
MXSCARNAとMULTIZそれぞれから得られるアラインメントにおけるデータ間の重複を示している。データの重複は30bp以上が重なっている配列対の数を示している。

３.２　RNAz の予測配列の既知機能性 RNA との重複と UTR 候補に含まれる配列の選別

RNAz の予測配列の検出感度を調べるために、予測配列を SGD による既知機能性 RNA のアノテーションとの重複する数を調べた(図 3)(数値は補足資料、補表 1 に記載)。その結果、RNAz で得られた予測配列では、SGD のアノテーションがある機能性 RNA の 474 個のうち 251 個(約 53%)の機能性 RNA において重複があり、特に、特徴的な構造を持つ tRNA ではアノテーション 299 個のうち 180 個(約 60%)を検出できていた。しかし、一方で、出芽酵母の機能性 RNA において大きなファミリーとなっている snoRNA については 77 個のうち 25 個(約 32%)しか検出できていなかった。これは、Washietl らが主にヒトの遺伝子間領域において RNAz を実行した場合においても、とりわけ C/D box 型の snoRNA は 19.5%のみしか検出できなかったことが報告されている[Washietl *et al.* 2005]。このことにより、snoRNA の中には 2 次構造が安定した領域を含まないものも多く存在すると考えられる。また、MULTIZ のアラインメントデータにおいて RNAz を実行したときに検出する数とは少しの違いはあったが、ほぼ同様の数の既知機能性 RNA を検出する結果となっていた(図 3)。

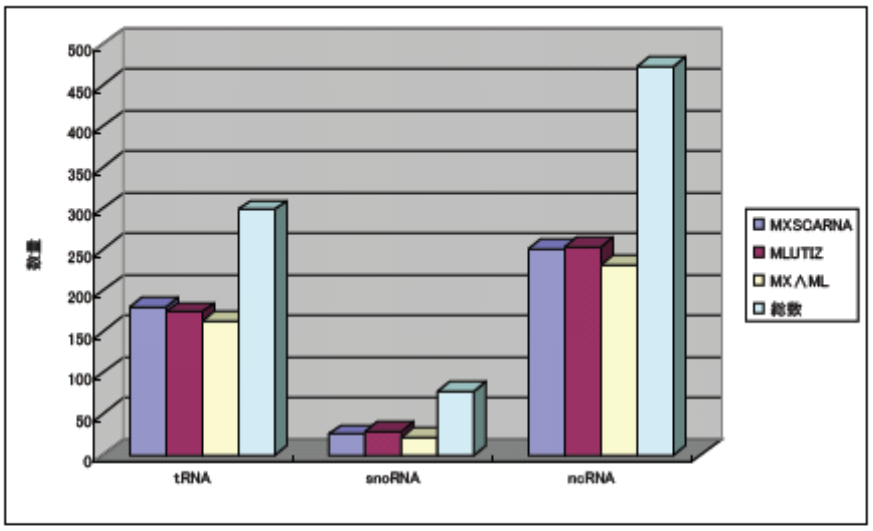


図3
RNAzが予測した領域が、既知機能性RNAに重複している数を図示している。アノテーションデータはSGDからダウンロードし、このとき、アノテーション領域の30%以上が含まれている場合を重複とした。総数はアノテーションデータに含まれる数を表しており、MX∧MLはMXSCARNAとMULTIZのアラインメントデータの両方から検出される数を示している。

また、予測配列が出芽酵母ゲノムの中でどのように位置しているかを、調べるために、既知のアノテーション(CDS、ncRNA、Intron、UTR 候補)に含まれる予測配列の数を調べた(補足資料、補表 2 に記載)。方法でも記したように、ここで、UTR の機能性 RNA 候補配列を調べるために、「UTR 候補」をイントロンに接しない CDS の両側 300bp 以内の領域と定義した。この定義の主な理由としては、多くの遺伝子は転写物の構造が明らかではなく、5'UTR と 3'UTR の領域がわからないことがある。また、先行研究において出芽酵母の UTR の平均の長さが約 260bp、遺伝子間領域の平均の長さが約 530bp であることなどを考慮に入れて決めた[Hurowitz *et al.* 2003]。

そして、RNAz により得られた予測配列は出芽酵母のゲノムにおいて、既知機能性 RNA に対しての重複だけではなく、CDS やイントロン、UTR 候補などのさまざまな領域において存在していることがわかった。ここで、導入や方法で示したいくつかの理由により、非タンパクコード領域の転写物の大きな要素である UTR 候補に 591 個の予測配列が含まれていることに注目した。まず、5'側、3'側の両方で UTR 候補に含まれる予測配列と、それらの予測配列が重なる場合に結合することで得た独立した予測領域を表 3 に示す(表 3)。

表 3 .UTR 候補における RNAzの予測配列と予測領域の数

	MXSCARNA	MULTIZ
5'UTR 候補の予測配列	304	283
3'UTR 候補の予測配列	287	270
5'UTR 候補の予測領域	250	229
3'UTR 候補の予測領域	238	220

UTR 候補はイントロンに接しない CDS の両側 300bp 以内の領域と定義している。アノテーションデータは SGD からダウンロードし、RNAz の予測配列が UTR のそれぞれの領域に含まれている場合を数えており、このとき、既知アノテーションに重複する場合はカウントをしていない。また、予測領域は予測配列が重なる場合にそれらを結合した数を示している。

3.3　UTR の機能性 RNA 候補配列の既知機能性 RNA との相同性検索

Rfam で定義される既知機能性 RNA ファミリーのプロファイルデータ(共分散モデル)を UTR 候補に存在する RNAz の予測配列に対して、INFERNAL のプログラム cmsearch を実行し、配列情報と 2 次構造情報の両方の観点から、既知機能性 RNA と相同性のある領域の数を調べた(表 4)。

Rfam は包括的な機能性 RNA のデータベースであり、機能性 RNA ファミリーを定義し、そのマルチプルアラインメントデータとプロファイルを管理している。また、それぞれのファミリーにおいて、INFERNAL のプログラム cmsearch から得られるスコアに関する閾値が与えられており、そこで定義されている TC(Trusted cutoff)、と NC(Noise cutoff)をカットオフの値として利用した。TC は、Rfam のデータベースにおいて、そのファミリーに含まれているアラインメントの配列の中で最も低いスコアを表しており、NC はそのファミリーのアラインメントには含まれない配列で最も高いスコアを表している。

表 4. INFERNAL による検出結果

	TC	NC
5'UTR 候補 MXSCARNA	8	16
3'UTR 候補 MXSCARNA	24	28
5'UTR 候補 MLUTIZ	8	17
3'UTR 候補 MLUTIZ	22	27

INFERNAL により検出された既知機能性 RNA と相同性のある領域の数を表している。表中の TC と NC のカラムは、それぞれスコアが TC 以上の値を示す領域の数、スコアが NC 以上の値を示す領域の数を表している。カットオフ値として採用した TC、NC は Rfam で定義されている閾値である。

これらの INFERNAL を実行した検出結果には、既知の ncRNA の逆鎖に位置しているものや、5'UTR 候補と 3'UTR 候補で互いに重複して含まれているもの多くあった。これらのことを考慮して、手作業により 10 個の候補を選定し、例として以下のような配列を検出した。その他の候補配列については補足資料の補表 3 に記した。

(i) U7 small nuclear RNA(RF00066)と相同性をもつ領域

U7 small nuclear RNA は動物においてヒストンの pre-mRNA のスプライシングに関与している RNA 分子として知られている。特定された領域はアノテーションデータの出芽酵母のゲノムにおいて chr14 の+鎖の 642411-642471 に位置する領域であり、INFERNAL のプログラム cmsearch によるスコアは 21.30 であった。これは、TC(20.52)の値を上回っているため Rfam の定義においては信頼できるスコアとされている。この領域から予測される 2 次構造を図 4 に示し、そのアラインメントデータを図 5 に示している。

この領域は遺伝子 YNR008W と YNR009W の間の領域に位置しており、

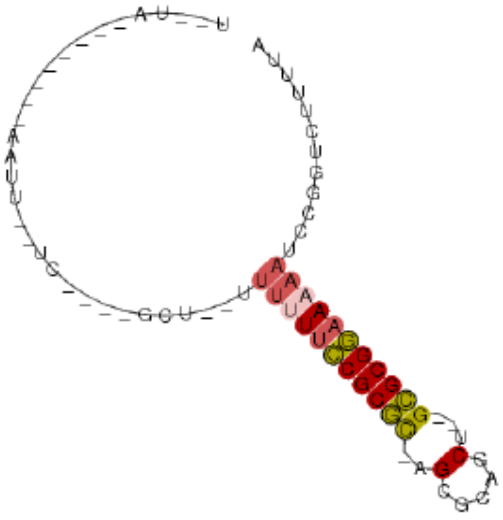


図4
U7 small nuclear RNA(RF00066)と相同性をもつ候補領域から予測された2次構造を示している。また、色づけられている箇所は2次構造の安定性に関与する塩基対を示している。色の違いについては図5において説明する。

この遺伝子間が 300bp 程度と近い領域であったことより、YNR008W の 3'UTR 候補と YNR009W の 5'UTR 候補の両方から検出されていた。また、MXSCARNA、MULTIZ のそれぞれのアラインメントデータにおいて検出された。

この領域がこの機能性 RNA ファミリーの特徴を有しているとすると、遺伝子の UTR において機能しているものとは考えにくく、独自の転写物として機能しているものと考えられる。cDNA 解析から得られている転写物(Y041_L18_F.ab1)はこの事実を示唆する結果を示している。この転写物は YNR008W にアノテートされているが、YNR008W の CDS の中に TSS を持ち、上記の領域を含んでいる転写物となっている。このことより、この転写物は YNR008W から得られるタンパク質をコードしているわけではなく、ここで示した機能性 RNA として働くための転写物である可能性が考えられる。

図 6 には、既知の機能性 RNA(ウニで発見されている 4 種)とのアラインメントを行った結果を示している(ウニの配列は目視での判断で相同性が比較的高いことから選んでいる)。

この結果、配列が保存されている箇所が 21 カラム(65 カラム中)、共通の塩基対が保存されている箇所が 9 ステム(13 ステム中)、さらにこれら機能性 RNA(ウニで発見されている 4 種)が mRNA との結合する際に重要な配列で保存されている箇所が 5 塩基(9 塩基中)で特定された。

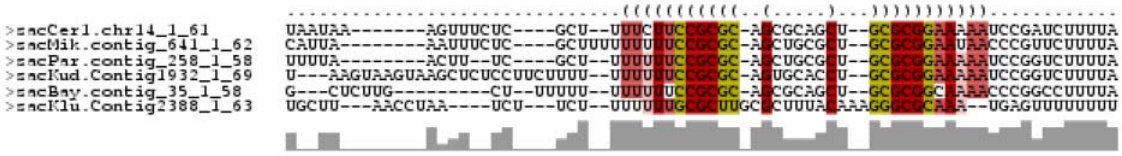


図5
U7 small nuclear RNA(RF00066)と相同性をもつ候補領域の近縁種とのアラインメントデータを示している。色づけられている箇所は2次構造の安定性に関与する塩基対を示している。色の違いは、2次構造の安定性に関与する塩基対のペアの種類を示しており(赤:1種類、黄:2種類、緑:3種類)、その透明度は塩基対が成立しない数を示す(0から2で順に薄くなっている)。(これは、図4,5,7,8において同様である)
アラインメントの上記に示す"("と")"は塩基対を形成することを示す。またアラインメントの下のグレーの棒は配列の保存度をその高さで表している。(これは図8で同様である)

図6
U7 small nuclear RNA(RF00066)と相同性をもつ候補領域(最上段)と既知の機能性RNA(ウニで発見されている4種:2段目から5段目)アラインメントデータを示している。
色づけられている箇所は2次構造の安定性に関与する塩基対を示している。Folding structureの行の"<"と">"は塩基対を形成することを示す。またAlignmentの行の"＊"は配列が保存が共通していることを示す。
(これらは図9でも同様である)

（ⅱ） Gurken localization signal(RF00626)と相同性を持つ領域

Gurken localization signal は Drosophila において多くの種で保存されている RNA を制御する因子であり、コード領域内でループを構成し、ダイニンが介在する RNA 輸送のシグナルとして役割を果たしている。

図7
Gurken localization signal(RF00626)と相同性を持つ候補領域から予測された2次構造を示している。

このファミリーと相同性を持つ領域は 2 箇所で特定され、アノテーションデータの出芽酵母ゲノムにおいて、chr10 の－鎖の 486311-486459 に位置する領域と、chr13 の+鎖の 158490-158633 に位置する領域であった。INFERNAL のプログラム cmsearch のスコアはそれぞれ、13.37 と 11.69 であり NC(10.21)以上の値となっていた。この 2 領域のうち後者の領域のから予測される 2 次構造を図 7 に示し、そのアラインメントデータを図 8 に示している。また図 9 には、既知の機能性 RNA(ハエで発見されている 2 種)とのアラインメントを行った結果を示している。(このファミリー(RF00626)には、この 2 配列のみ含まれている)

配列や塩基対において既知機能性 RNA との相同性はあまり見られていないが、近縁種では比較的長い領域で構造が保存されている。また、これらがゲノムの UTR 候補で発見されたことを考えると、同様の機能を有している可能性が考えられる。

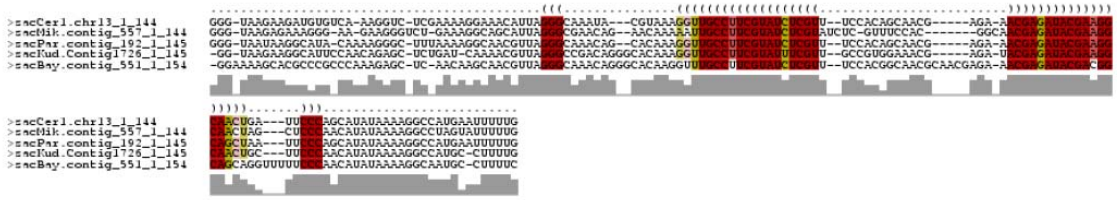


図8
Gurken localization signal(RF00626)と相同性をもつ候補領域の近縁種とのアラインメントデータを示している。
色のついたカラムとアラインメントの下のグレーの棒については図5の説明に記した。

