

食品衛生基準科学研究費補助金(食品安全科学研究事業)
「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推進のための研究」
令和 6 年度 分担研究報告書

ゲノム上の意図しない変化を網羅的に解析する手法の開発と整備に関する研究

研究分担者 曾我慶介 国立医薬品食品衛生研究所 生化学部 第二室長

研究要旨

本研究では、ゲノム編集技術応用食品等の安全性評価項目として挙げられている外来遺伝子の残存の確認やオフターゲット箇所の予測に関する既存手法を改良し、現在の科学技術に即した実行性の高い方法論を整備することを目的としている。今年度は、解析が困難な箇所として、同じ外来配列が繰り返して挿入されているケースをモデルに、残存が想定される T-DNA 情報を基に次世代シーケンサーを用いた全ゲノムシーケンス解析を検討した。繰り返しを有する外来配列はショートリードシーケンスを用いたアセンブリ解析は困難であったが、挿入されている外来配列より長いロングリードシーケンスを用いたアセンブリ解析が有効であることを示した。さらに、アセンブリ解析結果の妥当性を確認するために、想定挿入配列にマッピングしたリードを抽出し、リファレンスゲノム配列を用いた構造変異解析を行うことで、挿入箇所を推定することが有効であることを示した。オフターゲット予測では、*in vitro* 評価法である SITE-seq 法の偽陽性が多い問題点に着目し、より合理的な解釈を行うため、複数回実施することで蓋然性の高い箇所を絞り込める可能性があることを示した。今後、ゲノム編集技術等を応用し、様々な農作物が開発されることが想定されるが、合理的な安全性評価が可能な体制を構築するために、本研究で得られた知見や方法論が利活用されることを期待する。

研究協力者

成島純平 国立医薬品食品衛生研究所生化学部

次世代シーケンサー(NGS)を用いた確認方法がゲノムワイドに解析可能な方法として注目されている。

A. 研究目的

我が国では 2019 年よりゲノム編集技術応用食品(ゲノム編集食品)等の事前相談・届出制度が開始され、2024 年 10 月 16 日時点において食品衛生上の取扱要領に基づき 7 品目 9 件がゲノム編集食品として届出・公表され、一部はすでに流通している。この制度では、届出対象となるゲノム編集食品の届出情報の一部に①「外来遺伝子及びその一部の残存が無いことの確認に関する情報」および②「確認された DNA の変化がヒトの健康に悪影響を及ぼす新たなアレルゲンの産生及び含有する既知の毒性物質の増加を生じないことの確認に関する情報」が挙げられ、いずれも

届出情報①に関しては、これまでの研究において、NGS による全ゲノムシーケンス(WGS)データを用いたアセンブリ解析(以下、アセンブリ法)の有用性を示している(令和 3～5 年度厚生労働科学研究「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確保とリスクコミュニケーションのための研究」報告書)。一方で、2024 年 4 月には、欧州食品安全機関(EFSA)から EFSA Journal の Statement として「Technical Note on the quality of DNA sequencing for the molecular characterisation of genetically modified plants」が発刊され、NGS 解析の考慮点として挿入配列の性質(反復配列等)的に通常の解析では困難な場合、超ロングリード等を組み合わせる必要性に

ついて触れられている。

届出情報②に関しては、オフターゲットが起こる蓋然性の高いと推定される配列について、CRISPRdirect 等の *in silico* ツールを複数組み合わせ、アレルゲンや既知の毒性物質と相同性検索により照合することを求められる。先行研究において、NGS を活用した *in vitro* オフターゲット予測法 the selective enrichment and identification of tagged genomic DNA ends by sequencing (SITE-Seq) 法で予測される配列の一部に関して、*in silico* ツールのみを用いた場合は予測できないことが確認され、*in silico* および *in vitro* の予測法を組み合わせたスキームの重要性を示している (Narushima et al, Genes Cells, 2022)。一方、その SITE-Seq 法は偽陽性が多い点が問題として挙げられ、データの正確性の向上が求められる。

上述した通り、これら NGS 解析の有用性が挙げられる一方で、いくらか課題も明らかになっており、ゲノム編集食品の届出情報①または②に対応した安全性評価においてより洗練された方法論の整備が求められる。そこで本研究では、通常の NGS 解析では困難な反復配列等にも適用可能なユニバーサルな方法論を構築するために、ロングリードシーケンスの検討を行う。また、偽陽性をより低減化させた *in vitro* オフターゲット評価法の検討を行う。上記を実施することでゲノム上の意図しない変化を網羅的に解析する方法の開発と整備を目指す。

B. 研究方法

同じ外来配列を繰り返し有する作物ゲノムデータ作成

外来遺伝子が繰り返して残存する作物ゲノムモデルとして、カリフラワーモザイクウイルス 35S プロモーター (P35S)、ホスフィトリシンアセチル基転移酵素 (Pat) 遺伝子およびカリフラワーモザイクウイルス 35S ターミネーター (T35S) のコンスト

ラクト (P35S-pat-T35S) が単数体ゲノムに直列で 2 コピー挿入された遺伝子組換え (GM) 大豆 LibertyLinkSoybean (A2704-12 系統) (図 1 参照) をモデルとした。GM 作物の配列情報等が掲載されている公開データベース Nexplorer (<https://bioit-webapp-prod.sciensano.be/nexplorer/>) を参考に、P35S-pat-T35S (1,324 bp) を直列で繋いだ A2704-12 系統の GM コンストラクト (外来遺伝子全長 5,400 bp) を Fasta 形式で作成した (図 1)。この配列をリファレンスゲノムに挿入することで、疑似データを下記の通り作成した。

ケーススタディ①: P35S-pat-T35S 配列が直列で近位に繋がったモデル (P35S-pat-T35S 配列 2 コピー)

大豆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 に A2704-12 系統の GM コンストラクトを挿入。(5,400 bp 挿入)

ケーススタディ②: ケーススタディ①に加え、同 GM コンストラクトが別染色体にも挿入されたモデル (P35S-pat-T35S 配列 4 コピー)

大豆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 および 8 番染色体: chr8: 10,021,534 に A2704-12 系統の GM コンストラクトを挿入 (5,400 bp 挿入×2 箇所)

ショートリードシーケンスのシミュレーションデータ作成・アセンブリ法

概要を図 2A に示した。シミュレーター「ART」を用いて HiSeq X PCR free (150 bp) モード、ペアエンド式を指定して、シミュレーションデータを作成した。ツール「Trim-Galore!」によってトリミング後、アライナー「BWA」で A2704-12 系統の GM コンストラクトにマッピングし、ツール「Samtools」によってマッピングされたリードを抽出した。抽出リードを用いてアセンブラ「SPAdes」を用いてアセンブリを行った。マッピング状況の確認はゲノム可視化ソフトウェア IGV: Integrative Genomics Viewer

(<https://igv.org/>)を用いた。

ロングリードシーケンスのシミュレーションデータ作成・アセンブリ法

概要を図 2B に示した。シミュレーター「PBSIM」によって、全ゲノムシーケンスの qshmm QSHMM-ONT-HQ モデル(ナノポアシーケンサーモデル)、正確性 93.9288%としてシミュレーションデータを作成した。アライナー「minimap2」によって、A2704-12 系統の GM コンストラクトにマッピングし、「Samtools」によってマッピングされたリードを抽出した。抽出リードを用いてアセンブラ「Flye」を用いてアセンブリを行った。マッピング状況の確認は IGV を用いた。アセンブリ法の妥当性確認に GM 大豆リファレンスゲノムとして Williams 82 品種の「Glycine_max_v2.1.dna.toplevel.fa」、Zhonghuang 13 品種の「GWHAAEV000000000.genome.fasta」、Lee 品種の「glyma.Lee.gnm1.BXN C.genome_main.fna」を用いた。リファレンスゲノムに抽出リードを「minimap2」でマッピングし、SV コーラー「Sniffles」(オプション --minsupport 4, --minsvlen 50)を用いて構造変異を検出した。

k-mer 解析

k-mer 解析ツールは伊藤らが開発した「Kmer Analysis 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。そして、A2704-12 の GM コンストラクトを用いて外来遺伝子の検出を行った。

ゲノム DNA の抽出と精製

ヒト胎児腎由来 HEK293T 細胞は、10%FCS とカナマイシンを含む DMEM 培地を用いて、37°C、5% CO₂ の条件下で培養した。

75 cm² フラスコ 2 個分の HEK293T について、0.5%トリプシン-5.3 mM EDTA 溶液で剥離し、冷

却 PBS 緩衝液で 2 回洗浄した。細胞は 10 mL の冷却 PBS で再懸濁し、細胞濃度は約 $2 \times 10^6 \sim 1 \times 10^7$ cell/mLとした。培養細胞からのゲノム DNA の抽出と精製は、DNeasy Blood and Tissue Kit (QIAGEN)を用いて、付属のプロトコルに従って実施した。

ガイド RNA の調製

本検討での標的は、NGS を用いた *in vivo* オフターゲット予測法である Genome-wide, Unbiased Identification of DSBs Enabled by sequencing (GUIDE-Seq) 法の原著論文 (Tsai et al, Nat Biotechnol, 2015) で用いられていた vascular endothelial growth factor A (VEGFA) および Fanconi anemia group F protein (FANCF) の遺伝子配列とした。ガイド RNA の合成と精製には、SureGuide gRNA Synthesis Kit (Agilent)を用いた。付属プロトコルに従い、本検討では、長鎖型の (extended backbone) シングルガイド RNA を合成した。ガイド RNA の調製に使用したオリゴは「表 1」に示した。ガイド RNA 量は、Qubit RNA BR Assay Kit (Thermo Fisher Scientific) で測定し、濃度はその分子量 (108 bp) をもとに算出した。

SITE-Seq ライブラリの調製とシーケンス解析

SITE-Seq のシーケンスライブラリは、以前に報告した方法 (Narushima et al, Genes Cells, 2022) を一部改変して調製した。最終 Cas9 濃度が 64 nM、256 nM、または 1,024 nM となるよう調製し、Cas9 濃度に対して 5 倍のモル量のガイド RNA を添加して Cas9-ガイド RNA 複合体を形成した後、3 µg のゲノム DNA を添加して 37°C 16 時間インキュベートし、二本鎖切断を誘導した。また Cas9 は添加するが、ガイド RNA を添加せずにゲノム DNA とインキュベートする陰性コントロール条件も同時に調製した。調製したライブラリは Qubit DNA HS Assay Kit (Thermo Fisher Scientific) および Bioanalyzer 2100 (Agilent) を用いた DNA のクオリティチェック (QC) 後、NextSeq 1000/2000 P1

Reagents (300 Cycles) または MiniSeq Mid Output Kit (300 cycles) (Illumina) を用いて 150 bp のペアエンドシーケンスを実施した。なお、SITE-Seq の原著論文 (Cameron et al, Nat Methods, 2017) では同じヒトゲノムにおいて 62~246 万リードペアのシーケンスを実施していたため、各ライブラリにつき約 200 万リードペアを目安にシーケンスを実施した。

