

シロイヌナズナの短いコーディング遺伝子に特化した発現データベースと過剰発現体ライブラリーの構築

武田 智之・花田 耕介

九州工業大学 情報工学部 生命情報工学科

Transcriptome data and transgenic library of overexpression for small coding genes of *Arabidopsis thaliana*

Tomoyuki Takeda and Kousuke Hanada

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology

要旨: Functional roles of small coding genes tend to be ignored in 20 years because small coding genes have not been identified in even highly accurate genomes of model organisms. In plants, it is recognized that small coding genes play essential roles within a single cell as well as intercellular signal transductions like hormone-like peptides. However, many of small coding genes are not annotated in plant genomes even if novel hormone-like peptides encoded by small coding genes tend to be hidden in plant genomes. Here, we introduce transcriptome database (HanaDB-AT) and transgenic library of overexpression with focusing on Arabidopsis novel and annotated small coding genes. The database and transgenic library should be useful tools to discover novel hormone-like peptides as well as other functional peptides in small coding genes.

はじめに

約 20 年前に構築した高精度なゲノムを持つモデル生物においても、多くの研究者が 100 aa (アミノ酸) 以上のタンパク質に着目していたため、100 aa 以下の短いコーディング遺伝子は着目されていない。その後、網羅的に転写領域が同定されるトランスクリプトームが実施されると、遺伝子が存在しないと考えられている遺伝子間領域に数多くの転写領域が明らかになった。これらの転写領域は、100 aa 以下の Open Reading Frame (ORF) しか存在しないものが多い。そのため、これらの転写領域は、非コーディング RNA 遺伝子あるいは短いコーディング遺伝子であると予想される。しかしながら、ORF がコーディング領域あるいは非コーディング領域であるかを判断するために利用されている遺伝子予測プログラムの多くは、100 aa 以下の ORF がコーディング領域であることを判断することが難しい。そのため、遺伝子間領域に存在する領域の多くが、十分な検証をせずに非コーディング RNA 遺伝子としてアノテーションされている。このように、多くの場合、短いコーディング遺伝子は、植物ゲノムにおいてアノテーションされていない傾向にある (Andrews and Rothnagel 2014)。そこで、本研究ノートでは、近年の様々な解析によって機能的役割が明らかにされつつある「短いコーディング遺伝子の機能的役割」を解説すると共に、「新規に短いコーディング遺伝子を推定する方法」や「推定された短いコーディング遺伝子の機能解析の方法」について紹介する。

短いコーディング遺伝子の機能的役割

近年、遺伝子として登録されない傾向にある短いコーディング遺伝子に、様々な分子機能的役割が存在することが明らかになり、注目が集まりつつある。ここでは、植物で同定されつつある短いコーディング遺伝子を分子機構に基づいて 5 つのカテゴリーに分類した。表 1 には、短いコーディング遺伝子の分子機構カテゴリーおよび、モデル植物のシロイヌナズナにおけるそれぞれのカテゴリーに属する遺伝子数を記載している (Tavormina et al. 2015; Hsu and Benfey 2018)。短いコーディング遺伝子の最も大規模なカテゴリーは、upstream Open Reading Frames (uORFs) であり、現在アノテーションされている遺伝子の約 30% が、uORF を持つと予測されている (Andrews and Rothnagel 2014; Hellens et al. 2016)。uORF は、主要なコーディング遺伝子の 5 末端の非翻訳領域に存在し、自身がリボソームに翻訳されることで、下流に存在する主要なコーディング遺伝子の翻訳制御を担っている。しかしながら、uORFs から翻訳されたペプチド自体には、多くの場合で機能的役割は存在しない。一方で、micro RNA (miRNA) の上流に存在する一部の uORFs は、miRNA-encoded peptides (miPEP) として定義され、外因性のペプチド添加によって miRNA の発現を誘導することが検証されていることから、コーディング遺伝子の uORFs とは異なり、機能的役割を有するペプチドであると認識されている (Lauressergues et al. 2015)。miPEP は、バイオインフォマティ