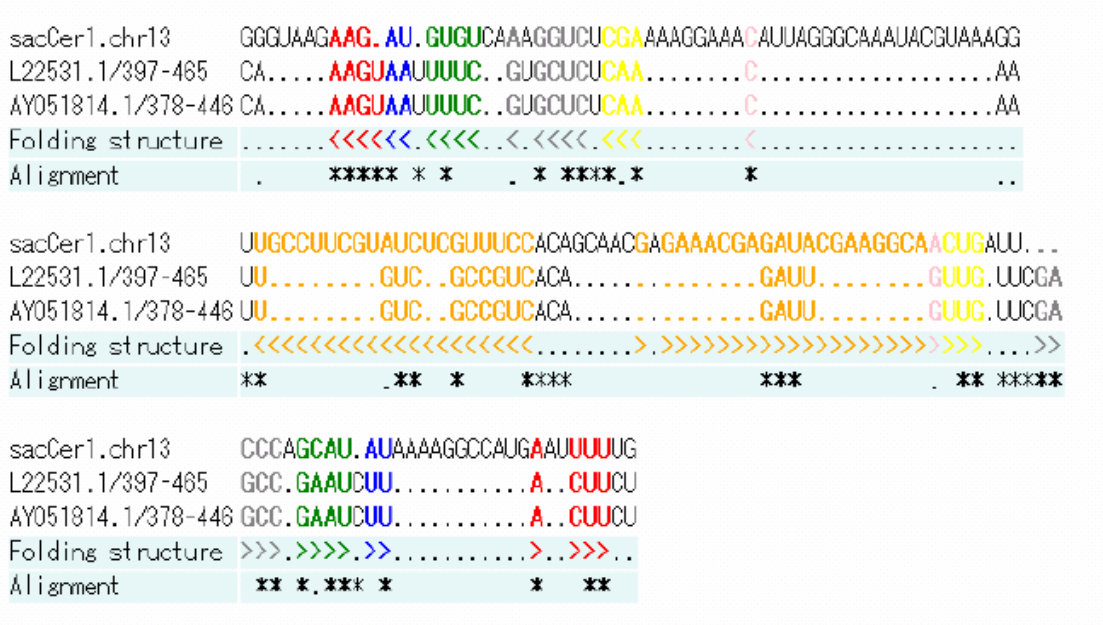


図9
Gurken localization signal(RF00626) と相同性をもつ候補領域(最上段)と既知の機能性RNA(ハエで発見されている2種:2段目、3段目)アラインメントデータを示している。

3.4　UTR 候補に存在する RNAz の予測領域の有意なクラスターの検出

出芽酵母の 3'UTR には特定の配列のモチーフがあり、それらの中には特定の機能に関連するモチーフがあることが先行研究により示されている[Shalgi *et al.* 2005]。このことから、それらの配列モチーフには特定の構造モチーフがあることが考えられるうえ、さらに、配列モチーフではなくとも構造モチーフを持ち、特定の機能に関連しているグループの存在が考えられる。

そこで、UTR 候補に存在する RNAz の予測領域に対して、LocARNA を用いて構造類似性を評価し、その結果として得られるスコアに基づいて WPGMA (weighted pair-group method algorithm)によりクラスタリングを行った。

INFERNAL による解析から、CDS から遠い位置にある予測領域では、既知の ncRNA の逆鎖に位置する領域や、5'UTR と 3'UTR で互いに重複して含まれているものなどのノイズが多く含まれていたことがわかった。よって、UTR 候補に存在した RNAz の予測領域のうち、CDS との境界から 100bp 以内に重複する領域を今回の解析の対象とした(表 5)。

表 5.解析の対象とした予測領域の数

	MXSCARNA	MULTIZ
5'UTR 候補	98	91
3'UTR 候補	98	102

UTR 候補に存在した予測領域のうち CDS との境界から 100bp 以内に重複がある領域の数を示している。

これらの配列それぞれにおいて、McCaskill のアルゴリズムにより塩基対確率行列を計算し、それらを入力として全ての 2 領域ずつの組み合わせにおいて LocARNA を実行することで、構造類似性のスコアを算出した。それらのスコアをもとに距離行列を作成し、WPGMA でクラスタリングした結果から樹形図を作成した。

例として MXSCARNA によるアラインメントデータにおける 3'UTR の予測領域のクラスタリングから得られた樹形図の一部を図 10 に示した。木においてそれぞれの葉は RNAz による予測領域の中で UTR に存在する領域を示している。また、得られた全ての樹形図と、それぞれの樹形図の葉に割り当てられた ID に対応する領域は補足資料(補図 1、補表 4-5)に記載している。

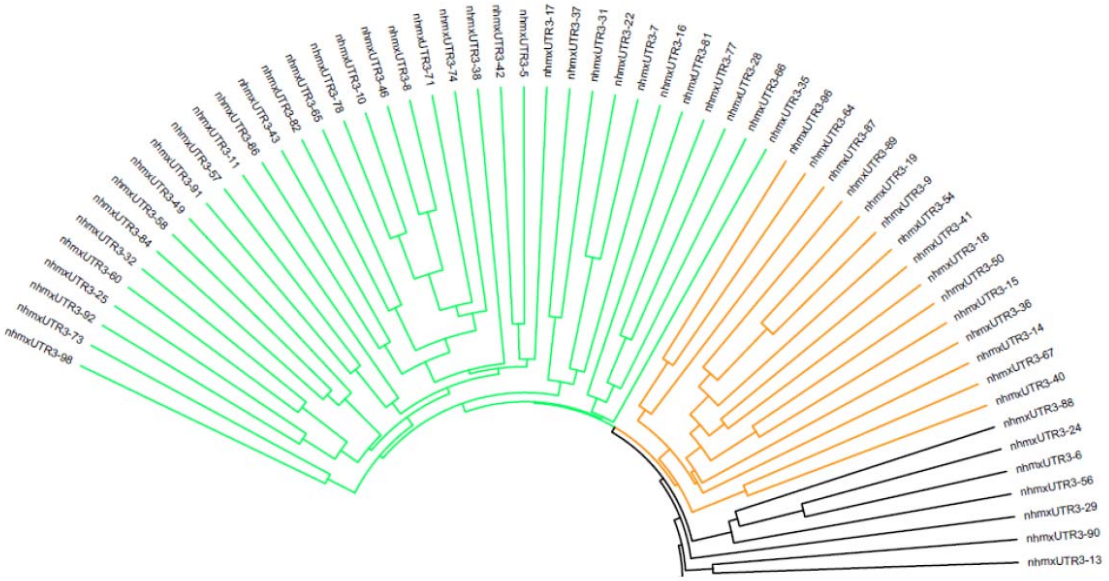


図10
MXSCARNAを実行したアラインメントデータの3’UTR候補に存在したRNAzの予測領域を
WPGMAでクラスタリングした結果の系統樹の一部を示している。
また、全体の系統樹と配列のIDと領域の位置については補足資料に示している。
緑色：mx3_cluster35,GO:0005830(cytosolic ribosome)において有意なクラスター
橙色：mx3_cluster49,GO:0031090(organelle membrane)において有意なクラスター

また、構造類似性クラスタリングと同様に、候補配列間の距離行列を BLAST のスコアに基づいて作成し、配列類似性に基づくクラスタリングも行った。ごく少数の配列の組み合わせにおいて、かなり高い類似性があり、これらは部分的な配列の重複の影響と考えられる。しかし、それら以外の配列群においては、非常に短い配列において類似性があるのみであった。これらのことより、これらの領域の配列では配列類似性はあまりないものと考えられる。

ここで、LocARNA に基づく構造類似性クラスタリングの結果の樹形図から有用な意味を導くために、得られた樹形図の全てのクラスター(全てのノード以下の葉の集まり)において、そのクラスターに含まれる領域に隣接した遺伝子のリストを作成した。そして、それを入力として GO::TermFinder を実行し、それぞれのクラスター内で有意な偏りを持つ GO タームを探索した。
GO::TermFinder の実行結果から 5’UTR 候補においては 34 個、3’UTR 候補では 34 個の GO タームが、特定のクラスターにおいて有意性のある GO タームとして検出された(表 6)。また、MULTIZ のアラインメントからはそれぞれ 34 個と 25 個であった。このとき、複数のクラスターで重複して現れる GO タームについては、P-value が最も低い場合のクラスターをその GO タームの有意なクラスターとした。

例として、図 10 に MXSCARNA によるアラインメントデータにおける 3'UTR 候補の予測領域のクラスタリングから得られた樹形図の中に含まれる有意な偏りを持つクラスターのうち、mx3_cluster35(緑色)と mx3_cluster49(橙色)を示している。これらのクラスターの遺伝子のリストには、それぞれ GO:0005830(cytosolic ribosome)、GO:0031090(organelle membrane)の GO タームに含まれる遺伝子が有意に多く含まれていた。

表 6.有意な GO タームの検出数

	MXSCARNA	MULTIZ
5'UTR 候補	34	34
3'UTR 候補	34	25

MXSCARNA、MULTIZ のそれぞれのアラインメントデータにおける 5'UTR 候補と 3'UTR 候補の予測領域のクラスターにおいて、有意な P-value を示した GO タームの数を示している。

有意と判断された GO タームの多くは、2、3 個程度の遺伝子による小さなクラスターとなっており、また、部分的な重複などの影響から検出されているものも多いと考えられる。しかし、共通の GO タームをもつ複数の遺伝子の UTR 候補の予測領域が、構造クラスタリングにより比較的近い位置に存在している例もいくつかあった。これらはそれらの配列において構造の類似性があり、その構造が機能に関係している可能性を示唆している。
有意な GO タームのうち、そのクラスターにおいて 5 個以上の遺伝子で共通の GO タームを持つものについて表 7 に示した。これらのクラスターに含まれる遺伝子と、4 個以下の遺伝子で共通の GO タームを持つクラスターを含めた表は補足資料の補表 6-9 に記載した。

表 7. 有意な GO タームを検出したクラスター

クラスター名	GO ターム	P-value	関連遺伝子数	総遺伝子数
mx5_cluster19	GO:0016070	0.043416	5	7
mx5_cluster68	GO:0006800	0.017271	5	51
mx5_cluster68	GO:0006979	0.016036	5	51
mx5_cluster68	GO:0050791	0.045315	14	51
mx3_cluster35	GO:0005830	0.041710	5	29
mx3_cluster49	GO:0031090	0.017712	6	13
ML5_cluster68	GO:0006351	0.037388	12	52
ML5_cluster68	GO:0006355	0.006787	11	52
ML5_cluster68	GO:0006357	0.026687	8	52
ML5_cluster68	GO:0019219	0.031635	11	52
ML5_cluster68	GO:0032774	0.041181	12	52
ML5_cluster68	GO:0045449	0.012382	11	52

有意に検出された GO タームのうち、そのクラスターにおいて 5 個以上の遺伝子で共通の GO タームを持つものを示している。関連遺伝子数は、それぞれのクラスターにおいて、有意な GO タームのアノテーションのあった遺伝子の数を示す。総遺伝子数は、それぞれのクラスターにおいてアノテーションのある遺伝子の数を示している。各クラスターに含まれる予測領域については補足資料(表 7 の補足)に示している。クラスター名のアンダーバー以前は、アラインメントのデータと UTR の領域の場所について示しており、mx5,mx3,ML5 はそれぞれ、MXSCARNA のデータにおける 5'UTR の予測領域の樹形図から得られるクラスター、MXSCARNA のデータにおける 3'UTR の予測領域の樹形図から得られるクラスター、MULTIZ のデータにおける 5'UTR の予測領域の樹形図から得られるクラスターを示している。

4 章　議論

マルチプルアラインメントデータは進化的な情報が得られることから、ゲノム解析において非常に有用なデータであり、RNAz による網羅的な機能性 RNA の探索の際に必要な入力データになる。出芽酵母が含まれるデータには 7 種の酵母のマルチプルアラインメントデータが存在するが、このデータは BLASTZ で作成されたアラインメントのセットをもとに MULTIZ により作成されているため、一次配列のみに基づいたデータであるという問題点がある。よって、2 次構造を考慮したアラインメントデータを作成することで、機能性 RNA の探索においてより信頼できる予測ができると考えた。

本研究では、MXSCARNA を用いてリアラインメントを行い、２次構造を考慮したアラインメントデータを得ることにより、RNAz による機能性 RNA 候補の予測において、それをしないときのアラインメントデータによる予測とは異なった予測配列が 1662 領域において検出されていた。1 次配列のみに基づいた MULTIZ によるアラインメントデータでは検出できなかった領域が多く存在していたことは、2 次構造を考慮したアラインメントデータの必要性を示唆している。また、既知の機能性 RNA を検出するという観点においても、MULTIZ によるアラインメントでは検出できなかった機能性 RNA で、MXSCARNA を実行したアラインメントデータからは検出されているものが 20 個あった。このことも、２次構造を考慮したアラインメントデータを利用する有用性を示している点といえる。

一方で、それぞれのアラインメントデータから共通に得られる結果も非常に多く存在していた(予測配列では 2983 領域、既知機能性 RNA との重複では 231 個)。この主な理由としては、MULTIZ により作成された既知のデータを利用して、MXSCARNA でリアラインメントしているために、アラインメントを修正できる部分が限られていたことがある。解決策の一つとして、MULTIZ のデータをリアラインメントするのではなく、7 種のそれぞれの酵母のゲノムにおいて２次構造を考慮したアラインメントを行ったうえでマルチプルアラインメントデータをつくることが挙げられる。これを行うことにより、2 次構造が関連する領域において、より正確性の高いアラインメントができることが考えられる。そして、それから得られたマルチプルアラインメントデータを利用し、RNAz による機能性 RNA の予測を行うことで検出感度が向上すると期待できる。