シーケンスにより得られた Fastq ファイルは、「Trim-Galore!」によりアダプタートリミングを実施した後、アライナー「bowtie2」または「BWA」によりヒトリファレンスゲノム「GRCh38p.13」にマッピングした。生成された SAM ファイルは「Samtools」を用いて BAM ファイルへ変換・ソートした後、IGV で可視化するのに必要なインデックスファイルを作成した。カットサイトの検出には SITE-Seq の原著論文 (Cameron et al, Nat Methods, 2017) にて公開されている Python プログラムを使用した。なお、初期ピーク検出の最小リード深度、カットサイト検出の最小リード数はいずれもデフォルト値である 5 リードとした。

C. 研究結果

1. 同じ外来配列がゲノム内に挿入された場合の解析に係る検討

同じ外来配列が複数挿入された場合、従来通りショートリードシーケンスで解析すると、同配列の挿入部位を識別できない可能性がある。本検討では、挿入塩基の全長を網羅するようなロングリードシーケンスを含め、NGS によるアセンブリ法を検討する。同じ配列が近位にある場合と、別染色体にあるケースの 2 パターンを想定し、解析シミュレーションを行った。

ケーススタディ①: A2704-12 (同配列が直列で近郊に繋がっている場合)

ショートリードシーケンスの WGS データを大

豆ゲノムあたり 60×としてシミュレーションにより取得し、アセンブリ法により挿入配列およびその周辺の取得を試みたところ、1 本のコンティグが形成された。しかし、形成されたコンティグは挿入配列 5,400 塩基に満たず、4,271 塩基であった (表 2)。形成コンティグについて確認するために、GM コンストラクト配列挿入箇所付近のマッピング状況を確認すると、2 つ目の P35S-pat-T35S (1,324 bp) 配列を境にコンティグ形成がされていないことが分かった (図 3)。二回繰り返す P35S-pat-T35S 部位は各リードがマルチマッピングされ、各 P35S-pat-T35S 配列の位置を識別できない状況であった。以上より、ショートリードシーケンスではリード長より長くかつ同じ配列が近郊に存在した場合、60×以上のデータを用いても同配列の境界付近のアセンブリに失敗することが示された。

次に、ロングリードシーケンスを検討した。ロングリードシーケンスの WGS データを大豆ゲノムあたり 60×、平均リード長約 13,000 塩基としてシミュレーションにより取得し、アセンブリ法により挿入配列およびその周辺の取得を行ったところ、1 本のコンティグが形成され、境界を含む挿入配列を正確に再現した (表 3)。また、GM コンストラクトの挿入塩基長が 5,400 bp であることからリード平均塩基長が同等の 5,400 塩基または約 2 倍長の 10,000 塩基でデータ量を 30×として 3 回の独立したシミュレーションデータにおいても同様に再現された (表 3)。得られたコンティグ長も数万塩基におよび、GM コンストラクトおよび周辺配列の情報を幅広く取得できていた。参考までに、アセンブリ法で再現されたコンティグおよびその際のマッピングリードのアライメント状況を IGV で示した (図 4)。ショートリードシーケンスと比較して、リードエラーが多いことも特徴として挙げられる。

ケーススタディ②: A2704-12 系統の GM コンストラクトが別染色体にも挿入された場合

ショートリードシーケンスの WGS データを大豆ゲノムあたり 60×としてシミュレーションにより取

得し、アセンブリ法により挿入配列およびその周辺の取得を試みたところ、1本のコンティグが形成された。しかし、形成されたコンティグは一か所のみでかつ挿入配列 5,400 塩基に満たず、4,272 塩基であった(表 4)。形成コンティグを確認するために、GM コンストラクトの挿入箇所付近のマッピング状況を確認すると、ケーススタディ①同様に、2 つ目の P35S-pat-T35S(1,324 bp) 配列を境にコンティグ形成がされていないことが分かった(データさず)。また、別染色体の情報も得られなかった。以上より、ショートリードシーケンスでは同配列が複数存在した場合、識別ができず、同配列の境界付近のアセンブリに失敗することが確認された。

ロングリードシーケンスの WGS データを大豆ゲノムあたり 30×以上でシミュレーションにより取得し、アセンブリ法により二か所の挿入配列およびその周辺の取得を検討した。リード長を検討するために、シミュレーション平均塩基長を複数検討したところ、平均塩基長が長いほど、挿入部位二か所を共に再構築するアセンブリの成功率が高くなった(表 5)。再現の成功条件を比較したところ、リード数は大きく変わらなかったが、マッピングリードの平均塩基長が $17,588 \pm 4,903$ (N50: 25,333) であったのに対し、再現できなかった条件のマッピングリードの平均塩基長は $12,272 \pm 5,034$ (N50: 18,052) とマッピングされたリードの平均長が明確に異なっていた(表 6)。正確に 2 箇所の配列を再現できない場合の失敗例は大きく分けて 2 つであった。1 つめのケースは複数本のコンティグが形成されるケースで 1 番あるいは 8 番染色体の P35S-pat-T35S あるいは GM コンストラクト配列を境にコンティグが分断されていた。この場合、挿入配列以外は正確に再現されており、周辺配列の情報は得られていた。2 つめのケースは、1 本のコンティグのみ形成されるケースで 1 番または 8 番染色体のどちらかが正確に再現されたが、もう片方の情報が得られてい

なかった。

ロングリードシーケンスのアセンブリ法の妥当性確認方法の検討

ケーススタディ②のように、ロングリードシーケンスを用いても染色体間で同配列が含まれる場合は正確に再現できない場合もあった。よって、得られたコンティグ情報の妥当性を確認する方法論が必要である。そこで、公開されているリファレンスゲノム配列を用いて、その妥当性を確認する方法を検討した。3 種類の大豆リファレンスゲノムに対して、アセンブリ法で失敗してしまった指定塩基長 5,400 ロングリードからの抽出リードをマッピングし、構造変異検出ツール「Sniffles」で変異検出を行ったところ、全てのリファレンスゲノムにおいて、挿入箇所を 1 および 8 番染色体であることを特定し、その位置情報の抽出に成功した(図 5)。よって、この位置情報とアセンブリで得られたコンティグ情報とを照合することで、その妥当性を確認することは有効とみられる。

k-mer 法による各ケーススタディの検証

外来遺伝子の検知に関しては、k-mer 法が利用されている。各ケーススタディにおいて、ショートリードの WGS データを用いて k-mer 法での検出を検証したところ、ケーススタディ①および②どちらの場合でも外来配列であることを検出した(図 6)。2 箇所ある P35S-pat-T35S(1,324 bp) 配列は他の部位と比べて G 値が約 2 倍大きかった。また、ケーススタディ②のように該当挿入箇所が 2 箇所に存在した場合は、挿入箇所が 1 箇所のケーススタディ①と比べて、G 値が約 2 倍大きかった。

2. SITE-Seq の正確性向上に係る検討

SITE-Seq 一度の施行では、多数のオフターゲット候補部位が検出されるため、実際に CRISPR/Cas システムを用いたゲノム編集で起こりうるオフターゲットの蓋然性が高い部位としての

解釈は難しい。本検討では、SITE-Seq を独立した実験として 2 回実施し、得られた結果の解釈について検討した。

ライブラリーの調製およびシーケンス

別々のフラスコで培養した細胞から、それぞれ別個で抽出したゲノム DNA をアガロースゲル電気泳動で確認した(図 7)。その結果、低分子の DNA はほとんど確認されず、概ね高分子の高品質なゲノム DNA が抽出できたことが示唆された。よってこれらのゲノム DNA を用いて SITE-Seq 解析を実施した。調製したライブラリーの QC を行ったところ、低分子のアダプターは認められなかった(図 8)。

シーケンスで得られたリードの統計値を表 7 に示す。おおむね 200 万リードペア以上のデータ量を取得したが、Rep. 2 の陰性コントロールおよび VEGFA では 200 万リードペアを下回った。一方で、データのクオリティを示す Q30(%)ではいずれのデータでも良好であったため、これらのデータを用いて SITE-Seq 解析を実施した。得られたペアエンドリードのうち、Cas9 による切断面をシーケンスしているリード 1 に関して「Trim Galore!」によってトリミングを実施し、解析に供した。トリミング後の各データの統計値については表 7 に示した。

VEGFA を標的した SITE-Seq 解析結果

トリミング後のリードは「BWA」を用いてヒトファレンスゲノム「GRCh38.p13」へアライメントし、カットサイト判定プログラムによるオフターゲット予測部位の検出を実施した。その結果、Rep. 1 および Rep. 2 とともに、いずれの最終 Cas9 濃度においても本来の標的である VEGFA が予測部位として検出されていた(図 9)。このことから、真のオフターゲット部位は 2 回の施行で共通して検出されると仮定し、2 施行間で検出された全予測部位を比較した。その結果、Rep. 1 では 1,947 箇所、Rep. 2 では 2,681 箇所の予測部位が検出され、2 施行

間で共通して検出された予測部位は 978 か所であった(図 10A)。続いて、除かれた予測部位の妥当性を確認するために、既報(Tsai et al, Nat Biotechnol, 2015)の GUIDE-Seq のオフターゲット予測結果と比較した。その結果、GUIDE-Seq の原著論文で示された 22 箇所の全予測部位が、SITE-Seq 解析を 2 回施行した際に共通して検出された予測部位に含まれていた(図 10A)。

本検討の解析部分ではアライナーに「BWA」を使用しているが、SITE-Seq の原著論文で使用されていた「bowtie2」を用いてリードをアライメントし、同様に解析を実施したところ、GUIDE-Seq で検出された 1 箇所については SITE-Seq のいずれの解析でも検出されなかった(図 10B)。当該予測部位は Chr5: 56,876,246 付近のオフターゲットであり、ガイド RNA のミスマッチは 4 塩基であった。IGV にてアライメント状況を確認すると、「BWA」を用いた場合では Rep. 1 および Rep. 2 の 256 および 1,024 nM のいずれでも十分なリードのマッピングが確認された(図 11A)。一方で、「bowtie2」では最長 70 塩基程度の短いリードがわずかにマッピングしているのみであった(図 11B)。よって「bowtie2」で当該予測部位が検出されなかった理由は、マッピングリードがカットサイト検出に必要な閾値である 5 リードを下回ったことが原因とみられる。このアライメントの差が生じた理由については、リード中に含まれる低クオリティの配列やリファレンスと異なる配列を除いて正確性の高い配列をアライメントさせる Soft-clipping 機能の有無によるものと考えられる。「BWA」は Soft-clipping 機能を有するが、「bowtie2」は有していない。実際に「BWA」でアライメントされたリードについて詳細を確認すると、いずれのリードにおいても 5' 上流側において Soft-clipping がみられたことから、用いたリファレンスゲノムと HEK293T 細胞との配列の相違がある可能性がある。

