厚生労働科学研究費補助金
食品の安全確保推進研究事業

# 新たなバイオテクノロジーを用いて得られた食品の安全性確保とリスクコミュニケーションのための研究

令和５年度　総括･分担研究報告書
【２１ＫＡ１００２】

研究代表者　近藤　一成

令和６（２０２４）年　３月

目　次

# Ⅰ．　総括研究報告書

厚生労働科学研究費補助金（食品の安全確保推進研究事業）
新たなバイオテクノロジーを用いて得られた食品の安全性確保と
リスクコミュニケーションのための研究

# 総括研究報告書

研究代表者　近藤 一成　（昭和女子大学）

**研究要旨：**

リスクコミュニケーション分野では、合成生物学やフードテックと呼ばれる方法で作られた食品が注目を集めてが、本年度は昆虫食、植物由来代替肉、培養肉、代替乳製品（人工乳、精密発酵乳）、微細藻類の５つの代替タンパク質に着目してアンケートを実施した。培養肉、精密発酵乳の認知度が低い事が分かった。５つの代替タンパク質に関するオンラインセミナーの結果を基に平易な冊子を作成した。代替タンパク質の日本での社会実装が進む際にリスクコミュニケーションの教材となることが期待される。

ゲノム解析分野では、ゲノム編集食品の安全性評価の一つである外来遺伝子配列のゲノム上における残存を網羅的に調べる方法として、次世代シークエンサーを利用して得られた全ゲノムシークエンスデータを用いた標準的解析手法の開発に取り組んだ。本年度は、実際に残存が想定されうる外来遺伝子配列（Cas9 配列など）について手法の妥当性を検討すると共に、このアセンブリ法によって解析可能な残存配列の長さや解析で必要とされる全ゲノムシークエンスにて取得すべきデータ量（シークエンスカバレッジ）を明らかにした。ゲノム編集食品の安全性評価の精緻化・向上のみならず、こうした新たなバイオテクノロジーを活用した食品に対する国民受容の向上にも役立つと期待される。

網羅的代謝物解析分野では、昨年度までの化合物単位での解析に加え、本年度は分析データからエンリッチメント解析を通じて代謝パスウェイにおける変動を明らかにする検討を行った。開発したスペクトル類似度計算に基づく未知化合物の構造解析および可視化を行う解析ツールを、ウェブブラウザ上で動作する Docker イメージとして配布、GitHub で公開した。これにより、未知化合物の迅速な解析と可視化という従来高度な質量分析とインフォマティクス技術が必要だった解析が広範な研究者および技術者にも利用可能となった。

アレルゲン性予測の分野では、既に開発したサポートベクターマシンを用いた手法である allerSTAT について、客観的性能評価のために F1 スコア、MCC で評価した結果、既存のツールよりも優れていることが確認できた。また、食物由来タンパクの主要組織適合性複合体 HLA への結合性を予測する手法を検討して、既存深層学習モデルをベースに追加の特徴量を組み合わせたトレーニングを行い、予測性能を比較、性能向上の可能性と有効な特徴量を検証した。

本研究班は、研究代表者を含む以下の６名から構成され各分担課題について研究を行った。

研究分担者　安達 玲子　（国立医薬品食品衛生研究所）
研究分担者　柴田 識人　（国立医薬品食品衛生研究所）
研究分担者　小泉　望　（大阪公立大学）
研究分担者　早川 英介　（沖縄科学技術大学院大学）
研究分担者　富井 健太郎　（産業技術総合研究所）

以下に、研究目的、方法、および研究成果の概要を記載する。研究内容の詳細については、各分担報告書に記載した。

## A. 研究目的

遺伝子改変技術を応用した食品開発は、技術的には外来遺伝子導入による遺伝子組換え食品（GM 食品）から生物自身が持つ内在性遺伝子改変で新たな形質を生み出すゲノム編集技術応用食品（ゲノム編集食品）へ、また、その生物が持たない多数の遺伝子を導入した酵母などから新規食品機能成分を産生させる合成生物学利用も諸外国を中心に進んでいる。ゲノム編集技術では DNA 2 本鎖切断を誘導するオリジナル手法から、1 本鎖切断から 1 塩基編集を行う塩基置換編集（Base editing)、これを発展させ数塩基の自由な組合せの塩基編集（prime editing)、さらに標的配列への制限をなくした PAM レス(PAM 配列を要求しない)編集、RNA 編集など非常に多様な手法が生み出され、そこから想定される意図しない変化も一様でないことが明らかになりつつある。したがって、配列に依存しない意図しない塩基変化やそこから生じる代謝成分の変化を網羅的に検出または予測し、その変化が与える影響を正確に評価することは、食品の安全性確保において急務の課題である。また、これらゲノム編集食品や合成生物学利用食品に加えて、細胞培養によって作成される肉、魚、乳、チョコレートなど細胞性食品の研究開発が急速に進んでいる。ますます、リスクコミュニケーションの重要性が増している。国民受容とともに製品開発や普及が並行して進むことで、規制精度の整備、国民理解と受容、イノベーション推進が進むことが望まれる。

本研究では、(１) 多様な遺伝子改変技術と開発に関する情報収集、(２) 一様でない意図しない変化の影響解析のための手法開発（ゲノム、代謝成分、アレルゲン性)、(３) ゲノム編集食品の理解の前段階として不可欠な国内 GM 食品利用の現状と審査届出制度の理解に重点したリスクコミュニケーション（若手研究と連携)、(４) リスク評価側の最新技術理解と能力向上、人材育成を柱に若手研究代表者とも連携して実施する。

## B. 研究方法、分担

本研究班構成では、意図しないゲノム DNA 配列の変化の解析手法開発と標準化を柴田が、意図しないタンパクの生成に伴うアレルゲン性の評価手法開発と実用化およびアレルゲンデータベース ADFS の維持更新を、深層学習も取り入れながら安達および富井が、また、意図しない代謝物変化の網羅的開発手法の開発と Web 環境で利用できるような実用化を早川が担当した。リスクコミュニケーションについては、ゲノム編集食品、合成生物学利用食品、特に代替タンパク質に重点を置きながら小泉が担当した。

研究方法の詳細は分担報告書を参照のこと。

## C. 研究成果

本研究では、新しいタイプの食品（ゲノム編集食品、細胞性食品等）に対するリスクコミュニケーション、および、意図しない変化を検出するための解析手法の開発を行っている。

リスクコミュニケーション分野では、一般消費者（モニター）を対象とした代替タンパク質に関するグループインタビューおよびオンラインセミナーを実施した。各代替タンパク質の認知度は昨年度の 5000 人を対象とした WEB アンケートを反映する結果となり、昆虫食や微細藻類もクロレラなどの認知度は高く、培養肉や代替乳（精密発酵）については初めて聞く、想像ができないといった意見が主流であった。アンケート調査では代替乳の認知度は 30%ぐらいあったが、豆乳やアーモンドミルクなどとの誤認があった可能性が高い。遺伝子組換えで牛乳成分となるタンパク質を生産

するといった説明については違和感が示された。しかし、オンラインセミナーなどを通じてサステナビリティへの貢献などが説明されると、大きな受容度は大きく上がる傾向があった。

ゲノム解析分野では、外来遺伝子残存の解析に必要な、遺伝子再構成について、NGS 測定時のシークエンスカバレッジの再検証を行った。アセンブリ法では、RRS2 系統にて挿入されている外来遺伝子全長（4,314 塩基）を 1 本で再構成するには 25×以上のカバレッジが必要であることが疑似データ（シュミレーション）で示唆されていたが、今回実際に 25×で測定したデータで検証した。その結果、ショートリードシークエンスによる WGS データであれば約 30×程度のシークエンスカバレッジがあれば 4,000 塩基程度の外来遺伝子配列を正確に検出できることが確認された。ゲノム編集に由来する外来遺伝子配列の検出では、疑似データを用いて検討した結果、Cas9 全長配列を挿入した配列をもとにした疑似 WGS データでは、10×カバレッジであれば全ての試行で検出に成功していたが、カナマイシン耐性遺伝子配列では 15×、Cas9 部分配列では 20×のカバレッジがそれぞれ全ての試行で検出できる最低カバレッジであった。また、配列長に関する要件検討です、20～30×のカバレッジが全ての試行で検出できる最低カバレッジであった。

網羅的代謝物解析分野では、ゲノム編集トマトのメタボローム解析では、特定のアミノ酸代謝パスウェイ、特に Phenylalanine, Tyrosine, Aspartate および Biotin metabolism が顕著にエンリッチされていることが明らかになった。GABA は植物のストレス応答に深く関与し、GABA の過剰生産はストレス応答による二次代謝物の生合成に影響を及ぼすことから、Phenylalanine や Tyrosine を前駆体とする stress defense 関連の二次代謝物の合成にも影響を与えると推測される。開発した解析ツールは、Docker を採用することで、プラットフォームに依存せず、インフォマティクスの専門知識がない研究者や技術者でも容易に利用できるユーザーフレンドリーなインターフェースを提供することができた。

アレルゲン性予測の分野では、遺伝子改変技術応用食品のアレルゲン性について、より高い精度での評価・予測を可能とすることを目的とした、アレルゲン性予測手法(allerStat)の性能検証及び機能拡充を行い、良好な予測性能を確認できた。アレルゲンデータベース ADFS の維持管理では、令和 4 年 6 月から令和 5 年 5 月までの 1 年間に NCBI PubMed に収載された論文から、エピトープ配列決定に関する 20 報のピアレビューを行い、22 種のアレルゲンについて、総数 87 のエピトープ情報を ADFS に追加した。HLA 結合性については、既存予測法の DeepSeqPanII をベースモデルとして、AAindex を特徴量に加えることで改良した分子間相互作用予測法を本予測法に活用するため、既存予測手法の一つである NetAllergen 中の分子間相互作用予測部分を改良した相互作用予測法で置き換え、予測性能に及ぼす影響をベンチマークデータセットを用いて評価し、オリジナルと同等の結果が得られることが分かった。

## D. 健康危険情報

該当なし

研究業績、知的財産権の出願などは、各分担報告書を参照。

# Ⅱ．　分担研究報告書

厚生労働科学研究費補助金（食品の安全確保推進研究事業）
「新たなバイオテクノロジーを用いて得られた食品の安全性確保と
リスクコミュニケーションのための研究」
分担研究報告書（令和５年度）

# リスクコミュニケーションに関する研究

研究分担者　小泉　望　　大阪公立大学　教授

研究要旨：
ゲノム編集食品とは別に合成生物学やフードテックと呼ばれる方法で作られた食品が注目を集めている。本研究ではこうした食品の中でも代替タンパク質に着目した。より具体的には昆虫食、植物由来代替肉、培養肉、代替乳製品（人工乳、精密発酵乳）、微細藻類の５つの代替タンパク質を対象とした。前年度までに国内外の研究開発動向の調査と消費者の量的、質的意識調査を行った。5,000 人を対象とした調査では昆虫食については最もネガティブに捉えられている。海外では目まぐるしく代替タンパク質の研究開発が進んでいるが、日本では産業界を中心とした取り組みが活性化しつつあるものの諸外国と比べると限定的である。グループインタビューによる定性調査の結果からは培養肉、精密発酵乳に関してはアンケート結果より認知度が低いと考えられた。５つの代替タンパク質に関するオンラインセミナーの結果を基に平易な冊子を作成した。代替タンパク質の日本での社会実装が進む際にリスクコミュニケーションの教材となることが期待される。日本での認知度が最も低く、社会実装に関する法整備も進んでいない精密発酵乳に特に焦点を当てて、海外の開発動向調査を実施するとともに一般モニターの主婦および食に関する高い専門知識を持つ主婦それぞれ３名が参加するグループインタビューを別々に実施した。前者はサステイナブルと言った言葉により行動変容を起こし易い。後者の議論では官民が関与する取り組みの重要性が示された。

## A. 研究目的

新たなバイオテクノロジーによって作られる食品としては、除草剤耐性や害虫抵抗性を持つ遺伝子組換え農作物に由来する食品（遺伝子組換え食品）が主流であり、1996 年に実用化が始まり既に 25 年以上利用されている。遺伝子組換え微生物由来の添加物（ビタミンや精製酵素など）も 2001 年から使用されている。遺伝子組換え技術を用いた食品と添加物を比較すると前者ではリスクコミュニケーションが困難な状況が続いている。諸外国でも同様で、単なる科学的な説明に加えて ELSI（倫理的・法的・社会的課題）が関与し、コミュニケーションを複雑にしている。一方、添加物については今のところ、そうした混乱は見られない。

2021 年秋に我が国においてゲノム編集技術応用食品（ゲノム編集食品）の実用化が始まった。具体的にはGABA高蓄積トマト、可食部増量マダイ、高成長トラフグである。2023 年には高成長ヒラメの上市も認可された。ゲノム編集食品に関するリスクコミュニケーションはその実用化の数年前から複数の機関により始まった。行政、国の研究機関、大学、NPO などである。開発者により設立されたベンチャー企業もコミュニケーション活動に積極的に関与した。遺伝子組換え食品では市場に導入されてからコミュニケーションが始まったのに対してゲノム編集食品では実用化の前に対応がとられたこともあり、現状では比較的冷静なリスクコミュニケーションが行われている。

既存の遺伝子組換え食品、ゲノム編集食品に続き新しいバイオテクノロジーを用いた食品が登場

しつつある。その製造法や性質、用途が多岐に渡るため一般的な呼称は定着していないが「合成生物学（Synthetic biology：Synbio）」、「フードテック」と言った用語が使われることが多い。しかし、Synbio、フードテックの概念はかなり漠然としており、分類のされ方も様々である。両方の要素を持つ食品も少なくないがイコールではない。遺伝子組換え技術あるいはゲノム編集食品技術が使われることもあるが、一概に遺伝子組換え食品と同列に扱うのは適当でない場合が多い。Synbio とフードテックの両方の概念を併せ持ち遺伝子組換え技術を使った食品の例として、米国の Perfect Day 社が開発した乳製品（実際に乳製品と呼べるかどうかは議論の余地があるが）が挙げられる。同社では、牛乳の主要なタンパク質（ホエー）を遺伝子組換え技術により微生物で生産し、代替乳（定着した用語ではない）を生産しアイスクリームを模した食品を製造、販売することに成功している。このような製造方法は「精密発酵」と呼ばれることが多い。精密発酵由来食品はすでに米国等で実用化されており、動物に由来しないのでビーガンの人にも受け入れられている。

この乳製品に限らず「ポスト新たなバイオテクノロジーによって作られた食品」とでもいうべき食品を本研究では「新たなコンセプトで作られた食品」と呼ぶ（世間一般に認知された用語ではないことに注意）。しかし、前述のように「新たなコンセプトで作られた食品」は本研究で扱おうとする食品は幅広いことから明確な定義づけは容易ではない。また海外では Novel food、Innovative food といった呼ばれ方をすることもある。今後の日本での呼称については検討の余地があろう。

新たなコンセプトで作られた食品の特徴として、全てでは無いが「アニマルフリー」、「生物資源の保護」などのいわゆる遺伝子組換え作物の特徴である効率性とは違うコンセプトが挙げられる。さらにアニマルフリーは環境負荷の低減、動物愛護などの異なる観点からもとらえられる。環境負荷の低減は家畜の飼育による温室効果ガス排出量、水使用量、エネルギー使用量の軽減を意味する。例えば牛のげっぷは温室効果ガスのかなりの部分を占める。生物資源の保護の例としては微生物でのバニリン（バニラの香気成分）生産などがあげられる。この場合、バニリンの合成に関わる複数の酵素の遺伝子が導入されており新たな代謝系が構築されたといえる。こうした方法は、合成生物学、Synbio と呼ばれるが狭義な合成生物学の厳密な定義には当てはまらない。

また、世界人口の増加と生活レベルの向上に伴う肉食の増加によるタンパク質クライシスが懸念されている。そのため代替肉の研究開発も盛んである。代替肉は大きくは主に大豆を原料としたダイズミート（エンドウなどが使われる場合もある）に代表される植物由来代替肉と細胞培養で牛などの細胞を培養し、それを成型する培養肉に大別される。植物由来代替肉は特にハンバーガーなどを中心にすでに国内外で実用化されている。米国では Impossible foods 社と Beyond meat 社が有名である。前者は遺伝子組換え技術を使いダイズのレグヘモグロビンを酵母で生産し、添加物として使用している。レグヘモグロビンの添加により、動物肉（主に牛肉）が含む血液の風味が加えられているとされる。国外（主に米国）ではすでに実用化されているが、国内では遺伝子組換え技術を用いていることから安全性審査が求められ実用化のハードルは高いと考えられる。

以上、例を挙げてきたような新たなコンセプトで作られた食品が主として国外で次々と開発され実用化されているが、その国内における認知度は低いと考えられる。近い将来こうした食品が国内でも流通する可能性は十分予想され、ホライズン・スキャニングの考え方に基づき新たなコンセプトで作られた食品の安全性について効果的なリスクコミュニケーション手法を開発することは円滑な厚生労働行政に資すると考えられる。

2021 年度は①新たなコンセプトで作られた食品の国内外の事例調査、②新たなコンセプトで作られた食品に対する多様なステークホルダーの受

け止め方の調査、③代替タンパク質に対する 5,000 人規模の意識調査をおこなった。

2022 年度はフードテック、Synbio 全てを対象とすると扱う食品の種類が膨大になることや現在世間で注目を集めていること、2021 年度に大規模意識調査の対象としたことなどから「代替タンパク質」に対象を絞り、そのリスクコミュニケーションに関する調査研究を行った。具体的には①代替タンパク質に関する国内外の事例調査、②代替タンパク質に対する 5,000 人の意識調査結果の解析（意識調査自体は 2021 年度に実施）、③代替タンパク質に関する消費者を対象としたグループインタビュー、④ゲノム編集食品の効率的な情報提供方法を検討した。

2023 年度は 2022 年度に引き続き代替タンパク質に焦点を絞り、①代替タンパク質に関する一般消費者を対象としたグループインタビュー、オンラインセミナー、②代替タンパク質に関する冊子作成、③食に対する知識が豊富な層を対象とした精密発酵に関するグループインタビュー④代替タンパク質に関する海外の事例調査（特に精密発酵に着目）、⑤冊子「ゲノム編集食品について話す」の発行とゲノム編集食品に関するリスクコミュニケーションを行った。尚、厳密には 2022 年度末に調査研究を実施したため 2022 年度（令和 4 年度）の報告書に記載出来なかった内容を含む。

**B. 研究方法**

<u>①一般消費者（モニター）を対象とした代替タンパク質に関するグループインタビューおよびオンラインセミナー</u>

2023 年 5 年 2 月および 3 月に一般消費者を対象とした代替タンパク質に関するグループインタビューを行った。参加者は楽天インサイトに登録している普段から料理をし食に関心を持つ 30 代から 50 代の主婦とし、毎回異なるパネラーが 3 名参加した。グループインタビューでは専門家、3 名のパネラーに加えて 1 名のファシリテーターが参加し、具体的には以下の 6 回を設定した。

2 月 27 日「新食材“代替タンパク質”はなぜ求められているのか」
専門家：五十嵐圭介（東北大学大学院 助教／日本細胞農業協会 代表理事）
3 月 1 日「細胞農業と培養肉」
専門家：杉崎麻友（Forsea Foods（イスラエル）研究員／日本細胞農業協会 理事）
3 月 2 日「タンパク質源としての微細藻類」
専門家：佐々木俊弥（(株）タベルモ COO）
3 月 3 日「消費者目線の精密発酵」
専門家：橋詰寛也（(株）Kinish CEO）
3 月 13 日「植物由来代替肉の現状と今後」
専門家：穴井豊昭（九州大学 教授）
3 月 14 日「地域課題解決と養殖昆虫食の共進化」
専門家：佐伯真二郎（NPO 法人食用昆虫科学研究会 理事長／JICA 草の根技術協力事業プロジェクトマネージャー）

グループインタビューの構成は
・趣旨説明、アイスブレイク（5 分程度）
・専門家による簡単な代替タンパク質の紹介（5 分程度）
・グループインタビュー（前半：70 分）
・休憩（10 分）
・専門家による情報提供（オンラインセミナーとして公開）（50 分）
・グループインタビュー（後半：40 分）
の計 180 分（3 時間）と設定した。

<u>②代替タンパク質に関するオンラインセミナーを基にした冊子の作成</u>

①で行ったオンラインセミナーを録画、文字起こしを行い、その内容を簡潔にまとめた冊子を作成した。

<u>③食に対する知識が豊富な層を対象とした精密発酵に関するグループインタビュー</u>

フードテックに詳しい石川伸一教授（宮城大学）が情報提供を行い、①のグループインタビューで

情報提供を行った橋詰寛也氏（Kinish CEO）がファシリテーターを務め、いずれも 50 代の主婦で食品安全や制度に詳しい消費生活アドバイザー、生協職員、大学教員が参加するグループインタビューを行った。

④代替タンパク質に関する海外の事例調査

主として「Foovo」（https://foodtech-japan.com/）から各種の情報を取得し整理した。特に精密発酵に着目して情報収集を行った。

⑤冊子「ゲノム編集食品について話す」の発行とゲノム編集食品に関するリスクコミュニケーション。

冊子「ゲノム編集食品について話す」を発行し、サイエンスカフェなどに利用した。リスクコミュニケーション活動の具体例は以下の通り。「ゲノム編集食品について」お米の勉強会、「ゲノム編集技術」（株）江崎グリコ、「遺伝子組換え食品とゲノム編集食品の最近の動向」大阪いずみ市民生活協同組合。

**C. 研究結果および考察**

①一般消費者（モニター）を対象とした代替タンパク質に関するグループインタビューおよびオンラインセミナー

グループインタビューについては主として精密発酵に関して記載する。

各代替タンパク質の認知度は 5000 人を対象とした WEB アンケートを反映する結果であり、植物由来代替肉は大豆ミートとしてよく知られていた。昆虫も受容の是非は別にして、2022 年 11 月に学校給食にコオロギパウダーが使用されたというニュースがあったこともあり、良く認知されいた。微細藻類もクロレラなどはかなり以前から耳にすることが多く、初めて聞くという参加者は居なかった。一方、培養肉や代替乳（精密発酵）については初めて聞く、想像ができないといった意見が主流であった。

アンケート調査では代替乳の認知度は 30%ぐらいあったが、豆乳やアーモンドミルクなどとの誤認があった可能性が高い。遺伝子組換えで牛乳成分となるタンパク質を生産するといった説明については違和感が示された。しかし、オンラインセミナーなどを通じてサステナビリティへの貢献などが説明されると、大きな受容度は大きく上がる傾向があった。例えば国の承認があるなど、安全性が担保されているのであれば、価格との相談もあるが試してみたいという意見もあった。オピニオンリーダー／インフルエンサーや知り合いの行動も購買行動に影響することも伺えた。