表1 短いコーディング遺伝子の機能カテゴリー

機能カテゴリー	役 割	ペプチドとしての機能	アノテーションされている数
uORF	翻訳の抑制	無し	コーディング遺伝子の約 30%
miPEP	miRNA のエンハンサー	有り	50 個
MicroProtein	タンパク質相互作用の抑制	有り	20 個
ホルモン様 ペプチド	細胞間のシグナル伝達	有り	200 個
ディフェンシン	病原菌に対する抗生機能	有り	298 個

クス解析によって、少なくとも 50 個のアノテーションされている miRNAs を発現誘導することが示唆されている。MicroProtein は、単独のドメインだけを持つペプチドであり、標的ペプチドと相互作用することによってドミナントネガティブ効果を与えるというタンパク質レベルで miRNAs に類似した役割を担っている。MicroProtein は、機能ドメインの類似性検索によってゲノム規模での同定が可能であり、シロイヌナズナのデータベースである TAIR10 では、約 20 個の MicroProtein がアノテーションされている (Guen et al. 2015)。このように、uORF、miPEP や MicroProtein として分類されている短いコーディング遺伝子は、細胞内のシグナル経路の制御に必要な不可欠な役割を果たしている。TAIR10 の詳細は、以下の URL (<https://www.arabidopsis.org/>) を参照されたい。

uORFs に次ぐ大規模なカテゴリーとして、ホルモン様ペプチド、あるいは、ディフェンシンがある。ディフェンシンは、バクテリア、真菌類、ウイルスに対して活性を持つ抗微生物ペプチドとして幅広い生物種に存在し、298 個もアノテーションされている。しかし、植物のディフェンシンには、ホルモン様ペプチドとして機能することが認識されているものが次々と見つかった。そのため、ディフェンシンとホルモン様ペプチドを明確に区別することは、困難である。固着生物である植物は、急激な変化や段階的な変化といった法則性のない環境ストレスを顕著に受けるため、細胞間のシグナル伝達が特に重要である (Takahashi and Shinozaki 2019)。ホルモン様ペプチドは、シグナル伝達を担う物質の一つであり、約 200 個もの植物ホルモン様ペプチドがシロイヌナズナでアノテーションされている。また、ホルモン様ペプチドの受容体の多くは、受容体キナーゼであることが判明している。シロイヌナズナでは、458 個の受容体キナーゼが同定されているが、現在では、33 個しかホルモン様ペプチドを受容する証拠が見つかっていない (Lehti-Shiu et al. 2009)。そのため、未同定なホルモン様ペプチドが存在すると予測されている。一方で、ホルモン様ペプチドは、高い配列の多様性を有するため、ホルモン様ペプチドをコードする可能性がある短いコーディング遺伝子を推定するだけでは、ホルモン様ペプチドの同定には繋がらない。このような点から、ホルモ

ン様ペプチドは、他のカテゴリーと比較して同定が困難である。

短いコーディング遺伝子を新規に探索する方法

短いコーディング遺伝子は、上記のような多種多様な役割を担っていることが明らかとなっているが、一般的な遺伝子予測プログラムでは、このような遺伝子は推定されにくい傾向にあり、機能解析に着手すること自体が困難であるという研究背景を持つ。ここでは、花田によって開発されたアルゴリズム (sORFfinder) によって短い遺伝子を推定する方法を紹介する (Hanada et al. 2010)。sORFfinder の利用は、以下の URL (<http://hanadb01.bio.kyutech.ac.jp/sORFfinder/>) を参照されたい (図 1)。

短い ORF 配列 (F) に対し、ベイズ法を用いた以下の式を用いて配列 F がコーディング領域、コーディングでない領域のそれぞれの事後確率を推定する。

$$P(\text{coding} | F) = \frac{P(F|\text{coding}) \cdot P(\text{coding})}{P(F)}$$

$$= \frac{P(F|\text{coding}) \cdot P(\text{coding})}{P(F|\text{coding})P(\text{coding}) + P(F|\text{intron})P(\text{intron})}$$