FANCF を標的した SITE-Seq 解析結果

次に、FANCF を標的とした SITE-Seq データについても同様に解析を実施した。その結果、Rep. 1、Rep. 2 ともにいずれの最終 Cas9 濃度においても本来の標的である FANCF が予測部位として検出された一方で、検出された予測部位の合計は Rep. 1 では 8,166 箇所、Rep. 2 では 53,367 箇所、検出された予測部位数は VEGFA と比較して十倍～数十倍程度大きく差が生じた(図 12)。各 Cas9 濃度データを合算した Rep 間で共通して検出された予測部位は 140 箇所となり、オフターゲットサイトとしての蓋然性の高い箇所をかなり絞りこむことができた(図 13A)。実際に GUIDE-Seq で検出された 9 箇所の予測部位のうち、8 箇所は SITE-Seq の 2 回の施行で共通して検出された予測部位に含まれていた。しかし、1 箇所のみ、SITE-Seq で検出されなかった部位が存在し、その 1 箇所の予測部位について詳細を確認すると、当該予測部位は Chr3: 196,144,387 付近のオフターゲットで、ガイド RNA とのミスマッチ数は 5 塩基であった。IGV 上でアライメント状況を確認したが、いずれのアライナーを用いた場合でも 1 リードもアライメントしていなかった(データ示さず)。

D. 考察

1. 同じ外来配列の繰り返し挿入および異なる染色体間での挿入の検出について

本研究において、従来の方法では難しいとされる同じ外来配列が繰り返して残存する場合の検出に関して、次世代シーケンサーを用いたシミュレーションによる検証を行った。同じ外来配列が長い場合はショートリードシークエンスではカバーできないことからアセンブリ法による検出は困難であった。しかし、ロングリードシークエンスにより、外来配列をカバーするリード長が得られれば、アセンブリ法により再現することができた。しかし、その同じ外来配列の挿入部位が異なる染色体に

複数存在する場合はロングリードシークエンスでも完全にアセンブリで再現できない場合があることが確認された(表 5)。成功と失敗時でそれぞれマッピングリードのパラメータを確認すると、リード数よりもリード長が関係していると考えられた(表 6)。繰り返して存在する外来配列の周辺を明らかにするには、その領域を網羅し、かつその外側のユニークな配列を含む長いリードが必要である。

実際に利用する場面は、何箇所挿入されているかは不明なため、ロングリードを用いて挿入箇所の当てを付け、得られたアセンブリの妥当性を確認できるような方法論も必要である。そこで本研究では、近年ロングリードシークエンスで活用される構造変異検出法に着目し、大豆リファレンスゲノム配列を利用して、挿入変異箇所を検出することで、アセンブリ法の結果の妥当性を確認した。アセンブリ法で失敗してしまった際の抽出リードを用いて、3 種類の公開されている大豆リファレンスゲノムから、染色体の特定箇所に挿入変異があることを確認することができた。この構造変異検出の閾値として今回デフォルトの「サポートリード4」を用いているが、30×以上のデータセットを用いているため、十分検出されたと考えられる。よって、アセンブリ法により、外来遺伝子の有無を検知および周辺配列情報を取得する際、まず T-DNA 等を有するベクター配列にマッピングしたリードを用いてアセンブリ法を行うと共に、その抽出リードを用いてリファレンスゲノムを参照して変異解析を行うことで、外来遺伝子の挿入されている位置情報を明らかにし、得られたアセンブリの妥当性を確認することが有効な方法と考えられる。一方で、リファレンスゲノムが存在しないような生物種で、アセンブリ法の妥当性を確認する方法論はまだ検討できていないため、今後の課題である。

2. k-mer 法でのケーススタディの検証

ショートリードシークエンスデータを用いて、k-mer 法での外来遺伝子の検出を行ったところ、ケ

ースタディ①②の両ケースにおいて、GM コンストラクト全長域で有意に検出された。また、二箇所挿入されるケーススタディ②ではケーススタディ①より有意に検出されたことから外来遺伝子の検出という点では、k-mer 法はユニークな挿入配列であれば同配列の検出能力が高いことが確認された。しかし、ショートリードシーケンスでは挿入箇所の特定制と周辺配列の再現が困難であることから、従来通り、k-mer 法で外来遺伝子の検出を行い、必要に応じてロングリードシーケンスを用いたアセンブリ法によって挿入様式を明らかにする方法論が適切と考えられる。近年 k-mer 法に加え、農研機構のグループから GenEditScan (<https://github.com/hirsakai/GenEditScan>) という新たなツールが開発された。このツールでは統計検定が補正される機能が追加され、偽陽性を抑えたとされる。

近年、挿入ベクターに宿主生物の内在プロモーター配列を組み込んだものが使われることが増えている。また、シスジェネシス等の議論もあることから、作物内在配列の意図せぬ挿入なども検知されることが望ましい。しかし、内在配列を検知することは、NGS においても組換えやゲノム編集の前後で同配列が多数含まれることから、これまでと同じスキームで検知することは困難と予想される。今後はそのような検出法を新たに開発・整備することを検討する予定である。

3. SITE-Seq の解釈

SITE-Seq はあらゆる生物種において、ゲノム DNA さえ抽出されれば実際に二本鎖切断されるオフターゲット箇所を予測できる点で有用な *in vitro* 予測法である。しかし、GUIDE-Seq のような *in vivo* 予測法と比べて検出される予測箇所が多く、オフターゲットの起こりうる蓋然性をそれぞれ検証する必要があり、結果の解釈が難しいケースが想定される。そこで、SITE-Seq を独立した施行で 2 回実施することで結果の解釈について検討した。VEGFA において、2 回の施行で共通して

検出された予測部位は単独で検出された予測部位と比べて約 50~65%減少した。FANCF においては、共通して検出された予測部位は単独で検出された予測部位と比べて約 98.3~99.7%減少した。これらの結果は、SITE-Seq 解析を複数回実施することでオフターゲットの起こりうる蓋然性の高い候補配列を絞り込めることを示唆している。実際に、GUIDE-Seq で予測されたほとんどのオフターゲット部位は、2 回の施行で絞り込まれた予測部位からも確認されており、その可能性を支持するものであった。GUIDE-Seq で検出されて SITE-Seq で検出されなかった 1 箇所に関しては、*in vivo* と *in vitro* での Cas9 のアクセシビリティの差、または用いたヒト細胞種の違い等の原因が挙げられるが、考察の域を出ていない。しかし、複数の方法に加え、SITE-Seq で複数回確認することでより正確な予測が可能となると考えられた。本検討では 2 回の施行で SITE-Seq 解析を実施したが、より多く実施した場合に更に絞り込みが可能か検証する予定である。また別のガイド RNA でも実施し、同様の傾向が得られるか確認する。

データ解析において「BWA」と「bowtie2」を比較すると、「BWA」は Soft-clipping 機能を有し、より多くのリードをマッピングするため検出される予測部位も多くなる傾向がみられた(図 9、12)。これはオフターゲット候補部位の検出数の増加につながる一方で、図 11 で示した通り、特に予測部位付近がリファレンスゲノム配列と実サンプルで配列に個体差がある場合でも検出ができる可能性があるため、偽陰性を減少させるのに有効と考えられる。アライナーの検討も引き続き進めていく予定である。

E. 結論

解析が難しいとされる反復配列の解析法を検討するため、同じ外来配列が直列で存在する場合を模して、NGS シミュレーションを行った。ショートリードを用いたアセンブリ法では同じ外来配

列が繰り返し挿入された箇所の周辺配列を正しく得ることは困難であったが、挿入配列より長いロングリードを用いることで、周辺配列をアセンブリすることに成功した。実際にアセンブリ法を利用する場面は、挿入箇所の情報が皆無なため、リファレンスゲノム配列を用いた変異検出により、挿入箇所の推定を予め実施しておくことの有効性を示した。今後は内在プロモーターを含む外来配列が挿入されたケース等のより難しい場면을想定した解析法の検討を行う。SITE-Seq を用いた *in vitro* オフターゲット予測では、多数の予測部位が検知され偽陽性が懸念されていたが、複数回実施することでその蓋然性の高い部位を絞り込める可能性が示唆された。今後その検証を行っていく予定である。

F. 研究発表

1. 論文発表

なし

2. 学会発表

なし

G. 知的財産権の出願・登録状況

特許取得

該当なし

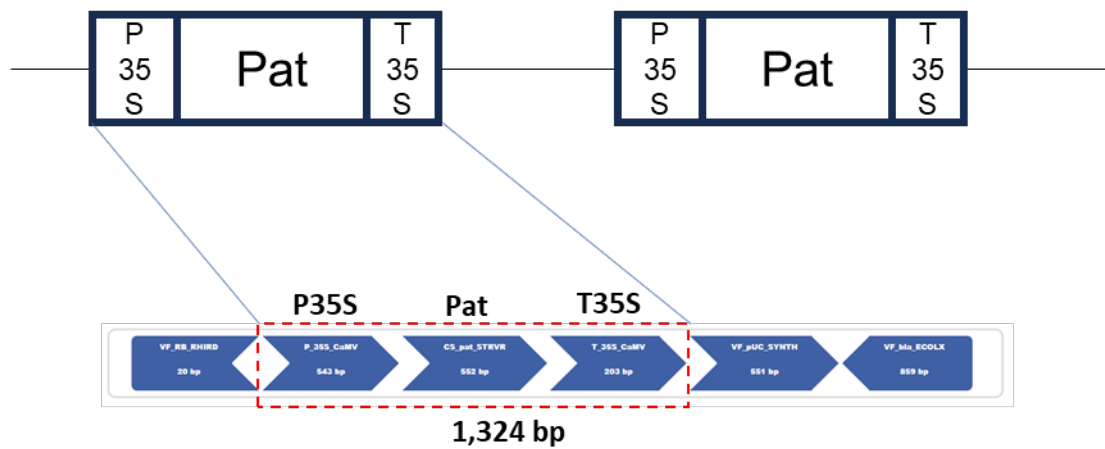
実用新案登録

該当なし

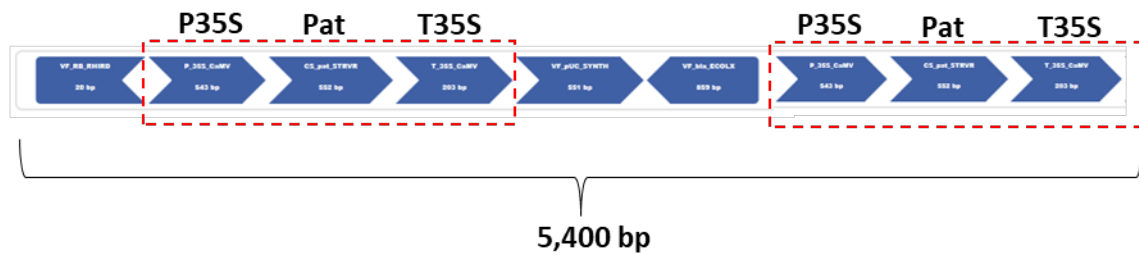
その他

該当なし

A2704-12系統のGM挿入配列



A2704-12系統のGM挿入配列の一部を
Nexplorerから取得



大豆リファレンス配列
Glycine-max-v2.1.dna.toplevel.faに挿入



A2704-12系統のゲノムを模した配列データ（Fasta形式）作成

図1 遺伝子組換え大豆 LibertyLinkSoybean(A2704-12 系統)のゲノムデータ作成

P35S: カリフラワーモザイクウイルス 35S プロモーター、Pat: ホスフィノトリシンアセチル基転移酵素の遺伝子、T35S:カリフラワーモザイクウイルス 35S ターミネーター。

