遺伝子発現および新規融合遺伝子に対する効率的な Exon-Aware RNA Capture

はじめに

トータル RNA シーケンシングは、細胞集団の転写状態を比較的偏りなく見ることができます。しかし、ほとんどのトータル RNA シーケンシング実験では、リボソーム RNA のように大多数を占める非コード転写産物や、mRNA 前駆体からのイントロンリードまたは混入ゲノム DNA など、遺伝子発現解析には有用ではない多数のリードを扱わなければなりません。ターゲットエンリッチメントは、ゲノムの有益な部分にシーケンシングを集中させる方法を提示し、低含量の転写産物の高感度な検出または特定の注目すべき遺伝子のみのプロファイリングを可能とします。

本資料では、新しいデザイン設計戦略を用いて Genecode v41 Basic に登録されるすべてのタンパク質コーディングアイソフォームを特異的にターゲットとした Twist の新 RNA Exome パネルのキャプチャーシーケンシング性能を紹介します。トランスクリプトームを標的としているものの、ここで示す設計戦略では設計バイアスを最小限に抑え、新たな融合遺伝子の発見を可能とするようプローブを配置しています。発現の定量性の評価において、ハイブリッドキャプチャ後でも相対的な転写産物量が保持されることを示しております。これにより、何桁にもわたる転写産物の正確かつ再現性のある定量化が可能となります。ターゲットキャプチャによる効率の向上と、がんに観察されることの多い RNA fusion の新規構造変異をキャプチャする性能をお示しします。さらに、RNA キャプチャパフォーマンスを評価するバイオインフォマティクスアプローチについて検討し、RNA シーケンシングの実験解析の特定の課題について論じます。要約すると、遺伝子発現ソリューション Twist ターゲットエンリッチメントは遺伝子発現を効率的にプロファイルし、遺伝子融合を検出する効果的な方法であるというエビデンスを示します。

結果

デザイン戦略と内容

この RNA Exome 開発の最初のステップは、コンテンツキュレーション戦略および転写産物用キャプチャプローブの設計方法に関する戦略の両方を決定することでした。GenCode 遺伝子定義（v41, hg38 使用）を用いてコンテンツキュレーションを実施しました。この手法の目的は、設計をタンパク質コード遺伝子のコーディング領域にフォーカスすることでした。このため、GenCode で定義されたコード配列（CDS）の総領域を、タンパク質をコーディングする遺伝子群と、特的の状況下でコーディングするという強い証拠を持つ遺伝子群に絞り込んでいきました。さらに、一部の遺伝子（再発性の融合遺伝子など）の 3’ および 5’ の非翻訳領域をカバーし、これらのイベントに対するパネルの感度が最大になるようにしました。これらの遺伝子から、多くの研究者が共通に関心のあるアイソフォームの大部分を自然とカバーするよう、より詳細に記述された転写産物モデルのセットを選択してタイリングしました。重要なポイントは、選択されたコンテンツが遺伝子群の中でも高い信頼性を持つエクソンセットで構成されている点です。

混入ゲノム DNA または mRNA 前駆体のキャプチャを避けるため、遺伝子の隣接するイントロン配列は標的から外しました。したがって、転写産物の成熟した mRNA を直接標的とするタイリング戦略を決定しました。これらの転写産物をカバーするための単純なアプローチは、端から端までプローブを用いてこれらをタイル表示にすることです **（図 1A）**。しかし、これでは既知のアイソフォームおよび融合転写産物に対するキャプチャのバイアスの問題が生じます。この方法の代わりに、エクソンとエクソンの境界にプローブを置くことを避ける新たな "exon-aware" デザイン戦略を採用しました。**（図 1B）**。これを行うことにより、プローブがエクソン間の既知のジャンクションを選択せずに、新規アイソフォームまたは融合転写産物を効率的に検出することが確実になります。

上記の exon-aware 戦略を用いてタイリング設計した後、重複配列を統合し、配列の複雑性が低いプローブおよび／または シーケンシング効率を低下させる非コード RNA に対する相同性を有するプローブ（ミトコンドリア、核リボソーム RNA、tRNA）を削除しました。この設計を確立した後、標準のターゲットエンリッチメントパネルプロセスを用い、Twist の DNA 合成技術を用いてプローブを合成しました。