次に、Washietlらの評価では、RNAzは$P_{SVM}=0.5$ において 84%の検出感度と報告されているにもかかわらず[Washietl *et al.* 2005]、今回の実験では既知アノテーションとの重複が 50%程度しかなかったことを考える。この原因の最も大きな理由として、入力とした 7 種の酵母のマルチプルアラインメントデータの正確性の問題が挙げられる。RNAzは入力となるマルチプルアラインメントデータが正確であることが前提となっているため、アラインメントデータの正確性がRNAzの予測結果に影響するからである。

先行研究において、17 種の脊椎動物の全ゲノムから MULTIZ で作成されたマルチプルアラインメントデータの機能性 RNA 領域についての正確性の評価が行われている[Wang *et al.* 2007]。この研究では、ヒトにおいて既知の機能性 RNA に対してアラインされている全ての配列で、同じ機能性 RNA が正確に整列されている割合はわずか 29.4%しかないことが示されている。70.6%の不完全なアラインメントには、ヒトの既知機能性 RNA に対して、部分的な配列のみがアラインメントされている場合や、異なる領域の複数の断片がアラインメントされている場合、あるいは、機能性 RNA ではない配列がアラインメントされている場合などがあることが示されており、このようにミスアラインメントされている配列は RNAz による予測において偽陰性を導く要因となる。
このことは本研究で利用した出芽酵母のマルチプルアラインメントデータにおいても、同様の問題があると考えられ、アラインメントデータには不正確な箇所が多数含まれていると考えられる。さらに、マルチプルアラインメントの対象となっている酵母 7 種のうち 6 種はゲノム上の位置がはっきり特定されていない断片配列のデータが利用されているため、アラインメントより前の問題として、配列データが不足している領域が多く存在している可能性も考えられ、これらも偽陰性を導いているかもしれない。
MXSCARNA を実行したデータにおいても MULTIZ により作成されたデータをもとにリアラインメントしているため、誤ってアラインされている配列は取り除くことはできず、同様の問題点が含まれている。既に上記した、7 種のそれぞれの酵母のゲノムにおいて、2 次構造を考慮したアラインメントを行ってマルチプルアラインメントデータをつくることはこれらの問題点を改善できる可能性がある。

UTR 候補に 591 個(5'側 304 個、3'側 287 個)の RNAz による予測配列が含まれていたことはとても興味深いことである。この理由としては、UTR は非たんぱく質コード領域の転写物の大きな要素であり、この UTR が特定の mRNA において転写後に重要な役割を果たしていることがいくつかの例で既に知られているということがある。また、出芽酵母ゲノムの約 12%は mRNA の UTR の配列であり、ゲノムにおいて多くの部分を占めていることが先行研究により示されていることから、機能を持つ UTR の配列が今後さらに多く発見されていく可能性が示唆されている[Batey *et al.* 2006]。
本研究において、RNAz から予測された機能性 RNA 候補のうち UTR 候補に含まれるものに対して INFERNAL のプログラム cmsearch によって既知機能性 RNA との相同性がある 44 配列が検出された。これらのなかには、U7 small nuclear RNA(RF00066)ファミリーに類似したように独立した機能性 RNA となりそうな領域も含まれていたが、Gurken localization signal(RF00626)ファミリーや Antizyme RNA frameshifting stimulation element (RF00381)ファミリーなど、他の生物種において、シスに機能する制御配列(cis-regulatory element)としての機能が特定されているファミリーと相同性がある領域も検出された。これらの検出された領域は、RNAz により熱力学的に安定で進化的に保存され

ている 2 次構造と予測されているうえに、INFERNAL により既知機能性と相同性がある領域と予測されている。このことから、相同性のあったファミリーと同様の役割を果たしていることや、あるいは出芽酵母において、その構造が独自の機能を有している可能性が考えられ、今後、実験面などでより詳細な解析を行うことで新たな発見が期待される。
また、RNAz から予測された機能性 RNA 候補のうち UTR 候補に含まれる配列において構造類似性に基づいてクラスタリングを行い、得られた全てのクラスターにおいて GO::TermFinder を実行することによって、68 個のクラスターが有意に特定の GO タームとの関わりを持っていたことがわかった。これらは、UTR に存在する類似した 2 次構造の RNA が特定の機能を持つ遺伝子において、共通の役割を持つ可能性を示唆するものである。結果における有意義な例としては、MXSCARNA によるアラインメントデータの 3'UTR に存在した予測配列におけるクラスターの mx3_cluster35 がある。このクラスターに含まれ GO のアノテーションがある 29 個の遺伝子のうち 5 個が GO:0005830（cytosolic ribosome）という共通の GO タームを有意な確率で持っていた。この GO ターム GO:0005830 では、3'UTR が有意に長いことがタイリングアレイ解析において報告されており[David *et al.* 2006]、この長い UTR は特定の機能を制御するための構造モチーフを持つために必要な領域だということが推測できる。
原核生物の 5'UTR においては、22 個の構造モチーフの候補が存在することが先行研究で示されている[Weinberg *et al.* 2007]。この研究では、相同性のある遺伝子における 5'UTR について、CMfinder[Yao *et al.* 2006]を用いてその領域に存在する構造を持つ RNA を発見するための情報科学的なパイプラインを提案している。マルチプルアラインメントを必要としない点などで、本研究とは異なるアプローチではあるが、本研究と同様に UTR の 2 次構造に着目した解析を行っている。この成果も本研究同様 UTR における構造が機能に関係している例を示しており、まだ機能のわかっていない多くの機能性 RNA の可能性を示している。また、この手法を出芽酵母やその他の生物にも適用することで、さらに別のいくつかのモチーフが発見されることも予想される。

本研究における新規機能性 RNA の探索プロセスにおいては今後さまざまな改善が考えられる。具体的には、配列情報の補填範囲の拡大や精度の向上、アラインメントの精度の向上、予測配列の特定の精度の向上、2 次構造類似度測定の精度の向上などである。今後、より多くの側面でより精度の高いデータを取得し、また、性能の良いツールが開発されることで、本手法の適用において、より意味のある結果が得られることが期待できる。また、本手法を他生物種のアラインメントデータに適用することで、より多くの生物学的な知見が得ることも期待できる。

5 章　結論

7 種の酵母のゲノムにおいて、2 次構造を考慮したアラインメントデータを作成し、それを入力として、出芽酵母のゲノム上に進化的に保存された 2 次構造を持つ RNA を UTR の可能性がある領域に 591 配列を特定した。

それらの RNAz において機能性 RNA の候補となった配列において、既知機能性 RNA ファミリーとの相同性に基づく解析を行い、44 配列を検出し、手作業により 10 個の候補を選定した。これらの配列の中には、他の生物種においてシスに機能する制御配列(cis-regulatory element)としての機能が特定されているファミリーと相同性がある領域も検出されており、相同性のあったファミリーと同様の役割を果たしていることや、あるいは出芽酵母において、その構造が独自の機能を有している可能性が考えられる。また、それらの配列において構造類似性に基づいたクラスタリングを行い、得られるクラスターの中で、そのクラスターに含まれる領域を UTR に持つ遺伝子が有意な数で共通の GO タームを持つ 68 個のクラスターを特定した。これらのクラスターの中には、先行研究による結果と関連が導かれたクラスターもあり、UTR に存在する類似した RNA の 2 次構造が、特定の機能（GO ターム）を持つ遺伝子において共通の役割を持つ可能性を示唆している。これらの配列は出芽酵母の UTR において新規機能性 RNA の存在の可能性を示唆している。

今後、配列情報の補填範囲の拡大や精度の向上、アラインメントの精度の向上、予測配列の特定の精度向上、2 次構造類似度測定の精度向上など、データやツールにおいて改善が進むことで、本手法の精度も向上すると考えられる。また、手法を他生物種のアラインメントデータに適用することでも、より多くの生物学的な知見が得ることが期待できる。

謝辞

本研究を進めるにあたり、常に暖かいご指導をして頂き、素晴らしい研究の場を提供してくださった浅井潔教授に深く感謝いたします。また、セミナーや日常生活の多くの場面で、金大真さん、木立尚孝さん、寺井悟朗さん、浜田道昭さん、佐藤健吾さんには研究を進めるうえで貴重なアドバイスを多くして頂きました。さらに、産業技術総合研究所生命情報工学研究センターの皆様には、大学院生活の中でさまざまな点でお世話になりました。そして、浅井研究室の加藤毅先生、岡田欣也さん、田部井靖生さん、芦田広樹君、八杉直樹君は研究のみならず日々の生活を非常に有意義なものとして頂きました。
多くの方々のご指導、ご協力等があったおかげで、この 2 年間の修士課程において非常に充実した大学院生活を送ることができました。本当にありがとうございました。

参考文献

Batey RT.
Structures of regulatory elements in mRNAs.
Curr Opin Struct Biol. 2006 Jun;16(3):299-306.

Bashirullah A, Cooperstock RL, Lipshitz HD.
Spatial and temporal control of RNA stability.
Proc Natl Acad Sci U S A. 2001 Jun 19;98(13):7025-8.

David L, Huber W, Granovskaia M, Toedling J, *et al.*
A high-resolution map of transcription in the yeast genome.
Proc Natl Acad Sci U S A. 2006 Apr 4;103(14):5320-5.

Eddy SR.
A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.
BMC Bioinformatics. 2002 Jul 2;3:18

Eddy SR.
Non-coding RNA genes and the modern RNA world.
Nat Rev Genet. 2001 Dec;2(12):919-29.

Ghaemmaghami S, Huh WK, Bower K, Howson RW, *et al.*
Global analysis of protein expression in yeast.
Nature. 2003 Oct 16;425(6959):737-41.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S,
Functional profiling of the Saccharomyces cerevisiae genome.
Nature. 2002 Jul 25;418(6896):387-91.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, *et al.*
Life with 6000 genes.
Science. 1996 Oct 25;274(5287):546, 563-7

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A.
Rfam: annotating non-coding RNAs in complete genomes.
Nucleic Acids Res. 2005 Jan 1;33(Database issue):D121-4.

Hurowitz EH, Brown PO.
Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae.
Genome Biol. 2003;5(1):R2.

Jansen RP.
mRNA localization: message on the move.
Nat Rev Mol Cell Biol. 2001 Apr;2(4):247-56.

Juneau K, Palm C, Miranda M, Davis RW.
High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing.
Proc Natl Acad Sci U S A. 2007 Jan 30;104(5):1522-7.

Kiss T.
Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions.
Cell. 2002 Apr 19;109(2):145-8. Review.

Kuersten S, Goodwin EB.
The power of the 3' UTR: translational control and development.
Nat Rev Genet. 2003 Aug;4(8):626-37. Review.

Lowe TM, Eddy SR.
A computational screen for methylation guide snoRNAs in yeast.
Science. 1999 Feb 19;283(5405):1168-71

Lowe TM, Eddy SR.
tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.
Nucleic Acids Res. 1997 Mar 1;25(5):955-64

McCutcheon JP, Eddy SR.
Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics.
Nucleic Acids Res. 2003 Jul 15;31(14):4119-28

Mignone F, Gissi C, Liuni S, Pesole G.
Untranslated regions of mRNAs.
Genome Biol. 2002;3(3):REVIEWS0004.

Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T.
A large-scale full-length cDNA analysis to explore the budding yeast transcriptome.
Proc Natl Acad Sci U S A. 2006 Nov 21;103(47):17846-51.

Pasquinelli AE, Ruvkun G.
Control of developmental timing by micrornas and their targets.
Annu Rev Cell Dev Biol. 2002;18:495-513. Epub 2002 Apr 2. Review.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, *et al.*
Identification and classification of conserved RNA secondary structures in the human genome.
PLoS Comput Biol. 2006 Apr;2(4):e33.

Pena-Castillo L, Hughes TR.
Why are there still over 1000 uncharacterized yeast genes?
Genetics. 2007 May;176(1):7-14.

Rivas E, Eddy SR.
Noncoding RNA gene detection using comparative sequence analysis.
BMC Bioinformatics. 2001;2:8.

Shalgi R, Lapidot M, Shamir R, Pilpel Y.
A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs.
Genome Biol. 2005;6(10):R86.

Steigele S, Huber W, Stocsits C, Stadler PF, Nieselt K.
Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions.
BMC Biol. 2007 Jun 18;5:25.

Torarinsson E, Yao Z, Wiklund ED, Bramsen JB *et al.*
Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions.
Genome Res. 2008 Feb;18(2):242-51. Epub 2007 Dec 20.

Storz G.
An expanding universe of noncoding RNAs.
Science. 2002 May 17;296(5571):1260-3

Tabei Y, Kiryu H, Kin T, Asai K.
A fast structural multiple alignment method for long RNA sequences.
BMC Bioinformatics. 2008 Jan 23;9(1):33

van der Velden AW, Thomas AA.
The role of the 5' untranslated region of an mRNA in translation regulation during development.
Int J Biochem Cell Biol. 1999 Jan;31(1):87-106

Wang AX, Ruzzo WL, Tompa M.
How accurately is ncRNA aligned within whole-genome multiple alignments?
BMC Bioinformatics. 2007 Oct 26;8(1):417

Washietl S, Hofacker IL, Stadler PF.
Fast and reliable prediction of noncoding RNAs.
Proc Natl Acad Sci U S A. 2005 Feb 15;102(7):2454-9.

Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, *et al.*
Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.
Nucleic Acids Res. 2007;35(14):4809-19.

Wilkie GS,Dickson KS,Gray NK.
Regulation of mRNA translation by 5'- and 3'-UTR-binding factors.
Trends Biochem Sci. 2003 Apr;28(4):182-8.

Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R.
Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering.
PLoS Comput Biol. 2007 Apr 13;3(4):e65.

Yao Z, Weinberg Z, Ruzzo WL.
CMfinder--a covariance model based RNA motif finding algorithm.
Bioinformatics. 2006 Feb 15;22(4):445-52.