オンラインセミナーについては②で作成した冊子にまとめて書いているのでごく簡潔にポイントを記載する。そもそもタンパク質は人間にとって重要な栄養素である。代替タンパク質が登場した背景としてはいわゆるタンパク質危機（プロテインクライシスがある）が挙げられる。植物由来代替タンパク質として主に使われている大豆はプロテインスコアの観点からも優れたタンパク源である。微細藻類は生産性には優れているがサプリメント等の利用が多くいわゆる食事としての利用は限定的である。昆虫食はラオスでは主要なタンパク源であり、日本とは食文化の違いが大きい。人工肉は細胞性食品とも呼ばれ、細胞を培養することで作る。代替乳もやはり細胞性食品であるが、培養肉が細胞をそのまま食するのと異なり、代替乳は細胞からタンパク質を精製する。また、培養肉は遺伝子組換え技術を使わないが代替乳は使う。もっとも、最終製品には組換え遺伝子は含まれない。

②代替タンパク質に関するオンラインセミナーを基にした冊子の作成

①で行った 6 回のオンラインセミナーをもとに A5 判、16 ページの図表をふんだんに盛り込んだ冊子を作成した（表紙のみ図 1）。

PDF として小泉の研究室のホームページからダウンロード可能である（図 1）。

https://www.omu.ac.jp/agri/pmb/assets/ap-pamphlet_web.pdf
関係者に冊子体を郵送したところ、Foovo セミナー（2024 年 1 月開催）で参加者に冊子体が配布された。また食の信頼向上をめざす会
https://sites.google.com/view/mezasukai/
から講演依頼があった。

今後、本冊子のリスクコミュニケーションへの活用が期待される。

③食に対する知識が豊富な層を対象とした精密発酵に関するグループインタビュー

精密発酵による代替乳自体の安全性に関しては官の積極的な関与が必要、社会受容に関しては民の自主的な行動が求められる。例えばカニカマは模倣食品だが受容されている。複数の企業が協力して取り組みことが重要。規制に関しては官民が協議していく必要がある。SDGs に代表されるサステイナブルという観点は一見、受容にポジティブに働くように見えるが酪農家とのせめぎあい等の問題点も多い。ニュージーランドは酪農先進国であるとともに環境保護を大切にしており精密発酵を推進している。日本でも国が推進するかどうかも重要なポイント。日本でもやはり、酪農との関連は複雑な問題。表示は重要な懸案事項。

④代替タンパク質（特に精密発酵に注目）に関する海外の事例調査

主として「Foovo」(https://foodtech-japan.com/) から各種の情報を取得した。
精密発酵以外では培養肉の認可、社会実装が 2023 年度の大きなニュースである。具体的には

- 国連食糧農業機関（FAO）と世界保健機関（WHO）による培養肉の安全性に関する新たなレポートの発表（2023 年 4 月）。
- 米国 GOOD Meat 及び UPSIDE Foods が、米国農務省から培養肉の販売許可を取得（2023 年 6 月）。
- UPSIDE Foods が米レストランで培養鶏肉の販売を実現（2023 年 7 月）。

こうした動きは培養肉が社会実装に向けて一歩進んだフェーズに移ったことを示す。

精密発酵（代替乳）に関しては以下が注目のニュースである。

- イスラエルの精密発酵企業 Remilk はシンガポール食品庁（SFA）から販売認可を取得したことを発。シンガポールでの認可取得と共に、アメリカ食品医薬品局（FDA）から「異議なし」のレター（No questions letter）を受け取ったことも発表（2023 年 2 月）。
- Remilk はカナダ保健省から「異議なし」のレターを受け取り、カナダで β-ラクトグロブリン（ホエータンパク質）の使用・販売が可能になったことを発表（2023 年 2 月）。
- ユニリーバは精密発酵乳タンパク質を使用したチョコレートアイスクリームの発売を発表。ユニリーバのアイスクリームブランド Breyers として発売され、米パーフェクトデイのホエイタンパク質を使用（2023 年 2 月）。
- Remilk は 27 日、イスラエル保健省から精密発酵タンパク質の認可を取得（2023 年 4 月）。

こうして見てくと Remilk の社会実装への動きが目立つ。米パーフェクトデイと並び他社とは一線を画しているように思われる。

⑤冊子「ゲノム編集食品について話す」の発行とゲノム編集食品に関するリスクコミュニケーション。

「ゲノム編集食品について話す」と名付けた冊子を発行した。3 名が会話しながらゲノム編集食品について学ぶ形式になっている。構成は、ゲノム編集食品とは何か？遺伝子の変化のさせ方、食品の種類、食品としての安全性、管理のありかたに関するポイントを平易にまとめている。小泉の研究室のホームページからダウンロード可能である。
https://www.omu.ac.jp/agri/pmb/assets/GEbooklet_web.pdf

また本冊子は日本植物生理学会のアウトリーチ活動のためのコンテンツとして当学会のホームページで紹介されている。
https://jspp.org/society/new_tech/newtech_link.html

2023年度はゲノム編集食品に関する3回の講演をおこなった。セグメントによって必要とする情報は異なっており、効果的なリスクコミュニケーションには十分なデータを保持しておく必要があることが改めて確認された。

## D. まとめ

2022 年度末に実施したオンラインセミナーをもとに5つの代替タンパク質を平易に説明する冊子を作成した。今後の代替タンパク質に関するリスクコミュニケーションに貢献できると考えている。日本で社会実装が進んでおらず認知度も低い代替乳（精密発酵乳）を中心に海外の研究開発動向の調査を行った。さらに調査会社にモニター登録している主婦あるいは食に関する専門性の高い知識を持つ主婦が参加する代替乳（精密発酵乳）に関するグループインタビューを行った。日本における代替乳（精密発酵乳）の社会受容性は小さく無いと考えられたが表示を含む規制に関してはクリアすべき問題点が多いことが示された。代替乳（精密発酵乳）の国内での社会実装には時間がかかるかもしれないが、リスクコミュニケーションのための材料を前もって準備しておくことが重要であると考えられる。

## E. 研究発表・業績

1. 論文発表
   1) 小泉望　ゲノム編集食品をどう伝えるか：生活協同組合研究．577, 34-41（2024）
   2) Shneha R., Takeda K.F., Yamaguchi Y., Koizumi N. A comparative analysis of attitudes towards genome-edited food among Japanese public and scientific community, PLOS ONE (in press)（2024）

2. 学会発表
   1) 標葉隆馬、武田浩平、小泉望「代替タンパク質をめぐる ELSI」社会科学技術論学会 2023 年 12 月 9 日、大阪大学（豊中市）
   2) 武田浩平、小泉望、標葉隆馬、「5種類の代替タンパク質に関する一般市民の態度」社会科学技術論学会　2023 年 12 月 10 日、大阪大学（豊中市）

3. 書籍
   1)小泉望 リスクコミュニケーションのために求められること：ゲノム編集技術 ～実験上のポイント／産業利用に向けた研究開発動向と安全性周知．p295-p301　情報機構（2023）
   2) 小泉望、四方雅人 ゲノム編集食品に関する取扱いルールの経緯とこれから：ゲノム編集技術を応用した製品開発とその実用化．p565-572　技術情報協会（2023）

4.講演
   1) 小泉望「代替タンパク質に対する消費者の姿勢」（一財）バイオインダストリー協会 2023 年 5 月 18 日（東京都）
   2) 小泉望「ゲノム編集食品について」お米の勉強会　2023 年 10 月 4 日（西宮市）
   3) 小泉望「ゲノム編集技術」（株）江崎グリコ　2023 年 11 月 14 日（大阪市）
   4) 小泉望「遺伝子組換え食品とゲノム編集食品の最近の動向」大阪いずみ市民生活協同組合　2023 年 12 月 8 日（オンライン）
   5) 小泉望「代替タンパク質について」食の信頼向上をめざす会　2024 年 1 月 22 日（オンライン）

## F．健康危険情報

該当なし

## G. 知的財産権の出願・登録状況

該当なし

厚生労働省科学研究費補助金（食品の安全確保推進研究事業）
「新たなバイオテクノロジーを用いて得られた食品の安全性確保と
リスクコミュニケーションのための研究」
分担研究年度終了報告書（令和 5 年度）

# 多様な遺伝子改変技術から生じる意図しない変化の網羅的解析手法の開発と環境整備に関する研究

研究分担者　柴田　識人　　国立医薬品食品衛生研究所

研究要旨：
本研究では、ゲノム編集食品の安全性評価の一つである外来遺伝子配列のゲノム上における残存を網羅的に調べる方法として、次世代シークエンサーを利用して得られた全ゲノムシークエンスデータを用いた標準的解析手法の開発に取り組んでいる。これまでのモデルサンプルを利用した検討で、残存が想定される配列が予め分かっていれば、全ゲノムシークエンスデータをアセンブリ解析（以下、アセンブリ法）に供することで、残存の有無、その場所や配列の内容を明らかにできることが示唆されている。今年度は、ゲノム編集食品で実際に残存が想定されうる外来遺伝子配列（Cas9 配列など）について残存性評価における本手法の妥当性を検討すると共に、このアセンブリ法によって解析可能な残存配列の長さや、解析で必要とされる全ゲノムシークエンスにて取得すべきデータ量（シークエンスカバレッジ）を明らかにした。また、解析環境の妥当性確認のための標準シークエンスデータを構築した。さらに同様に外来遺伝子配列の残存性を評価できる k-mer 法との性能比較を実施し、網羅的かつ効果的な評価スキームの提唱に至った。こうした全ゲノムシークエンスによる外来遺伝子配列の残存性を評価する一連のスキームを標準的な安全性評価の一つにすることは、ゲノム編集食品の安全性評価の精緻化・向上のみならず、こうした新たなバイオテクノロジーを活用した食品に対する国民受容の向上にも役立つと期待される。

研究協力者
曽我　慶介　（国立医薬品食品衛生研究所）
成島　純平　（国立医薬品食品衛生研究所）
杉野　御祐　（国立医薬品食品衛生研究所）

## A. 研究目的

我が国では 2019 年よりゲノム編集食品の事前相談・届出制度が開始されており、これまでに 6 品種がゲノム編集食品として届出・公表され、一部はすでに流通されている。この制度では、届出の対象となるゲノム編集食品の備えるべき要件の一つとして、ゲノム中に外来遺伝子を全く含まないことを求めているが、「ゲノム編集技術応用食品等の取扱いに関する留意事項」ではこうした外来遺伝子およびその一部が残存していないことを調べる方法として、次世代シークエンサー（NGS）による全ゲノムシークエンス（WGS）、サンガーシークエンス、サザンブロット法などを挙げている。このうち、得られるデータの網羅性や、ゲノム編集技術によって生じる意図しない遺伝子変化全般の解析にも適用できる汎用性などを考え、NGS による解析が着目されており、その解析結果の届出情報への活用が今後増加すると予想される。但し解析の必要要件や標準的手順が明確になっていないことから、NGS 解析を公定試験法とする上で検討すべき課題となっている。

本研究ではゲノム編集食品の外来遺伝子の有無を調べる標準的な NGS 解析手法の確立を試み、昨年度アセンブリ法の妥当性と必要要件などを報告した。そこで今年度はゲノム編集食品で実際に残存が想定される外来遺伝子およびその解析条件における本解析の妥当性などを検討した。

**B. 研究方法**

1. サンプル

外来遺伝子残存モデルサンプルとして、除草剤耐性遺伝子組換えダイズ RRS2 系統の認証標準物質である粉砕試料 Monsanto MON89788 Soybean Powder 994.0g 超/kg（AOCS; 0906-B）を購入した。

2. ゲノム抽出

ダイズ粉末試料約 1 g から NucleoBond® HMW DNA (MACHEREY-NAGEL) によりゲノム DNA を抽出。イソプロパノール沈殿後、滅菌蒸留水 100 µL で溶解したところ、ゲノム DNA 溶液が白濁していたため、DNeasy Blood & Tissue Kit (QIAGEN) で精製した。滅菌蒸留水で溶解し、50 µL のゲノム DNA 溶液（合計 95 µg ゲノム DNA 含有）を得た。

3. ショートリードシークエンサーによる WGS

精製したダイズゲノム DNA 200 ng を用いて Illumina DNA Prep, (M) kit（Illumina）により、シークエンスライブラリを調製。Qubit Fluorometer (Thermo Fisher Scientific) および Tape Station 4150 (Agilent) により定量・定性を行い、最終添加濃度は 750 pM とした。シークエンスは NextSeq 2000 を用いて、NextSeq 1000/2000 P1 Reagents (300 cycles) (Illumina) による 150 bp × 2 のペアエンドシークエンスを実施した。得られたリードデータのクオリティは「FastQC (ver. 0.11.9)」で確認し、リードデータに含まれるシークエンスアダプター配列および低クオリティリードデータは「Trim Galore! (ver. 0.6.7) 」を用いて除去した。

4. シミュレーションデータの生成

ゲノム編集技術による標的ゲノム配列の改変では部位特異的ヌクレアーゼが使用されることから、ゲノム編集食品で残存する可能性の高い外来遺伝子配列は、この部位特異的ヌクレアーゼおよびその導入に使用されたバックボーンプラスミドである。しかしながら、外来遺伝子が残存したゲノム編集食品の入手は困難であることから、こうした残存する可能性の高い配列を含むシミュレーションデータを作成することにした。現在開発が進むゲノム編集食品の大部分は化膿性連鎖球菌の Cas9 ヌクレアーゼ（SpCas9）を利用していることから、プラスミドバンク Addgene（https://www.addgene.org/）に寄託されている SpCas9 配列を含む植物向けゲノム編集ベクター pDGE64 (#79446) の全長またはその部分配列をダイズリファレンスゲノム (Glycine_max_v2.1.dna.toplevel.fa) の 1 番染色体（chr1: 24,617,123）へ挿入することで人工ゲノム配列を作成し、それをもとに疑似 WGS データを生成した。ショートリードシークエンスのシミュレーションには「ART (ver. 2.5.8) 」 (Huang W, et al. Bioinformatics. 28(4):593-4, 2012.)を使用し、HiSeqX PCR free (150 bp) ペアエンド、インサートサイズ 600 ± 10 bp のデータを約 60×カバレッジで生成した（12 併行）。ロングリードシークエンスのシミュレーションには「PBSIM3 (ver. 3.0.0)」を使用し、昨年度に Oxford Nanopore Technologies 社の PromethION を用いて遺伝子組換えダイズ RRS2 系統を WGS した際のシークエンスプロファイル（エラー率など）をもとに、約 60×の疑似 WGS データを生成した。生成データは全てクオリティ値 9 以上であった。

5. WGS データを用いた外来遺伝子の再構成

遺伝子組換えダイズ RRS2 系統で挿入されている遺伝子の配列情報は、遺伝子組換え食品の遺伝子配列データベース Nexplorer（https://bioit-webapp-prod.sciensano.be/nexplorer/）より取得した。pDGE64 の配列は Addgene より取得した。これらを参照配列と考え、ショートリードシークエンスデータでは「BWA (ver. 0.7.17-r1188)」の MEM アルゴリズムを、ロングリードシークエンスデータでは「Minimap2 (ver. 2.21-r1071)」を用いてマッピングを行った。その後「Samtools (ver. 1.13)」view コマンドの -F オプションを 4 に指定してマッピングされなかったデータを除くと同時に、BAM変換を行い、sortコマンドでソート、bam2fq

コマンドでBAMファイルからFASTQファイルへ変換することで参照配列にマッピングしたリードデータの抽出を行った。ショートリードシークエンスデータのアセンブリは、「SPAdes (ver. 3.15.5)」を用いた。ロングリードシークエンスデータのアセンブリには「Flye (ver. 2.9)」を用いた。Flye ではナノポアシーケンサーのエラー率 5%以下を想定した--nano-hqオプション、およびポリッシング回数を指定するオプション-i 5 にて実施した。マッピング状況は、Integrative Genomics Viewer (IGV; https://www.igv.org/）を用いて視覚的に確認した。

6. ランダムサンプリング

シークエンスリードデータのランダムサンプリングには「Seqkit (ver. 2.0.0)」の Sample コマンドを使用し、50, 40, 30, 25, 20, 15, 10, 5×シークエンスカバレッジになるよう-p のオプション値を設定した。各カバレッジとも 12 併行のリードデータをマッピングさせ、12 回全てで外来遺伝子の再構成が成功する最低カバレッジを検証した。

なおショートリードシークエンスデータについてはマッピングしたリード数のばらつきを調べるため、Smirnov-Grubbs 検定による外れ値検定を行った。ロングリードシークエンスデータについては各リードの長さが異なるため、外れ値検定は行わなかった。

7. k-mer 解析

k-mer 解析ツールは伊藤らが開発した「KmerAnalysis 2.3.1」（Itoh et al. Sci. Rep. 10, 4914, 2020.）を使用した。解析に当たっては、コントロールとして外来遺伝子の残存がない野生型の WGS データが必要であったため、ダイズリファレンスゲノム (Glycine_max_v2.1.dna.toplevel.fa) より ART で約 60×の疑似 WGS データをシミュレーションした後、「Seqkit」の Sample コマンドを使用してランダムサンプリングを実施した。

**C. 研究結果**

1. 外来遺伝子の再構成に必要なシークエンスカバレッジの再検証

<u>WGS データの概要</u>

前年度の検討から、ショートリードシークエンスによる WGS データを用いたアセンブリ法では、RRS2 にて挿入されている外来遺伝子の全長（4,314 塩基）を 1 本で再構成するには 25×以上のカバレッジが必要であることが示唆された。ただしこのアセンブリ法に供する WGS データの必要要件に関する検討では、約 60×の WGS データをランダムサンプリングすることで疑似的にシークエンスカバレッジを落としたデータでの検討であったため、最初から 25×程度のシークエンスカバレッジしか有しない WGS データでもアセンブリ法が本当に有効か不明である。そこで 25×程度のシークエンスカバレッジとなるよう改めて WGS を実施し、アセンブリ法に供した。シークエンスに当たっては、今年度新たに DNA ライブラリを用意した。表 1 に取得したシークエンスデータの統計値を示す。結果、全体的なリードクオリティは良好であり、アダプターおよび低クオリティリード除去後のシークエンスカバレッジは約 34×であった。なおダイズゲノムサイズは 1.1Gb としてシークエンスカバレッジを算出した。

<u>外来遺伝子配列へのマッピングとそのアセンブリ</u>

図 1 に RRS2 にて挿入されている外来遺伝子配列の概略図およびその配列へのマッピング様相を示した。また表 1 にマッピングしたリードの統計値およびマッピングリードを SPAdes でアセンブリして得られたコンティグの統計値を示す。アセンブリの結果、5 コンティグ（総計 5,546 塩基）が生成された。生成された 5 コンティグについて、挿入されている外来遺伝子配列に再度マッピング

した結果、1 コンティグ 4,540 塩基のみがマッピングされた（図 2）。このコンティグは外来遺伝子配列 4,314 塩基の全長をカバーしており、また配列も正確に再現していた。このことから、前年度のランダムサンプリングでの検討で示唆された通り、ショートリードシークエンスによる WGS データであれば約 30×程度のシークエンスカバレッジがあれば 4,000 塩基程度の外来遺伝子配列を正確に検出できることが確認された。

2. ゲノム編集に由来する外来遺伝子配列の検出

疑似 WGS データの概要

まず pDGE64（図 3）全長 16,159 塩基、Cas9 全長配列 4,104 塩基およびその部分配列 102 塩基、そして薬剤選抜に使用するカナマイシン耐性遺伝子配列 795 塩基をダイズリファレンス配列へ挿入し、この配列をもとに疑似 WGS データをシミュレーションした。

ショートリードシークエンスによる WGS データのシミュレーションには ART を使用し、各配列が挿入された 50×以上の疑似 WGS データを生成した。生成した疑似 WGS データの統計値を表 2 に示す。なおいずれの配列においても疑似データの生成は 12 併行で実施し、外れ値検定を実施した。外れ値として判定されたデータについては除外した上で、残りのデータセットで再度外れ値検定を実施し、外れ値として判定されないデータセットを用意した。

ロングリードシークエンスによる WGS データのシミュレーションには PBSIM3 を使用した。疑似データの生成は、pDGE64 全長配列と Cas9 全長配列では 4 併行で、カナマイシン耐性遺伝子配列と Cas9 部分配列では 1 併行で実施した。生成した疑似 WGS データの統計値を表 2 に示す。

各シークエンスカバレッジにおけるアセンブリ法の実施（ショートリードシークエンス）

生成した疑似 WGS データを pDGE64 配列へマッピングした様子を図 4 に示す。生成した各疑似 WGS データをそれぞれ 50～5×へとダウンサンプリングした。そのマッピングリード数と統計値は表 3 と図 5 に示す。カバレッジの減少とともにマッピングしたリード数も直線的に減少しており、相関係数の値はいずれも 0.99 以上であることから、ランダムサンプリングが適切に実施されたと考えられる。

次に、各カバレッジにおいて pDGE64 にマッピングしたリードを用いたアセンブリを試み、生成したコンティグのうちの 1 本で外来遺伝子配列全体を再現できるかを検出の指標として検討を行った。その結果、挿入した外来遺伝子配列が短くなるにつれ検出に必要なカバレッジは高くなる傾向が見られた。具体的には、pDGE64 全長配列や Cas9 全長配列を挿入した配列をもとにした疑似 WGS データでは、10×カバレッジであれば全ての試行で検出に成功していたが、カナマイシン耐性遺伝子配列では 15×、Cas9 部分配列では 20×のカバレッジがそれぞれ全ての試行で検出できる最低カバレッジであった（表 4）。

各シークエンスカバレッジにおけるアセンブリ法の実施（ロングリードシークエンス）

生成した疑似 WGS データを用い、pDGE64 全長配列、Cas9 全長配列、カナマイシン耐性遺伝子配列、Cas9 部分配列にそれぞれマッピングされたリード（図 6 左参照）を抽出し、pDGE64 全長配列に Minimap2 を用いてマッピングした（図 7）。ショートリードシークエンスの疑似 WGS データによる結果（図 4）と比較すると、ロングリードシークエンスによる疑似 WGS データでは正確性が低いことがわかる。マッピングリード数に関するデータを表 5 に示した。全リード（シークエンスカバレッジ 53.4×）を用いた場合、マッピングされたリードを用いてアセンブリ法を行うことでそれぞれの配列の全体を正確に再現するコンティグを生成することができた（表 6）。