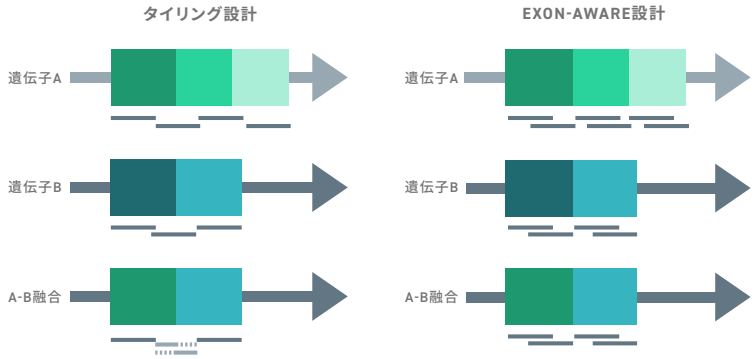


図 1（A）RNA 転写産物に単純なタイリング設計を使って融合遺伝子を検出する図。点線はミスマッチを示します。（B）exon-aware デザイン戦略を用いた融合遺伝子検出の図解。

エクソームキャプチャと WTS の比較および 3' カウンティング

ターゲットシーケンシングに加え、遺伝子発現を評価する一般的なワークフローには、リボソーム除去した RNA からランダムプライマーを用いて比較的バイアスの少ない転写産物のセットを選択する全トランスクリプトーム (WTS) と、オリゴ dT プライマーを使ってポリアデニル化 mRNA (主に mRNA) の 3' 末端を分離する 3' カウンティングがあります。WTS の利点は比較的バイアスのないトランスクリプトーム結果が見られるものの、イントロンや情報の少ないゲノム領域に相当数のリードが失われることです **(図 2A)**。同様に 3' カウンティングは、エクソン領域 (CDS および UTR) の選択においてより効率的ですが、転写産物の 3' 末端に対して強いバイアスを示します **(図 2B)**。長い遺伝子にとっては転写産物の一部しかプロファイリングされないために、異なるアイソフォームを検出する能力に影響すると予想されます。

これらの問題に対処するため、RNA エクソームパネルを設計し、タンパク質コード転写産物の CDS 全体をプロファイリングしました。これにより、WTS で得られたものと同様の 3' 末端バイアスおよび duplication 率の測定値を得ました **(図 2B)**。また、イントロン配列および高度に発現する非コード配列もこの設計から慎重に除外しました。これにより、WTS または 3' カウンティングのいずれよりも効率的にエクソンにリードを集中させることができました **(図 2A、2B)** 。ハイブリッドキャプチャによる成熟転写産物の選択には、他の利点もあり、3' カウンティングまたは WTS のいずれと比較しても、逆ストランドに由来するリードの割合が減少しました **(図 2B)**。

次いで、転写産物はおよそ 6 桁に及ぶ濃度で存在することから、RNA ハイブリッドキャプチャが低発現転写産物および高発現転写産物の両方に等しく効率的であったのかについて検討しました。この検討にあたり、WTS からのカウントと、同じサンプルタイプの RNA エクソームキャプチャから得られるカウントとの相関を見ました **(図 2C)**。発現の全範囲にわたってエンリッチメントが一貫しており、高発現の転写産物であってもキャプチャが飽和していないことを示しました。