$P(F|\text{coding})$ は、コーディング領域での F の生起確率であり、シロイヌナズナに存在するコーディング遺伝子の配列パターンを基にマルコフモデルを用いて推定される。 $P(F|\text{intron})$ は、イントロン領域での F の生起確率でイントロンの配列パターンを基にマルコフモデルを用いて推定される。ある ORF がコーディング領域であるかの判断には、実際のイントロンの配列パターンから仮想的な ORF を作り出し、擬陽性が 1% 以下になるように $P(\text{coding}|F)$ の閾値を設定している。

この方法がどの程度正確にコーディング遺伝子領域を推定できるかを確認するため、実際のコーディング遺伝子から 100 aa 以下の ORF またはイントロン内で偶然作られた 100 aa 以下の ORF を用いて検証している。その結果、アノテーションされている遺伝子の 81.5% が正確に同定されている一方で、イントロンに偶然に作られた ORF を 3% の割合で誤ってコーディング遺伝子と同定してしまうことが判明した。このように本方法は短いコーディング遺伝子を同定するためには、非常に有用であることが明らかにされて



sORFfinder(small Open Reading Frames finder)

- [Home](#)
- [Methods](#)
- [Web-application](#)
- [Download](#)
- [Manual for local analyses](#)
- [Authors](#)
- [Links](#)

sORFfinder is a web-based tool to identify small open reading frames (sORF) with high coding potential. However, web-application has some limitations. You can conduct more sophisticated analysis in your local computer downloading the perl scripts. The limitation of web-application is written in the [methods](#). Web-based program has two applications. First, you can identify sORFs which have similar nucleotide composition of coding genes by hidden markov model (Search_sORF). Then, the functionality of sORFs can be examined by comparative genomics (Examine_SP). Please see [Application Note](#) of this program package published in Bioinformatics.

1. Search_sORF can identify small open reading frames (sORF) with coding potential by hidden markov model. Web-application had already prepared studying models of 11 species (*S.cerevisiae*, *A.thaliana*, *O.sativa*, *C.elegans*, *D.melanogaster*, *A.gambiae*, *C.intestinalis*, *X.laavis*, *D.rerio*, *G.gallus* and *H.sapiens*). You should choose one of studying model from 11 species. Using the selected model, you can search sORF with high coding potential. However, if you want to use studying model for other organisms, you need to conduct the analysis in your local computer.
2. Examine_SP can infer functionality of identified sORF by comparative genomics. First, this program searches homologous sequences from available EST or genomic sequences and calculated synonymous (without amino acid replacements) and non-synonymous (with amino acid replacements) nucleotide substitutions. If sORF is functional, it is likely that synonymous substitutions have occurred more frequently than non-synonymous substitutions. Finally, in each sORF, you can get P-values which can be indicator for functionality.

Please send e-mail to me if you have problems or errors.

Kousuke Hanada

Kyushu Institute of Technology, Department of Bioscience and Bioinformatics

E-mail: kohanada (at) kyutech (dot) ac (dot) jp

図1 sORFfinderのWEBページ

sORFfinder (<http://hanadb01.bio.kyutech.ac.jp/sORFfinder/>) は、短いコーディング遺伝子を新規に推定する目的で作られたWEBアプリケーションである。二つのWEBアプリケーションを持っており、ひとつは、既知のコーディング配列パターンと非コーディング配列パターンを機械学習方法で学習し、ベイズ法を利用して small Open Reading Frame (sORF) と呼ばれる短いコーディング領域を探索するアプリケーション (Search_sORF) である。もう一つは、比較ゲノム解析の方法を利用して、sORF 領域と相同な配列を他の生物種ゲノムから探索し、アミノ酸を変化させる塩基置換とアミノ酸を変化させない塩基置換を比較するアプリケーションである。見出された sORF が、複数の生物種間において、機能的であるとする、アミノ酸を変化させない塩基置換率 (Ks) が変化させる塩基置換率 (Ka) より頻繁に起こっていると期待できる。さらに、Ka と Ks の頻度を統計的に比較することで、機能的な sORF を探索することができるアプリケーション (Examine_SP) がある。