次にダウンサンプリングによってカバレッジ数を落としたリードデータを用いて、pDGE64 全長配列、Cas9 全長配列、Cas9 部分配列のそれぞれにマッピング、アセンブリを行い、必要カバレッジの検討を行った。マッピングリード数に関するデータを表 5 に示した。pDGE64 全長配列では 4 併行のデータについて検討したが、12 回のダウンサンプリング全てで正確なアセンブリまで成功した最低カバレッジは 15×、20×（n=2）、40×とデータ間で差が大きかった（表 6）。Cas9 全長配列では 4 併行のデータについて検討したが、20×、25×、30×（n=2）であった（表 6）。Cas9 部分配列では、ダウンサンプリングしたリードデータを用いてアセンブリした場合、40×データを用いても 12 回のダウンサンプリングの一部で挿入した外来遺伝子配列を含むコンティグを生成することができなかった（検出率は10/12（表6））。また、10×以下のカバレッジではアセンブリが成功しても一部配列で正確に配列を再現できないこともあった。

3. ゲノム編集に由来する外来遺伝子配列の検出における配列長に関する必要要件の検討

疑似 WGS データの概要

先述の検討より、本アセンブリ法がゲノム編集で残存する可能性のある外来遺伝子配列（Cas9 配列など）の検出に有効であることが分かった。そこで配列長に関する検出限界を調べるため、Cas9 のヌクレアーゼ活性ドメインである HNH ドメインの部分配列 21, 31, 41, 51, 101 塩基の 5 パターンの配列をダイズリファレンス配列へ挿入し、この配列をもとに疑似 WGS データをシミュレーションした（ロングリードシークエンスによる疑似 WGS データのシミュレーションでは 31 塩基の配列の挿入は未実施）。

ショートリードシークエンスによる WGS データのシミュレーションには ART を使用し、約 54×の疑似 WGS データを生成した。生成した疑似 WGS データの統計値を表 7 に示す。いずれの配列においても疑似データの生成は 12 併行で実施し、一部外れ値として判定されたデータについては再度データ生成をやり直し、外れ値として判定されないデータセットを用意した。

ロングリードシークエンスによる WGS データのシミュレーションには PBSIM3 を使用した。生成した疑似 WGS データの統計値を表 7 に示す。

各シークエンスカバレッジにおけるアセンブリ法の実施（ショートリードシークエンス）

生成した疑似 WGS データを pDGE64 配列へマッピングした様子を図 8 に示す。生成した各疑似 WGS データはそれぞれ 50～5×へとダウンサンプリングし、そのマッピングリード数と統計値は表 8 と図 9 に示す。31～101 塩基を挿入したデータセットにおいては、カバレッジの減少とともにマッピングしたリード数も直線的に減少しており、相関係数の値はいずれも 0.99 以上であることから、ランダムサンプリングが適切に実施されたと考えられる。なお、21 塩基を挿入したデータにおいては、BWA の MEM アルゴリズムでは pDGE64 配列にマッピングすることができなかった。

次に各カバレッジにおいて pDGE64 にマッピングしたリードを用いたアセンブリを試み、生成したコンティグのうちの 1 本で外来遺伝子配列全体を再現できるかを検出の指標として検討を行った。その結果、マッピングリードが抽出できた 31～101 塩基が挿入された配列をもとにした疑似 WGS データでは、挿入された外来遺伝子配列長とその検出に必要な最低カバレッジの間に明確な相関関係は見られず、およそ 20～30×のカバレッジが全ての試行で検出できる最低カバレッジであった（表 9）。なお今回検討した中で、検出可能な最短の挿入配列長は 31 塩基であり、その検出には少なくとも 25×カバレッジ以上の疑似 WGS データが必要であった。

各シークエンスカバレッジにおけるアセンブリ法の実施（ロングリードシークエンス）

生成した疑似 WGS データを用い、pDGE64 全長配列にマッピングされたリード（図 6 右参照）を抽出し、pDGE64 全長配列に Minimap2 を用いてマッピングした（図 10）。51, 101 塩基の配列が挿入された疑似 WGS データでは pDGE64 全長配列にマッピングされたリードが得られたが、21, 41 塩基が挿入された配列をもとにした疑似 WGS データではマッピングリードが得られなかった（表 10）。なお、101 塩基が挿入された配列をもとにした疑似 WGS データでは、この 101 塩基の挿入配列に対してマッピングされたリードが抽出できたが、51 塩基が挿入された配列をもとにした疑似 WGS データでは、この 51 塩基の挿入配列に対してマッピングリードは得られなかった（図 6）。

次に各カバレッジにおいて pDGE64 にマッピングしたリードを用いたアセンブリ法を試み、生成したコンティグのうちの 1 本で挿入した外来遺伝子配列全体を再現できるかを指標に、挿入した外来遺伝子配列を検出できるか検討した。101 塩基が挿入された配列をもとにした疑似 WGS データでは、53.4×カバレッジのデータから抽出されたマッピングリード（図 10）を用いたアセンブリ法を試み、挿入配列を検出することができた。さらにこの疑似 WGS データをダウンサンプリングしたところ、25×カバレッジのデータでも 12 回の試行全てで挿入配列を検出することができた（表 11）。他方で、51 塩基が挿入された配列をもとにした疑似 WGS データでは、53.4×カバレッジのデータから抽出されたマッピングリード（図 10）を用いたアセンブリ法では挿入配列を検出することができたものの、この疑似 WGS データをダウンサンプリングした場合には、40×カバレッジのデータを用いても 12 回の試行中 9 回しか挿入配列を検出することができなかった（表 11）。

4. ゲノム編集に由来する外来遺伝子配列の k-mer 法による検出

Cas9 のヌクレアーゼ活性ドメインである HNH ドメインの部分配列をダイズリファレンス配列へ挿入し、この配列をもとにシミュレーションした疑似 WGS データ（ショートリードシークエンス）での検討より、本アセンブリ法で検出可能な最短の配列長は 31 塩基であった。同様に WGS データを用いた外来遺伝子の残存を解析する方法として、k-mer 法が報告されている。性能を比較するため、同じデータセットを用いて k-mer 法による解析を行った。解析には HNH ドメインの部分配列である 21 または 31 塩基が挿入された配列をもとにした疑似 WGS データ（それぞれ n=12）および野生型の疑似 WGS データ（n=1）を供した。解析ではどちらも 30×カバレッジにダウンサンプリングしたデータセットを使用した。その結果、21 または 31 塩基が挿入された配列をもとにした疑似 WGS データともに 12 試行全てで塩基配列の挿入を検知（G statistics > 6.634）した（図 11）。

**D. 考察**

本研究では、ゲノム編集食品に残存する外来遺伝子配列の有無を調べる標準的な NGS 解析手法の確立を目的としている。これまでの検討から、残存する可能性のある配列が予め分かっていれば、NGS による WGS データをアセンブリ法に供することで、残存の有無、その場所や配列の内容を明らかにできることを見出している。さらに昨年度の検討から、WGS データ（ショートリードシークエンスデータ）として 25×以上のシークエンスカバレッジデータが必要であることが示唆された。しかしこの検討は約 60×のシークエンスカバレッジを有する WGS データをダウンサンプリングしたデータセットを用いたシミュレーション実験の結果であるため、実サンプルとしてこの程度のシークエンスカバレッジを有する WGS データにお

いて本アセンブリ法による残存配列の検出が可能なのか、その妥当性は不明であった。今年度この点の検証を試みた。モデルサンプルとして遺伝子組換えダイズ RRS2（4,314 塩基が外来遺伝子配列として挿入されている）に対してショートリードシークエンスによる WGS を実施し、約 34×のシークエンスカバレッジを有する WGS データを得たが、このデータを用いたアセンブリ法により、挿入されている外来遺伝子配列の全長を正確に再現することができた。すなわち、ショートリードシークエンスによる WGS データを利用したアセンブリ法によって外来遺伝子配列を検出する際、約 30×以上のシークエンスカバレッジを有するデータが必要であることの妥当性が示された。なお、残存している外来遺伝子配列として宿主生物にも内在する配列が含まれる場合や、マルチコピーな配列が含まれる場合も想定されるが、リード長が 150 塩基程度のショートリードシークエンスではこうした配列を正しく区別して解析することができないことが示唆されている。ロングリードシークエンスによる WGS であれば 1 リードあたりのリード長が十分に長いので、こうした課題を解決できる可能性がある。今後の検証課題として挙げられる。

昨年までの検討では、外来遺伝子配列が残存しているモデルケースとして遺伝子組換え作物を利用していたが、こうしたモデルサンプルで挿入されている配列は除草剤耐性遺伝子や害虫殺虫活性タンパク質をコードする遺伝子などであり、ゲノム編集食品の作製過程において残存する可能性のある遺伝子配列ではない。ゲノム編集食品の開発では Cas9 のような部位特的ヌクレアーゼやその導入に使用されたバックボーンプラスミドが残存する可能性がある。そこで本研究課題で見出されたアセンブリ法のゲノム編集食品への適用を見据えて、こうした配列の全体または一部を疑似的に挿入した WGS データセットをシミュレーションして、アセンブリ法の有効性を検討した。その結果 Cas9 の塩基配列や抗生剤耐性遺伝子配列等を含むプラスミド配列であっても、アセンブリ法を用いることで、実際に挿入された配列を特定できることを確認できており、挿入されたゲノム上の位置ついても特定することが可能と考えられる。また検出に必要なシークエンスカバレッジは、前年度のモデルサンプルでの検討と同様に、30×程度であれば正確に検出できることを見出した。以上よりゲノム編集食品における外来遺伝子配列の残存性を評価する WGS データの解析法としてアセンブリ法が有効であることが分かった。

ゲノム編集技術による遺伝子改変では、部位特的ヌクレアーゼの細胞内導入に使用したプラスミド配列の一部がゲノムに残存されるケースも想定される。昨年度までの検討における課題の一つとして、このように残存する可能性のある配列の一部のみがゲノムに残存している場合でも、その残存部位に WGS で得られたリードはきちんとマッピングできるのか検証する必要がある。今年度の検討で、Cas9 のヌクレアーゼ活性ドメインである HNH ドメインの部分配列をゲノムに挿入し、これを利用したシミュレーション WGS データを作成したが、pDGE64（=残存する可能性がある外来遺伝子配列の全体）に対してマッピングしたデータをアセンブリ法に供したが、ショートリードシークエンスの場合は 31 塩基程度、ロングリードシークエンスの場合は 51 塩基程度の挿入を検出できることが分かった。すなわち、可能性のある配列の一部のみが残存している場合でも検出可能であることが示されたと共に、その検出限界が示された。この検出限界の要因として、これより短い塩基ではマッピングされたリードが検出されなかった点が挙げられる。今回の検討ではマッピングツールとして、ショートリードシークエンスでは BWA、ロングリードシークエンスでは Minimap2 を用いており、基本的なパラメータについてはデフォルトで解析を実施している。マッピングパラメータを調整することで、残存配列が

より短くてもマッピングさせることが可能かもしれないが、結果として全体的なマッピングの正確性が低下することが懸念される。具体的に今回用いたショートリードシークエンスでは 1 リードあたり 150 塩基の情報が記録されているが、このうち 31 塩基が外来遺伝子配列である場合、残り 119 塩基の情報を切り捨てて、31 塩基のみをマッピングに使用するクリッピングという処理が実行されている。より短い配列をマッピングさせようとするとその分だけマッピングにおける配列特異性が下がるため、リードの一部分のみを切り取って、正しくない場所に誤ってマッピングする可能性が高まる。一般的にマッピング率と正確性はトレードオフの関係にあると言われており、マッピングツールのパラメータ調整は全体のバランスを見ながら慎重に行う必要があるから、今後の検討課題である。

今回の結果からでは、ロングリードシークエンスデータとしての疑似 WGS データを用いたアセンブリ法では、検出に必要な最低のシークエンスカバレッジが 15〜40×と幅が大きい結果となっており、解析の必要要件を定めることが困難であった。リード長の大部分が 150 塩基となるショートリードシークエンスデータとは異なり、ロングリードシークエンスデータでは調製した DNA ライブラリに依存するため、同じシークエンスカバレッジの WGS データを取得しても、平均リード長やマッピングリード数が毎回大きく異なることが、ロングリードシークエンスデータを用いた解析手法の標準化における一つの障壁となっている。なお今回解析したデータでは、リードクオリティ 9 以上を用いているが、アセンブリ法の成否でリードクオリティ値の平均値に差が無かった（成功した場合は 12.6±0.4、失敗した場合は 12.7±0.7）ことから、リードクオリティ値はアセンブリ法の結果に大きな影響を及ぼしていないと考えられた。また、12 回のダウンサンプリングによる検証ではシークエンスカバレッジが小さいほどアセンブリに失敗する傾向がある（表 6, 11）が、マッピングリード数の相対標準偏差が大きくなっていることから、低シークエンスカバレッジではアセンブリに十分なマッピングリード数が得られていないと推察される（表 5, 10）。他方で、比較的マッピングリードが多く得られた場合においてもアセンブリに失敗するケースがあった。pDGE64 全長配列へのマッピング状況を確認したところ、一部の領域でわずかにリードデプスの窪みが生じていた（図 12A）。アセンブリの成功例（図 12B）ではそのような窪みが見られないことから、マッピングリード毎にそのリード長がバラつくことがこうした「窪み」の要因となる可能性が考えられる。アセンブリツールの条件設定も含め、ロングリードシークエンスによる WGS データを用いたアセンブリ法の必要要件に関しては今後も引き続き検証する必要がある。

ゲノム編集食品に残存した外来遺伝子配列を NGS によって解析する方法として、WGS データについて残存が想定される配列の断片と照合する k-mer 法と、残存が想定される配列の断片と相補的な DNA 断片を集めて NGS によって解析する capture-based target enrichment 法が報告されている。このうち、本研究課題で我々が検討しているアセンブリ法と同様に、k-mer 法は WGS データの取得までに特別な作業を必要とせず、同じデータセットを用いた性能比較が可能であった。30×のシークエンスカバレッジを有する WGS データ（ショートリードシークエンス）を用いた比較から、以下の点が明らかになった。

- 短い配列（21 塩基）の検出について、k-mer 法は正確かつ迅速に検出できたが、アセンブリ法ではこれを検出できなかった。
- アセンブリ法は残存配列の有無と内容そしてそのゲノム上の位置を明らかにできるが、k-mer 法のみではゲノム上の位置を特定することはできない。

- アセンブリ法はゲノム編集作物の WGS データのみで解析できるが、k-mer 法はバックグラウンドとなった野生型作物の WGS データも必要となる。よってシークエンスのコストが 2 倍になる。

以上の留意点を踏まえ、ゲノム中に残存した外来遺伝子配列の検出にあたっては、まず k-mer 法を実施し、検出された場合にはアセンブリ法を実施して具体的な配列やゲノム上の位置を特定する、こうしたスキームで解析することが望ましいと考えられる。

本研究課題で検討しているアセンブリ法では複数のバイオインフォマティクスツールを使用している。従って、この解析を各実験者が実施するには、各自の解析環境にこれらのツールを導入し、適切に解析環境が構築されているか確認する必要がある。こうした確認には解析のコントロールになるような標準サンプルが使用されるが、今回作成した疑似 WGS データが利用できる可能性がある。本研究によって、ショートリードシークエンスであれ、ロングリードシークエンスであれ、Cas9 部分配列（101 塩基）が挿入されたゲノムからシミュレートされた 30×疑似 WGS データより、Cas9 断片配列の検出が可能であることが分かっている。そこでこの疑似 WGS データより Cas9 断片配列の検出が検知可能かどうかを各実験者の解析環境で確認することで、この点の検証が可能であると考えられる。

## E. 結論

本年度の検討から、WGS データを用いたアセンブリ法によるゲノム編集食品での外来遺伝子配列の残存性評価について、実際に残存する可能性のある配列（Cas9 配列など）を実際に起こりうる状況（残存する可能性のある配列の一部のみがゲノムに残存している）で検出できる妥当な手法であること、および少なくともショートリードシークエンスによる WGS データを活用した解析では、検出可能な残存遺伝子配列の長さや解析に必要とされる WGS データ量を明らかにすることができた。また、アセンブリ法の実施環境の妥当性確認に必要な、標準サンプルについても準備ができている。さらに既知の解析法との比較検討から、「k-mer 法による外来遺伝子配列の有無の検討→（検出されれば）アセンブリ法による配列内容と位置の同定」という評価スキームを提唱するに至った。他方で現行のショートリードシークエンスによる WGS データを活用した解析では評価が難しい点もあることから、ロングリードシークエンスによる WGS データを用いたアセンブリ法の有用性や必要要件に関する検討は必要である。

## F. 研究発表・業績

1. 論文発表

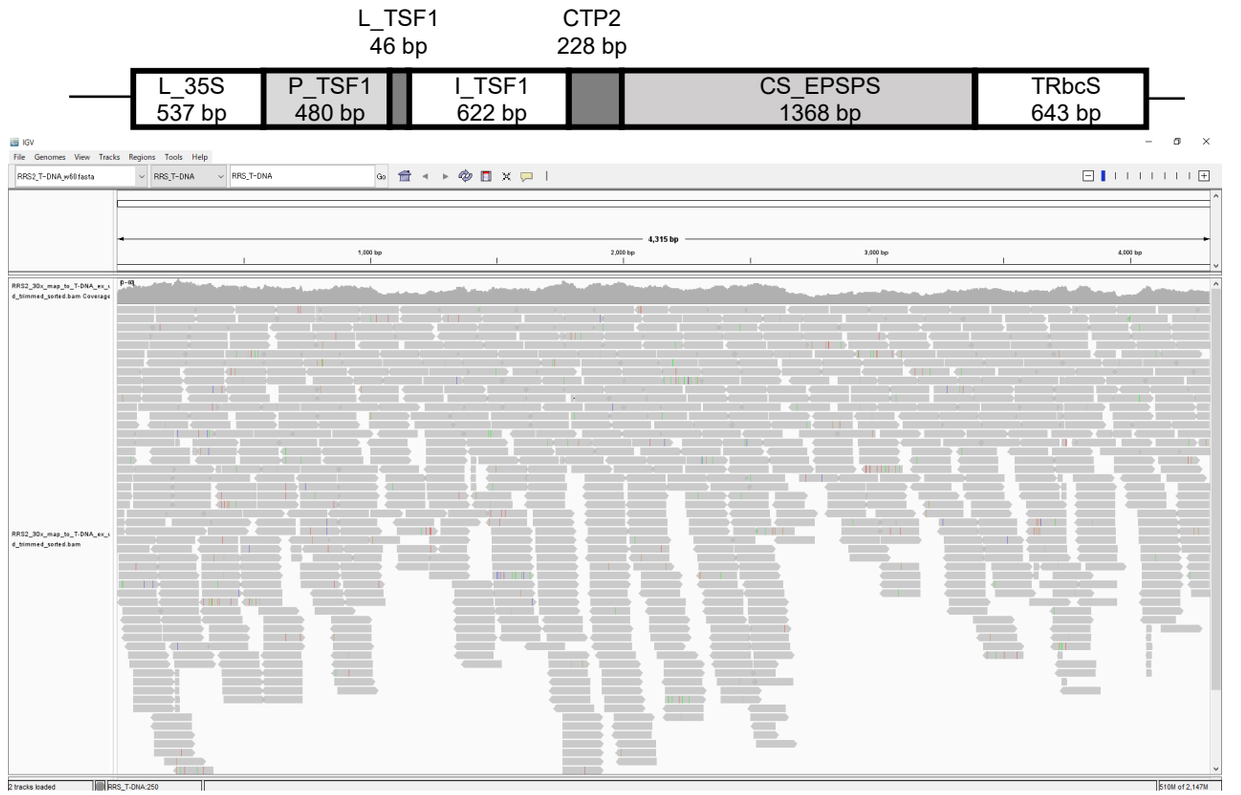
1) Soga K, Taguchi C, Sugino M, Egi T, Narushima J, Yoshiba S, Takabatake R, Kondo K, Shibata N: Investigation of genetically modified maize imported into Japan in 2021/2022 and the applicability of Japanese official methods. ***Food Hyg. Saf. Sci.*** 2023; 64: 218-225.