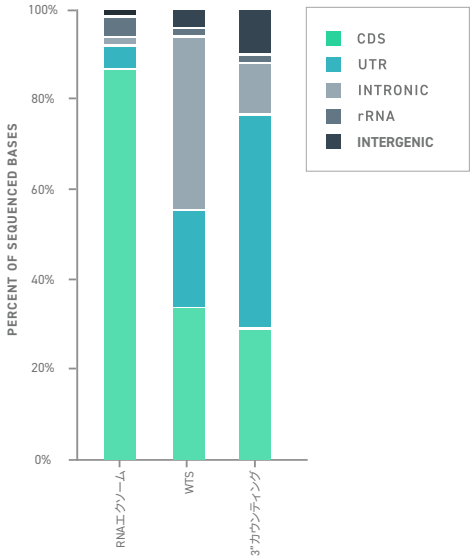


図 2A RNA エクソームキャプチャ、全トランスクリプトームシーケンシング (WTS) および 3' カウンティングにおけるリードのゲノム分布

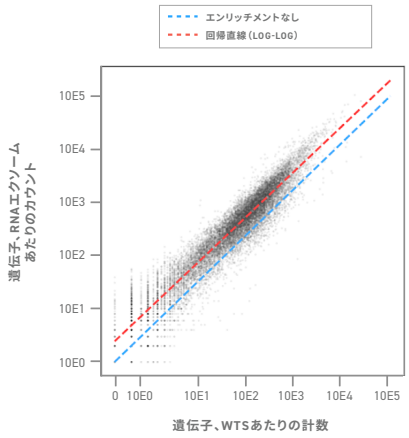


図 2C タンパク質コード遺伝子に関するキャプチャなしのカウント (x 軸) とキャプチャありのカウント (y 軸) の相関

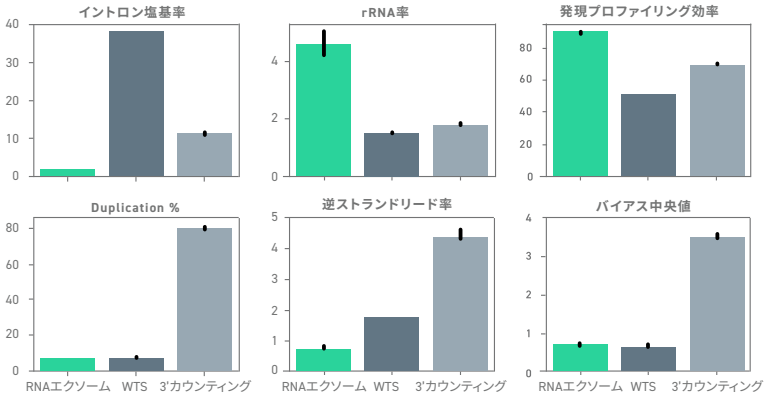


図 2B RNA エクソームキャプチャと WTS と 3' カウンティングとの間のシーケンシングメトリクスの比較

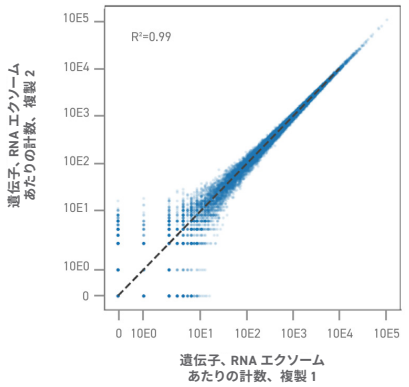


図 2D RNA エクソームで実施した 2 つのテクニカルレプリケートキャプチャの相関

少量インプット時および FFPE サンプルにおける TWIST RNA エクソームのパフォーマンス

ホルマリン固定パラフィン包埋（FFPE）組織は、組織学的検査用に保存された組織です。作製プロセスが核酸に損傷を与えますが、それでも FFPE 組織は臨床検体として容易に入手できるため、RNA-Seq にしばしば使用されます。FFPE サンプルに焦点を当てた RNA キャプチャの応用例の論文例もあるように（Jang et al 2021、Pennock et al 2019、Vahrenkamp et al 2019）、ここでは 3 つのインプット量（1 ng、10 ng および 100 ng）の FFPE サンプルを用いた RNA エクソームのパフォーマンスを評価しました。RNA エクソームではコーディング領域が効率的に選択されるため、WTS と RNA エクソームの両方で 10M から 30M リードの間の 5 段階のシークエンス量において、各ワークフローで特定の遺伝子数を検出するのに必要なリード数のおおよその数値を確認しました。アラインメントと標準的なカウンティングによって検出されたコーディング遺伝子数 **（図 3A）** と、k-mer ベースのアプローチによるアイソフォームの検出数 **（図 3B）** の両方を調べました。いずれの場合も、対応リード数 5 をカットオフとして検出数の定義としました。

この結果では、すべてのインプット量において、RNA エクソームが検出されたコーディング遺伝子および転写産物の数を劇的に改善することが示されています。最多量インプットの 15 M リードのコーディング遺伝子検出数は、30 M リードの WTS が示す検出数と同等であることが認められました。結果は、1 ng の FFPE インプットで特に顕著であり、ターゲットエンリッチメントではインプット量が多いほどコーディング遺伝子が増えるのに対し、WTS ワークフローでは 30M リードでもターゲットエンリッチメントで測定可能な約 1,000 の低発現遺伝子を検出することができませんでした。**（図 3A）**。検出された転写産物数のパターンは類似しており、少量インプットでは、すべてのレベルのリードサンプリングで測定可能な増加がより顕著でした。1ng の FFPE に関し、RNA エクソームでは、WTS を用いて 30M リードで検出された転写産物と同等数のものを 10M リードで検出することができました **（図 3B）**。