補足資料

補表 1.予測配列が重複する既知機能性 RNA 数

	MXSCARNA	MLUTIZ	MX∧ML	総数
tRNA	180	174	164	299
snoRNA	25	29	22	77
ncRNA	251	254	231	474

RNAz による予測配列が重複する既知機能性 RNA の数を示している。

アノテーションデータは SGD からダウンロードし、このとき、アノテーション領域の 30%以上が含まれている場合を重複とした。

総数はアノテーションデータに含まれる数を表しており、MX∧ML は MXSCARNA と MULTIZ のアラインメントデータの両方から検出される数を示している。

補表 2.各アノテーションに含まれる RNAz の予測配列の数

アノテーション名	MXSCARNA	MULTIZ
All	6211	5883
CDS	1955	1853
ncRNA	318	311
Intron	46	33
5'UTR 候補	304	283
3'UTR 候補	287	270

アノテーションデータは SGD からダウンロードし、RNAz の予測配列が既知のアノテーションに含まれている場合を数えている。UTR 候補はイントロンに接しない CDS の両側 300bp 以内の領域と定義している。

補表 3.INFERNAL から得られる候補領域

ID	ファミリー名	Chr	鎖	開始点	終端点	関連遺伝子	cDNA
RF00066	U7 small nuclear RNA	14	+	642411	642471	YNR009W,YNR008W	○
RF00194	Rubella virus 3' cis-acting element	7	−	384723	384879	YGL063C-A	？
RF00196	Alfalfa mosaic virus RNA 1 5' UTR stem-loop	10	−	486311	486459	YJR030C,YJR031C	？
RF00363	mir-BART1 microRNA precursor family	10	−	486311	486459	YJR030C,YJR031C	？
RF00365	mir-BHRF1-1 microRNA precursor family	13	+	158490	158633	YML058W-A	○
RF00366	mir-BHRF1-2 microRNA precursor family	13	+	158490	158633	YML058W-A	○
RF00381	Antizyme RNA frameshifting stimulation element	16	−	156265	156415	YPL210C,YPL209C	？
RF00385	Infectious bronchitis virus D-RNA	7	−	1061607	1061735	YGR284,YGR285C	○
RF00626	Gurken localisation signal	10	−	486311	486459	YJR030C,YJR031C	？
RF00626	Gurken localisation signal	13	+	158490	158633	YML058W-A	○

INFERNAL で検出された候補領域を手作業によって、候補を選定した結果を示している。ID は Rfam で定義されているファミリーの ID のことを示し、chr は染色体の番号を示す。関連遺伝子は UTR 候補を作成するために利用された CDS の遺伝子である（2 つある場合は 5'UTR、3'UTR のそれぞれの領域において検出されていることになる）。cDNA は cDNA：転写物の情報の有無を示す。[○]重複がある場合；[?]その遺伝子の転写物がない場合；[×]その遺伝子の転写情報 はあるが重複がない場合をそれぞれ示す。RF00381 のファミリーと相同性のある領域は MULTIZ によるアラインメントデータからしか検出されなかったが、残りの領域は MULTIZ と MXSCARNA の両方から検出されていた。また、RF00066 と相同性のあった領域のみが TC 以上のスコアであり、その他は NC 以上の領域である。

補図 1

UTR 候補に存在する予測領域のクラスタリングから得られる樹形図を示している。葉に示した ID に対応する領域は補表 3 と補表 4 に示している。

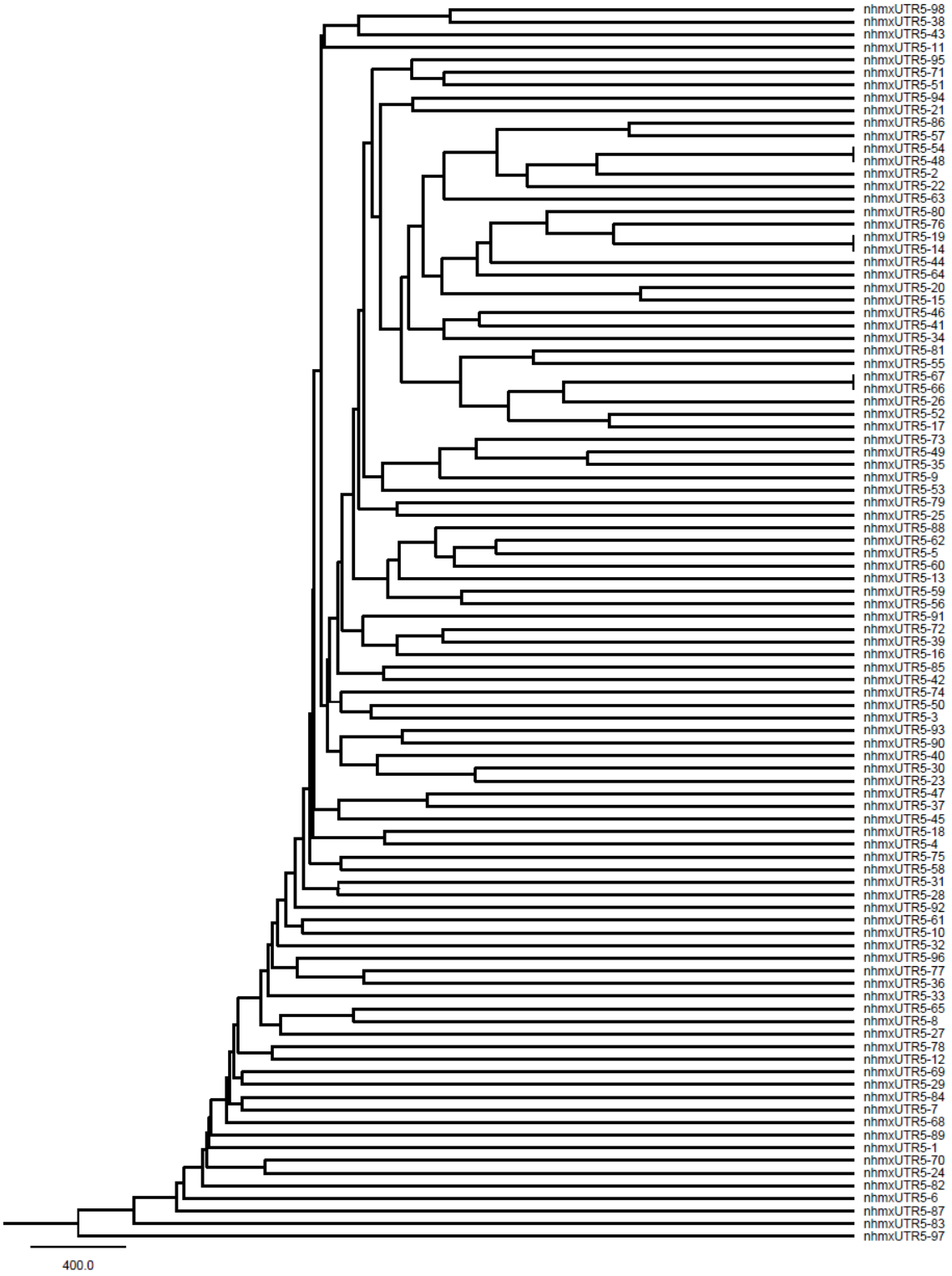
①MXSCARNA によるアラインメントデータの 5'UTR 候補に存在する予測領域

②MXSCARNA によるアラインメントデータの 3'UTR 候補に存在する予測領域

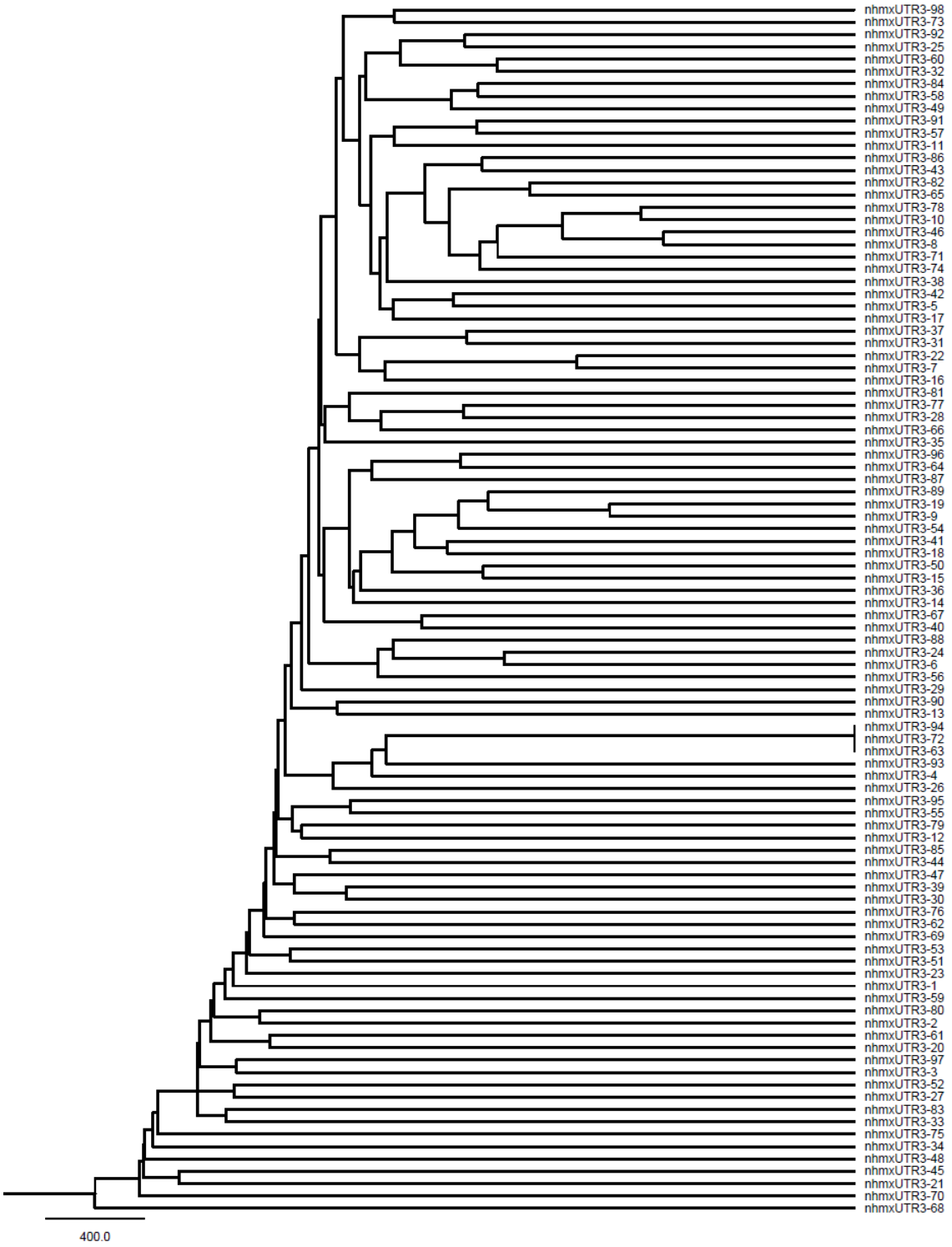
③MULTIZ によるアラインメントデータの 5'UTR 候補に存在する予測領域