いる (Cheng et al. 2011)。

さらに、花田らは、sORFfinder をシロイヌナズナゲノムに適用することで、7,159 個の新奇の短いコーディング遺伝子の候補を発見し、そのうち、2,376 個は機能を維持する方向で進化していることを見出した (Hanada et al. 2007)。この方法で見出された 7,159 個の短いコーディング遺伝子の候補のうち 844 個は、シロイヌナズナのゲノムデータベース (TAIR) において、研究当時の TAIR6 から現行のバージョンである TAIR10 へと更新される間に、短いコーディング遺伝子としてアノテーションされている。しかしながら、アノテーションされているコーディング遺伝子 (120 aa 以上) の 87% は機能的な情報が存在する一方で、短いコーディング遺伝子 (120 aa 以下) は 32% しか機能的な情報が存在しない。このことは、短いコーディング遺伝子の機能的役割が判明しつつあるが、一般的な遺伝子と比べて研究が進んでいない領域であることを強く示唆している (図2)。

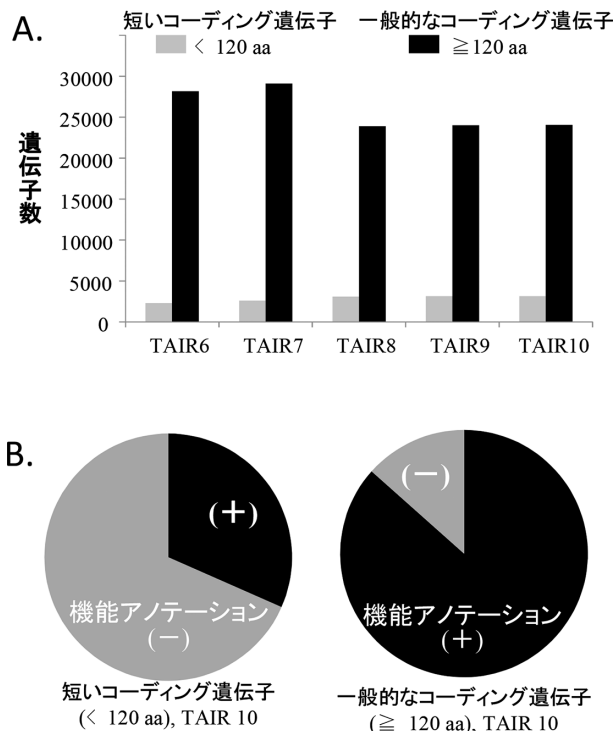


図2 The Arabidopsis Information Resource (TAIR) に登録されている短いコーディング遺伝子

シロイヌナズナのゲノムデータベース (TAIR) は、アップデートしており、TAIR6 は 2006 年 8 月、TAIR7 は 2007 年 4 月、TAIR8 は 2008 年 4 月、TAIR9 は 2009 年 7 月、TAIR10 は 2010 年 12 月に公開されている。現在は、TAIR10 のままで固定されており、ここで公開されている遺伝子情報をアノテーション情報と呼称している。A. TAIR6 から TAIR10 に至るまでの短いコーディング遺伝子 (120 aa 未満) と通常のコーディング遺伝子 (120 aa 以上) で、アノテーションされているコーディング遺伝子数を示している。B. TAIR10 でアノテーションされている遺伝子のうち、短いコーディング遺伝子 (120 aa 未満) と通常のコーディング遺伝子 (120 aa 以上) での機能アノテーションされている割合を示す。

推定された短いコーディング遺伝子の機能解析の方法 ・トランスクリプトームデータの利用

遺伝子機能を把握するためには、トランスクリプトームデータは必要不可欠である。例えば、環境ストレス耐性に関わる遺伝子を推定するためには、様々な環境ストレス条件下におけるトランスクリプトームデータを解析し、各環境ストレスで発現変動する遺伝子機能を調査することで、様々な分子シグナルに関与する遺伝子を効果的に見出すことが可能となる。実際に、著者らは、塩耐性に関わるホルモン様ペプチドを探索するために、高塩ストレス下で発現が上昇する短いコーディング遺伝子に着目し、複数の塩耐性を付与するホルモン様ペプチドを見出している (Nakaminami et al. 2018)。しかしながら、現在公開されているトランスクリプトームデータの多くは、一部の短いコーディング遺伝子の発現プロファイルしか観察できない。そのた