FFPE の RNA は高度に断片化される傾向があるため、ターゲット濃縮が、より分解の少ない配列の集合を選択的に濃縮している可能性があるかどうかを検討しました。全トランスクリプトームシーケンシングおよび RNA エクソームの両方を用いて FFPE サンプルをシーケンシングし、RNA 転写産物への直接的アライメントに基づき断片の推定サイズ分布をプロットしました。サイズ分布は、RNA エクソームサンプルに対して明らかな上方シフトを示しました **（図 3C）**。これは、RNA エクソームが実際に、よりインタクトな RNA を選択することを示しています。

最後に、品質の異なる FFPE サンプルに対する RNA エクソームキャプチャにどの程度の頑健性があるのかについて検討しました。そこで、市販されている 5 つの FFPE 標準品から RNA を抽出し、これらのサンプルを用いて全トランスクリプトームシーケンシングおよび RNA エクソームによるキャプチャの両方を実施しました。RNA エクソームが、分解レベルを問わずにテストしたサンプルすべてで遺伝子検出数を有意に増加させていることが認められました **（図 3D）**。

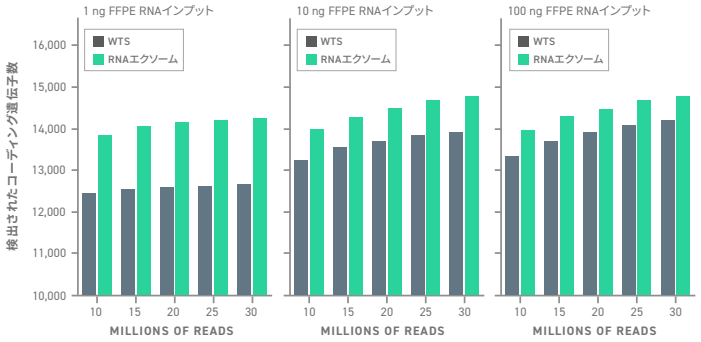


図 3A FFPE サンプルの 3 つの異なるインプット量における異なるレベルのダウンサンプリング（x 軸）で検出されたタンパク質コード遺伝子（y 軸）の数

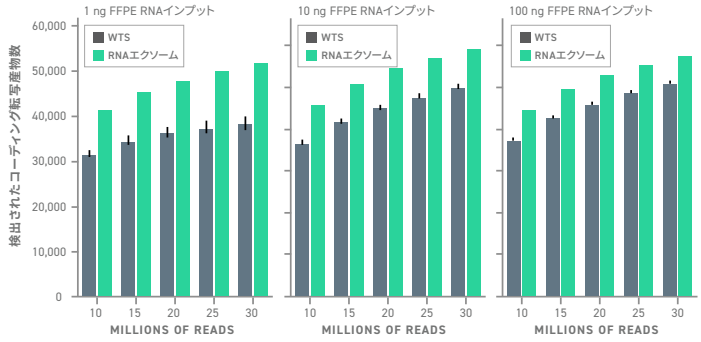


図 3B FFPE サンプルの 3 つのインプット量における異なるレベルのダウンサンプリング（x 軸）で検出されたタンパク質コード転写産物（同一遺伝子の異なるアイソフォームを含む）（y 軸）の数