2. 学会発表

1) 成島純平、杉野御祐、曽我慶介、吉場聡子、近藤一成、柴田識人：ゲノム編集イネにおける in vitro オフターゲット予測法 SITE-Seq を用いたオフターゲット予測性能の評価、日本ゲノム編集学会　第 8 回大会、東京、2023 年 6 月 6 日-8 日

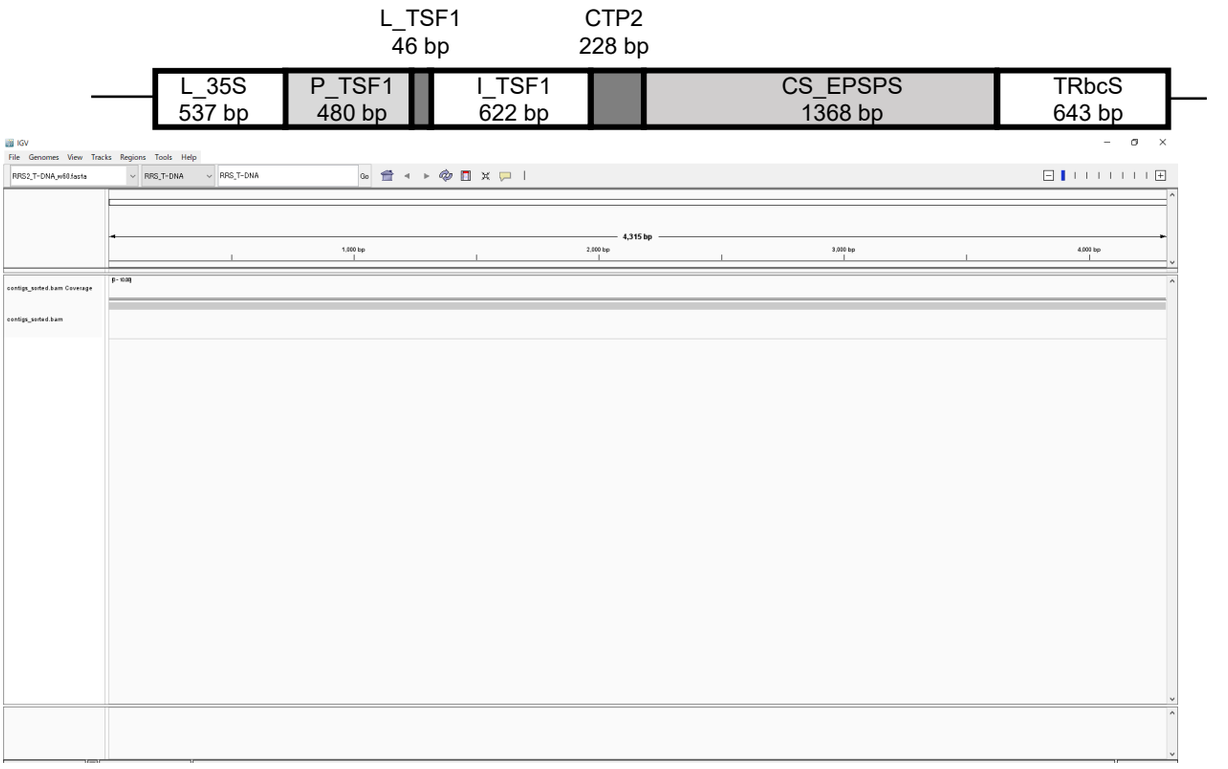
2) 成島純平、吉場聡子、細川葵、曽我慶介、杉野御佑、田口千恵、安達玲子、近藤一成、柴田識人：Cas9 targeted long-read sequencing による食品中の外来性遺伝子配列の同定、第 46 回日本分子生物学会年会、兵庫、2023 年 12 月 6 日-8 日

**G. 知的財産権の出願・登録状況**　該当なし



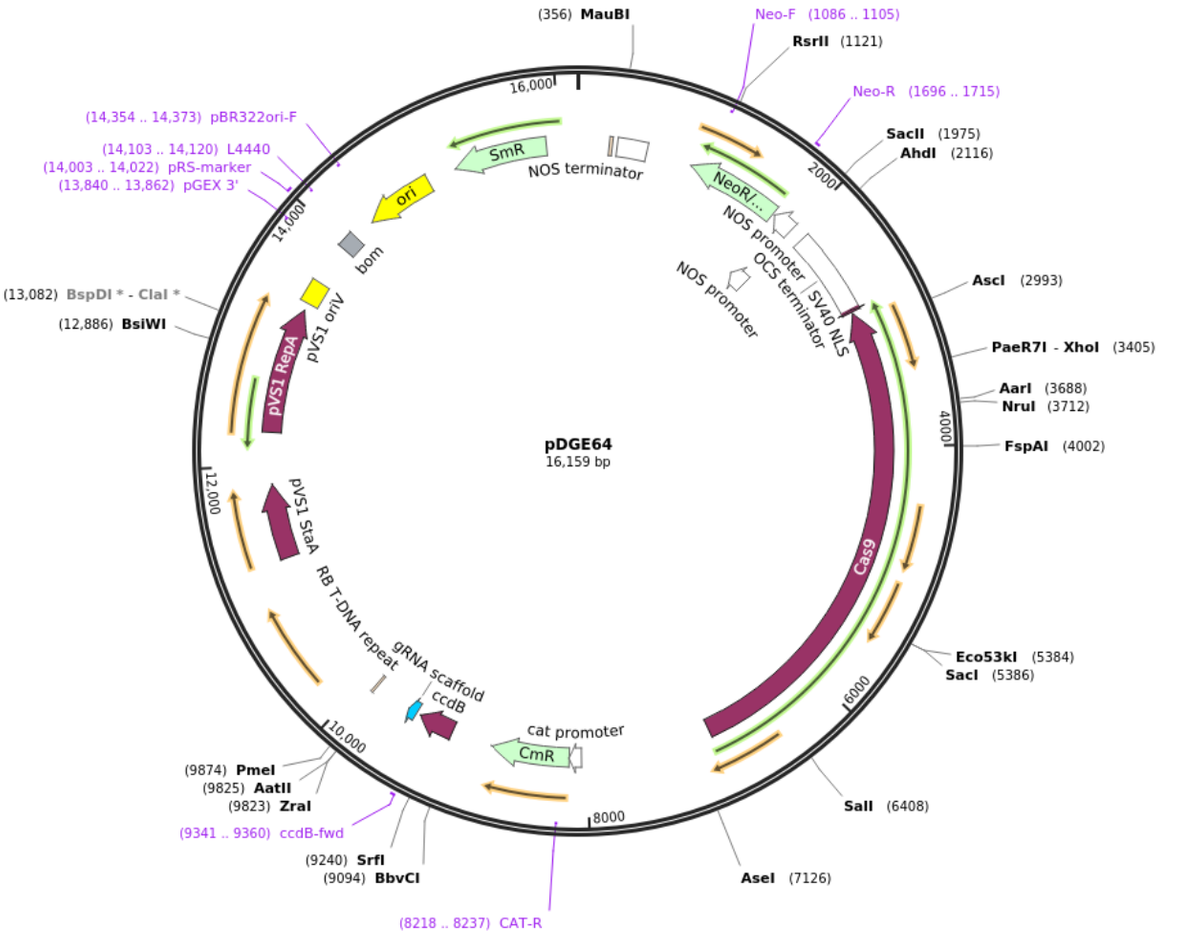

## 図1. RRS2において取得した約30×のWGSデータの外来遺伝子配列へのマッピング様相

遺伝子組換えダイズRRS2系統は1番染色体に約4,314 塩基の外来遺伝子が挿入されている。その外来遺伝子配列に対し、ショートリードシークエンスにて取得した約30×のシークエンスカバレッジを有するWGSリードデータをマッピングした。



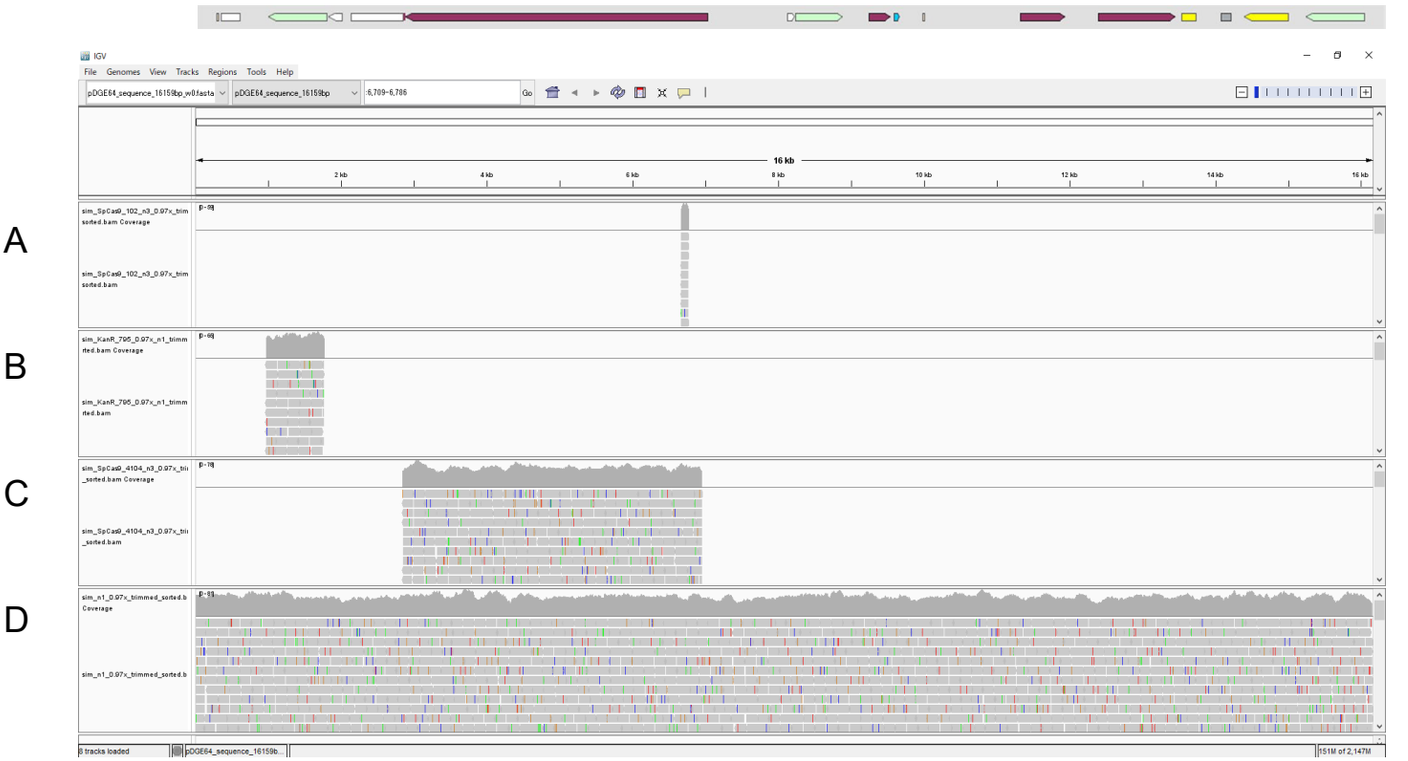

## 図2. 図1でマッピングした外来遺伝子配列リードをアセンブリし、生成されたコンティグ

外来遺伝子配列にマッピングしたリードはSPAdesによりアセンブリした。生成されたコンティグのうち、外来遺伝子配列全体にマッピングした1本のコンティグについて、そのマッピング様相を表す。




## 図3. pDGE64のベクターマップ

植物向けのゲノム編集ベクターであるpDGE64（#79446, 全長16,159 bp）のベクターマップ。AddgeneのWeb site (https://www.addgene.org/79446/)より。




## 図4. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータのマッピング様相（ショートリードシークエンス）

ゲノム編集ベクターpDGE64全長またはその部分配列をダイズリファレンスゲノムに挿入し、シミュレーターで疑似WGSデータを生成し、pDGE64の全長配列にマッピングした。

A. Cas9のArgドメインの遺伝子配列（102 塩基）

B. カナマイシン耐性遺伝子の全長（795 塩基）

C. Cas9遺伝子の全長（4,104 塩基）

D. pDGE64の全長（16,159 塩基）

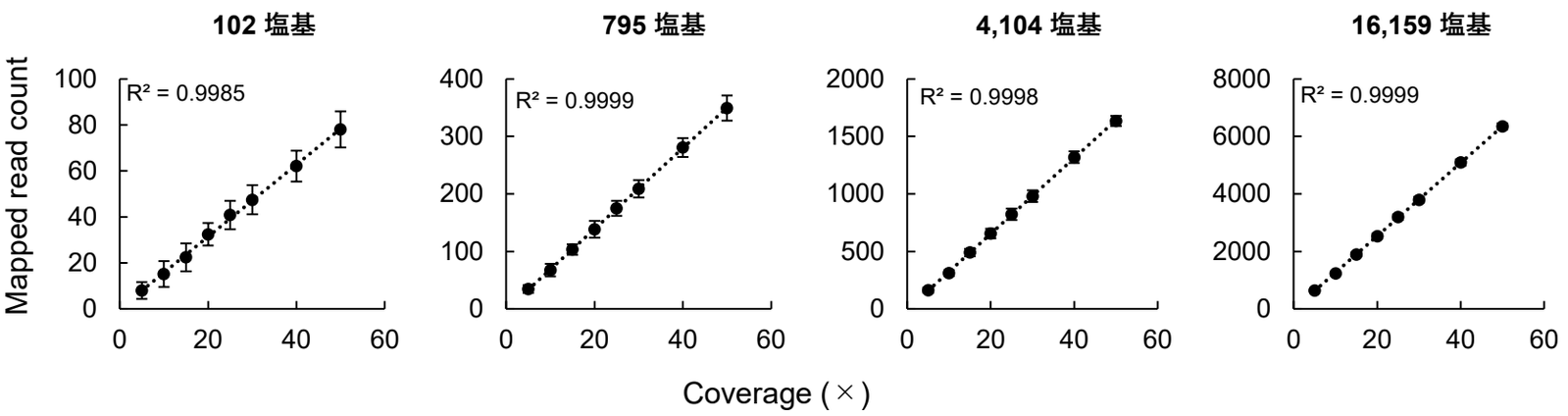



**図5. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ショートリードシークエンス）のマッピングリード数の併行精度**

Cas9のArgドメインの遺伝子配列（102 塩基）、カナマイシン耐性遺伝子（795 塩基）、Cas9遺伝子（4,104 塩基）およびpDGE64全長（16,159 塩基）を挿入した疑似WGSデータを約52×のシークエンスカバレッジで12併行分生成した。
この疑似WGSデータを50×～5×のシークエンスカバレッジになるようランダムサンプリングした。ランダムサンプリングしたリードデータをpDGE64全長配列へマッピングし、マッピングしたリード数の併行精度を算出した。なお、Cas9のArgドメイン（102 塩基）に関しては生成したデータにおいて外れ値が認められたため、外れ値となったn=1を除いた11併行の結果を示す。

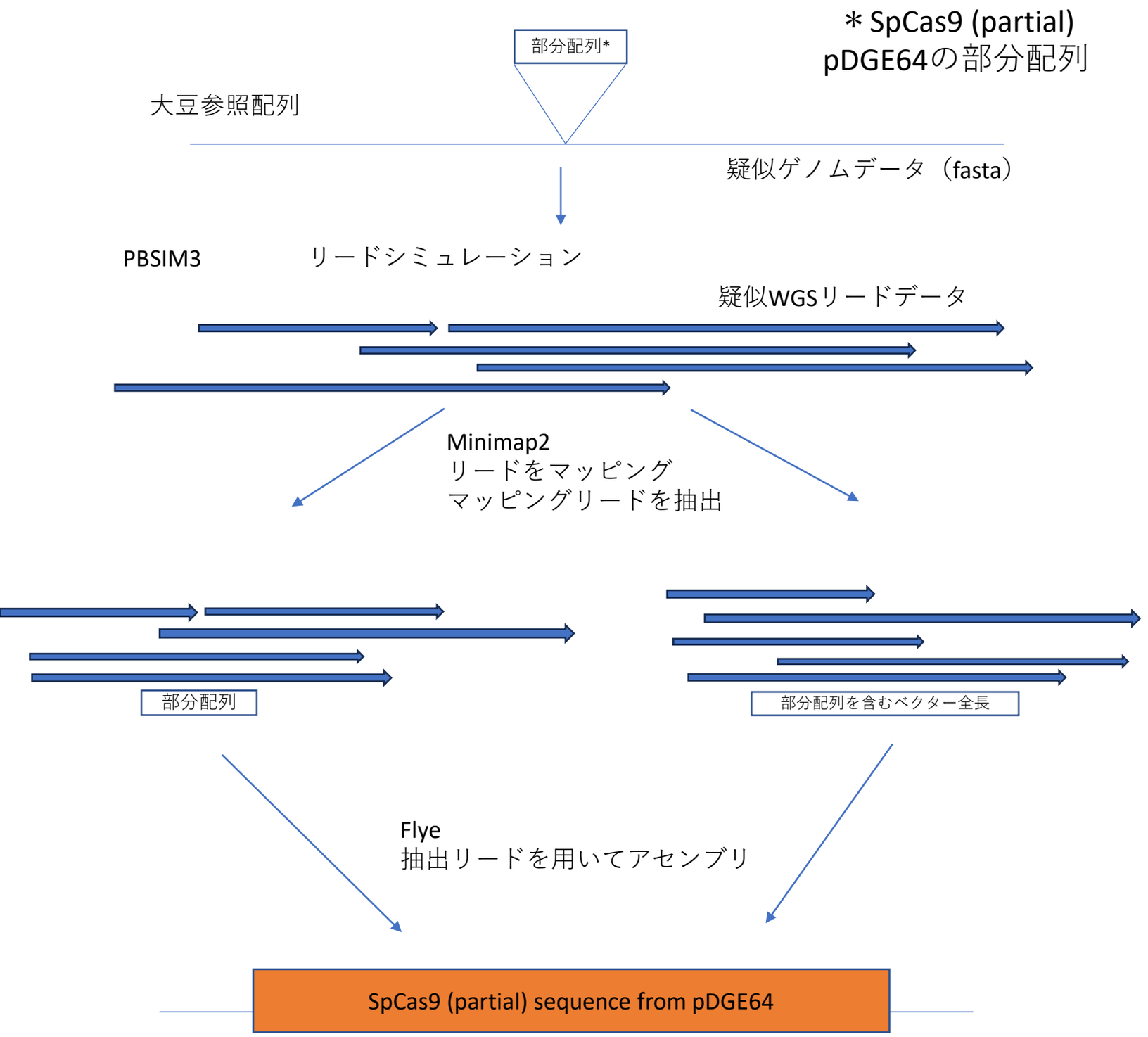
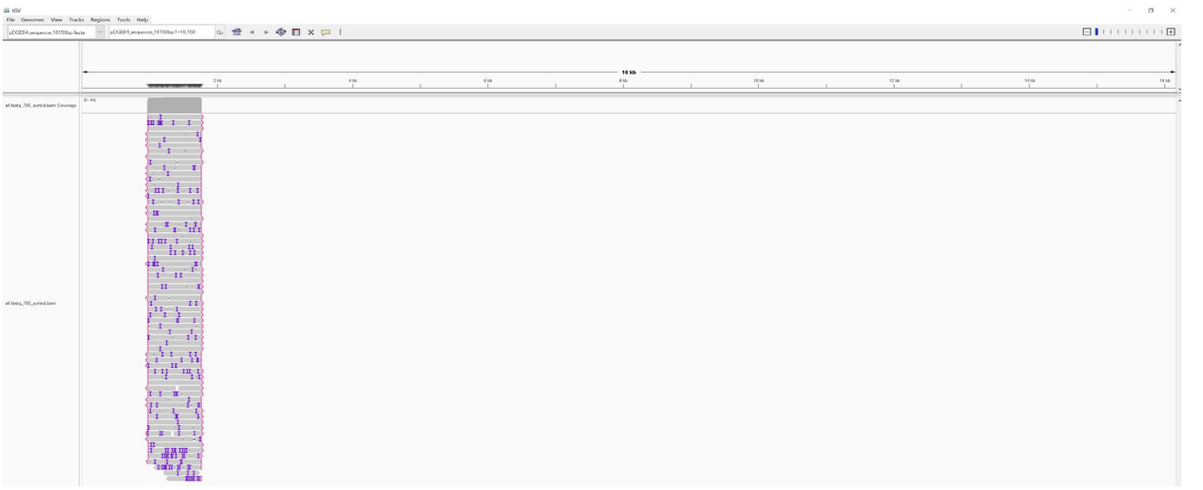



**図6. ロングリードシークエンスWGSデータを用いたアセンブリ法検証の概略**

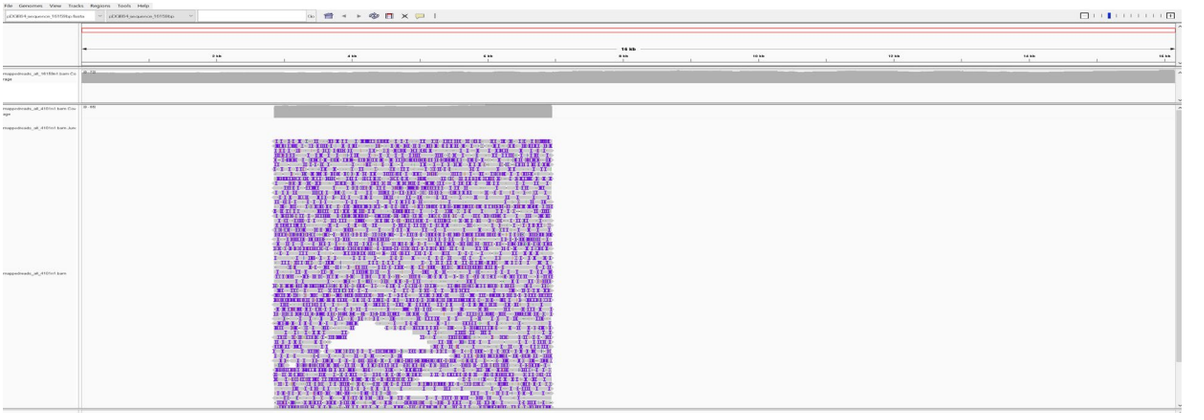
本研究では、リードをマッピングする際は、部分配列（左側）および部分配列を含むベクター全長（右側）を試している。Minimap2を用いた場合、短い部分配列（50 塩基程度）を含むベクター全長にマッピングしたリード（右側）でも、短い部分配列のみ（左側）にはマッピングしない場合があった。

A

B



C

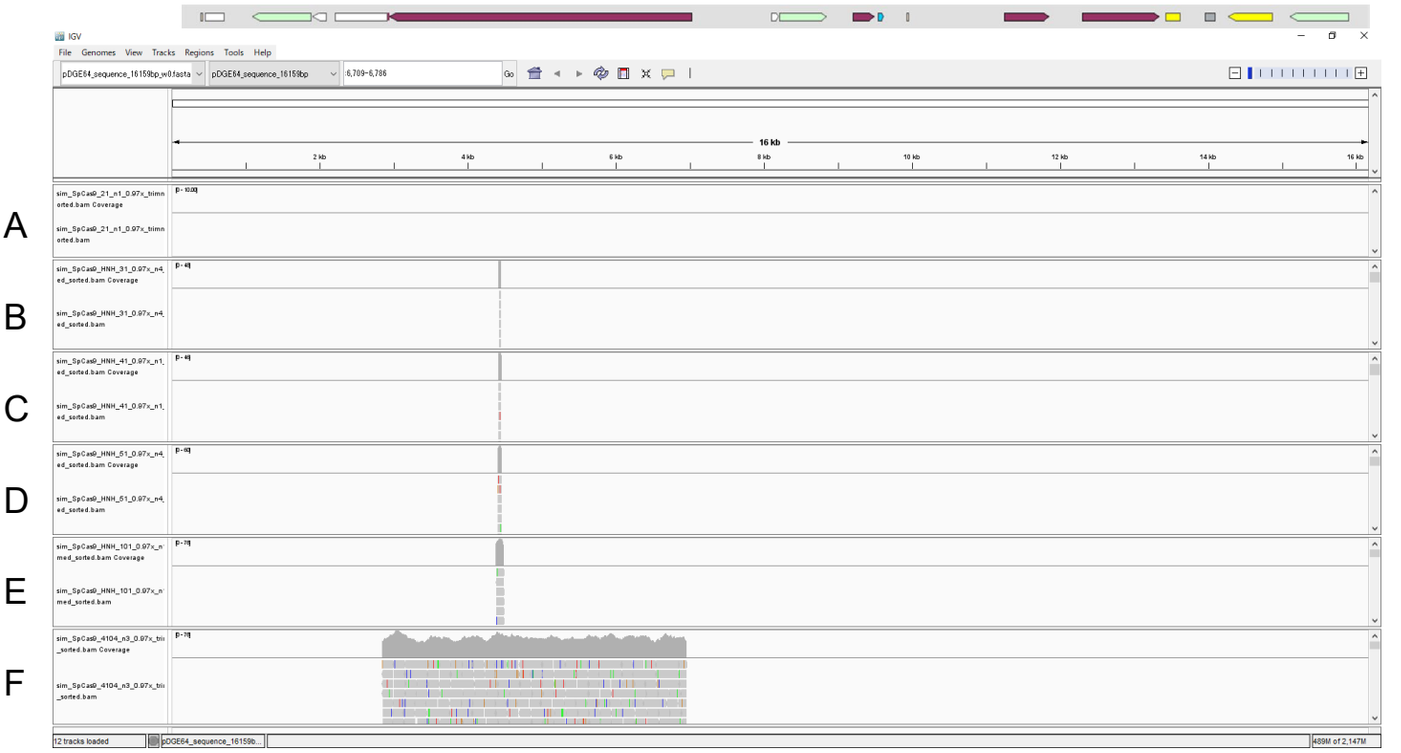


D

## 図7. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータのマッピング様相（ロングリードシークエンス）

ゲノム編集ベクターpDGE64全長またはその部分配列をダイズリファレンスゲノムに挿入し、シミュレーターで疑似WGSデータを生成し、pDGE64の全長配列にマッピングした。リードの紫色になっている部分は塩基の挿入があることを示す。