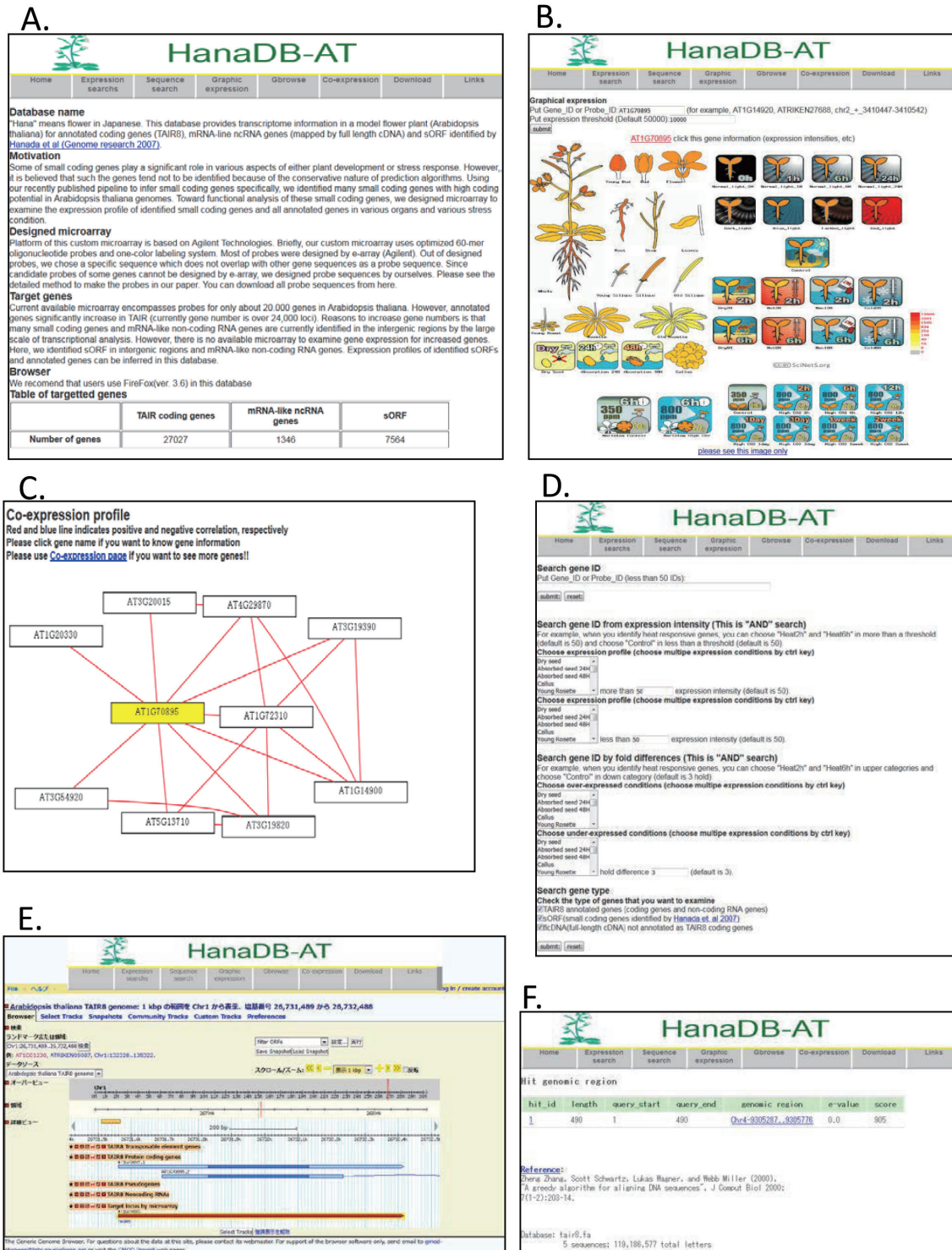


図3 短いコーディング遺伝子に特化したトランスクリプトームデータベース (HanaDB) の紹介

A. <http://hanadb01.bio.kyutech.ac.jp/hanadb-at> にアクセスすると現れるトランスクリプトームデータベースの入口, B. 発現強度や発現部位を視覚的に捉えることができる WEB アプリケーション, C. 共発現解析を行う WEB アプリケーション, D. 任意の発現プロファイルの検索する機能, E. ゲノムブラウザでのゲノム位置を確認する機能, F. 相同性検索の WEB アプリケーションが存在する.

め、花田らが推定した短いコーディング遺伝子の候補のうち、アノテーションされていないものから、様々な分子シグナルに関与するものを見出すには、自身でトランスクリプトーム解析を実施する必要がある。そこで、花田らは、推定した短いコーディング遺伝子の候補の発現変動も観察することが可能なカスタムマイクロアレイを利用したトランスクリプトームデータベースを構築した (Hanada et al. 2013)。短いコーディング遺伝子のデータベースは、HanDB-AT (<http://hanadb01.bio.kyutech.ac.jp/hanadb-at/>) を参照されたい (図 3A)。本データベースは、17 種の環境条件 (定常状態、低温 2/6 時間、高塩 2/6 時間、乾燥 2/6 時間、高温 2/6 時間、暗条件下からの光暴露 0/1/6/24 時間、青色光、赤色光、遠赤色光、暗所) および 16 種の組織 (0/24/48 浸水させた種、カルス、未成熟/成熟/老化したロゼット葉、茎生葉、茎、根、未成熟/成熟した蕾、花、未成熟/成熟/老化した莢) において、短いコーディング遺伝子および TAIR に登録されている遺伝子の発現強度や発現部位を視覚的に捉えることが可能となっている (図 3B)。さらに、共発現解析 (図 3C)、任意の発現プロファイルの検索 (図 3D)、ゲノムブラウザ (図 3E) および相同性検索 (図 3F) の機能もあり、発現情報から短いコーディング遺伝子の探索に特化したデータベースが利用できる。