④MULTIZ によるアラインメントデータの 3'UTR 候補に存在する予測領域

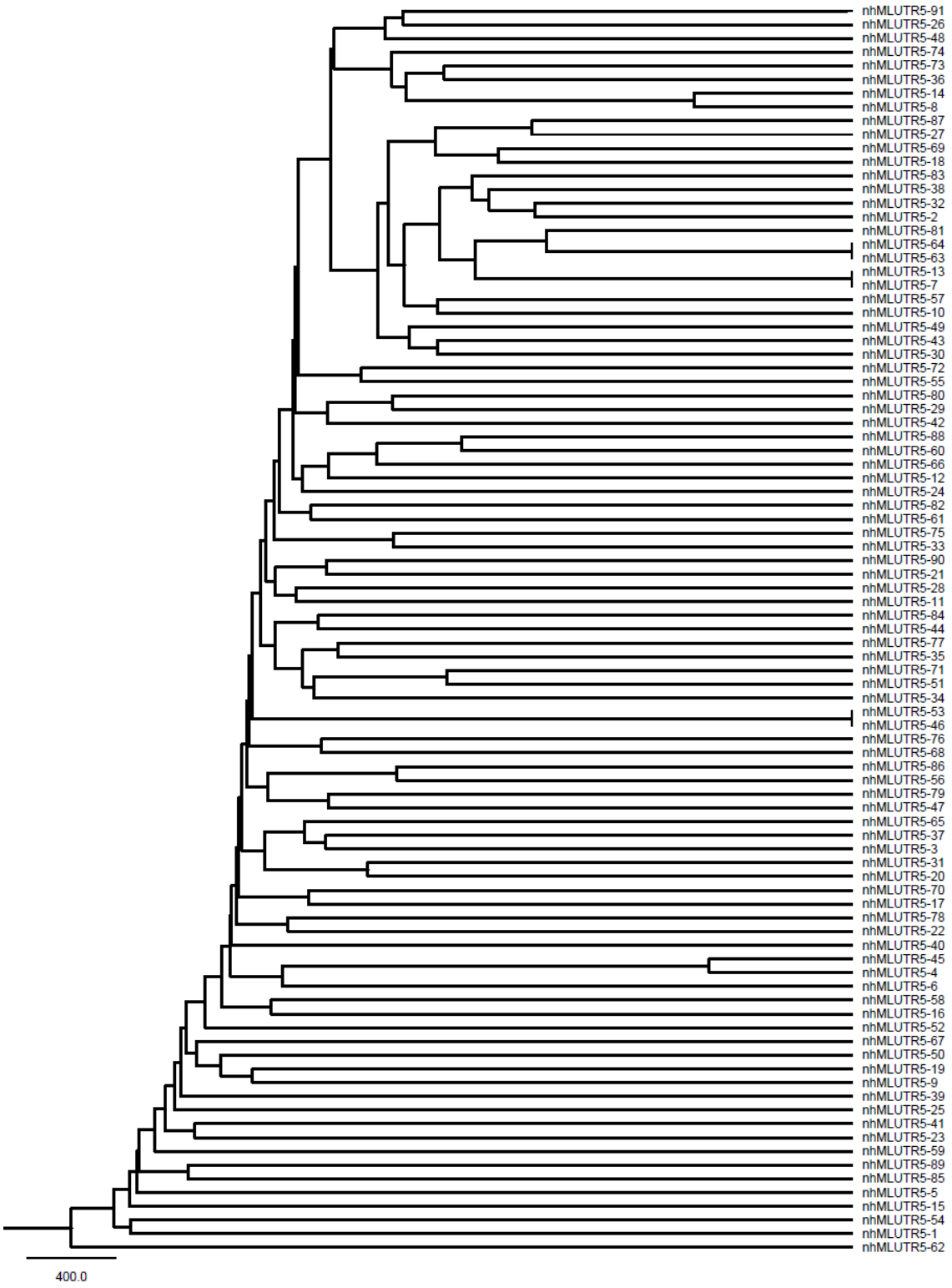
補図 1－①


nhmxUTR5-98
nhmxUTR5-38
nhmxUTR5-43
nhmxUTR5-11
nhmxUTR5-95
nhmxUTR5-71
nhmxUTR5-51
nhmxUTR5-94
nhmxUTR5-21
nhmxUTR5-86
nhmxUTR5-57
nhmxUTR5-54
nhmxUTR5-48
nhmxUTR5-2
nhmxUTR5-22
nhmxUTR5-63
nhmxUTR5-80
nhmxUTR5-76
nhmxUTR5-19
nhmxUTR5-14
nhmxUTR5-44
nhmxUTR5-64
nhmxUTR5-20
nhmxUTR5-15
nhmxUTR5-46
nhmxUTR5-41
nhmxUTR5-34
nhmxUTR5-81
nhmxUTR5-55
nhmxUTR5-67
nhmxUTR5-66
nhmxUTR5-26
nhmxUTR5-52
nhmxUTR5-17
nhmxUTR5-73
nhmxUTR5-49
nhmxUTR5-35
nhmxUTR5-9
nhmxUTR5-53
nhmxUTR5-79
nhmxUTR5-25
nhmxUTR5-88
nhmxUTR5-62
nhmxUTR5-5
nhmxUTR5-60
nhmxUTR5-13
nhmxUTR5-59
nhmxUTR5-56
nhmxUTR5-91
nhmxUTR5-72
nhmxUTR5-39
nhmxUTR5-16
nhmxUTR5-85
nhmxUTR5-42
nhmxUTR5-74
nhmxUTR5-50
nhmxUTR5-3
nhmxUTR5-93
nhmxUTR5-90
nhmxUTR5-40
nhmxUTR5-30
nhmxUTR5-23
nhmxUTR5-47
nhmxUTR5-37
nhmxUTR5-45
nhmxUTR5-18
nhmxUTR5-4
nhmxUTR5-75
nhmxUTR5-58
nhmxUTR5-31
nhmxUTR5-28
nhmxUTR5-92
nhmxUTR5-61
nhmxUTR5-10
nhmxUTR5-32
nhmxUTR5-96
nhmxUTR5-77
nhmxUTR5-36
nhmxUTR5-33
nhmxUTR5-65
nhmxUTR5-8
nhmxUTR5-27
nhmxUTR5-78
nhmxUTR5-12
nhmxUTR5-69
nhmxUTR5-29
nhmxUTR5-84
nhmxUTR5-7
nhmxUTR5-68
nhmxUTR5-89
nhmxUTR5-1
nhmxUTR5-70
nhmxUTR5-24
nhmxUTR5-82
nhmxUTR5-6
nhmxUTR5-87
nhmxUTR5-83
nhmxUTR5-97
400.0

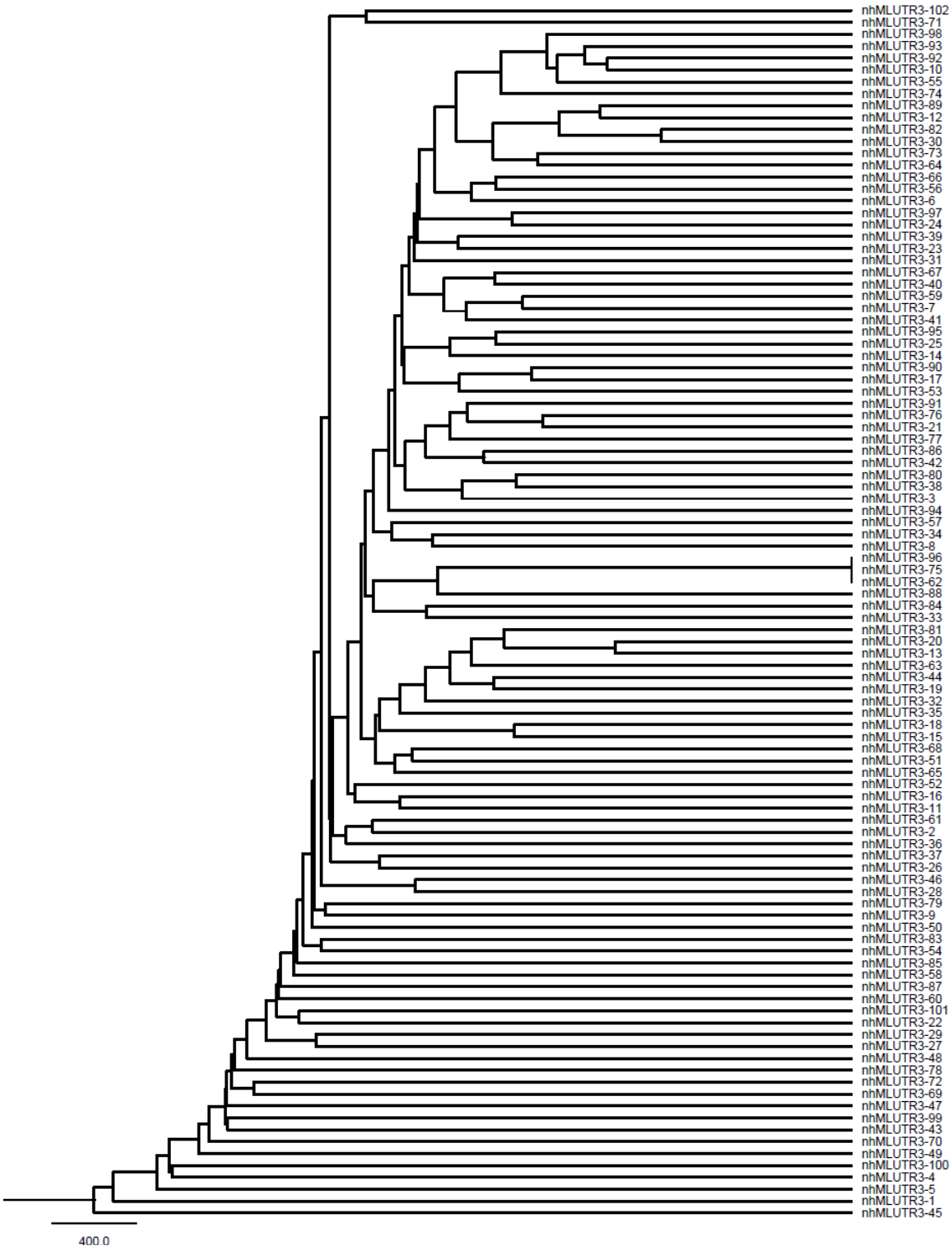
補図 1－②


nhmxUTR3-98
nhmxUTR3-73
nhmxUTR3-92
nhmxUTR3-25
nhmxUTR3-60
nhmxUTR3-32
nhmxUTR3-84
nhmxUTR3-58
nhmxUTR3-49
nhmxUTR3-91
nhmxUTR3-57
nhmxUTR3-11
nhmxUTR3-86
nhmxUTR3-43
nhmxUTR3-82
nhmxUTR3-65
nhmxUTR3-78
nhmxUTR3-10
nhmxUTR3-46
nhmxUTR3-8
nhmxUTR3-71
nhmxUTR3-74
nhmxUTR3-38
nhmxUTR3-42
nhmxUTR3-5
nhmxUTR3-17
nhmxUTR3-37
nhmxUTR3-31
nhmxUTR3-22
nhmxUTR3-7
nhmxUTR3-16
nhmxUTR3-81
nhmxUTR3-77
nhmxUTR3-28
nhmxUTR3-66
nhmxUTR3-35
nhmxUTR3-96
nhmxUTR3-64
nhmxUTR3-87
nhmxUTR3-89
nhmxUTR3-19
nhmxUTR3-9
nhmxUTR3-54
nhmxUTR3-41
nhmxUTR3-18
nhmxUTR3-50
nhmxUTR3-15
nhmxUTR3-36
nhmxUTR3-14
nhmxUTR3-67
nhmxUTR3-40
nhmxUTR3-88
nhmxUTR3-24
nhmxUTR3-6
nhmxUTR3-56
nhmxUTR3-29
nhmxUTR3-90
nhmxUTR3-13
nhmxUTR3-94
nhmxUTR3-72
nhmxUTR3-63
nhmxUTR3-93
nhmxUTR3-4
nhmxUTR3-26
nhmxUTR3-95
nhmxUTR3-55
nhmxUTR3-79
nhmxUTR3-12
nhmxUTR3-85
nhmxUTR3-44
nhmxUTR3-47
nhmxUTR3-39
nhmxUTR3-30
nhmxUTR3-76
nhmxUTR3-62
nhmxUTR3-69
nhmxUTR3-53
nhmxUTR3-51
nhmxUTR3-23
nhmxUTR3-1
nhmxUTR3-59
nhmxUTR3-80
nhmxUTR3-2
nhmxUTR3-61
nhmxUTR3-20
nhmxUTR3-97
nhmxUTR3-3
nhmxUTR3-52
nhmxUTR3-27
nhmxUTR3-83
nhmxUTR3-33
nhmxUTR3-75
nhmxUTR3-34
nhmxUTR3-48
nhmxUTR3-45
nhmxUTR3-21
nhmxUTR3-70
nhmxUTR3-68
400.0

補図 1－③


nhMLUTR5-91
nhMLUTR5-26
nhMLUTR5-48
nhMLUTR5-74
nhMLUTR5-73
nhMLUTR5-36
nhMLUTR5-14
nhMLUTR5-8
nhMLUTR5-87
nhMLUTR5-27
nhMLUTR5-69
nhMLUTR5-18
nhMLUTR5-83
nhMLUTR5-38
nhMLUTR5-32
nhMLUTR5-2
nhMLUTR5-81
nhMLUTR5-64
nhMLUTR5-63
nhMLUTR5-13
nhMLUTR5-7
nhMLUTR5-57
nhMLUTR5-10
nhMLUTR5-49
nhMLUTR5-43
nhMLUTR5-30
nhMLUTR5-72
nhMLUTR5-55
nhMLUTR5-80
nhMLUTR5-29
nhMLUTR5-42
nhMLUTR5-88
nhMLUTR5-60
nhMLUTR5-66
nhMLUTR5-12
nhMLUTR5-24
nhMLUTR5-82
nhMLUTR5-61
nhMLUTR5-75
nhMLUTR5-33
nhMLUTR5-90
nhMLUTR5-21
nhMLUTR5-28
nhMLUTR5-11
nhMLUTR5-84
nhMLUTR5-44
nhMLUTR5-77
nhMLUTR5-35
nhMLUTR5-71
nhMLUTR5-51
nhMLUTR5-34
nhMLUTR5-53
nhMLUTR5-46
nhMLUTR5-76
nhMLUTR5-68
nhMLUTR5-86
nhMLUTR5-56
nhMLUTR5-79
nhMLUTR5-47
nhMLUTR5-65
nhMLUTR5-37
nhMLUTR5-3
nhMLUTR5-31
nhMLUTR5-20
nhMLUTR5-70
nhMLUTR5-17
nhMLUTR5-78
nhMLUTR5-22
nhMLUTR5-40
nhMLUTR5-45
nhMLUTR5-4
nhMLUTR5-6
nhMLUTR5-58
nhMLUTR5-16
nhMLUTR5-52
nhMLUTR5-67
nhMLUTR5-50
nhMLUTR5-19
nhMLUTR5-9
nhMLUTR5-39
nhMLUTR5-25
nhMLUTR5-41
nhMLUTR5-23
nhMLUTR5-59
nhMLUTR5-89
nhMLUTR5-85
nhMLUTR5-5
nhMLUTR5-15
nhMLUTR5-54
nhMLUTR5-1
nhMLUTR5-62
400.0

補図 1－④


nhMLUTR3-102
nhMLUTR3-71
nhMLUTR3-98
nhMLUTR3-93
nhMLUTR3-92
nhMLUTR3-10
nhMLUTR3-55
nhMLUTR3-74
nhMLUTR3-89
nhMLUTR3-12
nhMLUTR3-82
nhMLUTR3-30
nhMLUTR3-73
nhMLUTR3-64
nhMLUTR3-66
nhMLUTR3-56
nhMLUTR3-6
nhMLUTR3-97
nhMLUTR3-24
nhMLUTR3-39
nhMLUTR3-23
nhMLUTR3-31
nhMLUTR3-67
nhMLUTR3-40
nhMLUTR3-59
nhMLUTR3-7
nhMLUTR3-41
nhMLUTR3-95
nhMLUTR3-25
nhMLUTR3-14
nhMLUTR3-90
nhMLUTR3-17
nhMLUTR3-53
nhMLUTR3-91
nhMLUTR3-76
nhMLUTR3-21
nhMLUTR3-77
nhMLUTR3-86
nhMLUTR3-42
nhMLUTR3-80
nhMLUTR3-38
nhMLUTR3-3
nhMLUTR3-94
nhMLUTR3-57
nhMLUTR3-34
nhMLUTR3-8
nhMLUTR3-96
nhMLUTR3-75
nhMLUTR3-62
nhMLUTR3-88
nhMLUTR3-84
nhMLUTR3-33
nhMLUTR3-81
nhMLUTR3-20
nhMLUTR3-13
nhMLUTR3-63
nhMLUTR3-44
nhMLUTR3-19
nhMLUTR3-32
nhMLUTR3-35
nhMLUTR3-18
nhMLUTR3-15
nhMLUTR3-68
nhMLUTR3-51
nhMLUTR3-65
nhMLUTR3-52
nhMLUTR3-16
nhMLUTR3-11
nhMLUTR3-61
nhMLUTR3-2
nhMLUTR3-36
nhMLUTR3-37
nhMLUTR3-26
nhMLUTR3-46
nhMLUTR3-28
nhMLUTR3-79
nhMLUTR3-9
nhMLUTR3-50
nhMLUTR3-83
nhMLUTR3-54
nhMLUTR3-85
nhMLUTR3-58
nhMLUTR3-87
nhMLUTR3-60
nhMLUTR3-101
nhMLUTR3-22
nhMLUTR3-29
nhMLUTR3-27
nhMLUTR3-48
nhMLUTR3-78
nhMLUTR3-72
nhMLUTR3-69
nhMLUTR3-47
nhMLUTR3-99
nhMLUTR3-43
nhMLUTR3-70
nhMLUTR3-49
nhMLUTR3-100
nhMLUTR3-4
nhMLUTR3-5
nhMLUTR3-1
nhMLUTR3-45
400.0