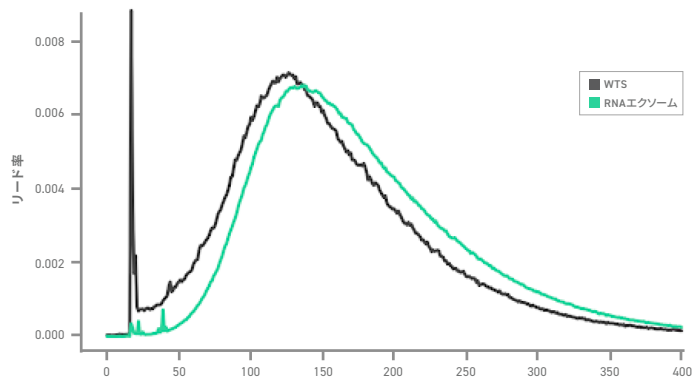


図 3C FFPE サンプルから得たキャプチャありのリード（RNA エクソーム）およびキャプチャなしのリード（WTS）のサイズ分布。

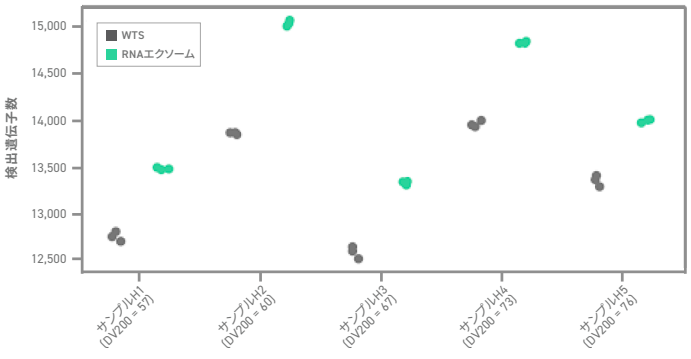


図 3D 品質の異なる様々な FFPE サンプルについて、キャプチャありのカウント数（RNA エクソーム）とキャプチャなしのカウント数（WTS）を比較。

TWIST RNA エクソームを用いた発現差異解析

RNA シーケンシングは、特定の刺激（薬物治療または遺伝子ノックアウトなど）が細胞の転写ランドスケープをどのように変化させるのかに関して生物学的洞察を得るためによく用いられます。この RNA エクソームは、全発現範囲にわたって一貫する濃縮度で転写産物を効率的に濃縮することができるため**（図 2C）（図 2D）**、RNA エクソームから得たカウントを直接用いて、2 つの条件間の発現の変化を評価することができるか検討しました。

これらの変化のモデルとして、我々は市販の対応する乳房腫瘍サンプルと正常な乳房組織サンプルを利用しました。全トランスクリプトームシーケンシングおよび Twist RNA エクソームによるキャプチャの両方を用い、2 つのインプット量条件（10ng 及び 100ng）で各サンプルを 3 回シーケンシングしました。遺伝子あたりのカウントを定量化して発現差異解析を実施した後、WTS で検出された Fold Change をキャプチャで検出された Fold Change と比較しました。これらの推定値 **（図 4A）** 間には良好な一致が認められ、遺伝子発現における Fold Change の条件間の推定にキャプチャから得たカウントを直接使用できることが示されました。RNA キャプチャにより遺伝子の全数が増加するので **（図 2C）**、この増加が発現差解析の検出力を向上させるのかを確認しました。予想されたように、キャプチャ済みサンプルのボルケーノプロットに全般的な上方シフトが認められ、全般的にほとんどのコールが条件間で統計学的に有意であることが示されました **（図 4B）**。個々の遺伝子について FDR 補正した p 値間の対比較を行ったところ、多くの遺伝子が高い検出力で検出されていることが認められました **（図 4C）**。したがって、Twist RNA エクソームは、発現に差異が認められる遺伝子間の既存の関係性を維持し、有用なリードの割合を増加させることにより、より大きな統計的検出力が差異の判定において可能になります。

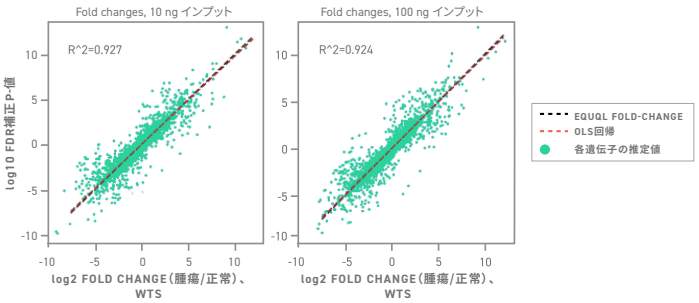


図 4A 2 つの異なる質量インプットで対応させた腫瘍ペアと正常ペアの間のキャプチャなし（WTS）とキャプチャあり（RNA エクソーム）の log2 Fold Change 推定値の相関。