・過剰発現体の利用

花田らが、推定した 7,159 個の新奇の短いコーディング遺伝子の候補の多くは、TAIR にアノテーションされていない。そこで、花田らは、7,159 個の候補のうち、473 個を無作為に選択し、過剰発現変異体を構築すると共に、表現型解析による機能スクリーニングを実施した (Hanada et al. 2013)。その結果、473 個のうち、約 10% (49/473 = 10.4%) が、ベクターコントロール (図 4A) と比較して

明白な形態異常を誘発することが判明した (図 4B, C)。一方で、完全長 cDNA として登録されているシロイヌナズナの遺伝子から無作為に過剰発現変異体を構築した場合では、1.4% しか明白な形態異常を誘発しない (Ichikawa et al. 2006)。つまり、花田らが推定した新奇の短いコーディング遺伝子の候補は、比較して、約 7 倍も形態異常をもたらすことを意味する。このことから、花田によって推定された短いコーディング遺伝子の候補は、形態形成に関与する遺伝子が多く含まれていることが示唆された。さらに、表現型効果を有する短いコーディング遺伝子候補の機能的役割を推論するために、花田らは、形態異常をもたらした短いコーディング遺伝子の候補と共発現する遺伝子の機能カテゴリーを調査した。その結果、新奇短いコーディング遺伝子の候補は受容体キナーゼと共発現する傾向があることが判明した。リガンドとなるホルモン様ペプチドおよび受容体キナーゼのペアが、共発現解析によって同定されたという例もあることから、花田が推定した新奇短いコーディング遺伝子の候補には、ホルモン様ペプチドが含まれている可能性が非常に高いことが予想される (Sugano et al. 2010)。実際に、花田らが推定した短いコーディング遺伝子の候補から塩耐性を植物体に付与するホルモン様ペプチドも発見されており、この可能性は尤もらしいといえる (Nakaminami et al. 2018)。