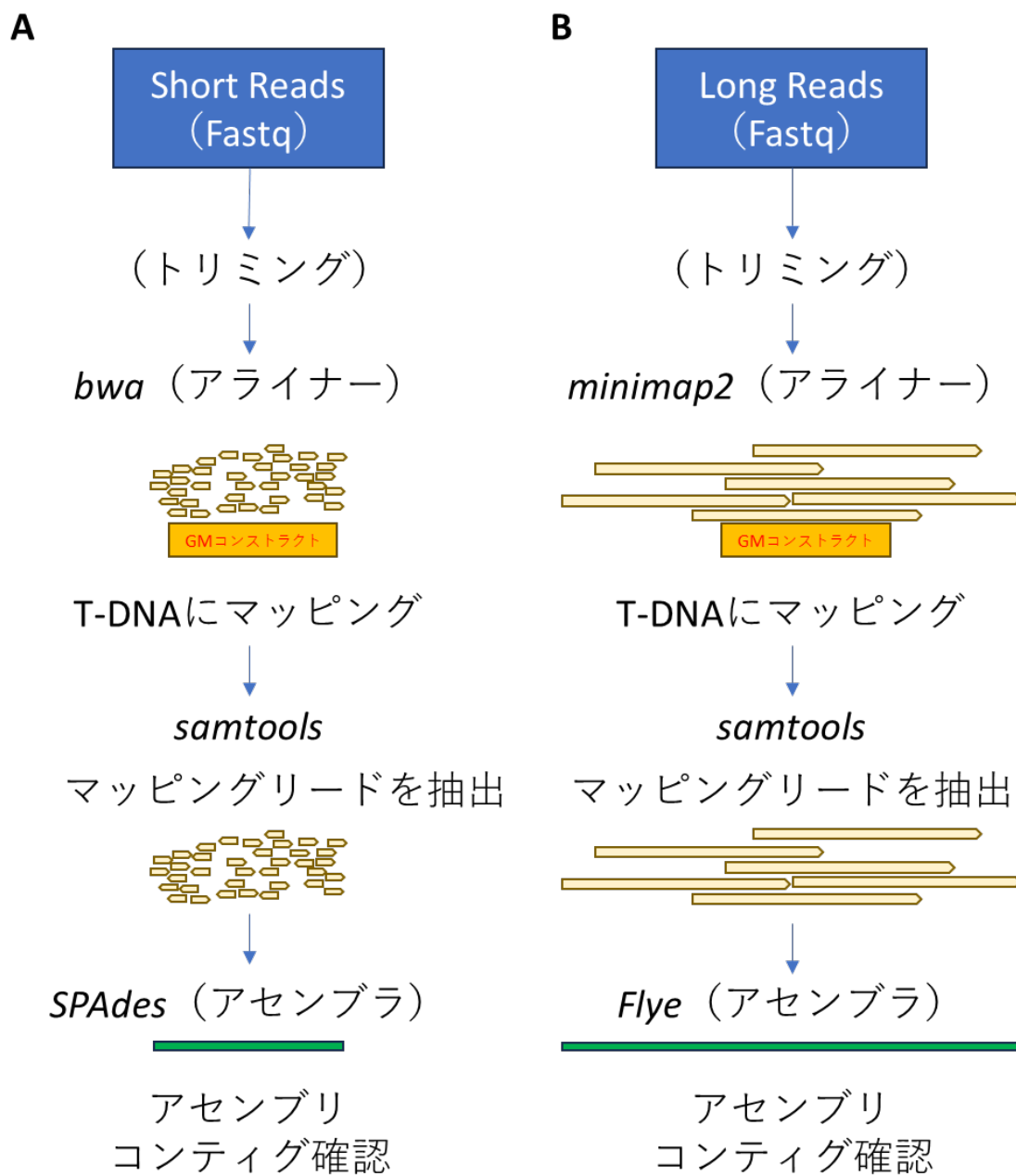


図2 アセンブリ法の概略図

斜体英字で示したものは利用したバイオインフォマティクスツール名を指す。

A) ショートリードシーケンスを用いた方法、B) ロングリードシーケンスを用いた方法

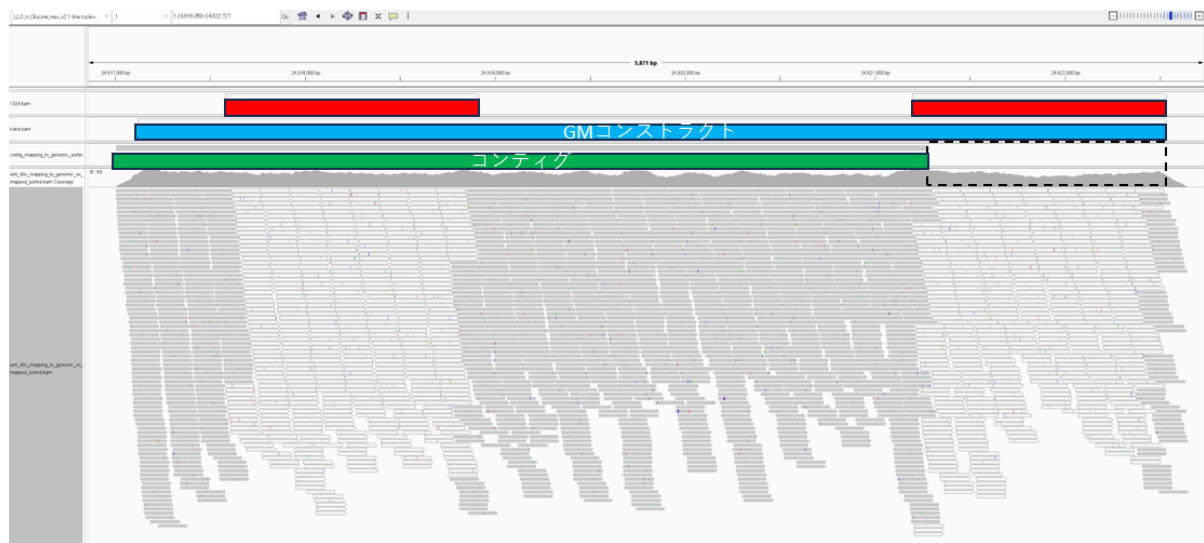
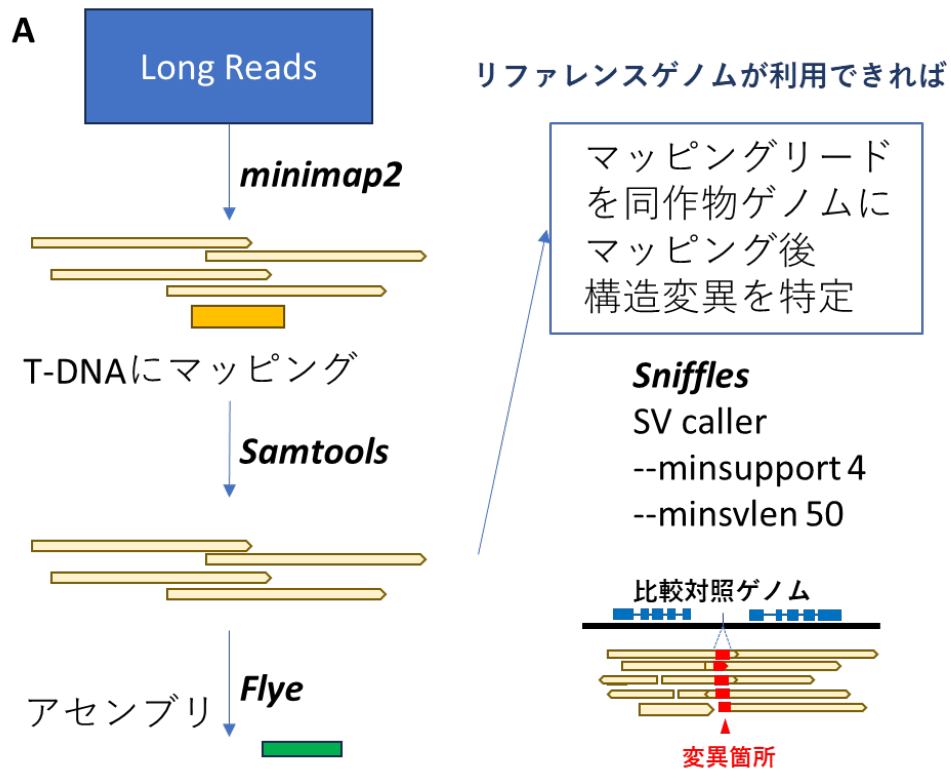


図3 ショートリードシーケンスを用いたアセンブリ法による A2704-12 の GM コンストラクトの検出
1 番染色体:chr1: 24,617,123 付近を示す。赤い領域に P35S-pat-T35S、青い領域に GM コンストラクト、
緑の領域に形成されたコンティグをそれぞれゲノムブラウザ IGV により可視化した。点線の領域は挿入配
列を網羅できなかった箇所を指す。また、シーケンスリードをそれぞれマッピングした様子を示した。
P35S-pat-T35S 箇所はマルチマッピング (白抜き:MAPQ=0) となった。



図4 ロングリードシーケンスを用いたアセンブリ法による A2704-12 の GM コンストラクトの検出
1 番染色体:chr1: 24,617,123 付近を示した。赤い領域に P35S-pat-T35S、青い領域に GM コンストラクト、
緑の領域に形成されたコンティグを各々ゲノムブラウザー IGV により可視化した。リードの中にところどころ
見られる紫の箇所はロングリードシーケンスエラーに起因したリファレンスゲノムと異なる配列を含むこと
を示している。



B

コンティグ名	座標	変異
①Glycine_max_v2.1.dna.toplevel.fa /Williams 82 (cultivar)		
1	24,617,126	INS
8	10,021,534	INS
②GWHA AEV00000000.genome.fasta/Zhonghuang 13(cultivar)		
GWHA AEV000000001	26,327,564	INS
GWHA AEV000000008	10,171,644	INS
③glyma.Lee.gnm1.BXNC.genome_main.fna /Lee (cultivar)		
glyma.Lee.gnm1.Gm01	25,388,371	INS
glyma.Lee.gnm1.Gm08	10,238,127	DEL
glyma.Lee.gnm1.Gm08	10,246,308	INS
glyma.Lee.gnm1.Gm08	10,273,673	DEL