A. Cas9のArgドメインの遺伝子配列（102 塩基）
B. カナマイシン耐性遺伝子の全長（795 塩基）
C. Cas9遺伝子の全長（4,104 塩基）
D. pDGE64の全長（16,159 塩基）




## 図8. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータのマッピング様相（ショートリードシークエンス）

Cas9のヌクレアーゼ活性ドメインであるHNHドメインの部分配列を挿入した疑似WGSデータをpDGE64配列へマッピングした際のマッピング様相を表す。**A**は21塩基、**B**は31塩基、**C**は41塩基、**D**は51塩基、 **E**は101塩基のHNHドメインの部分配列を挿入した。また参考までにCas9全長（4,104塩基）を挿入した疑似WGSデータのマッピング様相を**F**に示す。

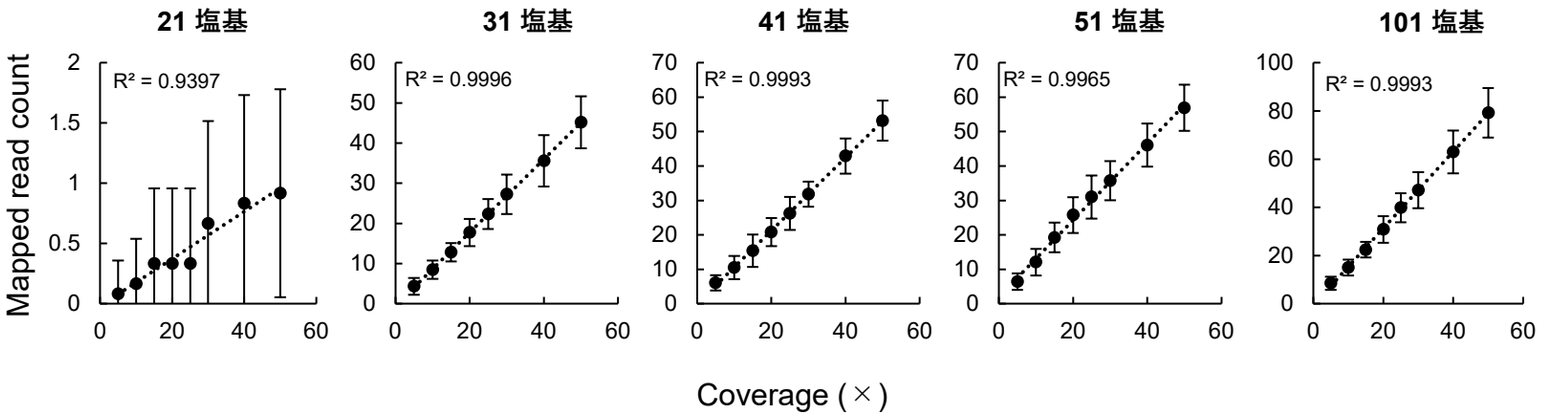



**図9. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ショートリードシークエンス）のマッピングリード数の併行精度**

Cas9の活性ドメインであるHNHドメインの部分配列をダイズの1番染色体に挿入し、シミュレーターにより疑似WGSデータを生成した。疑似データは約52×のシークエンスカバレッジで12併行生成し、それらを50×～5×のシークエンスカバレッジになるようランダムサンプリングした。ランダムサンプリングしたリードデータをpDGE64のベクター配列へマッピングし、マッピングしたリード数の統計値を算出した。

A

B

C

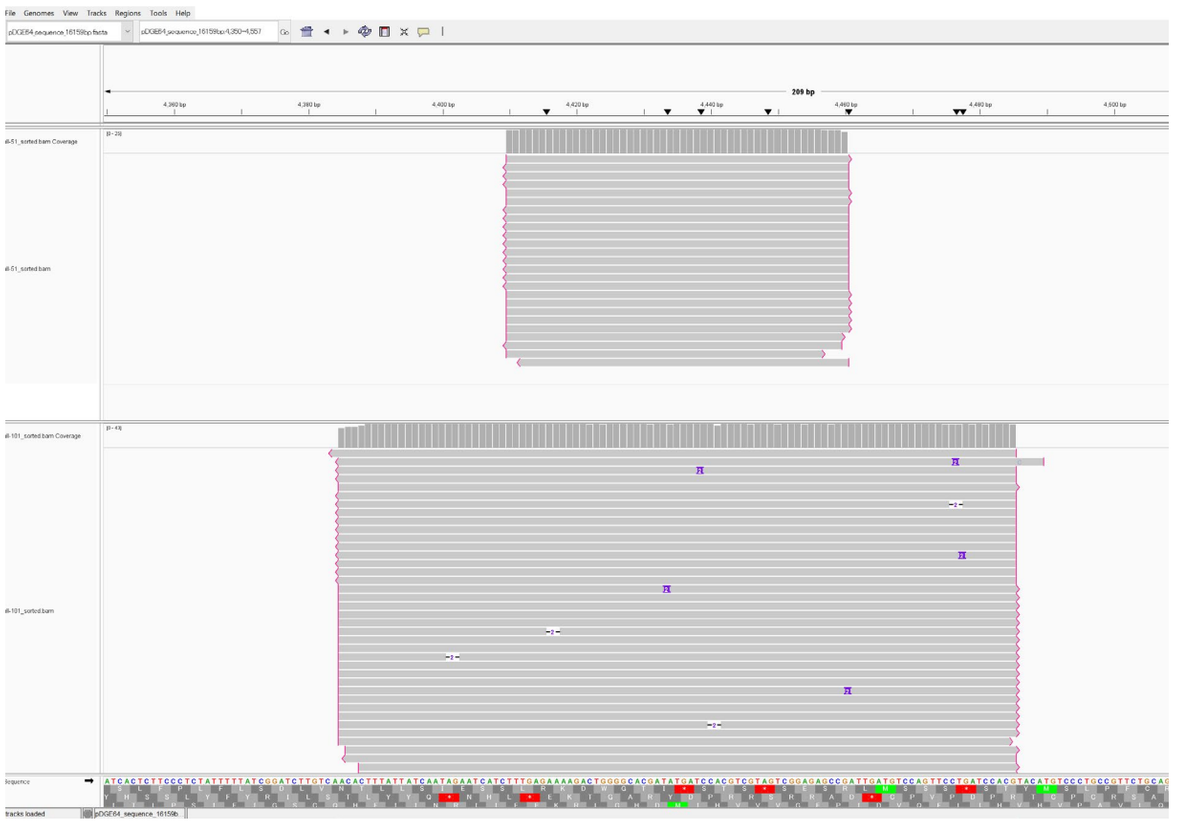
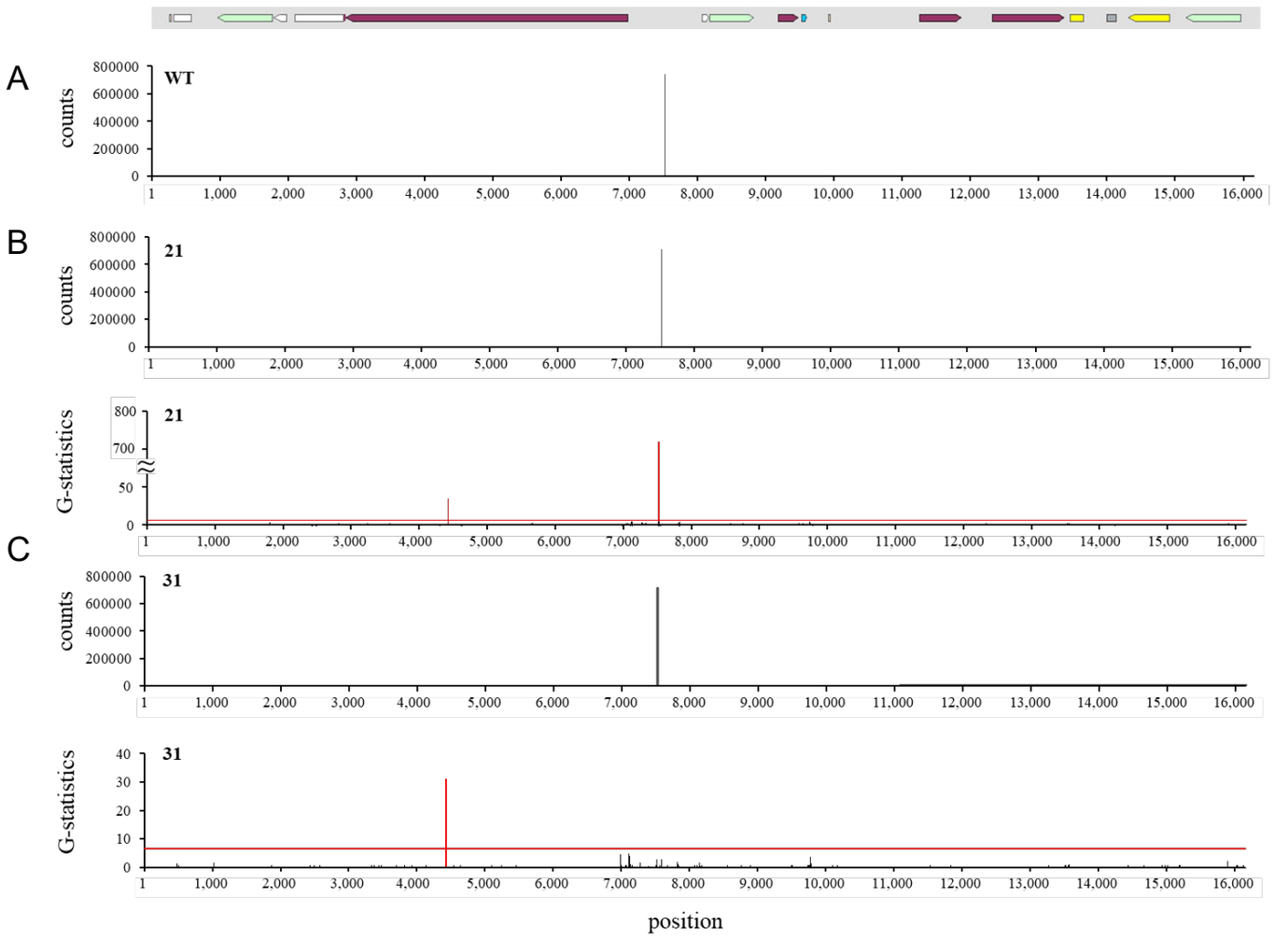


**図10. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとに　シミュレーションしたWGSデータ（ロングリードシークエンス）のマッピング様相**

Cas9のヌクレアーゼ活性ドメインであるHNHドメインの部分配列を挿入したゲノムデータから生成した疑似WGSデータをpDGE64配列へマッピングした際のマッピング様相を表す。**A**は上から21塩基（マッピングリードが得られなかった）および41塩基、**B**は上から51塩基および101塩基のHNHドメインの部分配列を挿入した。**C**はBの配列挿入箇所を拡大表示した。




## 図11. k-mer解析の結果

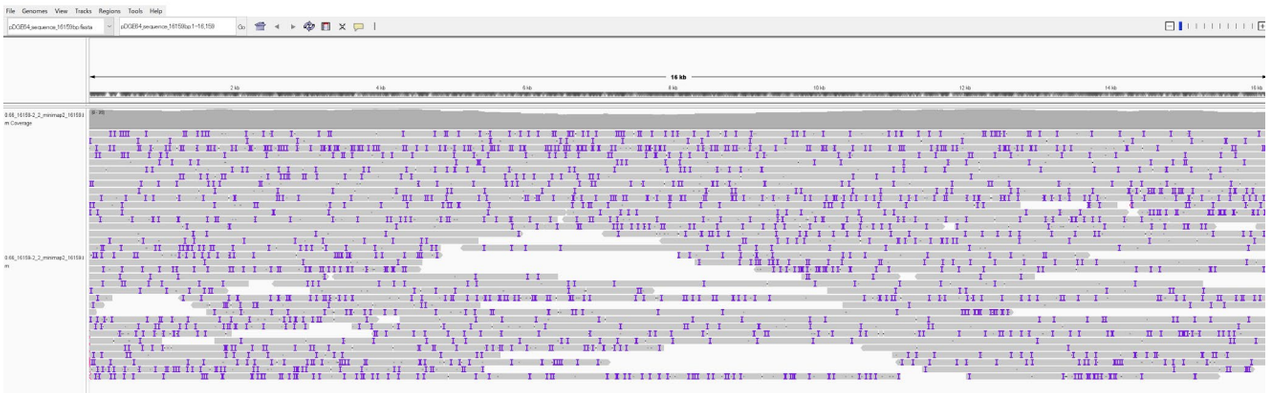
Cas9のヌクレアーゼ活性ドメインであるHNHドメインの21, 31 bpを挿入した疑似WGSデータセットを用いて、k-mer解析を実施した。k-mer解析にあたってはG-staticsを実施し、WTと比較して有意に（G > 6.634、Thr. lineは赤色で表示）外来遺伝子の挿入が認められる領域は赤色で示した。

A. WT （野生型）、データ数は n=1

B. 21 塩基 挿入データセット、n=12のデータについて解析。図はそのうちの代表的なデータ。

C. 31 塩基 挿入データセット、n=12のデータについて解析。図はそのうちの代表的なデータ。

A



B

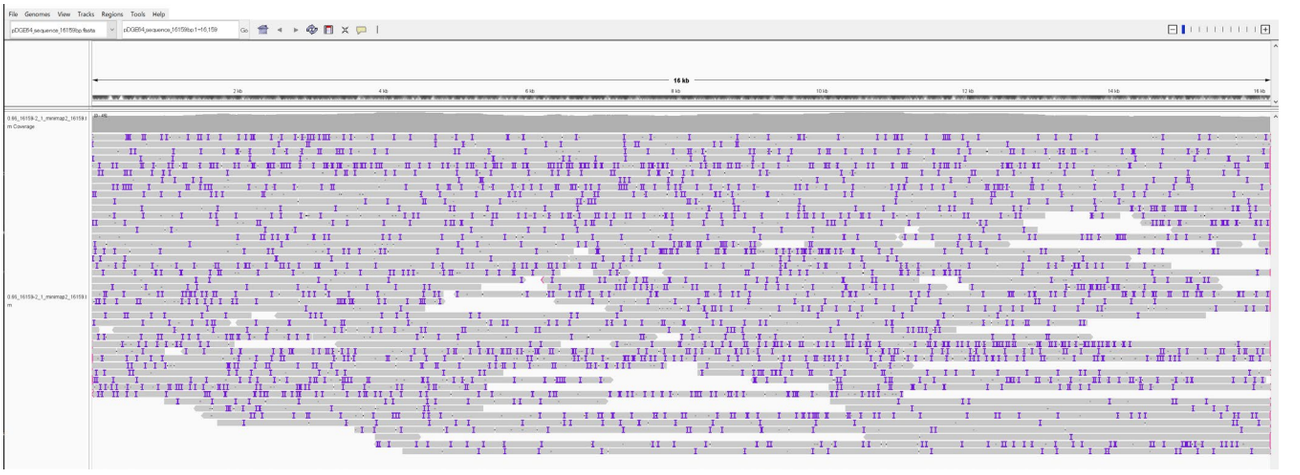


**図12. 疑似WGSデータ（ロングリードシークエンス）を用いたアセンブリの失敗例と成功例のマッピング様相**

A）高カバレッジでの失敗例（16159 塩基-data2, カバレッジ×35, リード数:69, N50:27,686）
B）高カバレッジでの成功例（16159 塩基-data2, カバレッジ×35, リード数:83, N50:29,914）

**表1. RRS2における約30×のWGSデータ（ショートリードシークエンス）を取得した際の統計値**

| | | RRS2 | |
|---|---|---|---|
| | | 前年度 | 今年度 |
| トリミング前 | Total reads | 455.2 M | 254.1 M |
| | Total bases | 68.3 G | 38.3 G |
| | Coverage | 62.1 x | 34.9 x |
| トリミング後 | Total reads | 454.7M | 254.0 M |
| | Total bases | 66.9 G | 37.7 G |
| | Coverage | 60.8 x | 34.2 x |
| マッピング後 | Total reads | 1.5 K | 1.1 K |
| | Total bases | 218.7 K | 156.0 K |
| | Coverage | 50.7 x | 36.3 x |
| アセンブリ後 | Contigs | 1 | 5 |
| | Total bases | 4.6 K | 5.5 K |
| | Longest | 4.6 K | 4.5 K |
| | Accuracy (%) | 100 | 100 |

遺伝子組換えダイズRRS2系統において、約30×シークエンスカバレッジとなるようショートリードシークエンサーによるWGSを実施した。トリミングの前後、RRS2にて挿入された外来遺伝子配列へマッピングしたリード、およびそれらを用いてアセンブリした際の統計値を示す（参照「今年度」欄）。参考までに前年度に実施した約62×のWGSの統計値も示す。

**表2. pDGE64全長配列、Cas9全長配列、カナマイシン耐性遺伝子配列、Cas9部分配列を挿入したゲノム配列をもとにシミュレーションされたWGSデータの統計値**

| | | Target (塩基) | 16,159 | 4,104 | 795 | 102* |
|---|---|---|---|---|---|---|
| ショートリード (平均値, n=12) | トリミング前 | Total reads | 378.9 M | 378.9 M | 378.9 M | 378.9 M |
| | | Total bases | 56.8 G | 56.8 G | 56.8 G | 56.8 G |
| | | Coverage | 51.7 x | 51.7 x | 51.7 x | 51.7 x |
| | トリミング後 | Total reads | 378.8 M | 378.8 M | 378.8 M | 378.8 M |
| | | Total bases | 56.6 G | 56.6 G | 56.6 G | 56.6 G |
| | | Coverage | 51.5 x | 51.5 x | 51.5 x | 51.5 x |
| ロングリード (平均値) | | Total reads | 4.5 M | 4.5 M | 4.5 M | 4.5 M |
| | | Total bases | 58.7 G | 58.7 G | 58.7 G | 58.7 G |
| | | N50 | 24,847 bp | 24,841 bp | 24,841bp | 24,830 bp |
| | | Coverage | 53.4 x | 53.4 x | 53.4 x | 53.4 x |

疑似WGSデータ（ショートリードシークエンス）は12併行で生成し、解析に供した。なおCas9のArgドメイン102 塩基が挿入されたデータセットについては、外れ値検定において外れ値が認められたため、n=11で解析を行っている。

疑似WGSデータ（ロングリードシークエンス）は、Cas9のArgドメイン102 塩基またはカナマイシン耐性遺伝子795 塩基が挿入されたデータセットについては n=1、Cas9全長配列4,104 塩基またはpDGE64全長配列16,159 塩基が挿入されたデータセットについては n=4の平均値を示している。

## 表3. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ショートリードシークエンス）のマッピングリード数

| | Target (塩基) | 50× | 40× | 30× | 25× | 20× | 15× | 10× | 5× |
|---|---|---|---|---|---|---|---|---|---|
| Mapped read count (Average ± SD, n=12) | 102 | 77.1±8.2 | 61.8±6.5 | 47.1±6.1 | 40.7±5.9 | 32.8±4.8 | 22.8±6.0 | 15.5±5.4 | 7.8±3.6 |
| | 795 | 349.2±22 | 280.8±16.5 | 208.8±15.2 | 175.0±13.1 | 138.5±14.6 | 103.3±8.9 | 67.6±10.9 | 34.7±6.6 |
| | 4,104 | 1,634.7±44.8 | 1,319.2±50.9 | 980.6±49.8 | 823.2±48.9 | 655.8±40.5 | 490.3±32.1 | 311.7±26.0 | 163.5±22.1 |
| | 16,159 | 6,347.8±101.8 | 5,091.8±106.2 | 3,796.9±103.5 | 3,194.6±94.4 | 2,533.9±91.9 | 1,894.7±68.8 | 1,230.5±61.5 | 642.8±32.5 |

Cas9のArgドメインの遺伝子配列（102 塩基）、カナマイシン耐性遺伝子（795 塩基）、Cas9遺伝子（4,104 塩基）およびpDGE64全長（16,159 塩基）を挿入した疑似WGSデータを約52×のシークエンスカバレッジで12併行分生成した。この疑似WGSデータを50×～5×のシークエンスカバレッジになるようランダムサンプリングした。ランダムサンプリングしたリードデータをpDGE64全長配列へマッピングし、マッピングしたリード数の平均値及び標準偏差を示す。なお、 Cas9のArgドメイン（102 塩基）に関しては生成したデータにおいて外れ値が認められたため、外れ値となったn=1を除いた11併行の結果を示す。

**表4. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ショートリードシークエンス）を用いたアセンブリ結果**

| | Target (塩基) | 50× | 40× | 30× | 25× | 20× | 15× | 10× | 5× |
|---|---|---|---|---|---|---|---|---|---|
| Contig 1本で全体をカバー | 102 | 11/11 | 11/11 | 11/11 | 11/11 | 11/11 | 8/11 | 3/11 | 0/11 |
| | 795 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 6/12 | 0/12 |
| | 4,104 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 0/12 |
| | 16,159 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 0/12 |

pDGE64全長配列にマッピングしたリードを抽出し、SPAdesにてアセンブリすることで、挿入させた外来遺伝子配列を正確に再現できるかを検証した。アセンブリで生成されたコンティグのうち、1本のコンティグのみで元の外来遺伝子配列全長を再現できたかを基準に12併行（102 塩基のみ11併行）での成功数を表す。