補表 4 には MXSCARNA によるアラインメントデータから得られる UTR 候補の予測領域の構造類似性に基づいて作成した樹形図に示した ID とその配列のゲノム上の位置と関連遺伝子を示している。ゲノム上の位置と関連遺伝子のカラムには、順に、染色体、ストランド、開始点、終止点、関連遺伝子(その UTR 候補に隣接する遺伝子)が示している。
表の左半分には 5'UTR 候補における領域について、また右半分には 3'UTR 候補における領域についてそれぞれ示している。

補表 4.MXSCARNA によるアラインメントデータから得られる UTR の予測領域の ID と領域の対応表

ID	ゲノム上の位置と関連遺伝子	ID	ゲノム上の位置と関連遺伝子
nhmxUTR5-1	chr1:−:69608:69663:YAL039C	nhmxUTR3-1	chr1:+:126812:126881:YAL016W
nhmxUTR5-2	chr1:−:82026:82179:YAL034C	nhmxUTR3-2	chr2:−:36850:36913:YBL100C
nhmxUTR5-3	chr1:−:99764:99864:YAL026C	nhmxUTR3-3	chr2:+:43129:43187:YBL097W
nhmxUTR5-4	chr1:−:201800:201897:YAR047C	nhmxUTR3-4	chr2:−:336982:337094:YBR050C
nhmxUTR5-5	chr2:+:5611:5717:YBL109W	nhmxUTR3-5	chr2:+:680562:680661:YBR230W-A
nhmxUTR5-6	chr2:+:23995:24053:YBL102W	nhmxUTR3-6	chr3:+:216478:216574:YCR052W
nhmxUTR5-7	chr2:+:43129:43187:YBL095W	nhmxUTR3-7	chr3:+:249276:249384:YCR075W-A
nhmxUTR5-8	chr2:+:101759:101854:YBL063W	nhmxUTR3-8	chr4:+:167291:167469:YDL163W
nhmxUTR5-9	chr2:−:132124:132228:YBL047C	nhmxUTR3-9	chr4:−:403518:403638:YDL027C
nhmxUTR5-10	chr2:−:159756:159846:YBL033C	nhmxUTR3-10	chr4:−:432070:432238:YDL011C
nhmxUTR5-11	chr2:+:217296:217403:YBL005W	nhmxUTR3-11	chr4:+:433094:433200:YDL010W
nhmxUTR5-12	chr2:−:245676:245743:YBR004C	nhmxUTR3-12	chr4:−:602074:602164:YDR078C
nhmxUTR5-13	chr2:−:251057:251147:YBR007C	nhmxUTR3-13	chr4:−:976575:976676:YDR260C
nhmxUTR5-14	chr3:−:13066:13193:YCL067C	nhmxUTR3-14	chr4:−:1111868:1111958:YDR322C-A
nhmxUTR5-15	chr3:+:13066:13193:YCL066W	nhmxUTR3-15	chr4:+:1125719:1125832:YDR327W
nhmxUTR5-16	chr3:−:50656:50747:YCL041C	nhmxUTR3-16	chr4:+:1301078:1301178:YDR416W
nhmxUTR5-17	chr3:−:56589:56742:YCL038C	nhmxUTR3-17	chr4:−:1301329:1301459:YDR417C
nhmxUTR5-18	chr3:+:199143:199244:YCR038W-A	nhmxUTR3-18	chr4:−:1445671:1445772:YDR498C
nhmxUTR5-19	chr3:−:200229:200384:YCR039C	nhmxUTR3-19	chr5:+:101272:101382:YEL027W
nhmxUTR5-20	chr3:+:200267:200384:YCR040W	nhmxUTR3-20	chr5:+:117022:117082:YEL021W
nhmxUTR5-21	chr4:−:125097:125211:YDL188C	nhmxUTR3-21	chr6:+:114055:114100:YFL011W
nhmxUTR5-22	chr4:+:242396:242518:YDL122W	nhmxUTR3-22	chr6:+:115287:115424:YFL010W-A
nhmxUTR5-23	chr4:−:264977:265066:YDL110C	nhmxUTR3-23	chr6:+:199077:199162:YFR022W
nhmxUTR5-24	chr4:−:478246:478306:YDR015C	nhmxUTR3-24	chr7:+:73768:73875:YGL226W
nhmxUTR5-25	chr4:−:1149551:1149649:YDR338C	nhmxUTR3-25	chr7:−:80499:80598:YGL222C
nhmxUTR5-26	chr4:−:1490023:1490224:YDR524C-A	nhmxUTR3-26	chr7:−:274581:274688:YGL124C
nhmxUTR5-27	chr4:−:1491625:1491722:YDR526C	nhmxUTR3-27	chr7:+:591973:592046:YGR049W

nhmxUTR5-28	chr5:+:85469:85584:YEL034W
nhmxUTR5-29	chr5:+:121378:121459:YEL018W
nhmxUTR5-30	chr5:+:188114:188216:YER016W
nhmxUTR5-31	chr5:+:213309:213399:YER030W
nhmxUTR5-32	chr5:−:461813:461893:YER145C
nhmxUTR5-33	chr6:+:90869:90958:YFL023W
nhmxUTR5-34	chr6:−:259510:259629:YFR054C
nhmxUTR5-35	chr7:−:384723:384879:YGL063C-A
nhmxUTR5-36	chr7:−:738252:738345:YGR123C
nhmxUTR5-37	chr7:−:784896:784991:YGR146C
nhmxUTR5-38	chr7:+:823950:824057:YGR162W
nhmxUTR5-39	chr7:+:945047:945131:YGR225W
nhmxUTR5-40	chr7:+:976299:976402:YGR242W
nhmxUTR5-41	chr7:−:1061607:1061735:YGR284C
nhmxUTR5-42	chr8:+:117678:117761:YHR006W
nhmxUTR5-43	chr8:+:180174:180252:YHR036W
nhmxUTR5-44	chr8:−:236596:236732:YHR070C-A
nhmxUTR5-45	chr8:+:436484:436593:YHR165W-A
nhmxUTR5-46	chr8:+:442006:442109:YHR169W
nhmxUTR5-47	chr8:+:512618:512723:YHR206W
nhmxUTR5-48	chr9:+:636:755:YIL177W-A
nhmxUTR5-49	chr9:−:60776:60882:YIL151C
nhmxUTR5-50	chr9:−:165780:165889:YIL107C
nhmxUTR5-51	chr9:+:277612:277702:YIL040W
nhmxUTR5-52	chr9:+:334680:334821:YIL010W
nhmxUTR5-53	chr9:−:394588:394693:YIR020C
nhmxUTR5-54	chr10:+:619:738:YJL225W-A
nhmxUTR5-55	chr10:+:85417:85562:YJL181W
nhmxUTR5-56	chr10:+:180018:180135:YJL127W-A
nhmxUTR5-57	chr10:−:419569:419688:YJL010C
nhmxUTR5-58	chr10:+:424965:425057:YJL005W
nhmxUTR5-59	chr10:−:572148:572252:YJR072C
nhmxUTR5-60	chr10:+:638463:638580:YJR114W
nhmxUTR5-61	chr10:+:703727:703821:YJR146W
nhmxUTR5-62	chr11:−:166030:166172:YKL151C
nhmxUTR5-63	chr11:+:355821:355957:YKL044W

nhmxUTR3-28	chr7:+:727776:727869:YGR118W
nhmxUTR3-29	chr7:−:1057653:1057754:YGR282C
nhmxUTR3-30	chr8:+:17899:18007:YHL041W
nhmxUTR3-31	chr8:−:33009:33123:YHL034C
nhmxUTR3-32	chr8:−:35082:35190:YHL033C
nhmxUTR3-33	chr8:−:92521:92576:YHL008C
nhmxUTR3-34	chr8:−:98201:98244:YHL006C
nhmxUTR3-35	chr8:−:119889:120004:YHR007C
nhmxUTR3-36	chr8:+:180174:180252:YHR035W
nhmxUTR3-37	chr8:−:213063:213178:YHR054C
nhmxUTR3-38	chr8:−:268297:268444:YHR082C
nhmxUTR3-39	chr8:+:358473:358573:YHR124W
nhmxUTR3-40	chr8:−:481841:481943:YHR188C
nhmxUTR3-41	chr9:+:247213:247331:YIL058W
nhmxUTR3-42	chr9:+:300180:300299:YIL029W-A
nhmxUTR3-43	chr10:+:85417:85562:YJL183W
nhmxUTR3-44	chr10:+:121764:121854:YJL159W
nhmxUTR3-45	chr10:+:223417:223460:YJL106W
nhmxUTR3-46	chr10:−:486311:486493:YJR031C
nhmxUTR3-47	chr10:+:526742:526834:YJR048W
nhmxUTR3-48	chr10:+:570793:570848:YJR071W
nhmxUTR3-49	chr10:−:572148:572252:YJR073C
nhmxUTR3-50	chr10:−:663833:663933:YJR129C
nhmxUTR3-51	chr11:+:15768:15845:YKL219W
nhmxUTR3-52	chr11:+:579391:579461:YKR074W
nhmxUTR3-53	chr11:−:617509:617583:YKR094C
nhmxUTR3-54	chr12:−:210376:210482:YLR034C
nhmxUTR3-55	chr12:+:243360:243441:YLR048W
nhmxUTR3-56	chr12:+:323203:323312:YLR091W
nhmxUTR3-57	chr12:+:456105:456264:YLR154W-F
nhmxUTR3-58	chr12:+:489994:490113:YLR162W
nhmxUTR3-59	chr12:+:628230:628317:YLR246W
nhmxUTR3-60	chr12:−:673667:673776:YLR264C-A
nhmxUTR3-61	chr12:+:679989:680053:YLR267W
nhmxUTR3-62	chr12:−:752058:752160:YLR310C
nhmxUTR3-63	chr13:+:2247:2366:YML133W-A

nhmxUTR5-64	chr11:−:566914:567006:YKR066C	nhmxUTR3-64	chr13:+:57525:57620:YML106W
nhmxUTR5-65	chr11:+:635380:635454:YKR099W	nhmxUTR3-65	chr13:−:209186:209338:YML034C-A
nhmxUTR5-66	chr12:+:452141:452340:YLR154W-A	nhmxUTR3-66	chr13:−:347330:347441:YMR038C
nhmxUTR5-67	chr12:+:490114:490313:YLR162W-A	nhmxUTR3-67	chr13:+:426125:426215:YMR079W
nhmxUTR5-68	chr12:−:797178:797249:YLR334C	nhmxUTR3-68	chr13:+:514040:514072:YMR123W
nhmxUTR5-69	chr13:−:79747:79811:YML097C	nhmxUTR3-69	chr13:+:559949:560047:YMR147W
nhmxUTR5-70	chr13:+:79747:79811:YML096W	nhmxUTR3-70	chr13:+:646804:646854:YMR191W
nhmxUTR5-71	chr13:−:243062:243156:YML015C	nhmxUTR3-71	chr13:−:777379:777530:YMR254C
nhmxUTR5-72	chr13:+:243062:243156:YML014W	nhmxUTR3-72	chr14:+:2154:2273:YNL339W-A
nhmxUTR5-73	chr13:−:251028:251131:YML009C-A	nhmxUTR3-73	chr14:−:58693:58808:YNL305C
nhmxUTR5-74	chr13:−:346596:346679:YMR037C	nhmxUTR3-74	chr14:+:68500:68648:YNL299W
nhmxUTR5-75	chr13:+:353701:353802:YMR043W	nhmxUTR3-75	chr14:+:283023:283074:YNL190W
nhmxUTR5-76	chr13:+:627974:628088:YMR184W	nhmxUTR3-76	chr14:+:283093:283193:YNL190W
nhmxUTR5-77	chr13:+:818687:818795:YMR276W	nhmxUTR3-77	chr14:+:352108:352214:YNL146W
nhmxUTR5-78	chr13:+:856833:856902:YMR294W	nhmxUTR3-78	chr14:+:500275:500482:YNL067W
nhmxUTR5-79	chr14:+:137558:137676:YNL269W	nhmxUTR3-79	chr14:+:532269:532359:YNL052W
nhmxUTR5-80	chr14:+:182689:182780:YNL247W	nhmxUTR3-80	chr14:+:642411:642471:YNR008W
nhmxUTR5-81	chr14:+:210035:210178:YNL234W	nhmxUTR3-81	chr14:+:675909:676027:YNR027W
nhmxUTR5-82	chr14:+:276372:276420:YNL192W	nhmxUTR3-82	chr14:−:759900:760040:YNR068C
nhmxUTR5-83	chr14:+:276451:276488:YNL192W	nhmxUTR3-83	chr15:+:14750:14811:YOL160W
nhmxUTR5-84	chr14:+:606210:606269:YNL014W	nhmxUTR3-84	chr15:−:61172:61266:YOL138C
nhmxUTR5-85	chr15:−:426184:426280:YOR051C	nhmxUTR3-85	chr15:+:238203:238304:YOL051W
nhmxUTR5-86	chr15:−:427252:427360:YOR052C	nhmxUTR3-86	chr15:+:254653:254767:YOL039W
nhmxUTR5-87	chr15:−:467670:467723:YOR074C	nhmxUTR3-87	chr15:+:259060:259155:YOL036W
nhmxUTR5-88	chr15:+:480457:480564:YOR084W	nhmxUTR3-88	chr15:+:265185:265284:YOL033W
nhmxUTR5-89	chr16:−:14404:14467:YPL279C	nhmxUTR3-89	chr15:+:413570:413674:YOR044W
nhmxUTR5-90	chr16:−:28237:28350:YPL272C	nhmxUTR3-90	chr15:−:426657:426752:YOR052C
nhmxUTR5-91	chr16:−:95199:95291:YPL242C	nhmxUTR3-91	chr15:+:480457:480564:YOR083W
nhmxUTR5-92	chr16:+:370805:370877:YPL093W	nhmxUTR3-92	chr15:−:507835:507942:YOR098C
nhmxUTR5-93	chr16:+:415612:415710:YPL074W	nhmxUTR3-93	chr15:−:738750:738842:YOR211C
nhmxUTR5-94	chr16:+:482670:482775:YPL036W	nhmxUTR3-94	chr16:+:2063:2182:YPL283W-A
nhmxUTR5-95	chr16:−:503052:503148:YPL025C	nhmxUTR3-95	chr16:−:95199:95291:YPL241C
nhmxUTR5-96	chr16:+:642100:642162:YPR035W	nhmxUTR3-96	chr16:−:278218:278318:YPL145C
nhmxUTR5-97	chr16:+:643743:643771:YPR036W	nhmxUTR3-97	chr16:−:491853:491916:YPL031C
nhmxUTR5-98	chr16:+:939737:939856:YPR201W	nhmxUTR3-98	chr16:+:649998:650102:YPR041W