図5 アセンブリ法結果の妥当性確認スキーム案

斜体英字で示したものはバイオインフォマティクスツール名を指す。

A) スキーム案、B) 各大豆リファレンスゲノムを用いて「Sniffles」で構造変異検出を行った結果を示した。

INS は挿入変異、DEL は欠失変異を指す。

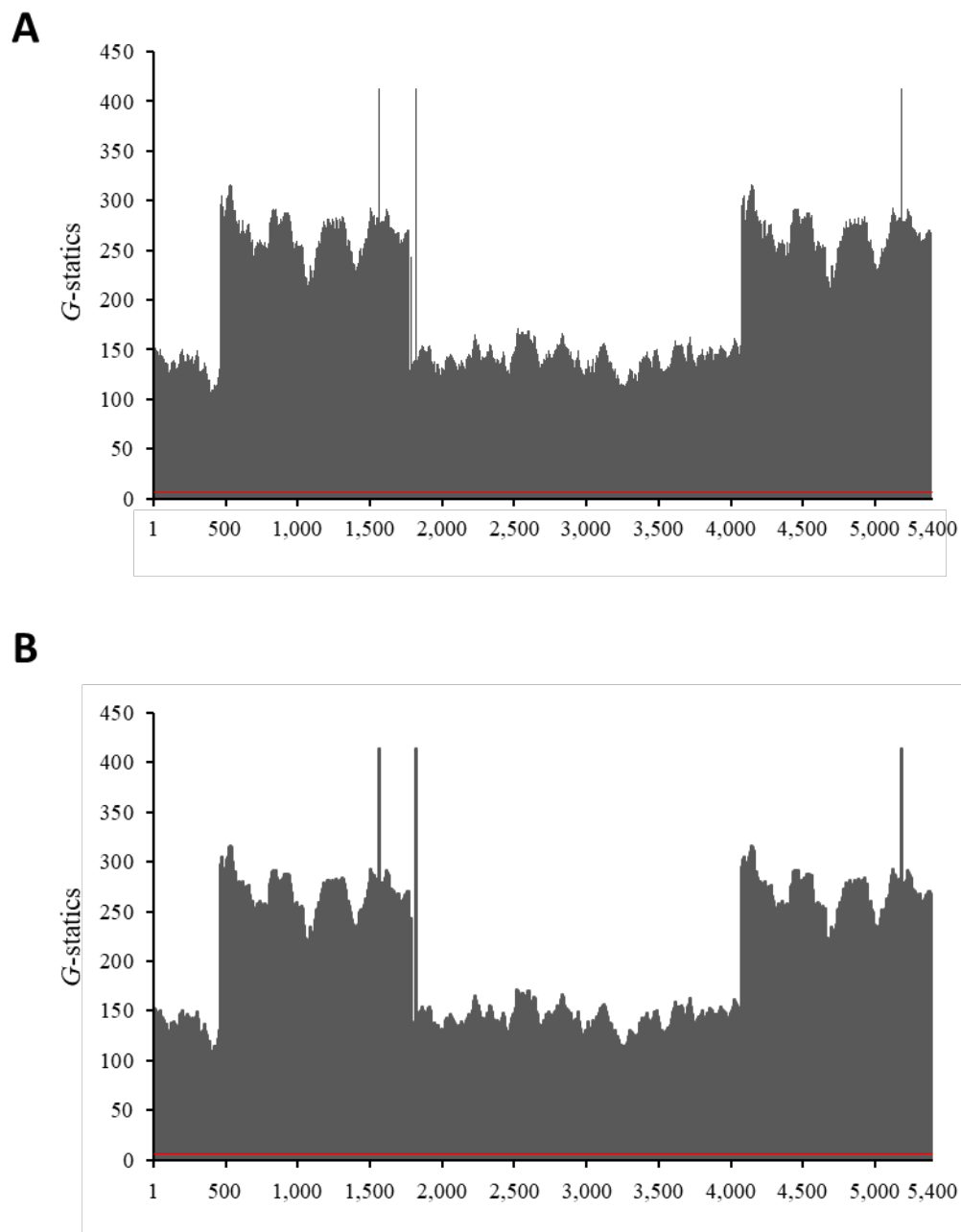


図6 A2704-12 における GM コンストラクトの k-mer 法による検出

GM コンストラクト配列上の各位置 (横軸) で検出された k-mer の数に基づく G 検定結果を示した。

1%有意水準(6.634)は赤線で示した。A) ケーススタディ①、B) ケーススタディ②

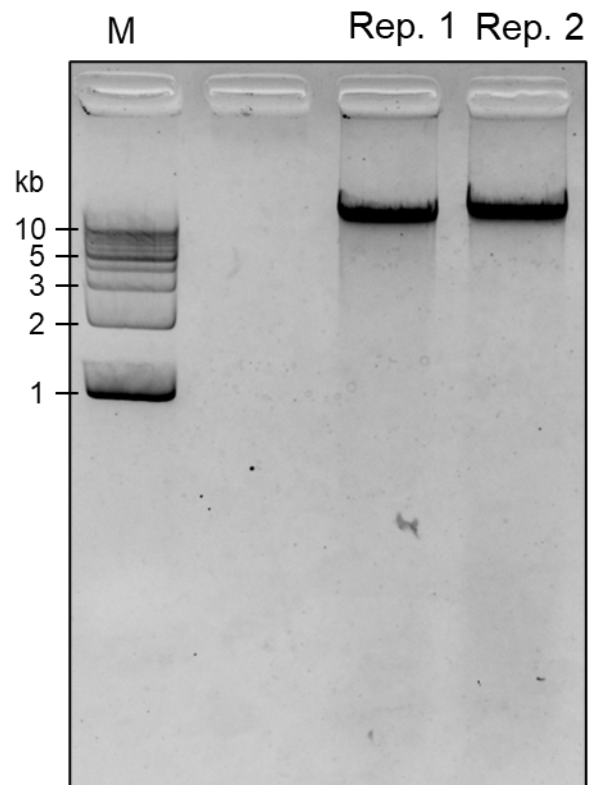
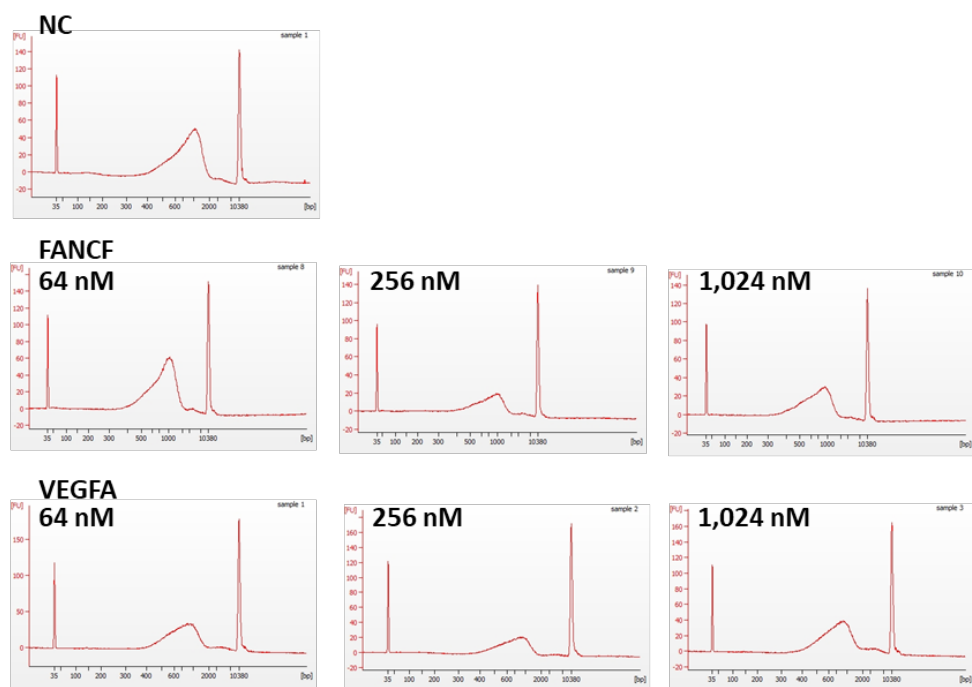


図7 SITE-Seq に供したゲノム DNA の品質確認

HEK293T 細胞 (n=2) から抽出したゲノム DNA を 1%アガロースゲルにて電気泳動した。

M: 1 kb Ladder Marker

A



B

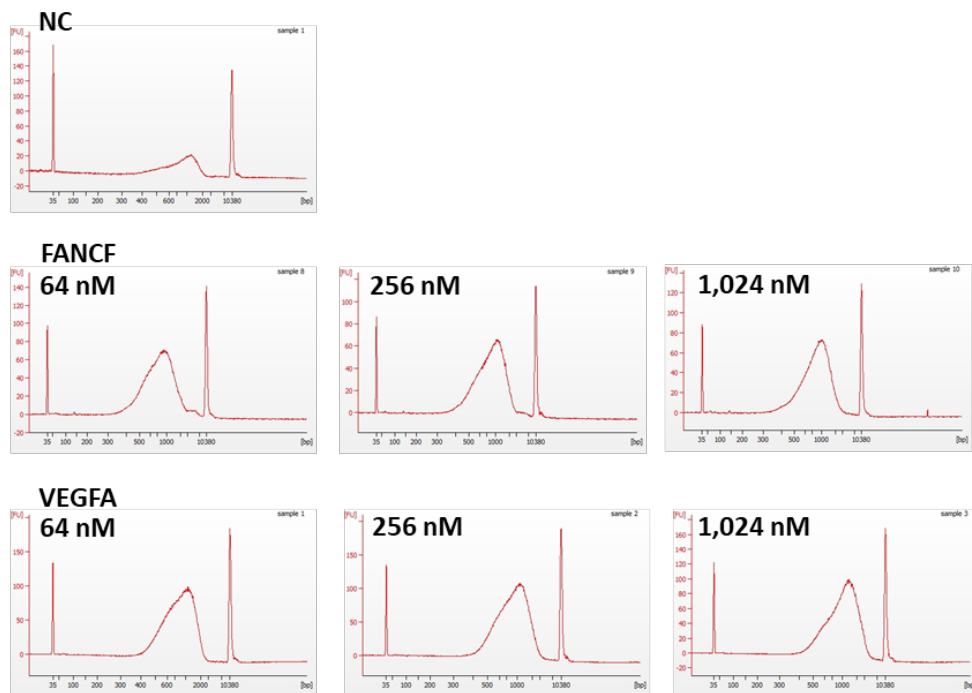


図8 SITE-Seq ライブラリの QC 結果

調製したライブラリは Bioanalyzer 2100 にて泳動した。内部マーカーとして Lower marker は 35 bp、Upper marker は 10,380 bp を示す。A) Rep1、B) Rep2