**表5. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ロングリードシークエンス）を各挿入配列にマッピングして得られたリード数**

| Target (塩基) | 53.4× (all) | 40× | 35× | 30× | 25× | 20× | 15× | 10× | 5× | 3× |
|---|---|---|---|---|---|---|---|---|---|---|
| 102 | 38 | 30.7±2.8（9.2） | - | 23.1±4.4（19.0） | - | 17.2±3.0（17.7） | - | - | - | - |
| 795 | 67 | - | - | 37.6±3.8（10.2） | - | 25.9±4.1（15.7） | 19.3±3.5（18.2） | 11.3±4.1（36.2） | - | - |
| 4,104 Data1 | 79 | - | - | 41.8±4.6（11.1） | 34.9±3.9（11.1） | 28.6±3.8（13.2） | - | - | - | - |
| 4,104 Data2 | 80 | - | - | 43.9±5.4（12.4） | 36.1±5.5（15.1） | 27.9±4.2（15.2） | 21.4±3.4（16.1） | - | - | - |
| 4,104 Data3 | 58 | - | - | - | 25.9±3.1（12.1） | 20.9±2.7（13.1） | - | - | - | - |
| 4,104 Data4 | 93 | - | - | - | - | 36.2±5.4（14.9） | 25.7±4.2（16.2） | - | - | - |
| 16,159 Data 1 | 137 | - | - | - | - | 53.3±4.6（8.7） | - | 25.6±5.3（20.6） | 13.4±3.2（23.7） | 7.0±2.3 (32.8) |
| 16,159 Data 2 | 117 | 88.8±4.4（4.9） | 78.3±5.5（7.0） | 65.3±5.3（8.1） | - | 43.8±5.3（12.1） | - | 23.0±3.9（16.9） | - | - |
| 16,159 Data 3 | 149 | - | - | - | - | 59.0±5.2（8.8） | 42.4±5.4（12.8） | 28.7±3.3（11.4） | - | - |
| 16,159 Data 4 | 128 | - | - | - | - | - | 34.9±5.2（14.9） | 22.8±4.2（18.4） | 12.1±3.3（27.7） | - |

40×以下のデータはダウンサンプリングをそれぞれ12回実施し、その際のマッピングリード数の平均値及び標準偏差を示す。括弧内は相対標準偏差（%）を示す。

- は未実施。

**表6. pDGE64全長またはその部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ロングリードシークエンス）を各挿入配列にマッピングして得られたリードを用いたアセンブリ法の結果**

| | Target (塩基) | 53.4× (all) | 40× | 35× | 30× | 25× | 20× | 15× | 10× | 5× | 3× |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Contig 1本で全体をカバー | 102 | 1/1 (100) | 10/12 (100) | - | 9/12 (100) | - | 8/12 (100) | - | - | - | - |
| | 795 | 1/1 (100) | - | - | 12/12 (100) | - | 12/12 (100) | 9/12 (100) | 6/12 (100) | - | - |
| | 4,104 data1 | 1/1 (100) | - | - | 12/12 (100) | 11/12 (100) | 10/12 (100) | - | - | - | - |
| | 4,104 data2 | 1/1 (100) | - | - | 12/12 (100) | 12/12 (100) | 12/12 (100) | 11/12 (100) | - | - | - |
| | 4,104 data3 | 1/1 (100) | - | - | - | 12/12 (100) | 11/12 (100) | - | - | - | - |
| | 4,104 data4 | 1/1 (100) | - | - | - | - | 12/12 (100) | 11/12 (100) | - | - | - |
| | 16,159 data 1 | 1/1 (100) | - | - | - | - | 12/12 (100) | - | 12/12 (99.9) | 11/12 (95.1) | 3/12 (66.2) |
| | 16,159 data 2 | 1/1 (100) | 12/12 (100) | 11/12 (100) | 10/12 (100) | - | 8/12 (100) | - | 10/12 (99.9) | - | - |
| | 16,159 data 3 | 1/1 (100) | - | - | - | - | 12/12 (100) | 11/12 (100) | 11/12 (99.9) | - | - |
| | 16,159 data 4 | 1/1 (100) | - | - | - | - | - | 12/12 (100) | 12/12 (99.9) | 5/12 (96.9) | - |

各疑似WGSデータについて、各挿入配列にマッピングし、抽出されたリードをアセンブリすることで、挿入させた外来遺伝子配列を正確に再現できるかを検証した。アセンブリで生成されたコンティグのうち、1本のコンティグのみで元の外来遺伝子配列全長を再現できたかを基準に12併行での成功数を表す。括弧内数字はアセンブリ成功時の挿入塩基配列の正確性の平均値（%）を示す。

- は未実施。

**表7. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータの統計値**

| | | Target (塩基) | 101 | 51 | 41 | 31 | 21 |
|---|---|---|---|---|---|---|---|
| ショートリード (平均値, n=12) | トリミング前 | Total reads | 378.9 M | 378.9 M | 378.9 M | 378.9 M | 378.9 M |
| | | Total bases | 56.8 G | 56.8 G | 56.8 G | 56.8 G | 56.8 G |
| | | Coverage | 51.7 x | 51.7 x | 51.7 x | 51.7 x | 51.7 x |
| | トリミング後 | Total reads | 378.8 M | 378.8 M | 378.8 M | 378.8 M | 378.8 M |
| | | Total bases | 56.6 G | 56.6 G | 56.6 G | 56.6 G | 56.6 G |
| | | Coverage | 51.5 x | 51.5 x | 51.5 x | 51.5 x | 51.5 x |
| ロングリード (n=1) | | Total reads | 4.5 M | 4.5 M | 4.5 M | - | 4.5 M |
| | | Total bases | 58.7 G | 58.7 G | 58.7 G | - | 58.7 G |
| | | N50 | 24,824 bp | 24,825 bp | 24,849 bp | - | 24,836 bp |
| | | Coverage | 53.4 x | 53.4 x | 53.4 x | - | 53.4 x |

- は未実施

**表8. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ショートリードシークエンス）のマッピングリード数**

| | Target (塩基) | 50× | 40× | 30× | 25× | 20× | 15× | 10× | 5× |
|---|---|---|---|---|---|---|---|---|---|
| Mapped read count (Average ± SD, n=12) | 21 | 0.9±0.9 | 0.8±0.9 | 0.7±0.8 | 0.3±0.6 | 0.3±0.6 | 0.3±0.6 | 0.2±0.4 | 0.1±0.3 |
| | 31 | 45.2±6.4 | 35.6±6.4 | 27.3±4.9 | 22.3±3.7 | 17.8±3.4 | 12.8±2.3 | 8.5±2.3 | 4.3±2.1 |
| | 41 | 52.6±6.2 | 41.9±5.6 | 31.2±4.6 | 25.8±4.4 | 20.3±3.4 | 15.4±3.7 | 10.3±3.1 | 6.0±2.0 |
| | 51 | 56.9±6.7 | 46.1±6.2 | 35.8±5.7 | 31.0±6.3 | 25.8±5.2 | 19.3±4.3 | 12.1±3.8 | 6.4±2.4 |
| | 101 | 79.3±10.3 | 63.0±8.9 | 47.2±7.5 | 39.8±6.0 | 30.8±5.5 | 22.4±3.2 | 15.0±3.3 | 8.5±2.8 |

Cas9の活性ドメインであるHNHドメインの部分配列をダイズの1番染色体に挿入し、シミュレーターにより疑似WGSデータを生成した。疑似データは約52×のシークエンスカバレッジで12併行生成し、それらを50×～5×のシークエンスカバレッジになるようランダムサンプリングした。ランダムサンプリングしたリードデータをpDGE64のベクター配列へマッピングし、マッピングしたリード数の平均値及び標準偏差を示す。

**表9. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ショートリードシークエンス）を用いたアセンブリ結果**

| | Target (塩基) | 50× | 40× | 30× | 25× | 20× | 15× | 10× | 5× |
|---|---|---|---|---|---|---|---|---|---|
| Contig 1本で全体をカバー | 21 | - | - | - | - | - | - | - | - |
| | 31 | 12/12 | 12/12 | 12/12 | 12/12 | 8/12 | 2/12 | 0/12 | 0/12 |
| | 41 | 12/12 | 12/12 | 12/12 | 11/12 | 11/12 | 6/12 | 3/12 | 1/12 |
| | 51 | 12/12 | 12/12 | 12/12 | 12/12 | 12/12 | 9/12 | 3/12 | 1/12 |
| | 101 | 12/12 | 12/12 | 12/12 | 12/12 | 11/12 | 6/12 | 1/12 | 0/12 |

pDGE64全長配列にマッピングしたリードを抽出し、SPAdesにてアセンブリすることで、挿入させた外来遺伝子配列を正確に再現できるかを検証した。アセンブリされた1コンティグで元の外来性遺伝子配列全長を再現できたかを基準に12併行での成功数を示す。21 塩基に関してはリードが抽出できなかったため、アセンブリは実施しなかった（-と表記）。

**表10. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ロングリードシークエンス）をpDGE64全長配列にマッピングして得られたリード数**

| Target (塩基) | 53.4× (all) | 40× | 30× | 25× | 20× |
|---|---|---|---|---|---|
| 21 | 0 | - | - | - | - |
| 41 | 0 | - | - | - | - |
| 51 | 25 | 19.3±2.5 （12.9） | - | - | - |
| 101 | 43 | 31.9±3.4 （10.8） | 24.0±4.5 （18.6） | 19.6±3.8 （19.5） | 16.5±3.3 （19.9） |

40× 以下のデータはダウンサンプリングをそれぞれ12回実施し、その際のマッピングリード数の平均値及び標準偏差を示す。括弧内は相対標準偏差（%）を示す。

- は未実施。

**表11. Cas9ヌクレアーゼ活性ドメインの部分配列を挿入したゲノム配列をもとにシミュレーションしたWGSデータ（ロングリードシークエンス）をpDGE64全長配列にマッピングして得られたリードを用いたアセンブリ法の結果**

| | Target (塩基) | 53.4× (all) | 40× | 35× | 30× | 25× | 20× |
|---|---|---|---|---|---|---|---|
| Contig 1本で全体をカバー | 21 | 0/1 | - | - | - | - | - |
| | 41 | 0/1 | - | - | - | - | - |
| | 51 | 1/1 (100) | 9/12 (100) | - | - | - | - |
| | 101 | 1/1 (100) | 12/12 (100) | - | 12/12 (100) | 12/12 (100) | 11/12 (100) |

各疑似WGSデータについて、pDGE64全長配列にマッピングし、抽出されたリードをアセンブリすることで、挿入させた外来遺伝子配列を正確に再現できるかを検証した。アセンブリで生成されたコンティグのうち、1本のコンティグのみで元の外来遺伝子配列全長を再現できたかを基準に12併行での成功数を表す。

21及び41 bpに関してはリード自体の抽出ができなかったため、アセンブリは実施しなかった。

括弧内数字はアセンブリ成功時の挿入塩基配列の正確性の平均値（%）を示す。

- は未実施。

別紙３

厚生労働科学研究費補助金（食品の安全確保推進研究事業）

「新たなバイオテクノロジーを用いて得られた食品の安全性確保と

リスクコミュニケーションのための研究」

分担報告書

# メタボロームインフォマティクスによる未知化合物推定

研究分担者　早川英介　（国立研究開発法人理化学研究所 環境資源科学研究センター・メタボローム情報研究チーム・研究員）

**研究要旨**

令和５年度の研究では、昨年度までの化合物単位での解析に加え、分析データからエンリッチメント解析を通じて代謝パスウェイにおける変動を明らかにする検討を行った。さらに、これまでに開発したスペクトル類似度計算に基づく未知化合物の構造解析および可視化を行う解析ツールに関して、ウェブブラウザ上で動作する Docker イメージとして配布、GitHub で公開し、さらにチュートリアルや様々なドキュメントを整備した。これにより、未知化合物の迅速な解析と可視化という従来高度な質量分析とインフォマティクス技術が必要だった解析が広範な研究者および技術者にも利用可能となった。本研究の成果は、食品衛生学会で発表され、学術誌への論文投稿を準備中である。

## A. 研究目的

本年度の研究目的は、ゲノム編集などの新しいバイオテクノロジーによって開発された食品に含まれる未知化合物の迅速な検出と構造推定を行うシステムに関して、代謝パスウェイレベルでの俯瞰的な解析の検討と、さらに広範なユーザーに向けたツール構築と公開である。具体的には、エンリッチメント解析を統合し、これにより代謝系に及ぼす影響を明らかにする手法を検討し、ゲノム編集食品の安全性評価への応用を目指す。さらに、開発したスペクトルデータの類似度計算および構造解析・可視化を行うツールを、様々な研究者が容易にアクセスできる解析環境として提供することで、広範な研究分野での利用を促進し、食品安全性評価の新たな手法の確立を目指した。

## B. 研究方法

本研究では、食品試料の質量分析データを基に、未知化合物の検出及び構造推定、そしてエンリッチメント解析を通じた代謝系への影響評価を目指した。エンリッチメント解析の手法的な側面では、質量分析データから得られるスペクトル情報を利用して、試料中で変動した化合物が属する代謝パスウェイを特定することで代謝パスウェイがゲノム編集によってどのように影響を受けるかを解析した。実例としてゲノム編集および通常トマトとの比較定量データをもとにして、ゲノム編集で変動する代謝パスウェイの特定を行うとともに、ゲノム編集による変異が代謝系に及ぼす具体的な影響を評価した。

データ解析・可視化ツールに関しては、様々なユーザーの質量分析データに柔軟に対応するために質量分析データの解析と可視化を行うための二つの主要なコンポーネント、Spectral Network Generator と Spectral Network Visualizer を個別にパッケージ化して構築した（図１）。

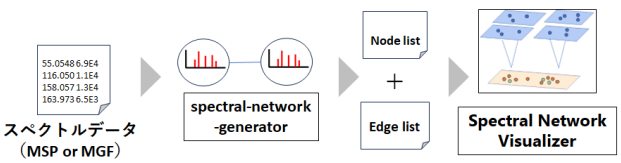



図１　　解析・可視化ツールのフロー

Generator は、質量スペクトルデータとに質量スペクトル間の類似度を計算し、未知化合物と既知化合物間の関係性を明らかにするためのネットワークデータを生成する。この過程では、Cosine score や modified cosine score をはじめとする複数の類似度計算アルゴリズムを実装し、幅広いスペクトルデータに対応できるようにした。また、公共のデータフォーマットに対応させることで、ユーザーが容易にデータを取り込み、解析を行えるように配慮した。Visualizer は、Generator によって提供される類似度データをもとに、未知化合物のスペクトル類似度のネットワークを直感的に可視するとともに、多層レイヤー構造により化合物クラスや毒性情報といったスペクトルライブラリにもとづいた構造解析を可能にするとともに、ネットワーク構造を活用した探索的な可視化・解析を可能にする。

このツールは Django プラットフォームを利用することで、ユーザーがブラウザ上で簡単に操作できるインターフェースを提供する。Docker イメージとしての配布により、異なるオペレーティングシステム上での利用障壁を低減し、幅広い研究者に利用されることを目指した。これらのツールは Docker イメージとして GitHub 上で配布され、詳細なドキュメントとチュートリアルの公開も行った。

## C. 研究結果および考察

本研究におけるゲノム編集トマトのメタボローム解析では、特定のアミノ酸代謝パスウェイ、特に Phenylalanine, Tyrosine, Aspartate および Biotin metabolism が顕著にエンリッチされていることが明らかになった。GAD 酵素が中心的な役割を果たす GABA 合成経路が活性化されることで、窒素配分に変化が生じ、「窒素プール」への影響が考えられる。この結果、Aspartate, Phenylalanine, Tyrosine の生合成にも影響を及ぼす可能性が示唆される。さらに、GABA は植物のストレス応答に深く関与し、GABA の過剰生産はストレス応答による二次代謝物の生合成に影響を及ぼすことから、Phenylalanine や Tyrosine を前駆体とする stress defense 関連の二次代謝物の合成にも影響を与えると推測される。また、シグナル伝達物質としての GABA の増加は、シグナル伝達経路を介して他のアミノ酸代謝にも影響を及ぼす可能性がある。このようなパスウェイ解析結果を、未知化合物の解析プロセスに統合することで、未知化合物の検出とその背景にあるメカニズムが明らかになるデータ解析の体制が構築できると考えられる。代謝パスウェイの変動が示す生物学的意義を、未知化合物の解析に反映させることで、ゲノム編集による影響の全体像をより詳細に捉えることが可能となると期待できる（図２）。

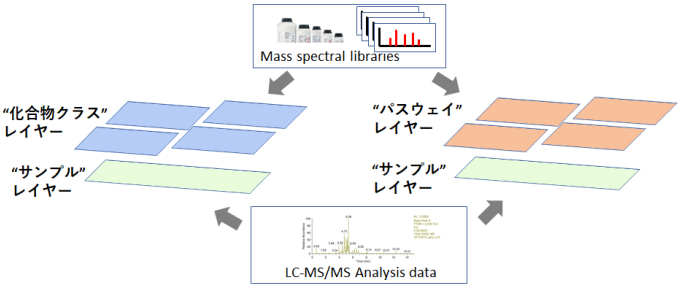



**図２　化合物・パスウェイ解析連携のイメージ　　解析・可視化ツールのフロー**

開発した解析ツールは、Docker を採用することで、プラットフォームに依存せず、インフォマティクスの専門知識がない研究者や技術者でも容易に利用できるユーザーフレンドリーなインターフェースを提供している。このアプローチにより、ゲノム編集食品に限らず、広範囲の食品安全での分析において、ツールの適用が期待される。今後はネットおよび論文のみならず、様々な場面でこのツールの周知を進めるとともに、ユーザーのフィードバックを取り入れながら機能拡張を行うことが望まれる。

## E. 結論

本年度の研究によりゲノム編集食品の解析におけるエンリッチメント解析の有効性が示され、これまでの未知化合物解析との連携の重要性が明らかとなった。また解析ツールはDockerイメージとして構築され、チュートリアルと共に一般に公開され、今後本ツールが広く食品分析のデータ解析に活用されることが期待される。

## E. 研究発表・業績

- 日本食品衛生学会第119回学術講演会(2023年10月12日-13日）タワーホール船堀「未知化合物解析支援のための質量スペクトルネットワーク可視化ツール」 早川英介、平田香南子、近藤一成、有田正規（ポスター発表）

- 開発された解析ツールのDockerイメージ配布レポジトリ　https://github.com/mass-spec-info/SNVis
- 解析・可視化ツールの投稿論文準備中

**F. 健康危険情報**

該当なし

**G. 知的財産権の出願・登録状況**

該当なし

**H. 知的財産権の出願・登録状況**

1. 特許取得

該当なし

2. 実用新案登録

該当なし

厚生労働科学研究費補助金（食品の安全確保推進研究事業）
「新たなバイオテクノロジーを用いて得られた食品の安全性確保とリスクコミュニケーション」
分担研究報告書（令和 5 年度）

# 新規アレルゲン性予測手法開発のための基盤的研究

研究分担者： 安達玲子　　国立医薬品食品衛生研究所生化学部　第二室長
研究協力者： 爲廣紀正　　国立医薬品食品衛生研究所生化学部　第三室長

**研究要旨**

本研究では、遺伝子改変技術応用食品のアレルゲン性について、より高い精度での評価・予測を可能とすることを目的として、アレルゲン性予測手法(allerStat)に関して詳細な性能検証を行った。令和 4 年度の検証項目に加え、今年度は比較対照のアレルゲン性予測ツールとして AlgPred2 を、また評価指標として AUC-10%、F1 スコア、MCC を追加した。その結果、allerStat は、全ての評価指標において他の 5 種の予測ツールよりも高い値を示した。今後バイオテクノロジー技術応用食品のリスク評価手法としての実用化を目指す。また、国立医薬品食品衛生研究所にて運用・公開しているアレルゲンデータベース(Allergen Database for Food Safety, ADFS)に関して、令和 4 年 6 月から令和 5 年 5 月までの 1 年間に NCBI PubMed に収載された論文から、エピトープ配列決定に関する 20 報のピアレビューを行い、22 種のアレルゲンについて、総数 87 のエピトープ情報を 追加した。これらの情報更新により、ADFS のアレルゲン及びイソアレルゲンのアミノ酸配列情報は 2,408、エピトープ既知のアレルゲン数は 286 となり、遺伝子改変技術応用食品のアレルゲン性評価に有用なデータベースである ADFS を充実させることができた。

## A. 研究目的

遺伝子改変技術を応用した食品開発は、技術的には、外来遺伝子導入による遺伝子組換え食品から、内在性遺伝子の改変を行うゲノム編集技術応用食品へ、また、酵母等に多数の外来遺伝子を導入し新規食品機能成分を産生させる合成生物学の利用へと変化している。現在、ゲノム編集技術では多様な手法が生み出されており、これらの手法による意図しない塩基変化も一様ではないことが明らかになりつつある。従って、このような意図しない変化、及びそこから生じる代謝成分の変化を検出または予測し、その変化が与える影響を正確に評価することは、食品の安全性確保において急務の課題である。

バイオテクノロジー技術を用いて開発された食品のリスクの 1 つに、アレルゲン性増大の可能性がある。本研究では、国立医薬品食品衛生研究所生化学部にて管理・公開している、アレルゲン性予測機能（FAO/WHO 法等）を装備したアレルゲン・エピトープ情報データベース(Allergen Database for Food Safety, ADFS, https://allergen.nihs.go.jp/ADFS/)について、新規エピトープ情報の収集・解析等によりアレルゲン性評価に関する検討を行い、遺伝子改変技術応用食品のリスク評価に資するデータベースとなるよう、情報を更新し内容を充実させる。

また、人工知能（AI）を活用した新規高精度アレルゲン性予測手法の開発を進める。令和 2 年度までの先行研究班では、アレルゲン及び非アレルゲンタンパク質から抽出した特徴的なアミノ酸配列パターンを利用して機械学習によりアレルゲン性を予測する手法(allerStat)を開発してきた。本研究班では、この予測システムの詳細な性能検証を行い、高精度アレルゲン性予測法としての実用化を進める。

## B．研究方法

<u>アレルゲン性予測手法（allerStat)の性能検証</u>

allerStat の予測性の検証においては、これまで Allerdictor、Allertop、及び MEME の 3 種の予測法を比較対照とし、Leave-Category-Out Cross-Validation における ROC 曲線の AUC を評価指標としてきた。今年度はさらに、比較対照に AlgPred2 を、評価指標に AUC-10%、F1 スコア、MCC (Matthews Correlation Coefficient)を追加し、より詳細な性能検証を行った。また、allerStat 開発時に使用したアレルゲンデータは COMPARE アレルゲンデータベース（https://comparedatabase.org/）2020 年版のデータであった。今年度はその後の 3 年間に更新された COMPARE アレルゲンデータを確認し、allerStat の学習データに反映させた。