補表 5 には MULTIZ によるアラインメントデータから得られる UTR 候補の予測領域の構造類似性に基づいて作成した樹形図に示した ID とその配列のゲノム上の位置と関連遺伝子を示している。ゲノム上の位置と関連遺伝子のカラムには、順に、染色体、ストランド、開始点、終止点、関連遺伝子(UTR が含まれる遺伝子)が示している。
表の左半分には 5'UTR 候補における領域について、また右半分には 3'UTR 候補における領域についてそれぞれ示している。

補表 5. MULTIZ によるアラインメントデータから得られる UTR 候補の予測領域の ID と領域の対応表

ID	ゲノム上の位置と関連遺伝子	ID	ゲノム上の位置と関連遺伝子
nhMLUTR5-1	chr1:-:69609:69663:YAL039C	nhMLUTR3-1	chr1:-:87751:87770:YAL029C
nhMLUTR5-2	chr1:-:82029:82184:YAL034C	nhMLUTR3-2	chr2:-:162887:162985:YBL030C
nhMLUTR5-3	chr1:-:201821:201911:YAR047C	nhMLUTR3-3	chr2:+:216946:217065:YBL006W-A
nhMLUTR5-4	chr1:+:220873:220971:YAR066W	nhMLUTR3-4	chr2:-:344561:344598:YBR055C
nhMLUTR5-5	chr2:-:245676:245743:YBR004C	nhMLUTR3-5	chr2:+:466142:466173:YBR113W
nhMLUTR5-6	chr2:+:256176:256276:YBR010W	nhMLUTR3-6	chr2:+:680567:680661:YBR230W-A
nhMLUTR5-7	chr3:-:13066:13245:YCL067C	nhMLUTR3-7	chr2:-:685302:685416:YBR234C
nhMLUTR5-8	chr3:+:13094:13208:YCL066W	nhMLUTR3-8	chr3:+:213764:213871:YCR048W
nhMLUTR5-9	chr3:-:50656:50747:YCL041C	nhMLUTR3-9	chr3:+:227500:227584:YCR061W
nhMLUTR5-10	chr3:-:56595:56742:YCL038C	nhMLUTR3-10	chr3:+:249246:249362:YCR075W-A
nhMLUTR5-11	chr3:-:74082:74192:YCL026C-B	nhMLUTR3-11	chr4:-:60674:60773:YDL222C
nhMLUTR5-12	chr3:+:199143:199244:YCR038W-A	nhMLUTR3-12	chr4:+:167290:167477:YDL163W
nhMLUTR5-13	chr3:-:200229:200385:YCR039C	nhMLUTR3-13	chr4:-:403518:403638:YDL027C
nhMLUTR5-14	chr3:+:200267:200385:YCR040W	nhMLUTR3-14	chr4:+:1021698:1021793:YDR280W
nhMLUTR5-15	chr4:-:147627:147679:YDL174C	nhMLUTR3-15	chr4:+:1125719:1125832:YDR327W
nhMLUTR5-16	chr4:+:242396:242477:YDL122W	nhMLUTR3-16	chr4:-:1288050:1288160:YDR408C
nhMLUTR5-17	chr4:-:264977:265066:YDL110C	nhMLUTR3-17	chr4:-:1301347:1301462:YDR417C
nhMLUTR5-18	chr4:-:356761:356876:YDL055C	nhMLUTR3-18	chr4:+:1311674:1311767:YDR420W
nhMLUTR5-19	chr4:-:478246:478306:YDR015C	nhMLUTR3-19	chr4:-:1445663:1445771:YDR498C
nhMLUTR5-20	chr4:+:746568:746713:YDR145W	nhMLUTR3-20	chr5:+:101272:101387:YEL027W
nhMLUTR5-21	chr5:-:41967:42083:YEL060C	nhMLUTR3-21	chr5:-:226689:226807:YER038C
nhMLUTR5-22	chr5:+:85468:85584:YEL034W	nhMLUTR3-22	chr5:-:382451:382519:YER111C
nhMLUTR5-23	chr5:+:121378:121454:YEL018W	nhMLUTR3-23	chr5:-:454978:455093:YER142C
nhMLUTR5-24	chr5:-:461813:461894:YER145C	nhMLUTR3-24	chr6:+:115280:115431:YFL010W-A
nhMLUTR5-25	chr5:+:510245:510323:YER165W	nhMLUTR3-25	chr7:-:80494:80600:YGL222C
nhMLUTR5-26	chr5:-:513078:513208:YER165C-A	nhMLUTR3-26	chr7:-:274584:274691:YGL124C
nhMLUTR5-27	chr5:-:522727:522889:YER168C	nhMLUTR3-27	chr7:+:591973:592031:YGR049W

nhMLUTR5-28	chr6:+:56215:56332:YFL037W
nhMLUTR5-29	chr6:−:259511:259630:YFR054C
nhMLUTR5-30	chr7:−:49763:49882:YGL239C
nhMLUTR5-31	chr7:+:190964:191082:YGL166W
nhMLUTR5-32	chr7:−:384723:384881:YGL063C-A
nhMLUTR5-33	chr7:+:465997:466116:YGL014W
nhMLUTR5-34	chr7:+:945045:945134:YGR225W
nhMLUTR5-35	chr7:+:976292:976402:YGR242W
nhMLUTR5-36	chr7:+:1021513:1021632:YGR265W
nhMLUTR5-37	chr7:−:1061632:1061718:YGR284C
nhMLUTR5-38	chr7:−:1080427:1080620:YGR293C
nhMLUTR5-39	chr8:−:38547:38624:YHL032C
nhMLUTR5-40	chr8:+:117678:117773:YHR006W
nhMLUTR5-41	chr8:+:180174:180252:YHR036W
nhMLUTR5-42	chr8:+:236831:236946:YHR071W
nhMLUTR5-43	chr8:+:441976:442119:YHR169W
nhMLUTR5-44	chr8:−:498897:499001:YHR199C-A
nhMLUTR5-45	chr8:+:541484:541582:YHR214W
nhMLUTR5-46	chr9:+:636:755:YIL177W-A
nhMLUTR5-47	chr9:−:60779:60889:YIL151C
nhMLUTR5-48	chr9:−:75820:75935:YIL146C
nhMLUTR5-49	chr9:−:117729:117861:YIL127C
nhMLUTR5-50	chr9:−:137930:138003:YIL119C
nhMLUTR5-51	chr9:−:394588:394697:YIR020C
nhMLUTR5-52	chr9:−:421866:421958:YIR035C
nhMLUTR5-53	chr10:+:619:738:YJL225W-A
nhMLUTR5-54	chr10:−:53222:53266:YJL204C
nhMLUTR5-55	chr10:−:73503:73613:YJL192C
nhMLUTR5-56	chr10:+:180018:180135:YJL127W-A
nhMLUTR5-57	chr10:+:205072:205226:YJL112W
nhMLUTR5-58	chr10:−:419606:419688:YJL010C
nhMLUTR5-59	chr10:+:578710:578780:YJR078W
nhMLUTR5-60	chr11:+:355830:355943:YKL044W
nhMLUTR5-61	chr11:−:477756:477868:YKR019C
nhMLUTR5-62	chr12:−:390306:390342:YLR121C
nhMLUTR5-63	chr12:+:452141:452340:YLR154W-A

nhMLUTR3-28	chr7:+:644644:644723:YGR082W
nhMLUTR3-29	chr7:+:985780:985839:YGR247W
nhMLUTR3-30	chr8:−:34980:35195:YHL033C
nhMLUTR3-31	chr8:+:98122:98240:YHL006W-A
nhMLUTR3-32	chr8:−:98122:98240:YHL006C
nhMLUTR3-33	chr8:−:109881:109970:YHR003C
nhMLUTR3-34	chr8:−:119882:119997:YHR007C
nhMLUTR3-35	chr8:+:180174:180252:YHR035W
nhMLUTR3-36	chr8:+:207226:207319:YHR049W
nhMLUTR3-37	chr8:−:252266:252362:YHR077C
nhMLUTR3-38	chr8:−:268299:268415:YHR082C
nhMLUTR3-39	chr8:+:289731:289846:YHR093W
nhMLUTR3-40	chr8:+:358473:358575:YHR124W
nhMLUTR3-41	chr8:+:441976:442119:YHR168W
nhMLUTR3-42	chr9:−:99696:99804:YIL132C
nhMLUTR3-43	chr9:−:187846:187887:YIL093C
nhMLUTR3-44	chr9:+:247213:247331:YIL058W
nhMLUTR3-45	chr9:+:350010:350022:YIL003W
nhMLUTR3-46	chr10:+:193499:193574:YJL117W
nhMLUTR3-47	chr10:+:223417:223460:YJL106W
nhMLUTR3-48	chr10:+:570793:570852:YJR071W
nhMLUTR3-49	chr10:−:580341:580376:YJR080C
nhMLUTR3-50	chr10:−:604457:604537:YJR094C
nhMLUTR3-51	chr10:−:663849:663933:YJR129C
nhMLUTR3-52	chr11:+:15768:15845:YKL219W
nhMLUTR3-53	chr11:+:281279:281381:YKL083W
nhMLUTR3-54	chr11:+:579391:579470:YKR074W
nhMLUTR3-55	chr12:+:456102:456261:YLR154W-F
nhMLUTR3-56	chr12:+:489994:490113:YLR162W
nhMLUTR3-57	chr12:+:607239:607349:YLR232W
nhMLUTR3-58	chr12:+:628230:628317:YLR246W
nhMLUTR3-59	chr12:−:673663:673774:YLR264C-A
nhMLUTR3-60	chr12:+:679989:680053:YLR267W
nhMLUTR3-61	chr12:−:752045:752147:YLR310C
nhMLUTR3-62	chr13:+:2247:2366:YML133W-A
nhMLUTR3-63	chr13:+:70381:70498:YML100W-A

nhMLUTR5-64	chr12:+:490114:490313:YLR162W-A
nhMLUTR5-65	chr12:+:660577:660666:YLR258W
nhMLUTR5-66	chr12:−:758865:758983:YLR312C
nhMLUTR5-67	chr12:−:797178:797249:YLR334C
nhMLUTR5-68	chr12:+:954862:954970:YLR417W
nhMLUTR5-69	chr13:+:158488:158668:YML058W-A
nhMLUTR5-70	chr13:−:251027:251135:YML009C-A
nhMLUTR5-71	chr13:−:318516:318630:YMR021C
nhMLUTR5-72	chr13:+:353705:353805:YMR043W
nhMLUTR5-73	chr13:+:627977:628093:YMR184W
nhMLUTR5-74	chr14:−:23354:23456:YNL328C
nhMLUTR5-75	chr14:+:137558:137676:YNL269W
nhMLUTR5-76	chr14:+:187868:187952:YNL243W
nhMLUTR5-77	chr14:+:210048:210154:YNL234W
nhMLUTR5-78	chr14:+:413477:413573:YNL112W
nhMLUTR5-79	chr14:−:759111:759213:YNR067C
nhMLUTR5-80	chr15:+:310132:310244:YOL008W
nhMLUTR5-81	chr15:−:426176:426334:YOR051C
nhMLUTR5-82	chr15:+:480461:480575:YOR084W
nhMLUTR5-83	chr15:−:796820:796972:YOR246C
nhMLUTR5-84	chr15:−:813807:813909:YOR259C
nhMLUTR5-85	chr16:−:14404:14467:YPL279C
nhMLUTR5-86	chr16:+:75597:75699:YPL250W-A
nhMLUTR5-87	chr16:−:156265:156415:YPL210C
nhMLUTR5-88	chr16:−:503045:503144:YPL025C
nhMLUTR5-89	chr16:+:642100:642162:YPR035W
nhMLUTR5-90	chr16:+:739920:740028:YPR106W
nhMLUTR5-91	chr16:+:939737:939856:YPR201W