おわりに

花田は、モデル植物のシロイヌナズナにおいて、遺伝子推定が困難な短いコーディング遺伝子の候補を 7,159 個見出した。これらの候補は、機能を維持する方向で進化していると共に、過剰発現変異体を構築した際に顕著な形態異常をもたらすことから、本推定法は高精度で短いコーディング遺伝子を見出すことが可能であることが期待でき

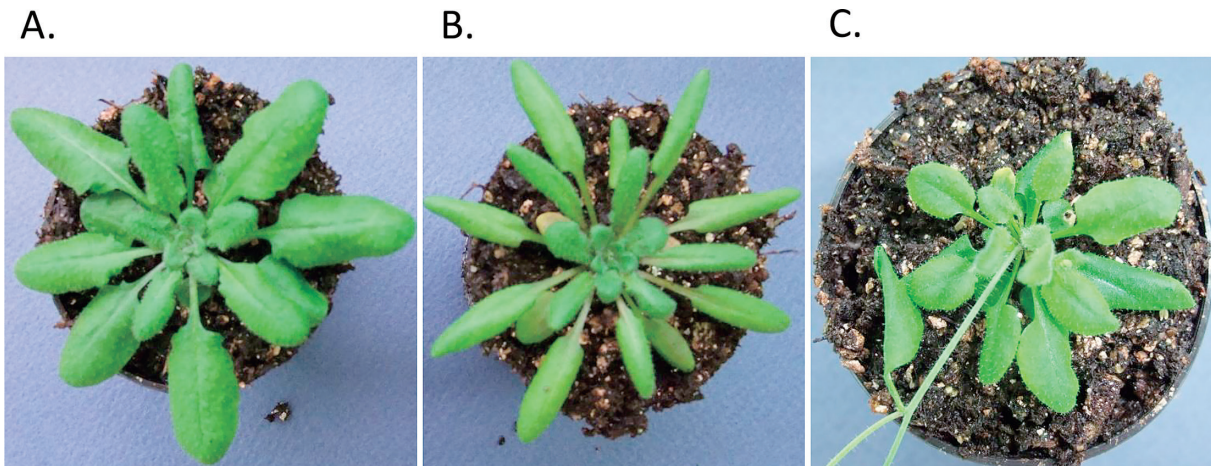


図 4 過剰発現変異体を用いた網羅的な機能スクリーニングの例
新奇短い遺伝子の CDS 領域をカリフラワーモザイクウイルス由来 35S promoter 制御による過剰発現変異体を構築し、定常条件下 (明条件: 16 h, 暗条件: 8 h, 温度: 22 度一定) にて生育した際における形態異常の例。A. Vector control. B. 葉形状の形態異常。C. 矮小化の形態異常。

る。さらに、本推定法によって見出した短いコーディング遺伝子にホモログな配列がイネに存在することも判明しており、実際には、幅広い植物種に短いコーディング遺伝子が未だ未同定な状態で存在していることが予想される (Okamoto et al. 2014)。また、過剰発現変異体にて形態異常を示した短いコーディング遺伝子の多くは、受容体キナーゼと共発現していることや、これまでの報告にあるホルモン様ペプチドをコードする短い遺伝子は、変異体の構築により形態異常をもたらす傾向にあることを踏まえると、過剰発現体を用いた網羅的な機能スクリーニングは、同定が困難なホルモン様ペプチドの探索手法として非常に有用である可能性が高い。

現在、花田によって推定された短いコーディング遺伝子の候補の過剰発現変異体は、継続して構築を行っている。さらに、本研究室に連絡を取ることによって変異体植物の分与も開始しており、誰でも短いコーディング遺伝子の研究にとりかかることが可能となっている。

謝辞

本研究で紹介した著者らの研究の多くは、理化学研究所の樋口美恵子博士、中南健太郎博士、清水みなみ氏、宇都宮大学の岡本昌憲博士、高崎健康福祉大学の吉積 毅博士とともに生物系特定産業技術研究支援センターの研究支援の補助を得て行いました。改めて、御礼を申し上げます。