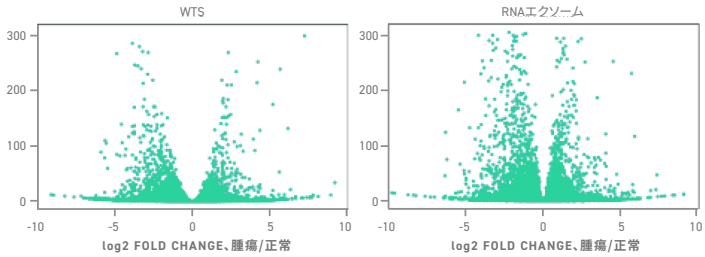


図 4B キャプチャサンプルなし（WTS）およびキャプチャありサンプル（RNA エクソーム）に関する推定 log2 Fold Change 推定値および log10 p 値を示すボルケーノプロット

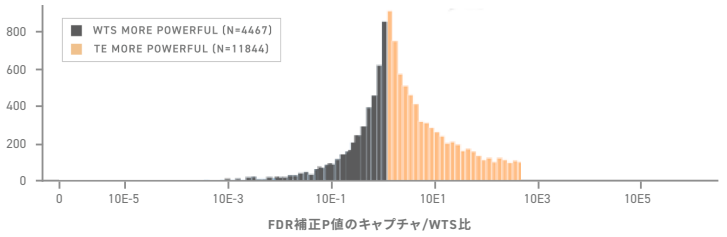


図 4C キャプチャありサンプル（RNA エクソーム）とキャプチャなしサンプルなし（WTS）の FDR 補正 p 値の比率分布を示すヒストグラム オレンジ色のビンは RNA エクソームでより有意な p 値を示し，青色のビンは WTS でより有意な p 値を示します。

TWIST RNA エクソームによる融合転写産物の検出

RNA シーケンシングが成熟した転写産物を検出することから、細胞に機能的影響を及ぼす新規の構造変異体を特異的に検出できます。これらのイベントには融合遺伝子が含まれており、2 つの異なる遺伝子の間に新規のエクソン間ジャンクションが形成される作用機序であり、がんによく見られます。Twist RNA エクソームは、設計スペースで用いられる転写産物の検出を改善するため、設計スペースに含まれなかったこれらの新規転写産物を、Twist RNA エクソームが効率的に検出できるかどうかについても検討しました。

これを調べるため、融合イベントが既に検証済みである細胞系統由来の FFPE 標準品を利用しました。GenCode v41 から組み合わせた予想される融合配列にインデックスを作成し、この転写産物のセットに対して、WTS 実験または RNA エクソームキャプチャのいずれかを用いてサンプリングした 10M リードを分類しました。RNA エクソームが両方の融合転写産物の検出イベント数を有意に増加させることを認めました **（図 4A）**。腫瘍サンプルでは、これらのイベントの発現は低頻度となる可能性があるので、これらのデータにより、RNA キャプチャが重要なイベントの検出力を改善させる効率的な方法であることを立証しました。RNA エクソームが融合ジャンクションにわたるリードを正確に検出することを確保するために、サンプルから得た融合転写産物の配列にアラインされたリードを同様に検討し、多数のリードが転写産物スペースで予想されるジャンクションを超えたことを認めました **（図 5A、図 5B、図 5C）**。これらの結果から、ターゲットリードの改善は、新規のエクソン間ジャンクションをプロファイリングする能力を犠牲にするものではなく、実際にはターゲットリード数が増加することにより、これらの重要なイベントの検出感度の向上を可能にすることが示されました。