nhMLUTR3-64	chr13:−:209175:209330:YML034C-A
nhMLUTR3-65	chr13:+:298455:298544:YMR013W-A
nhMLUTR3-66	chr13:−:347306:347423:YMR038C
nhMLUTR3-67	chr13:−:367888:368007:YMR049C
nhMLUTR3-68	chr13:+:426126:426215:YMR079W
nhMLUTR3-69	chr13:+:514016:514072:YMR123W
nhMLUTR3-70	chr13:−:557357:557401:YMR146C
nhMLUTR3-71	chr13:+:559944:560044:YMR147W
nhMLUTR3-72	chr13:+:646805:646854:YMR191W
nhMLUTR3-73	chr13:−:777384:777533:YMR254C
nhMLUTR3-74	chr13:−:796411:796511:YMR265C
nhMLUTR3-75	chr14:+:2154:2273:YNL339W-A
nhMLUTR3-76	chr14:−:58688:58808:YNL305C
nhMLUTR3-77	chr14:+:68521:68639:YNL299W
nhMLUTR3-78	chr14:+:283023:283074:YNL190W
nhMLUTR3-79	chr14:+:283089:283171:YNL190W
nhMLUTR3-80	chr14:+:352108:352216:YNL146W
nhMLUTR3-81	chr14:+:379303:379418:YNL131W
nhMLUTR3-82	chr14:+:500275:500482:YNL067W
nhMLUTR3-83	chr14:−:505687:505776:YNL064C
nhMLUTR3-84	chr14:+:532270:532363:YNL052W
nhMLUTR3-85	chr14:+:642411:642471:YNR008W
nhMLUTR3-86	chr14:−:759910:760027:YNR068C
nhMLUTR3-87	chr15:+:14750:14811:YOL160W
nhMLUTR3-88	chr15:−:61179:61283:YOL138C
nhMLUTR3-89	chr15:−:91742:91930:YOL121C
nhMLUTR3-90	chr15:+:220474:220602:YOL058W
nhMLUTR3-91	chr15:+:254653:254772:YOL039W
nhMLUTR3-92	chr15:+:413560:413677:YOR044W
nhMLUTR3-93	chr15:+:462052:462159:YOR072W-A
nhMLUTR3-94	chr15:+:480461:480575:YOR083W
nhMLUTR3-95	chr15:+:956691:956790:YOR337W
nhMLUTR3-96	chr16:+:2063:2182:YPL283W-A
nhMLUTR3-97	chr16:+:132283:132455:YPL222W
nhMLUTR3-98	chr16:−:156265:156415:YPL209C
nhMLUTR3-99	chr16:−:278263:278311:YPL145C

	nhMLUTR3-100	chr16:−:337357:337393:YPL112C
	nhMLUTR3-101	chr16:−:711299:711363:YPR088C
	nhMLUTR3-102	chr16:+:876048:876151:YPR165W

補表 6－9 には、MXSCARNA、MULTIZ のそれぞれのアラインメントデータから得られる UTR 候補に存在する予測領域から作成された樹形図において、有意に検出される GO タームを記した。関連遺伝子数は、それぞれのクラスターにおいて有意な GO タームのアノテーションのあった遺伝子の数を示す。総遺伝子数は、それぞれのクラスターにおいてアノテーションのある遺伝子の数を示している。

補表 6.MXSCARNA によるアラインメントデータから得られる 5’UTR 候補の予測領域の樹形図において検出される有意なクラスター

クラスター名	GO タームム	P-value	関連遺伝子数	総遺伝子数
mx5_cluster6	GO:0016787	2.72E-02	2	2
mx5_cluster13	GO:0003714	4.91E-05	2	2
mx5_cluster13	GO:0007533	3.80E-04	2	2
mx5_cluster13	GO:0007535	1.38E-05	2	2
mx5_cluster13	GO:0016564	3.04E-04	2	2
mx5_cluster18	GO:0003713	3.37E-05	2	2
mx5_cluster18	GO:0016563	4.26E-04	2	2
mx5_cluster19	GO:0003712	3.22E-07	4	7
mx5_cluster19	GO:0005515	1.14E-02	4	7
mx5_cluster19	GO:0006350	3.75E-02	4	7
mx5_cluster19	GO:0006351	2.64E-02	4	7
mx5_cluster19	GO:0006355	6.80E-03	4	7
mx5_cluster19	GO:0006357	1.01E-03	4	7
mx5_cluster19	GO:0006366	4.73E-03	4	7
mx5_cluster19	GO:0007275	1.41E-02	4	7
mx5_cluster19	GO:0007530	3.96E-07	4	7
mx5_cluster19	GO:0007531	3.96E-07	4	7
mx5_cluster19	GO:0007532	1.02E-09	4	7
mx5_cluster19	GO:0008134	4.41E-07	4	7
mx5_cluster19	GO:0016070	4.34E-02	5	7
mx5_cluster19	GO:0019219	1.32E-02	4	7
mx5_cluster19	GO:0019222	2.58E-02	4	7
mx5_cluster19	GO:0030528	1.90E-03	4	7
mx5_cluster19	GO:0031323	2.06E-02	4	7
mx5_cluster19	GO:0032774	2.75E-02	4	7
mx5_cluster19	GO:0045449	8.79E-03	4	7
mx5_cluster30	GO:0004601	1.66E-02	2	19

mx5_cluster30	GO:0016209	2.31E-02	2	19
mx5_cluster30	GO:0016684	1.66E-02	2	19
mx5_cluster38	GO:0005575	2.98E-02	2	2
mx5_cluster59	GO:0008150	4.98E-02	3	4
mx5_cluster68	GO:0006800	1.73E-02	5	51
mx5_cluster68	GO:0006979	1.60E-02	5	51
mx5_cluster68	GO:0050791	4.53E-02	14	51

補表 7.MXSCARNA によるアラインメントデータから得られる 3' UTR 候補の予測領域の樹形図において検出される有意なクラスター

クラスター名	GO タームム	P-value	関連遺伝子数	総遺伝子数
mx3_cluster2	GO:0005488	3.34E-02	2	2
mx3_cluster4	GO:0005198	3.29E-02	2	4
mx3_cluster18	GO:0000137	8.64E-04	2	6
mx3_cluster18	GO:0005795	1.74E-03	2	6
mx3_cluster18	GO:0031984	1.74E-03	2	6
mx3_cluster18	GO:0031985	1.74E-03	2	6
mx3_cluster28	GO:0003674	4.69E-02	3	3
mx3_cluster35	GO:0005830	4.17E-02	5	29
mx3_cluster40	GO:0005774	2.50E-02	2	4
mx3_cluster40	GO:0006873	2.10E-02	2	4
mx3_cluster40	GO:0006875	7.67E-03	2	4
mx3_cluster40	GO:0006879	1.67E-03	2	4
mx3_cluster40	GO:0008324	8.58E-03	2	4
mx3_cluster40	GO:0015075	1.16E-02	2	4
mx3_cluster40	GO:0019725	2.55E-02	2	4
mx3_cluster40	GO:0030003	1.62E-02	2	4
mx3_cluster40	GO:0030005	5.66E-03	2	4
mx3_cluster40	GO:0042592	2.82E-02	2	4
mx3_cluster40	GO:0044437	2.84E-02	2	4
mx3_cluster40	GO:0046916	4.30E-03	2	4
mx3_cluster40	GO:0050801	2.26E-02	2	4
mx3_cluster42	GO:0005215	8.52E-03	3	5
mx3_cluster46	GO:0006818	9.09E-03	2	8
mx3_cluster46	GO:0015672	1.43E-02	2	8
mx3_cluster46	GO:0015992	9.09E-03	2	8

mx3_cluster48	GO:0006629	1.42E-02	2	2
mx3_cluster48	GO:0006643	3.86E-03	2	2
mx3_cluster48	GO:0006644	2.25E-03	2	2
mx3_cluster48	GO:0006650	8.47E-04	2	2
mx3_cluster48	GO:0030384	4.29E-04	2	2
mx3_cluster48	GO:0044255	1.27E-02	2	2
mx3_cluster49	GO:0031090	1.77E-02	6	13
mx3_cluster85	GO:0007049	2.74E-02	2	2
mx3_cluster85	GO:0051276	3.64E-02	2	2

補表 8.MULTIZ によるアラインメントデータから得られる 5’ UTR 候補の予測領域の樹形図において検出される有意なクラスター

クラスター名	GO タームP-value	関連遺伝子数	総遺伝子数
ML5_cluster4	GO:0003713	3.37E-05	2	2
ML5_cluster4	GO:0016563	4.26E-04	2	2
ML5_cluster10	GO:0016779	8.06E-03	2	4
ML5_cluster15	GO:0003674	4.69E-02	3	3
ML5_cluster15	GO:0008150	1.53E-02	3	3
ML5_cluster16	GO:0003714	4.91E-05	2	2
ML5_cluster16	GO:0005515	4.41E-02	2	2
ML5_cluster16	GO:0007533	3.80E-04	2	2
ML5_cluster16	GO:0007535	1.38E-05	2	2
ML5_cluster16	GO:0016564	3.04E-04	2	2
ML5_cluster25	GO:0003712	7.80E-05	4	21
ML5_cluster25	GO:0007530	9.81E-05	4	21
ML5_cluster25	GO:0007531	9.81E-05	4	21
ML5_cluster25	GO:0007532	2.62E-07	4	21
ML5_cluster25	GO:0008134	1.06E-04	4	21
ML5_cluster27	GO:0008301	1.59E-02	2	23
ML5_cluster49	GO:0009056	2.73E-02	3	5
ML5_cluster49	GO:0009057	1.41E-02	3	5
ML5_cluster49	GO:0030163	1.95E-03	3	5
ML5_cluster49	GO:0043285	7.22E-03	3	5
ML5_cluster53	GO:0000723	1.97E-02	2	2
ML5_cluster53	GO:0032200	1.97E-02	2	2

ML5_cluster61	GO:0003702	6.99E-04	2	2
ML5_cluster61	GO:0006352	1.13E-03	2	2
ML5_cluster61	GO:0006366	3.45E-02	2	2
ML5_cluster61	GO:0006367	7.26E-04	2	2
ML5_cluster61	GO:0030528	4.98E-03	2	2
ML5_cluster68	GO:0006351	3.74E-02	12	52
ML5_cluster68	GO:0006355	6.79E-03	11	52
ML5_cluster68	GO:0006357	2.67E-02	8	52
ML5_cluster68	GO:0019219	3.16E-02	11	52
ML5_cluster68	GO:0032774	4.12E-02	12	52
ML5_cluster68	GO:0045449	1.24E-02	11	52
ML5_cluster69	GO:0005575	2.98E-02	2	2

補表 9.MULTIZ によるアラインメントデータから得られる 3’ UTR 候補の予測領域の樹形図において検出される有意なクラスター

クラスター名	GO ターム	P-value	関連遺伝子数	総遺伝子数
ML3_cluster6	GO:0008150	4.98E-02	3	4
ML3_cluster8	GO:0005842	4.09E-03	2	2
ML3_cluster8	GO:0015934	8.92E-03	2	2
ML3_cluster9	GO:0005198	3.19E-04	3	3
ML3_cluster9	GO:0005830	3.60E-04	3	3
ML3_cluster9	GO:0005840	1.75E-03	3	3
ML3_cluster9	GO:0030529	1.15E-02	3	3
ML3_cluster11	GO:0003735	8.45E-05	3	3
ML3_cluster11	GO:0005829	1.58E-02	3	3
ML3_cluster11	GO:0044445	5.07E-04	3	3
ML3_cluster26	GO:0003676	2.06E-02	2	2
ML3_cluster27	GO:0006401	5.77E-03	2	3
ML3_cluster27	GO:0006402	4.21E-03	2	3
ML3_cluster27	GO:0016071	4.55E-02	2	3
ML3_cluster42	GO:0006629	1.31E-02	2	2
ML3_cluster42	GO:0008202	4.55E-04	2	2
ML3_cluster42	GO:0044255	1.16E-02	2	2
ML3_cluster43	GO:0005783	2.92E-02	2	2
ML3_cluster43	GO:0006066	6.82E-03	2	2

ML3_cluster43	GO:0016125	4.34E-04	2	2
ML3_cluster55	GO:0005215	4.26E-03	3	5
ML3_cluster63	GO:0045021	2.37E-02	2	46
ML3_cluster84	GO:0005488	3.36E-02	2	2
ML3_cluster86	GO:0007049	1.57E-02	2	2

表 7 の補足

以下には、表 7 の各クラスターに含まれる領域について、有意なクラスターに含まれる領域の ID を示している。クラスター名以下の {} に含まれる数は ID の数値の部分を示している。クラスター名に含まれる mx、ML はそれぞれ MXSCARNA、MULTIZ によるアラインメントデータであることを示し、それにつづく 5、3 は 5'UTR 候補、3'UTR 候補の領域のデータであることを示している。

また {} 内に含まれる数字は候補領域の ID の数値部分の数である。つまり、mx5_cluster19 の場合、{nhmxUTR5_80, nhmxUTR5_76, nhmxUTR5_19,・・・}ということを示している。

mx5_cluster19

{80,76,19,14,44,64,20,15,}

mx5_cluster68

{98,38,43,11,95,71,51,94,21,86,57,54,48,2,22,63,80,76,19,14,44,64,20,15,46,41,34,81,55,67,66,26,52,17,73,49,35,9,53,79,25,88,62,5,60,13,59,56,91,72,39,16,85,42,74,50,3,93,90,40,30,23,47,37,45,18,4,75,58,}

mx3_cluster35

{98,73,92,25,60,32,84,58,49,91,57,11,86,43,82,65,78,10,46,8,71,74,38,42,5,17,37,31,22,7,16,81,77,28,66,35,}

mx3_cluster49

{96,64,87,89,19,9,54,41,18,50,15,36,14,67,40,}

ML5_cluster68

{91,26,48,74,73,36,14,8,87,27,69,18,83,38,32,2,81,64,63,13,7,57,10,49,43,30,72,55,80,29,42,88,60,66,12,24,82,61,75,33,90,21,28,11,84,44,77,35,71,51,34,53,46,76,68,86,56,79,47,65,37,3,31,20,70,17,78,22,40,}

ML3_cluster89

{102,71,98,93,92,10,55,74,89,12,82,30,73,64,66,56,6,97,24,39,23,31,67,40,59,7,41,95,25,
14,90,17,53,91,76,21,77,86,42,80,38,3,94,57,34,8,96,75,62,88,84,33,81,20,13,63,44,19,32
,35,18,15,68,51,65,52,16,11,61,2,36,37,26,46,28,79,9,50,83,54,85,
58,87,60,101,22,29,27,48,78,}