<u>ADFS エピトープ情報の追加</u>

令和 4 年 6 月から令和 5 年 5 月までの 1 年間に NCBI PubMed に収載された論文から、キー

ワード検索により、エピトープ決定に関するものを抽出した。キーワードとしては、IgE、epitope、linear、conformational、sequence、recognition等々のワードを使用し、これらを複数組み合わせて 6 通りの検索式を作成して検索を行った。この検索により抽出されてきた論文についてピアレビューを行った。その結果エピトープ情報を報告していると判断された論文について、そのエピトープ情報を整理し、ADFSのデータに追加した。

ADFS FAO/WHO 法アレルゲン性予測ツールの改修

ADFS に搭載している FAO/WHO 法アレルゲン性予測ツールについて、従来使用していた FAO/WHO 法の改変法から本来の FAO/WHO 法に改修し、アレルゲン性予測ツールとしてより精密な解析情報を得られるよう改良した。

## C.研究結果

アレルゲン性予測手法（allerStat)の性能検証

allerStat の予測性の検証においては、これまで Allerdictor（固定長 k の連続したアミノ酸配列解析（k=6))、Allertop（アミノ酸の物理化学的性質に基づくアプローチ)、及び MEME（Multiple Expectation maximizations for Motif Elicitation、期待値最大化を利用した配列パターンマイニング法）の 3 種の予測法を比較対照とし、Leave-Category-Out Cross-Validation における ROC 曲線の AUC を評価指標としてきた。今年度は、比較対照として AlgPred2（固定長 k の連続したアミノ酸配列解析（k=1, 2))を追加し、allerStat の予測性を検証した。図 1 には Leave-Category-Out Cross-Validation における各カテゴリー（食品）の ROC 曲線を、表 1 にはそれぞれの予測法の ROC-AUC を示す。カテゴリーによって違いはあるが、AUC 平均値は allerStat で最も大きく、他の 5 種の予測法と比較して高い予測性能を有することが示された。

続いて、ROC-AUC 以外の評価指標として、AUC-10%、F1 スコア、MCC を追加し、各予測法の性能評価を行った。

・AUC-10%：ROC 曲線横軸の False positive rate が 10%までの AUC を 10 倍した値。

・F1 スコア：Recall（再現率：真の陽性のうち陽性と判断された割合、感度）と Precision（適合率：陽性と判断されたうち真の陽性である割合）との調和平均

・MCC：真陽性(TP)、偽陽性(FP)、偽陰性(FN)、真陰性(TN)の全てを考慮する精度指標。次の式で表される。

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

結果を表 2 に示す。AUC-10%、F1 スコア、MCC に関しても、AUC の場合と同様に allerStat で最も大きな値となり、やはり allerStat が他の予測法と比較して高い予測性能を有することが示された。

allerStat 開発にあたり当初使用したアレルゲンデータは、COMPARE アレルゲンデータベース 2020 年版のデータであった。そこで、その後の COMPARE データの更新状況を確認したところ、2021 年に 104 個、2022 年に 118 個、2023 年に 171 個、合計 393 個の新規データが追加されており、10 個のデータが削除されていた。この結果を allerStat の学習データに反映させ、更新後のアレルゲンデータは 2,631 個（383 個の増）となった。この学習データを用いて再度学習及び予測性能検証を行ったところ、予測性能の顕著な向上は特に見られなかった。

ADFS エピトープ情報の追加

令和 4 年 6 月から令和 5 年 5 月までの 1 年間で、キーワード検索により抽出された論文は 33 報であった。その中からエピトープ情報が記載されていると思われる 20 報を選択し（表 3)、ピアレビューを行った。その結果、11 報の論文から 22 種のアレルゲンについて、総数 87 種のエピトープ情報を抽出し、ADFS に追加した（表 4)。

上記のエピトープ情報更新作業により、最終的に、ADFS のアレルゲン及びイソアレルゲンのアミノ酸配列情報は 2,408、エピトープ既知のアレルゲン数は 286、構造既知のアレルゲン数は 194、糖鎖付加アレルゲン数は 127 となった。

ADFS FAO/WHO 法アレルゲン性予測ツールの改修

FAO/WHO のアレルゲン性予測法は、1)クエリタンパク質を N 末端側から 80 残基のウインドウで順に区切ってゆき（sliding window 方式)、FASTA アラインメントにて既知アレルゲンと 35%以上のアミノ酸が一致する場合、あるいは、2)クエリタンパク質の連続する 6-8 アミノ酸が既知アレルゲンと完全に一致する場合にアレルゲン性が疑われる、とするものである。このうち 1)の方法は、ウインドウ単位に細分化された配列を大量に処理する必要があり計算速度の遅延を招くと予想されたことから、ADFS ではこれまで改変法を用いてきた。改変法では、クエリタンパク質全長に対して FASTA アラインメントにより既知アレルゲンとの相同性比較を行い、80 残基以上の領域でオーバーラップが認められ、かつその 35%以上のアミノ酸が一致する場合を陽性と判定する。しかし、オーバーラップ長全体では一致率＜35%であっても、80 残基 sliding window では一致率＞35%となる場合が見られることが分

かった。そこで、現在の ADFS サーバは十分なスペックを有していることも考慮し、FAO/WHO 法の 1)を本来の方法に改修し公開した。この改修により、ADFS の FAO/WHO 法は、80 残基 sliding window 毎のより詳細な解析情報を得られるアレルゲン性予測ツールとなった。

## D. 考察

本研究では、AI を活用した新規アレルゲン性予測手法開発に向けて、これまでよりも詳細な予測性能の検証を行った。比較対照のアレルゲン性予測ツールとして AlgPred2（固定長 k の連続したアミノ酸配列解析（k=1, 2））を追加し、また、評価指標として AUC-10%、F1 スコア、MCC を追加した。その結果、allerStat は、全ての評価指標において他の 5 種の予測ツールよりも高い値を示し、非常に高性能のアレルゲン性予測ツールであることが示された。今後、バイオテクノロジー技術応用食品のリスク評価手法としての実用化・公開に向けて整備を進める。

ADFS に関しては、22 種のアレルゲンについて総数 87 のエピトープ情報を追加した。また、FAO/WHO 法のアレルゲン予測ツールをこれまでの改変法から本来の方法に改修して公開し、アレルゲンデータベースとしての継続的な充実を進めた。

## E．結論

本研究では、遺伝子改変技術応用食品のアレルゲン性について、より高い精度での評価・予測を可能とすることを目的として、アレルゲン性予測手法(allerStat)の性能検証及び機能拡充を行った。また、令和 4 年 6 月から令和 5 年 5 月までの 1 年間に NCBI PubMed に収載された論文から、エピトープ配列決定に関する 20 報のピアレビューを行い、22 種のアレルゲンについて、総数 87 のエピトープ情報を ADFS に追加した。さらに、FAO/WHO 法のアレルゲン予測ツールをこれまでの改変法から本来の方法に改修した。これらの更新により、遺伝子改変技術応用食品のアレルゲン性評価に有用なデータベースである ADFS を充実させることができた。

## F. 研究発表

1. 論文発表

1) Goto K, Tamehiro N, Yoshida T, Hanada H, Sakuma T, Adachi R, Kondo K, Takeuchi I. AllerStat: Finding Statistically Significant Allergen-Specific Patterns in Protein Sequences by Machine Learning. J Biol Chem. 2023 Jun;299(6):104733.

2. 学会発表

1) 爲廣紀正. 新開発食品の安全性評価～アレルギー誘発性～. 2023 年 12 月 1 日、オンライン開催、令和 5 年度東京農業大学総合研究所研究会【食の安全と安心部会】第 6 回シンポジウム

## H. 知的財産権の出願・登録状況

1. 特許取得
   なし

2. 実用新案登録
   なし

3. その他
   なし

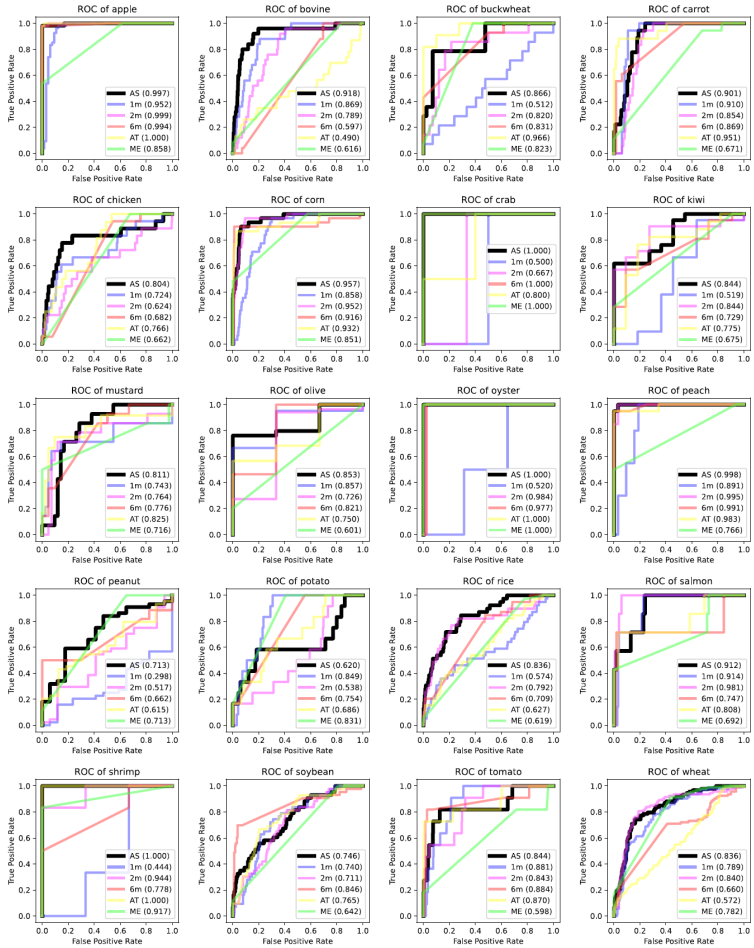



図1　leave-category-out cross-validationによる評価結果：ROC曲線
AS: allerStat, 1m: AlgPred2 1mer, 2m: AlgPred2 2mer, 6m: Allerdictor, AT: Alllertop, ME: MEME.

表1　leave-category-out cross-validationによる評価結果：ROC-AUC

| Category | allerStat | AlgPred2 | | Allerdictor | Allertop | MEME |
|---|---|---|---|---|---|---|
| | | 1mer | 2mer | 6mer | | |
| Apple | 0.9969 | 0.9524 | 0.9991 | 0.9943 | 0.9997 | 0.8575 |
| Bovine | 0.9182 | 0.8689 | 0.7894 | 0.5966 | 0.4898 | 0.6159 |
| Buckwheat | 0.8656 | 0.5119 | 0.8197 | 0.8308 | 0.9659 | 0.8231 |
| Carrot | 0.9014 | 0.9099 | 0.8543 | 0.869 | 0.9515 | 0.6712 |
| Chicken | 0.8037 | 0.7241 | 0.6242 | 0.6817 | 0.7662 | 0.6616 |
| Corn | 0.9569 | 0.8583 | 0.9522 | 0.9165 | 0.9321 | 0.8505 |
| Crab | 1 | 0.5 | 0.6667 | 1 | 0.8 | 1 |
| Kiwi | 0.8442 | 0.5195 | 0.8442 | 0.7294 | 0.7754 | 0.6753 |
| Mustard | 0.8112 | 0.7432 | 0.7636 | 0.7755 | 0.8254 | 0.716 |
| Olive | 0.8532 | 0.8571 | 0.7262 | 0.8214 | 0.75 | 0.6012 |
| Oyster | 1 | 0.5195 | 0.9844 | 0.9766 | 1 | 1 |
| Peach | 0.9984 | 0.8906 | 0.9953 | 0.9906 | 0.9828 | 0.7656 |
| Peanut | 0.7126 | 0.2981 | 0.5174 | 0.6618 | 0.6151 | 0.7132 |
| Potato | 0.6201 | 0.8492 | 0.5376 | 0.754 | 0.6859 | 0.8308 |
| Rice | 0.8362 | 0.5744 | 0.7915 | 0.7086 | 0.6265 | 0.6186 |
| Salmon | 0.9124 | 0.9143 | 0.9806 | 0.7475 | 0.8079 | 0.6922 |
| Shrimp | 1 | 0.4444 | 0.9444 | 0.7778 | 1 | 0.9167 |
| Soybean | 0.7463 | 0.7395 | 0.7107 | 0.8461 | 0.7648 | 0.6416 |
| Tomato | 0.8439 | 0.8807 | 0.8432 | 0.8842 | 0.8697 | 0.5976 |
| Wheat | 0.8364 | 0.7889 | 0.8401 | 0.6604 | 0.5723 | 0.7823 |
| Average | 0.8729 | 0.7173 | 0.8092 | 0.8111 | 0.809 | 0.7516 |

表2　leave-category-out cross-validationにおける種々の指標による評価結果

| Criterion | Category | allerStat | AlgPred2 | | Allerdictor | Allertop | MEME |
|---|---|---|---|---|---|---|---|
| | | | 1mer | 2mer | 6mer | | |
| AUC | Average | 0.8729 | 0.7173 | 0.8111 | 0.8111 | 0.809 | 0.7516 |
| AUC-10% | Average | 0.7697 | 0.58399 | 0.6859 | 0.7404 | 0.7561 | 0.6787 |
| F1 | Average | 0.5174 | 0.0066 | 0.4051 | 0.4939 | 0.4577 | 0.299 |
| MCC | Average | 0.4495 | 0.0007 | 0.316 | 0.4388 | 0.4249 | 0.3068 |

# 表3　令和5年度にエピトープ情報収集のためピアレビューした論文

| | |
|---|---|
| | Zabel M, Weber M, Kratzer B, Köhler C, Jahn-Schmid B, Gadermaier G, Gattinger P, Bidovec-Stojkovič U, Korošec P, Smole U, Wurzinger G, Chen KW, Panaitescu CB, Klimek L, Pablos I, Niespodziana K, Neunkirchner A, Keller W, Valenta R, Pickl WF. |
| | Art v 1 IgE epitopes of patients and humanized mice are conformational |
| | ***J Allergy Clin Immunol*** 2022 Oct;150(4):920-930 |
| 1. | PMID：35738928 |
| | Sharma A, Vashisht S, Gaur SN, Batra JK, Arora N. |
| | Immuno-informatic Prediction of B and T cell Epitopes of Cysteine Protease Allergen from Phaseolus vulgaris with Cross-reactive Potential and Population Coverage |
| | ***Curr Protein Pept Sci*** 2022;23(7):475-494 |
| 2. | PMID：35927799 |
| | Caballero LR, Treudler R, Delaroque N, Simon JC, Kern K, Szardenings M. |
| | Peptide epitopes as biomarkers of soya sensitization in rBet v 1 immunotherapy of birch-related soya allergy |
| | ***Clin Exp Allergy*** 2023 Mar;53(3):316-326 |
| 3. | PMID：36102274 |
| | Li MS, Xia F, Liu Q, Chen Y, Yun X, Liu M, Chen GX, Wang L, Cao MJ, Liu GM. |
| | IgE Epitope Analysis for Scy p 1 and Scy p 3, the Heat-Stable Myofibrillar Allergens in Mud Crab |
| | ***J Agric Food Chem*** 2022 Sep 28;70(38):12189-12202 |
| 4. | PMID：36110087 |
| | Huan F, Gao S, Han TJ, Liu M, Li MS, Yang Y, Chen YY, Lai D, Cao MJ, Liu GM. |
| | Identification of the Immunoglobulin E Epitope of Arginine Kinase, an Important Allergen from Crassostrea angulata |
| | ***J Agric Food Chem*** 2022 Oct 19;70(41):13419-13430 |
| 5. | PMID：36205062 |
| | Pi X, Liu J, Sun Y, Ban Q, Cheng J, Guo M. |
| | Heat-induced changes in epitopes and IgE binding capacity of soybean protein isolate |
| | ***Food Chem*** 2023 Mar 30:405(Pt A):134830 |
| 6. | PMID：36370556 |
| | Hendrich JM, Wangorsch A, Rödel K, Jacob T, Mahler V, Wöhrl BM. |
| | Allergenicity and IgE Recognition of New Dau c 1 Allergens from Carrot |
| | ***Mol Nutr Food Res*** 2023 Feb;67(3):e2200421 |
| 7. | PMID：36458641 |
| | Zhang Y, Bhardwaj SR, Vilches A, Breksa A, Lyu SC, Chinthrajah S, Nadeau KC, Jin T. |
| | IgE binding epitope mapping with TL1A tagged peptides |
| | ***Mol Immunol*** 2023 Jan;153:194-199 |
| 8. | PMID：36527758 |
| | Chen Y, Jin T, Li M, Yun X, Huan F, Liu Q, Hu M, Wei X, Zheng P, Liu G. |
| | Crystal Structure Analysis of Sarcoplasmic-Calcium-Binding Protein: An Allergen in Scylla paramamosain |
| | ***J Agric Food Chem*** 2023 Jan 18;71(2):1214-1223 |
| 9. | PMID：36602420 |
| | Huang Y, Li Z, Wu Y, Li Y, Pramod S, Chen G, Zhu W, Zhang Z, Wang H, Lin H. |
| | Comparative analysis of allergenicity and predicted linear epitopes in α and ß parvalbumin from turbot (Scophthalmus maximus) |
| | ***J Sci Food Agric*** 2023 Mar 30;103(5):2313-2324 |
| 10. | PMID：36606403 |
| | Ferrari E, Breda D, Spisni A, Burastero SE. |
| | Component-Resolved Diagnosis Based on a Recombinant Variant of Mus m 1 Lipocalin Allergen |
| | ***Int J Mol Sci*** 2023 Jan 7;24(2):1193 |
| 11. | PMID：36674705 |
| | Kronfel CM, Cheng H, McBride JK, Nesbit JB, Krouse R, Burns P, Cabanillas B, Crespo JF, Ryan R, Simon RJ, Maleki SJ, Hurlburt BK. |
| | IgE epitopes of Ara h 9, Jug r 3, and Pru p 3 in peanut-allergic individuals from Spain and the US |
| | ***Front Allergy*** 2023 Jan 9:3:1090114 |
| 12. | PMID：36698378 |
| | Li T, Han K, Feng G, Guo J, Wang J, Wan Z, Wu X, Yang X. |
| | Bile Acid Profile Influences Digestion Resistance and Antigenicity of Soybean 7S Protein |
| | ***J Agric Food Chem*** 2023 Feb 15;71(6):2999-3009 |
| 13. | PMID：36723618 |
| | Foo ACY, Nesbit JB, Gipson SAY, DeRose EF, Cheng H, Hurlburt BK, Kulis MD, Kim EH, Dreskin SC, Mustafa S, Maleki SJ, Mueller GA. |
| | Structure and IgE Cross-Reactivity among Cashew, Pistachio, Walnut, and Peanut Vicilin-Buried Peptides |
| | ***J Agric Food Chem*** 2023 Feb 15;71(6):2990-2998 |
| 14. | PMID：36728846 |
| | Zhao S, Pan F, Cai S, Yi J, Zhou L, Liu Z. |
| | Secrets behind Protein Sequences: Unveiling the Potential Reasons for Varying Allergenicity Caused by Caseins from Cows, Goats, Camels, and Mares Based on Bioinformatics Analyses |
| | ***Int J Mol Sci*** 2023 Jan 27;24(3):2481 |
| 15. | PMID：36768806 |
| | Figo DD, Cordeiro Macedo PR, Gadermaier G, Remuzgo C, Castro FFM, Kalil J, Galvão CES, Santos KS. |
| | IgE and IgG4 Epitopes of Dermatophagoides and Blomia Allergens before and after Sublingual Immunotherapy |
| | ***Int J Mol Sci*** 2023 Feb 20;24(4):4173 |
| 16. | PMID：36835585 |
| | Zhang A, Zhao H, Pei S, Chi Y, Fan X, Liu J. |
| | Identification and Structure of Epitopes on Cashew Allergens Ana o 2 and Ana o 3 Using Phage Display |
| | ***Molecules*** 2023 Feb 16;28(4):1880 |
| 17. | PMID：36838874 |
| | Wang Q, Ju D, Gao J, Tong P, Chen H. |
| | Epitope Mapping of Lysozyme Using the Chinese Egg-Allergic Sera at Both Pooled and Individual Levels |
| | ***J Agric Food Chem*** 2023 Apr 26;71(16):6458-6467 |
| 18. | PMID：37053565 |
| | Terán MG, García-Ramírez B, Mares-Mejía I, Ortega E, O'Malley A, Chruszcz M, Rodríguez-Romero A. |
| | Molecular Basis of Plant Profilins' Cross-Reactivity |
| | ***Biomolecules*** 2023 Mar 28;13(4):608 |
| 19. | PMID：37189355 |
| | Zhang Y, Song M, Xu J, Li X, Yang A, Tong P, Wu Z, Chen H. |
| | IgE Recognition and Structural Analysis of Disulfide Bond Rearrangement and Chemical Modifications in Allergen Aggregations in Roasted Peanuts |
| | ***J Agric Food Chem*** 2023 Jun 14;71(23):9110-9119 |
| 20. | PMID：37256970 |