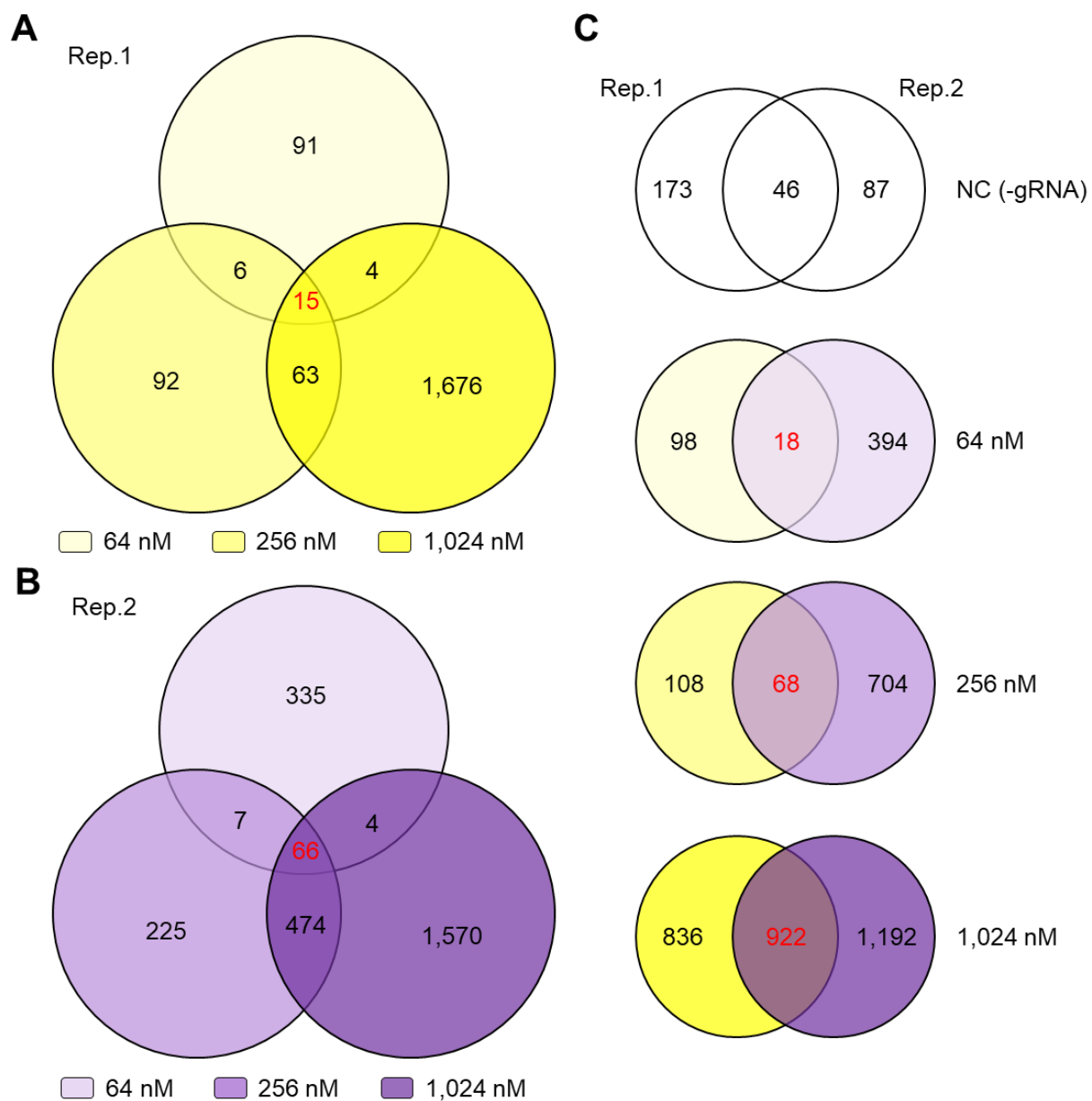


図9 VEGFA を標的とした SITE-Seq 結果

SITE-Seq は最終 Cas9 濃度 64 nM, 256 nM, 1,024 nM で実施した。(A) は Rep. 1、(B) は Rep. 2 の各濃度間で共通して検出されたカットサイト数を表す。(C) Rep. 1、2 の各濃度間で共通して検出されたカットサイト数を表す。本来の標的である VEGFA が検出された区分は赤字で示した。なお各濃度において検出されたカットサイトのうち、ネガティブコントロール (NC (-gRNA)) 条件で検出されたカットサイトは差し引いて表示した。

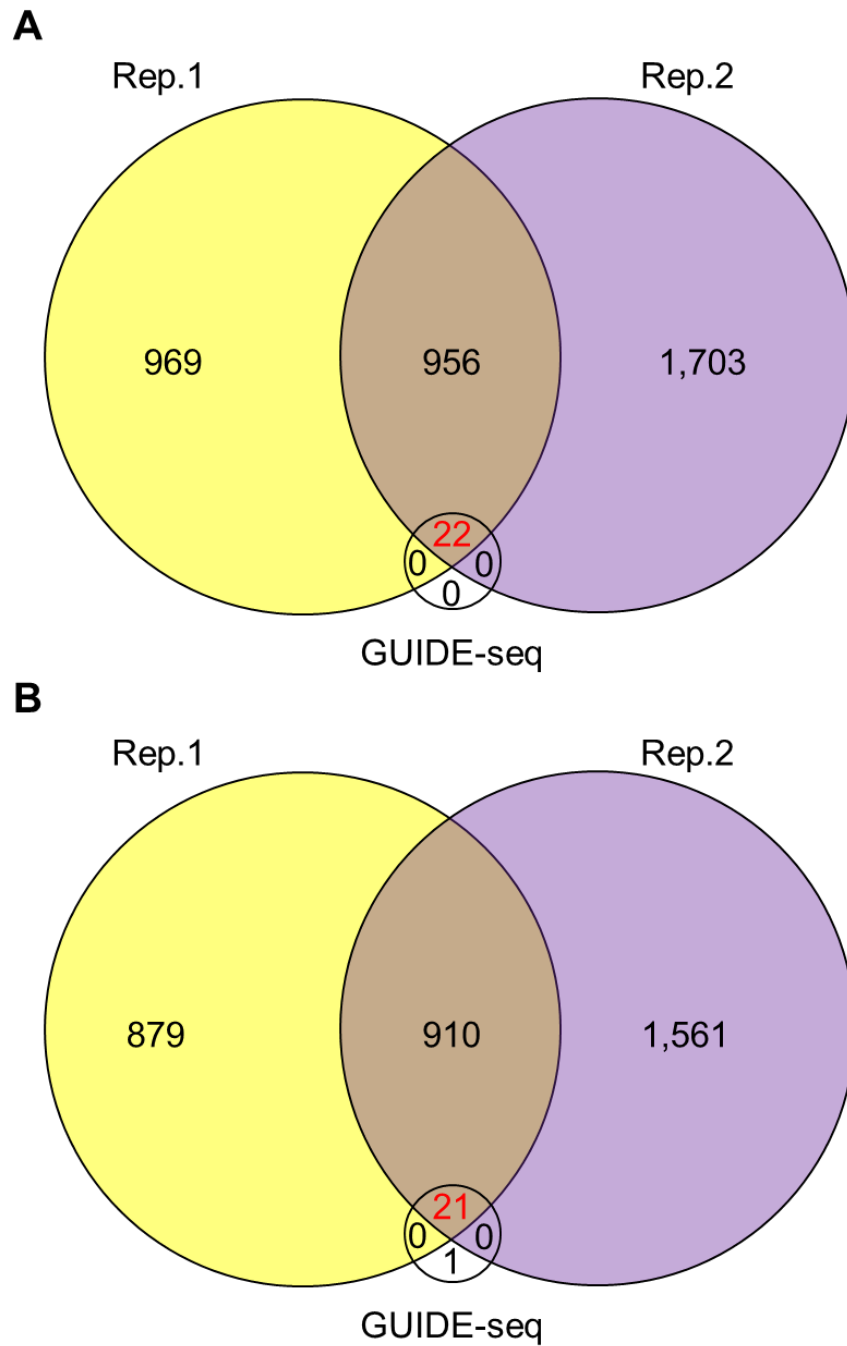


図10 VEGFA を標的とした SITE-Seq と GUIDE-Seq 結果の比較

最終 Cas9 濃度 64、256、1,024 nM において検出された全てのカットサイトを、Rep 間および GUIDE-seq の結果と比較した。(A) は BWA、(B) は bowtie2 を用いてリードのアライメントを実施した。本来の標的である VEGFA が検出された区分は赤字で示した。

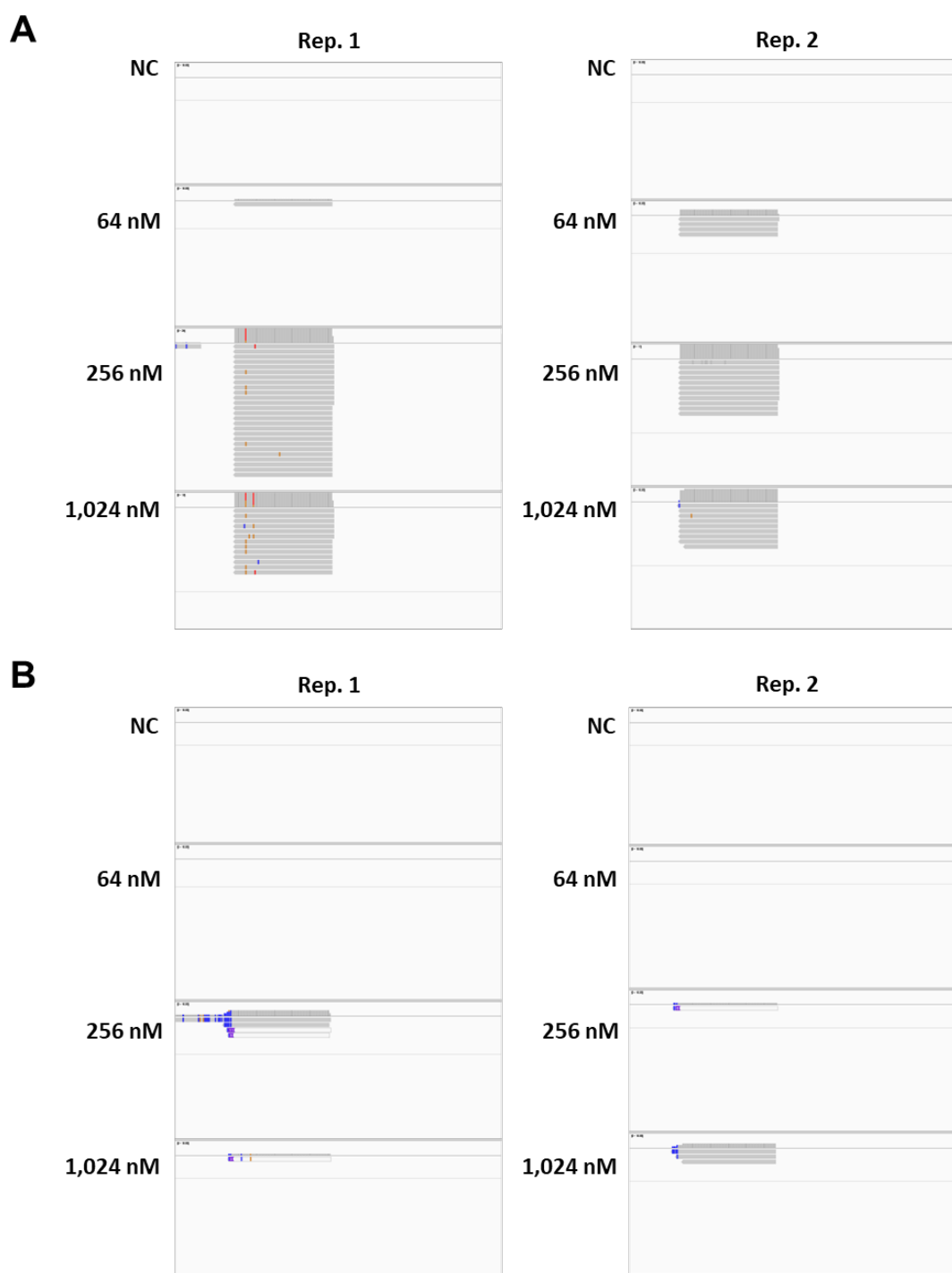


図11 bowtie2 で検出されなかったオフターゲット予測部位付近の IGV 図

図 10B において GUIDE-seq のみで検出されたカットサイト (Chr5: 56,876,246 付近) について、IGV にてリードアライメント状況を示す。(A) は BWA、(B) は bowtie2 でアライメントした IGV 図を示す。リード表示の window 幅はいずれも 218 bp とした。