文 献

- Andrews, SJ and Rothnagel, JA (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* **15**: 193-204.
- Cheng, H, Chan, WS, Li, Z, Wang, D, Liu, S and Zhou, Y (2011) Small open reading frames: current prediction techniques and future prospect. *Curr Protein Pept Sci* **12**: 503-507.
- Eguen, T, Straub, D, Graeff, M and Wenkel, S (2015) MicroProteins: small size-big impact. *Trends Plant Sci* **20**: 477-482.
- Hanada, K, Zhang, X, Borevitz, JO, Li, WH and Shiu, SH (2007) A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. *Genome Res* **17**: 632-640.
- Hanada, K, Akiyama, K, Sakurai, T, Toyoda, T, Shinozaki, K and Shiu, SH (2010) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**: 399-400.
- Hanada, K, Higuchi-Takeuchi, M, Okamoto, M, Yoshizumi, T, Shimizu, M, Nakaminami, K, Nishi, R, Ohashi, C, Iida, K, Tanaka, M, Horii, Y, Kawashima, M, Matsui, K, Toyoda, T, Shinozaki, K, Seki, M and Matsui, M (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc Natl Acad Sci U S A* **110**: 2395-2400.
- Hellens, RP, Brown, CM, Chisnall, MAW, Waterhouse, PM and Macknight, RC (2016) The emerging world of small ORFs. *Trends Plant Sci* **21**: 317-328.
- Hsu, PY and Benfey, PN (2018) Small but mighty: Functional peptides encoded by small ORFs in plants. *Proteomics* **18**: e1700038.
- Ichikawa, T, Nakazawa, M, Kawashima, M, Iizumi, H, Kuroda, H, Kondou, Y, Tsubura, Y, Suzuki, K, Ishikawa, A, Seki, M, Fujita, M, Motohashi, R, Nagata, N, Takagi, T, Shinozaki, K and Matsui, M (2006) The FOX hunting system: an alternative gain-of-function gene hunting technique. *Plant J* **48**: 974-985.
- Lauressergues, D, Couzigou, JM, Clemente, HS, Martinez, Y, Dunand, C, Bécard, G and Combier, JP (2015) Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**: 90-93.
- Lehti-Shiu, MD, Zou, C, Hanada, K and Shiu, SH (2009) Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. *Plant Physiol* **150**: 12-26.
- Nakaminami, K, Okamoto, M, Higuchi-Takeuchi, M, Yoshizumi, T, Yamaguchi, Y, Fukao, Y, Shimizu, M, Ohashi, C, Tanaka, M, Matsui, M, Shinozaki, K, Seki, M and Hanada, K (2018) AtPep3 is a hormone-like peptide that plays a role in the salinity stress tolerance of plants. *Proc Natl Acad Sci U S A* **115**: 5810-5815.
- Okamoto, M, Higuchi-Takeuchi, M, Shimizu, M, Shinozaki, K and Hanada, K (2014) Substantial expression of novel small open reading frames in *Oryza sativa*. *Plant Signal Behav* **9**: e27848.
- Sugano, SS, Shimada, T, Imai, Y, Okawa, K, Tamai, A, Mori, M and Hara-Nishimura, I (2010) Stomagen positively regulates stomatal density in Arabidopsis. *Nature* **463**: 241-244.
- Takahashi, F and Shinozaki, K (2019) Long-distance signaling in plant stress response. *Curr Opin Plant Biol* **47**: 106-111.
- Tavormina, P, De Coninck, B, Nikonorova, N, De Smet, I and Cammue, BP (2015) The plant peptidome: An expanding repertoire of structural features and biological functions. *Plant Cell* **27**: 2095-2118.

連絡先: 〒 820-8502 福岡県飯塚市川津 680-4
九州工業大学 情報工学部 生命情報工学科
花田 耕介
TEL: 0948-29-7842
FAX: 0948-29-7801
E-mail: kohanada@bio.kyutech.ac.jp