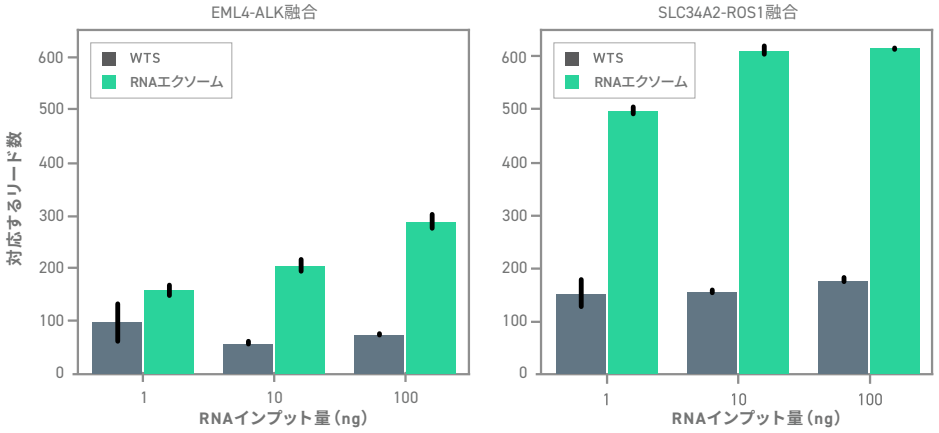
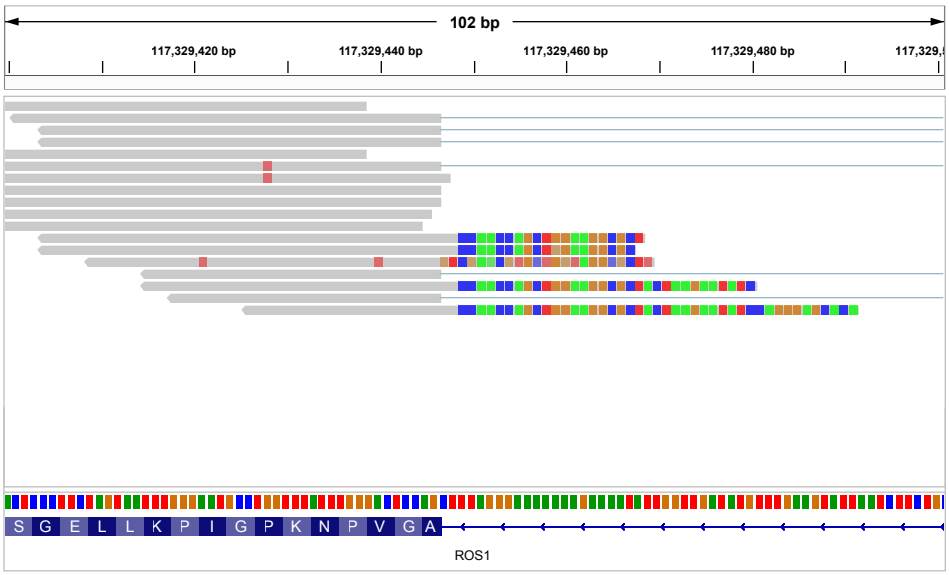


図 5A キャプチャありサンプル（RNA エクソーム）およびキャプチャなしサンプル（WTS）の様々なインプット量に対する 2 つの融合転写産物（EML4-ALK および SLC34A2-ROS1）を検出するリード数

SLC34A2-ROS1 融合



EML4-ALK 融合

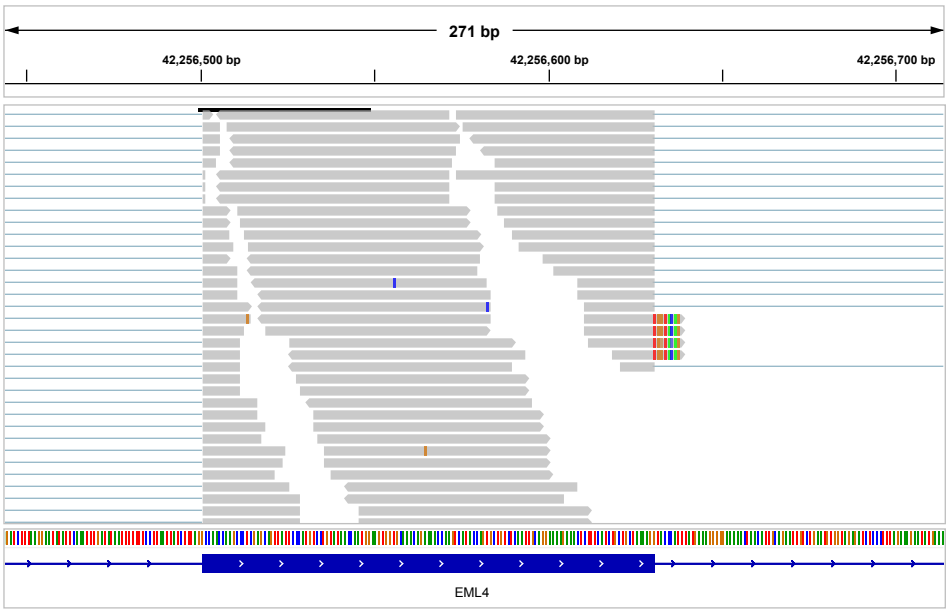


図 5B および図 5C ROS1（上）および EML4（下）のブレイクポイントにおける融合の検出ゲノムブラウザの表示 リードの末尾に認められる色付きブロックは、参照配列にアラインする塩基ではなく、融合パートナーに対応する塩基を示しています。