# 表4　令和5年度に新たにADFSに追加したエピトープ情報

| | Name | start | end | Sequence | Method | CTYPE | Reference | UniProt acc.No |
|---|---|---|---|---|---|---|---|---|
| 001 | Art v 1 | | | | ELISA/Cell based assay | C | PMID 35738928 | Q84ZX5 |
| 002 | Pha v ? | 130 | 146 | DWRLKGAVGPIKDQGNC | ELISA/Cell based assay | L | PMID 35927799 | Q41110 |
| | Pha v ? | 173 | 183 | SEQELVDCDR | ELISA/Cell based assay | L | PMID 35927799 | Q41110 |
| | Pha v ? | 210 | 225 | EDYPTYQGIDGTCDETK | ELISA/Cell based assay | L | PMID 35927799 | Q41110 |
| 003 | Api g 1 | 92 | 101 | LGFIESIENH | ıge display/peptide microa | L | PMID 36102274 | P49372 |
| 004 | Gly m 4 | 57 | 66 | TFLEDGETKF | ıge display/peptide microa | L | PMID 36102274 | P26987 |
| 005 | Scy p 1 | 60 | 71 | AQEQLSAANTKL | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 77 | 90 | ALQNAEGEVAALNR | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 101 | 110 | RSEERLNTAT | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 113 | 125 | LAEASQAADESER | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 126 | 139 | MRLVLENRSLSDEE | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 140 | 162 | RMDALENQLKEARFLAEEADRKY | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 163 | 175 | DEVARKLAMVEAD | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 176 | 188 | LERAEERAESGES | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 189 | 210 | KIVELEEELRVVGNNLKSLEVS | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | 211 | 231 | EEKANQREETYKEQIKTLANK | Phage display/Dot blot/C | L | PMID 36110087 | A7L5V2 |
| | Scy p 1 | | | K12, E164, Y267 | Phage display/Dot blot/C | C | PMID 36110087 | A7L5V2 |
| 006 | Scy p 3 | 7 | 24 | ARDVERAKFAFSIYDFEG | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | 35 | 50 | DCLRALNLNPTLAVIE | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | 51 | 61 | KVGGKTKKKEK | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | 66 | 87 | DDFLPIFAQVKKDKDAGSFEDF | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | 96 | 106 | KTENGTMLYAE | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | 107 | 125 | LEHILLSLGERLEKSELEP | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | 135 | 147 | DEDGFIPYEPFLK | Phage display/Dot blot/C | L | PMID 36110087 | A0A514C9K9 |
| | Scy p 3 | | | C36, Y52, D66、K77、N99,I110、P141 | Phage display/Dot blot/C | C | PMID 36110087 | A0A514C9K9 |
| 007 | Cra a 2 | 105 | 118 | PLDATGEFIVSTRV | Phage display/Dot blot/El | L | PMID 36205062 | |
| | Cra a 2 | 236 | 247 | YKRLVSAIKQLE | Phage display/Dot blot/El | L | PMID 36205062 | |
| | Cra a 2 | | | L7, M187, Q227, H277 | Phage display/Dot blot/El | C | PMID36205062 | |
| 008 | Ara h 2 | 90 | 105 | DRRGAGSSQHQERCCN | ELISA/Cell based assay | L | PMID 36527758 | Q6PSU2 |
| | Ara h 2 | 111 | 126 | ENNQRCMCEALQQIME | ELISA/Cell based assay | L | PMID 36527758 | Q6PSU2 |
| 009 | Sco m ? | 17 | 28 | EECKKPESFCHK | Western Blot | L | PMID 36606403 | A0A2U9CSM9 |
| | Sco m ? | 51 | 65 | IDQDKSGFIEEDELK | Western Blot | L | PMID 36606403 | A0A2U9CSM9 |
| | Sco m ? | 73 | 86 | AGARSLTDTETKNL | Western Blot | L | PMID 36606403 | A0A2U9CSM9 |
| 010 | Sco m ? | 34 | 45 | LAKKSADDVKA | Western Blot | L | PMID 36606403 | A0A2U9CS92 |
| | Sco m ? | 72 | 83 | AGARALTDKETA | Western Blot | L | PMID 36606403 | A0A2U9CS92 |
| | Sco m ? | 91 | 104 | ADGDGKIGIEEFAN | Western Blot | L | PMID 36606403 | A0A2U9CS92 |
| 011 | Ara h 9.0201 | 1 | 15 | LSCGQVNSALAPCIT | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 6 | 20 | VNSALAPCITFLTKG | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 11 | 25 | APCITFLTKGGVPSG | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 16 | 30 | FLTKGGVPSGPCCSG | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 21 | 35 | GVPSGPCCSGVRGLL | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 26 | 40 | PCCSGVRGLLGAAKT | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 46 | 60 | AACNCLKAAAGSLHG | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 66 | 80 | AAALPGRCGVSIPYK | Peptide array | L | PMID 36698378 | B6CG41 |
| | Ara h 9.0201 | 71 | 85 | GRCGVSIPYKISTST | Peptide array | L | PMID 36698378 | B6CG41 |
| 012 | Jug r 3.0101 | 26 | 40 | AVITCGQVASSVGSC | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 36 | 50 | SVGSCIGYLRGTVPT | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 41 | 55 | IGYLRGTVPTVPPSC | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 51 | 65 | VPPSCCNGVKSLNKA | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 56 | 70 | CNGVKSLNKAAATTA | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 66 | 80 | AATTADRQAACECLK | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 96 | 110 | GLPGKCGVSVPYKIS | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 101 | 115 | CGVSVPYKISTSTNC | Peptide array | L | PMID 36698378 | C5H617 |
| | Jug r 3.0101 | 106 | 119 | PYKISTSTNCKAVK | Peptide array | L | PMID 36698378 | C5H617 |
| 013 | Pru p 3.03 | 31 | 45 | IQAGLAPCLGYLQRG | Peptide array | L | PMID 36698378 | B6CQU7 |
| | Pru p 3.03 | 46 | 60 | GVPAGGCCPGIKRLV | Peptide array | L | PMID 36698378 | B6CQU7 |
| | Pru p 3.03 | 91 | 105 | AAALPSLCGVKIPYK | Peptide array | L | PMID 36698378 | B6CQU7 |
| | Pru p 3.03 | 96 | 110 | SLCGVKIPYKISAST | Peptide array | L | PMID 36698378 | B6CQU7 |
| 014 | Der p 1 | 149 | 163 | RNQSLDLAEQELVDC | Peptide array | L | PMID 36835585 | P08176 |
| | Der p 1 | 296 | 310 | DNGYGYFAANIDLMM | Peptide array | L | PMID 36835585 | P08176 |
| 015 | Der p 2 | 24 | 38 | DCANHEIKKVLVPGC | Peptide array | L | PMID 36835585 | Q1H8P8 |
| 016 | Der p 10 | 110 | 124 | TAKLEEASQSADESE | Peptide array | L | PMID 36835585 | O18416 |
| | Der p 10 | 146 | 160 | NQLKEARMMAEDADR | Peptide array | L | PMID 36835585 | O18416 |
| | Der p 10 | 149 | 163 | KEARMMAEDADRKYD | Peptide array | L | PMID 36835585 | O18416 |
| | Der p 10 | 170 | 184 | AMVEADLERAEERAE | Peptide array | L | PMID 36835585 | O18416 |
| | Der p 10 | 194 | 208 | EEELRVVGNNLKSLE | Peptide array | L | PMID 36835585 | O18416 |
| | Der p 10 | 212 | 226 | EKAQQREEAHEQQIR | Peptide array | L | PMID 36835585 | O18416 |
| | Der p 10 | 257 | 271 | EDELVHEKEKYKSIS | Peptide array | L | PMID 36835585 | O18416 |
| 017 | Blo t 5 | 19 | 33 | EHKPKKDDFRNEFDH | Peptide array | L | PMID 36835585 | O96870 |
| | Blo t 5 | 25 | 39 | DDFRNEFDHLLIEQA | Peptide array | L | PMID 36835585 | O96870 |
| 018 | Blo t 6 | 187 | 201 | NLQVGELKIVSQEEC | Peptide array | L | PMID 36835585 | A1KXI3 |
| 019 | Blo t 12 | 22 | 36 | DEQTTRGRHTEPDDH | Peptide array | L | PMID 36835585 | Q17282 |
| | Blo t 12 | 97 | 111 | EEGPIHIQEMCNKYI | Peptide array | L | PMID 36835585 | Q17282 |
| 020 | Ana o 2 | 109 | 113 | QQGRQ | Phage display | L | PMID 36838874 | Q8GZP6 |
| | Ana o 2 | 114 | 120 | QQQSGRF | Phage display | L | PMID 36838874 | Q8GZP6 |
| | Ana o 2 | 182 | 187 | PKDVFQ | Phage display | L | PMID 36838874 | Q8GZP6 |
| | Ana o 2 | 219 | 225 | IKQLKSE | Phage display | L | PMID 36838874 | Q8GZP6 |
| 021 | Ana o 3 | 10 | 24 | AFAVLLLVANASIYR | Phage display | L | PMID 36838874 | Q8H2B8 |
| | Ana o 3 | 13 | 27 | VLLLVANASIYRAIV | Phage display | L | PMID 36838874 | Q8H2B8 |
| | Ana o 3 | 40 | 48 | QRQFEEQQR | Phage display | L | PMID 36838874 | Q8H2B8 |
| | Ana o 3 | 66 | 71 | YNQRQE | Phage display | L | PMID 36838874 | Q8H2B8 |
| | Ana o 3 | 101 | 107 | QQEQIKG | Phage display | L | PMID 36838874 | Q8H2B8 |
| | Ana o 3 | 108 | 115 | EEVRELYE | Phage display | L | PMID 36838874 | Q8H2B8 |
| | Ana o 3 | 116 | 123 | TASELPRI | Phage display | L | PMID 36838874 | Q8H2B8 |
| 022 | Gal d 4 | 42 | 52 | SLGNWVCAAKF | Peptide array | L | PMID 37053565 | P00698 |
| | Gal d 4 | 99 | 109 | SALLSSDITAS | Peptide array | L | PMID 37053565 | P00698 |
| | Gal d 4 | 120 | 130 | GNGMNAWVAWR | Peptide array | L | PMID 37053565 | P00698 |

厚生労働科学研究費補助金（食品の安全確保推進研究事業）
「新たなバイオテクノロジーを用いて得られた食品の安全性確保と
リスクコミュニケーションのための研究」
分担研究報告書（令和 5 年度）

# 新規アレルゲン性評価手法開発の基盤研究と AI のリスク評価への応用

研究分担者　富井　健太郎　　産業技術総合研究所

研究要旨：
アレルギー発症機構には、アレルゲン性タンパク質由来ペプチド－HLA クラス II 分子間相互作用が大きく関与しているものと考え、近年提案された深層学習モデルベースのアレルゲン性予測パイプラインの一つである NetAllergen に、新たな特徴量としてアミノ酸の物理化学的インデックスを組み合わせて改良した DeepSeqPanII 由来の相互作用予測モデルを組み込むことで、アレルゲン性の予測性能を比較、検証した。今回用いたアレルゲン性タンパク質および非アレルゲン性タンパク質の双方を含む 2,780 タンパク質からなるデータセットに対する AUC 値による性能評価結果では、改良した DeepSeqPanII により計算される分子間相互作用に関する特徴量を用いることで、NetAllergen と同等のアレルゲン性予測を達成可能であることが明らかとなった。

研究協力者
坂無　英徳　　産業技術総合研究所
猪浦　裕子　　産業技術総合研究所

## A. 研究目的

世界の人口増加、温暖化による砂漠化などによる農地面積の減少、作物収穫量の低下、農業従事者の減少などの諸問題によって世界的な食糧不足への懸念が強まっている。こうした問題を背景に、害虫抵抗性や除草剤耐性をもたせることで収量増を見込める遺伝子組み換え作物（GMO）の実用化が進んでいる。日本では遺伝子組み換え作物規制条例で栽培を規制しているが、家畜飼料用がほとんどではあるものの、輸入に依存しているトウモロコシ、ダイズ、菜種などは半量が既に遺伝子組換え作物であると推定されている。

GMO のアレルゲン性に関して、遺伝子改変食品の安全性が問われているが、すべての遺伝子組換え食品のアレルゲン性を実験的に評価するのはコスト面からの困難さが想定される。このため、科学的根拠をもつ信頼性のある評価方法の確立が求められている。適切なリスク管理対策の適用により、遺伝子改変食品のアレルゲン性リスクを低減することが出来るかもしれない。

組換え DNA 技術により導入された新規遺伝子産物（タンパク質）や形質転換による意図しない新規タンパク質のアレルゲン性予測方法としては、FAO（国連食糧農業機関）/WHO（世界保健機関）が提唱しているデータベースに登録済みのアレルゲンタンパク質との類似性比較（[1] 80 個の連続したアミノ酸配列について 35%以上の同一残基率、[2] 6～8 個の連続したアミノ酸配列の完全一致）が標準的に使用されている。

しかし、配列長が短い既知アレルゲン性ペプチドとの類似性に基づくため偽陽性の割合が高いことが指摘されている。また、進化に基づくアミノ酸置換行列を用いる配列類似性比較は、オフターゲット効果による変異をもつ新規タンパク質に対するアレルゲン性の判定には十分ではない可能性が考えられる。

そこで本研究では、標的配列と類似した配列の

オフターゲット検索しかできない点を克服すべく、人工知能を活用して配列類似性が明瞭ではないアレルゲン性タンパク質由来ペプチド－HLA クラス II 分子間結合予測法の開発を行うことを目的に取り組んでいる。今年度は、これまでに開発したアミノ酸物理化学的インデックスを新たな特徴量として用い、分子間相互作用予測法 DeepSeqPanII を改良したモデルを利用し、アレルゲン性予測に活用するためのパイプラインを開発した。また既存データセットを用いて開発パイプラインの予測性能を検証した。

## B. 研究方法

（1）パイプラインの構築

予測性能の検証のため、以下の 3 種類のパイプラインを用意した。パイプラインの基盤には、近年提案された NetAllergen を用いた。NetAllergen は、MHC 提示傾向を導入した新しいアレルゲン予測方法である。ここでは、MHC 提示傾向計算部分に、これまでに開発した分子間相互作用予測法 DeepSeqPanII を改良したモデルを利用した。

① MHC 提示傾向を示す特徴量として、NetAllergen で用いられている NetMHCIIpan4.0 から生成された 2 件の特徴量をそのまま使用した。

②MHC 提示傾向を示す特徴量として、NetMHCIIpan4.0 から生成された 2 件の特徴量を除外し、開発した分子間相互作用予測モデルで生成した 1 件の特徴量を使用した。

③MHC 提示傾向を示す特徴量として、NetMHCIIpan4.0 から生成された 2 件の特徴量に加え、さらに開発した分子間相互作用予測モデルで生成した 1 件の特徴量を追加で使用した。

（2）特徴量

今回、予測法に組み込む MHC 提示傾向を示す特徴量は、対象タンパク質のアミノ酸配列中の連続した 15 残基長に基づき計算される。

NetMHCIIpan 4.0 では、バインダー1 および 2 の二種類の特徴量が計算される。アレルギーに積極的に関連するバインダー1 として、HLA-DRB1*04:01, HLA-DQA1*02:01-DQB1*02:02, HLA-DQA1*04:01-DQB1*03:01, HLA-DQA1*01:03-DQB1*06:01, HLA-DQA1*03:01-DQB1*03:02, と HLA-DQA1*05:01-DQB1*02:01 が考慮される。陰性/陽性に関連するバインダー2 では HLA-DRB1*15:01 が考慮される。

DeepSeqPanII を改良したモデルでは、以下の特徴量を用いて分子間相互作用が予測される（One-hot encoding と BLOSUM62 は、DeepSeqPanII において採用されている特徴量である）。

- One-hot encoding：20 種類のアミノ酸を 20 次元の bit ベクトルで表現。
- BLOSUM62：アミノ酸置換行列の値を 23 次元のベクトルで表現。
- AAindex 物理化学的インデックス：20 種類のアミノ酸に AAindex データベースに登録されている全 566 種類のインデックスを、類似したインデックスを相関係数に基づいて非冗長な 62 種類に削減したうえで次元圧縮したものを特徴量として追加。

（3）アレルゲン性予測

近年提案されたアレルゲン予測法である NetAllergen をベースモデルとして、前述した特徴量を組み合わせ、3 種類の予測パイプラインを構築し、それぞれ交差検証テストを行った。

DeepSeqPanII を改良したモデルは、$\alpha$ 鎖、$\beta$ 鎖、ペプチドを入力として $IC_{50}$ を予測する。$\alpha$ 鎖、$\beta$ 鎖の組み合わせとしては、NetAllergen における NetMHCpan4.0 の予測を参考に、HLA-DQA1*02:01-DQB1*02:02, HLA-DQA1*04:01-DQB1*03:01, HLA-DQA1*01:03-DQB1*06:01, HLA-DQA1*05:01-DQB1*02:01, HLA-DQA10301-DQB10302 の 5 つを用いた。

ここで、DeepSeqPanII をベースモデルに用いた利点は、第一に、DeepSeqPanII が再帰的ニューラルネットワーク（Recurrent Neural Network；RNN）の一種である LSTM（long short-term memory）を用いていることである。LSTM の強みは、時系列データの学習や予測（回帰・分類）にあり、HLA クラス II の $\alpha$ 鎖、$\beta$ 鎖、ペプチドで構成される 3 組の一次元配列間における線形での結合状態を考慮する上で、モデルが複雑になり過ぎないことが期待される。第二に、注意機構（attention mechanism）を採用しており、RNN が記憶しきれない過去の情報を記憶にキャッシュすることによって、ニューラルネットワークの内部を可視化することができる長所をもつ。第三に、一連のプログラムが GitHub にて公開されており、MIT License として再利用が認められていることである。したがって、新たに別の特徴量を組み込んで利用することは問題にならない。

（4）交差検証テストと AUC の算出

NetAllergen の論文で使用されたデータセットの内、各タンパク質のアミノ酸配列に対する正解ラベルを確認できた APV_2780.fa のデータセットを予測性能の検証に用いた。タンパク質のアミノ酸配列数は 2,780 件(正例[アレルゲンタンパク質]765 件、負例[非アレルゲンタンパク質]2015 件)であった。

DeepSeqPanII を改良したモデルを利用する場合、計算量削減のため、ペプチドは、APV_2780.fa の各配列について、構造ベースのアレルゲン性ペプチペプチド結合予測法 MHCII3D によって結合の可能性が高いと予測された重複含む 12 箇所のペプチド配列(各配列最大長 15 残基)を用いた。各配列について上述の $\alpha$ 鎖、$\beta$ 鎖、ペプチドの組み合わせ 60 通り(アレル 5 通り×「PDB-HB データセット」のうち DRA*01:01 と対を成す 12 の DRB 鎖につき、12 通り)の予測を実施した。予測値の中で最良の値(最小の値)をその配列の新規の特徴量とした。

開発結合予測モデルには LOMO 検証で高い精度であり本データセットにも使用可能である AAindex 10 次元次元圧縮を用いた。またその中でも最も精度が高かった LOMO DRA01:01-DRB104:04(学習データ IEDB2022)を使用した。

各パイプラインについて 10 分割交差検証による学習・及び予測を行い、

$$\mathrm{FPR} = \mathrm{FP} / (\mathrm{FP} + \mathrm{TN})$$

$$\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN})$$

を算出し、横軸に FPR（偽陽性率）、縦軸に TPR（真陽性率）をとった時に描かれる ROC 曲線下側の面積（AUC：area under the curve）の値に基づき予測性能を評価した。

## C. 研究結果および考察

APV_2780.fa のデータセットに対する交差検証の結果、AUC による評価では、①～③の 3 種類のパイプラインはほぼ同等の予測性能を示した(表 1)。

| パイプライン | 最大 AUC | 平均 AUC | 標準偏差 |
|---|---|---|---|
| ① | 0.9539 | 0.8830 | 0.03709 |
| ② | 0.9554 | 0.8829 | 0.03902 |
| ③ | 0.9558 | 0.8834 | 0.03577 |

表 1 3 種類のパイプラインの AUC

NetAllergen は、既存のアレルゲン性予測ツール(AlgPred 2)を上回る性能を示すため、今回、これまでに開発した DeepSeqPanII を改良したモデルを利用することで更なる予測性能向上を期待した。しかし、今回用いたデータセットでは、元々の問題設定でも 0.883 という高い AUC の値を示していることもあってか、有意な性能向上を示すまでには至らなかった。ただし、NetMHCIIpan4.0 から生成された 2 件の特徴量に加え、さらに開発した分子間相互作用予測モデルで生成した 1 件の特徴量を追加で使用したでは、微弱ながら改善傾向を示しており、予期したように、アレルゲン性タンパク質由来ペプチド－HLA クラス II 間の結合

親和性をより精緻に考慮することが、今後の更なる予測精度改善につながるかもしれない。

また今回の研究では、計算時間削減のため、入力とするペプチドの位置を事前に計算したが、代わりに、一気通貫に計算を行うパイプライン構築も、予測精度改善に向け検討されるべき課題であると考えられる。

## D．結論

深層学習を用いた DeepSeqPanII をベースモデルに、新たな特徴量（アミノ酸物理化学的インデックス、）を組み合わせることで、アレルゲン性タンパク質由来ペプチド－HLA クラス II 間の結合予測モデルの改良を進めてきた。今回、この予測モデルを MHC 提示傾向に関する特徴量を導入した新しいアレルゲン予測法 NetAllergen に組み合わせることで、実際に食物アレルギーの発症が確認されているアレルゲンペプチドデータを用いた予測と評価を行なった。今回の条件下では、明瞭な向上を示すまでには至らなかったものの、同等以上の性能を示すパイプラインを構築することができた。今後、考察欄で述べた課題の解決や他データセットでの検証などを通した改良が期待される。

## E. 研究発表・業績

1. 論文発表

   無し

2. 学会発表

   無し

## F．健康危険情報

該当なし

## G. 知的財産権の出願・登録状況

なし

# III.　研究成果の刊行に関する一覧表

研究成果の刊行に関する一覧表
（令和 5 年度）

書籍

| 著者氏名 | 論文タイトル名 | 書籍全体の編集者名 | 書籍名 | 出版社名 | 出版地 | 出版年 | ページ |
|---|---|---|---|---|---|---|---|
| 小泉　望 | リスクコミュニケーションのために求められること | | ゲノム編集技術～実験上のポイント／産業利用に向けた研究開発動向と安全性周知 | 情報機構 | 東京 | 2023 | p295 -p301 |
| | | | | | | | |

雑誌

| 発表者氏名 | 論文タイトル名 | 発表誌名 | 巻号 | ページ | 出版年 |
|---|---|---|---|---|---|
| 小泉望 | ゲノム編集食品をどう伝えるか | 生活協同組合研究 | 577 | 34-41 | 2024 |
| 小泉望、四方雅人 | ゲノム編集技術 ～実験上のポイント／産業利用に向けた研究開発動向と安全性周知 | 技術情報協会 | | 565 | 2023 |
| Shneha R., Takeda K.F., Yamaguchi Y., Koizumi N. | A comparative analysis of attitudes towards genome-edited food among Japanese public and scientific community | PLOS ONE | In press | | 2024 |
| Goto K, Tamehiro N, Yoshida T, Hanada H, Sakuma T, Adachi R, Kondo K, Takeuchi I. | Novel machine learning method allerStat identifies statistically significant allergen-specific patterns in protein sequences | *J Biol Chem.* | 296 | 104733 | 2023 |
| Soga K, Taguchi C, Sugino M, Egi T, Narushima J, Yoshiba S, Takabatake R, Kondo K, Shibata N | Investigation of genetically modified maize imported into Japan in 2021/2022 and the applicability of Japanese official methods. | *Food Hyg. Saf. Sci.* | 64 | 218-225 | 2023 |