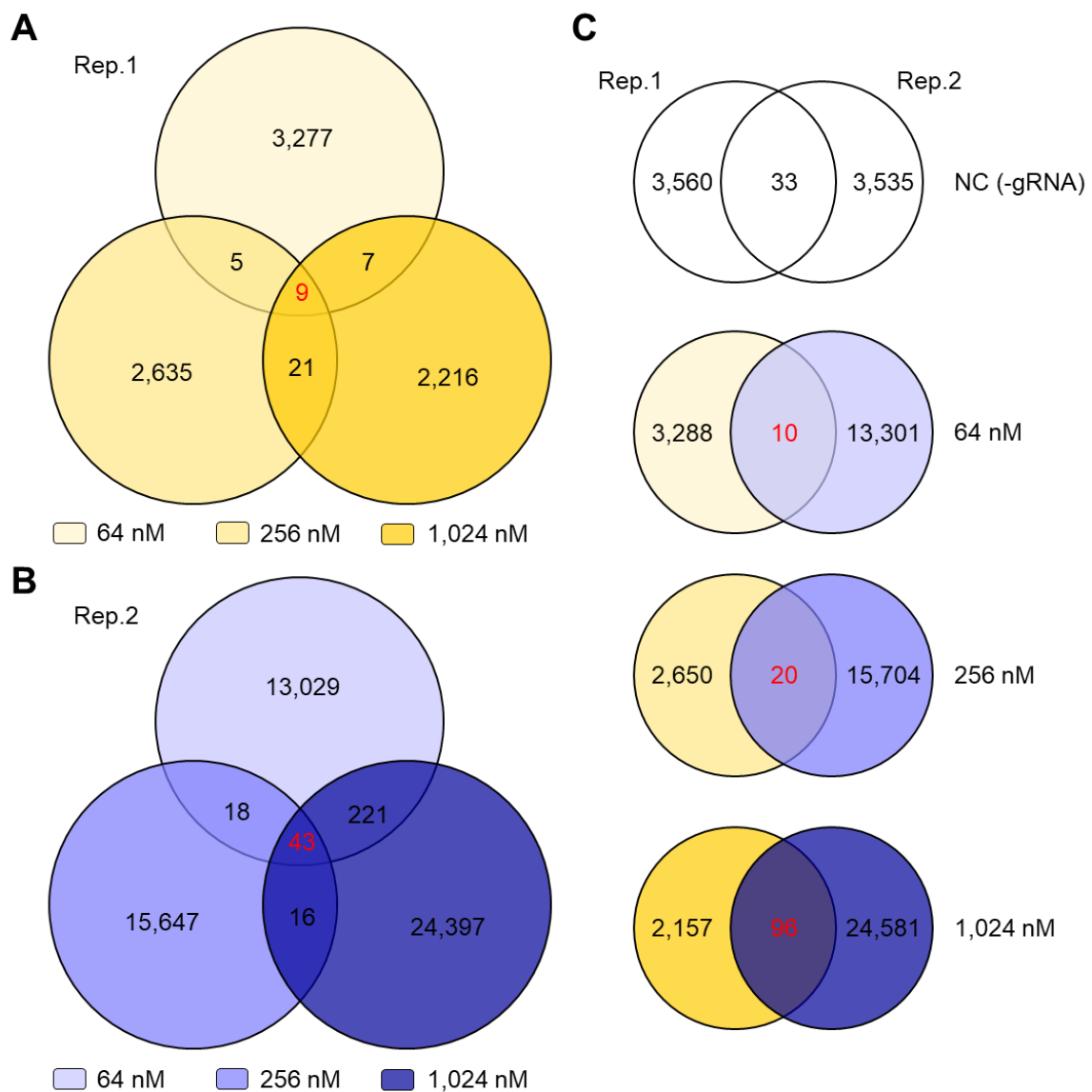


図12 FANCFを標的とした SITE-Seq 結果

SITE-Seq は最終 Cas9 濃度 64 nM, 256 nM, 1,024 nM で実施した。(A) は Rep. 1、(B) は Rep. 2 の各濃度間で共通して検出されたカットサイト数を表す。(C) Rep. 1、2 の各濃度間で共通して検出されたカットサイト数を表す。本来の標的である VEGFA が検出された区分は赤字で示した。なお各濃度において検出されたカットサイトのうち、ネガティブコントロール (NC (-gRNA)) 条件で検出されたカットサイトは差し引いて表示した。

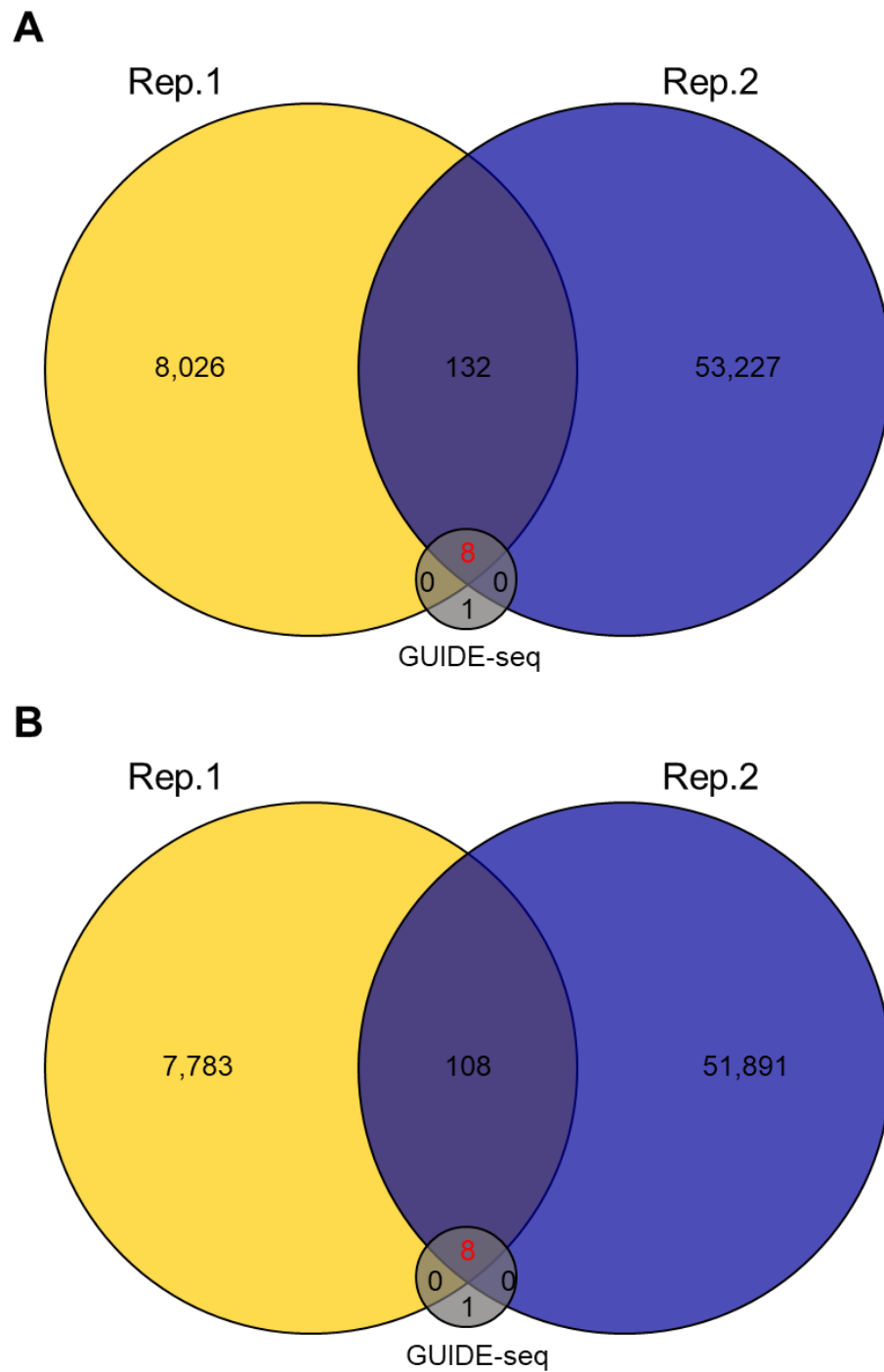


図13 FANCFを標的とした SITE-Seq と GUIDE-Seq 結果の比較

最終 Cas9 濃度 64、256、1,024 nM において検出された全てのカットサイトを、Rep 間および GUIDE-Seq の結果と比較した。(A) は BWA、(B) は bowtie2 を用いてリードのアライメントを実施した。本来の標的である FANCF が検出された区分は赤字で示した。