概要

我々の結果は、Twist RNA エクソームを用いた RNA キャプチャが、転写プロファイリングの強力なアプローチであることを示しています。RNA エクソームが、成熟した転写産物のプロファイリングにおいて 3’ 末端計数または全トランスクリプトームシーケンシングよりも効率的であることを示しています **(図 2)**。この効率の向上から期待されるシークエンス量の節約を定量化し、分解サンプルや低インプット量に対して RNA エクソームが有用であることを実証しました **(図 3)**。最後に、エクソームキャプチャにより、バイアスが生じることなく発現頻度に差異が認められるコールの統計的検出力を高めることができ **(図 4)**、RNA エクソームが融合転写産物に対する感受性を高めることを示しています **(図 5)** 。

材料および方法

Twist RNA エクソームパネルをテストするために、1ng、10ng、 または 100ng の Universal Human Reference RNA（Agilent P/N 740000）または FFPE RNA Fusion Reference Standards（Horizon Discovery P/N HD784）を、Twist RNA シーケンシングライブラリ調製キットに追加しました。ライブラリ作成の前に、Qiagen RNeasy®FFPE キットを用いて FFPE 検体を抽出しました。ターゲットエンリッチメントは、500ng のライブラリおよびハイブリダイゼーション反応時間 16 時間の Twist Target Enrichment Standard Hybridization v2 プロトコルを用いて実施しました。シーケンシングは、Illumina NextSeq プラットフォームおよび 76 bp のペアエンドリードを用いて実施しました。

解析は、FASTQ ファイルを一定のリード数（特に指定されていない限り、10M ペア／20M リード）までサンプリングすることにより実施しました。STAR (Dobin et al 2013) を用いて hg38 に対してアラインメントを実施し、GenCode v41 遺伝子アノテーションと共に FeatureCounts（Liao et al 2014）を用いて遺伝子を定量化しました。メトリクスは、Picard CollectRnaSeqMetrics を用いて算出しました。DESeq2（Love et al 2014）を用いて発現の差異を評価しました。データの処理および可視化は、カスタム Python スクリプトを用いて Pandas および Seaborn で実施しました。IGV を用いてゲノムブラウザを可視化しました。融合転写産物は、融合転写産物配列に連結した GenCode v41 転写産物配列から構築したインデックスと共に Salmon（Patro et al 2017）を用いて定量化しました。

参考文献

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25 PMID: 23104886; PMCID: PMC3530905.

Jang JS, Holicky E, Lau J, McDonough S, Mutawe M, Koster MJ, Warrington KJ, Cuninngham JM. Application of the 3' mRNA-Seq using unique molecular identifiers in highly degraded RNA derived from formalin-fixed, paraffin-embedded tissue. BMC Genomics. 2021 Oct 24;22(1):759. doi: 10.1186/s12864-021-08068-1. PMID: 34689749; PMCID: PMC8543821

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014 Apr 1;30(7):923-30. doi: 10.1093/bioinformatics/btt656. Epub 2013 Nov 13 PMID: 24227677.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PMID: 25516281; PMCID: PMC4302049.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417-419. doi: 10.1038/nmeth.4197. Epub 2017 Mar 6. PMID: 28263959; PMCID: PMC5600148.

Pennock ND, Jindal S, Horton W, Sun D, Narasimhan J, Carbone L, Fei SS, Searles R, Harrington CA, Burchard J, Weinmann S, Schedin P, Xia Z. RNA-seq from archival FFPE breast cancer samples: molecular pathway fidelity and novel discovery. BMC Med Genomics. 2019 Dec 19;12(1):195. doi: 10.1186/s12920-019-0643-z. PMID: 31856832; PMCID: PMC6924022.

Vahrenkamp JM, Szczotka K, Dodson MK, Jarboe EA, Soisson AP, Gertz J. FFPEcap-seq: a method for sequencing capped RNAs in formalin-fixed paraffin-embedded samples. Genome Res. 2019 Nov;29(11):1826-1835. doi: 10.1101/gr.249656.119. Epub 2019 Oct 24 PMID: 31649055; PMCID: PMC6836741.

さらに詳しく知る

TWISTBIOSCIENCE.COM/NGS
JSALESCUSTOMER@TWISTBIOSCIENCE.COM