表1 本研究で使したオリゴ一覧

オリゴ名	配列 (5'-3')	備考
VEGFA_gRNA_Fwd	CGATGTAATACGACTCACTATAGGGGGTGGGGGAGTTTGCTCC†GTTT TAGAGCTATGCTGAAA	sgRNA synthesis
FANCF_gRNA_Fwd	CGATGTAATACGACTCACTATAGGGGAATCCCTTCTGCAGCACCG†GTTT TAGAGCTATGCTGAAA	
sgRNA_G_Rev	AAGCACCGACTCGGTGCCACTTTTCAAGTTGATAACGGA CTAGCCTTATTTAACTTGCTATGC TTTTCAGCATAGCTCTAAACG‡	
Adapter1_Fwd	[BioOn]GTTGACATGCTGGATTGAGACTTCCTACACTCTTCCCTACACGACGCTCTTCCGATCT	Adapter1 oligo
Adapter1_Rev	GATCGGAAGAGCGTCGTAGGGAAAGAGTGTAGGAAGTCTCAATCCAGCATGTCAAC	
Adapter2_Fwd	[PHO]GATCGGAAGAGCACACGTCTGAACTCC	Adapter2 oligo
Adapter2_Rev	GGAGTTCAGACGTGTGCTCTTCCGATCT	
Recovery_PCR_Fwd	GGAGTTCAGACGTGTGCTCTTC	Recovery PCR primer
Recovery_PCR_Rev	GTTGACATGCTGGATTGAGACTTCC	
Index-UDI0001-Fwd	AATGATACGGCGACCACCGAGATCTACACAGCGCTAGACACTCTTCCCTACACGACG	i5 Indexing PCR primer
Index-UDI0002-Fwd	AATGATACGGCGACCACCGAGATCTACACGATATCGAACACTCTTCCCTACACGACG	
Index-UDI0003-Fwd	AATGATACGGCGACCACCGAGATCTACACGCGAGACGACACTCTTCCCTACACGACG	
Index-UDI0004-Fwd	AATGATACGGCGACCACCGAGATCTACACTATGAGTAACACTCTTCCCTACACGACG	
Index-UDI0007-Fwd	AATGATACGGCGACCACCGAGATCTACACACATAGCGACACTCTTCCCTACACGACG	
Index-UDI0008-Fwd	AATGATACGGCGACCACCGAGATCTACACGTGCGATAACACTCTTCCCTACACGACG	
Index-UDI0016-Fwd	AATGATACGGCGACCACCGAGATCTACACCTACAGTTACACTCTTCCCTACACGACG	
Index-UDI0017-Fwd	AATGATACGGCGACCACCGAGATCTACACATATTCACACACTCTTCCCTACACGACG	
Index-UDI0018-Fwd	AATGATACGGCGACCACCGAGATCTACACGCGCCTGTACACTCTTCCCTACACGACG	
Index-UDI0069-Fwd	AATGATACGGCGACCACCGAGATCTACACCATGCCATACACTCTTCCCTACACGACG	
Index-UDI0070-Fwd	AATGATACGGCGACCACCGAGATCTACACTGCATTGCACACTCTTCCCTACACGACG	
Index-UDI0005-Rev	CAAGCAGAAGACGGCATACTAGATCAGTGGATGTGACTGGAGTTCAGACGTGTGCTC	i7 Indexing PCR primer
Index-UDI0013-Rev	CAAGCAGAAGACGGCATACTAGATAGACTTGGGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0014-Rev	CAAGCAGAAGACGGCATACTAGATGAGTCCAAGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0039-Rev	CAAGCAGAAGACGGCATACTAGATACCACTTAGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0040-Rev	CAAGCAGAAGACGGCATACTAGATGTTGTCCGGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0051-Rev	CAAGCAGAAGACGGCATACTAGATAACCTGTTGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0052-Rev	CAAGCAGAAGACGGCATACTAGATGGTTACCGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0063-Rev	CAAGCAGAAGACGGCATACTAGATCTAGGCAAGTGACTGGAGTTCAGACGTGTGCTC	
Index-UDI0064-Rev	CAAGCAGAAGACGGCATACTAGATTCGAATGGGTGACTGGAGTTCAGACGTGTGCTC	

†下線部は標的配列を表す

‡太字は標的配列の 3'末端の塩基に対して相補的な塩基を表す

表2 ケーススタディ①におけるショートリードデータとアセンブリ結果

	リードデータ					マッピングされたリードデータ				コンティグ		
	カバレッジ	リード数	全塩基数	N50	Q30 (%)	リード数	全塩基数	N50	Q30 (%)	SPAdes	塩基長	コンティグ
Read 1	60	189,642,461	28,363,589,009	150	95.6	2,249	335,954	150	91.9	○	4,271	1
Read 2		189,642,461	28,338,265,020	150	88.2							

表3 ケーススタディ①におけるロングリードデータとアセンブリ結果

PBSIM 条件 平均塩基長	リードデータ							マッピングされたリードデータ					コンティグ		
	カバレッジ	リード数	全塩基数	平均リード長	N50	平均Q		リード数	全塩基数	平均リード長	N50	平均Q	Flye	塩基長	コンティグ
—	60	4,453,839	58,714,329,016	13,183	24,847	12.4		85	2,391,023	28,130	42,003	12	○	120,379	1
5,400	30	5,291,879	29,354,951,877	5,547	9,136	11		60	484,384	8,073	12,041	10.8	○	50,959	1
5,400	30	5,290,340	29,354,947,101	5,549	9,144	11		50	382,684	7,654	11,369	10.5	○	32,480	1
5,400	30	5,285,892	29,354,949,025	5,553	9,147	11		59	531,895	9,015	15,335	11.5	○	42,641	1
10,000	30	2,892,635	29,354,964,988	10,148	16,882	11		43	967,296	22,495	31,648	10.7	○	92,620	1
10,000	30	2,894,368	29,354,963,029	10,142	16,882	11		42	763,870	18,187	22,636	10.5	○	72,470	1
10,000	30	2,892,438	29,354,963,340	10,148	16,915	11		46	763,780	16,604	25,776	10.8	○	67,083	1

表4 ケーススタディ②におけるショートリードデータとアセンブリ結果

	リードデータ					マッピングされたリードデータ				コンティグ		
	カバレッジ	リード数	全塩基数	N50	Q30 (%)	リード数	全塩基数	N50	Q30 (%)	SPAdes	塩基長	コンティグ
Read 1	51.55	189,642,403	28,363,612,742	150	95.62	4,427	661,443	150	91.89	○	4,272	1
Read 2		189,642,403	28,338,196,891	150	88.17							

表5 ケーススタディ②におけるロングリードデータとアセンブリ結果

PBSIM 指定塩基長	平均塩基長	N50	マッピングリード 平均塩基長	マッピングリード N50	成功率
5,400	5,550	9,100	8,530	12,451	0/3
6,000	6,160	10,150	9,330	13,872	1/3
7,000	7,150	11,800	11,170	16,863	2/3
8,000	8,155	13,500	14,090	20,957	3/3
9,000	9,150	15,200	14,030	23,019	2/3
10,000	10,145	16,900	16,490	17,658	3/3
11,000	11,130	18,600	19,773	29,225	3/3
12,000	12,130	20,300	20,274	29,732	1/3
13,000	13,130	22,000	22,127	32,770	3/3
—	13,170	24,850	24,770	34,940	3/3

表6 リード長とアセンブリ成功との関係

	成功時 マッピングリード				失敗時 マッピングリード			
	リード長	N50	Q 値	リード数	リード長	N50	Q 値	リード数
平均	17,588	25,334	11.4	97.1	12,272	18,052	11.4	107.1
SD	4,903	7,438	0.6	22.8	5,034	7,442	0.5	20.6

表7 SITE-Seq 解析で用いたリードデータの統計値

トリミング前

リード名	リード数	総塩基数	最小塩基長	平均塩基長	最大塩基長	Q20(%)	Q30(%)
Rep. 1							
陰性コントロール__Read1	3,277,963	494,972,413	151	151	151	93.26	92.08
陰性コントロール__Read2	3,277,963	494,972,413	151	151	151	91.06	89.55
FANCF_64nM_Read1	3,039,985	459,037,735	151	151	151	96.71	92.45
FANCF_64nM_Read2	3,039,985	459,037,735	151	151	151	93.00	86.44
FANCF_256nM_Read1	2,601,806	392,872,706	151	151	151	96.88	92.79
FANCF_256nM_Read2	2,601,806	392,872,706	151	151	151	93.24	86.78
FANCF_1024nM_Read1	2,947,411	445,059,061	151	151	151	96.89	92.76
FANCF_1024nM_Read2	2,947,411	445,059,061	151	151	151	93.26	86.77
VEGFA_64nM_Read1	2,728,815	412,051,065	151	151	151	92.87	91.66
VEGFA_64nM_Read2	2,728,815	412,051,065	151	151	151	90.68	89.16
VEGFA_256nM_Read1	3,077,165	464,651,915	151	151	151	93.40	92.26
VEGFA_256nM_Read2	3,077,165	464,651,915	151	151	151	91.14	89.67
VEGFA_1024nM_Read1	2,638,928	398,478,128	151	151	151	92.94	91.72
VEGFA_1024nM_Read2	2,638,928	398,478,128	151	151	151	90.40	88.81
Rep. 2							
陰性コントロール__Read1	1,826,890	275,860,390	151	151	151	93.79	92.38
陰性コントロール__Read2	1,826,890	275,860,390	151	151	151	90.34	87.79
FANCF_64nM_Read1	3,378,558	510,162,258	151	151	151	96.77	92.52
FANCF_64nM_Read2	3,378,558	510,162,258	151	151	151	92.97	86.37
FANCF_256nM_Read1	3,432,299	518,277,149	151	151	151	96.85	92.7
FANCF_256nM_Read2	3,432,299	518,277,149	151	151	151	93.15	86.71
FANCF_1024nM_Read1	3,028,504	457,304,104	151	151	151	96.77	92.54
FANCF_1024nM_Read2	3,028,504	457,304,104	151	151	151	92.85	86.16
VEGFA_64nM_Read1	1,521,549	229,753,899	151	151	151	93.33	91.85
VEGFA_64nM_Read2	1,521,549	229,753,899	151	151	151	89.52	86.90
VEGFA_256nM_Read1	1,456,968	220,002,168	151	151	151	93.24	91.79
VEGFA_256nM_Read2	1,456,968	220,002,168	151	151	151	88.45	85.71
VEGFA_1024nM_Read1	1,827,450	275,944,950	151	151	151	93.15	91.77
VEGFA_1024nM_Read2	1,827,450	275,944,950	151	151	151	86.71	83.81

トリミング後

file	リード数	総塩基数	最小塩基長	平均塩基長	最大塩基長	Q20(%)	Q30(%)
Rep. 1							
陰性コントロール__Read1	3,277,866	488,809,832	20	149.1	151	93.86	92.69
FANCF_64nM_Read1	3,039,906	456,452,774	20	150.2	151	96.82	92.6
FANCF_256nM_Read1	2,601,729	390,727,711	20	150.2	151	96.99	92.93
FANCF_1024nM_Read1	2,947,359	442,502,742	20	150.1	151	97.00	92.90
VEGFA_64nM_Read1	2,728,627	406,539,837	20	149.0	151	93.55	92.35
VEGFA_256nM_Read1	3,076,976	458,573,990	20	149.0	151	94.02	92.88
VEGFA_1024nM_Read1	2,638,776	393,376,044	20	149.1	151	93.57	92.35
Rep. 2							
陰性コントロール__Read1	1,826,061	272,546,377	20	149.3	151	94.48	93.10
FANCF_64nM_Read1	3,378,347	507,296,933	20	150.2	151	96.88	92.66
FANCF_256nM_Read1	3,431,988	515,081,237	20	150.1	151	96.95	92.84
FANCF_1024nM_Read1	3,028,352	453,878,712	20	149.9	151	96.89	92.69
VEGFA_64nM_Read1	1,520,441	226,043,900	20	148.7	151	94.12	92.68
VEGFA_256nM_Read1	1,455,493	213,062,960	20	146.4	151	94.13	92.73
VEGFA_1024nM_Read1	1,826,748	257,126,872	20	140.8	151	94.38	93.08