

アプリケーションノート

*Kinnex full-length RNA kit*を用いた アイソフォームシーケンシング

はじめに

真核生物の選択的スプライシング（AS）は、同じ遺伝子内で発現するエクソンの異なる組合せで機能的な多様性を産みだします。AS で生成された全長転写産物アイソフォームの正確なキャラクタライゼーションは、生物学と疾患の研究において大変重要です。ショートリードを用いたバルク RNA-Seq では、複雑な AS をアセンブリで完全には復元できないため、最新のコンピューターツールでもアイソフォームの完全な解明ができません（[Stark et al., 2019](#)）。PacBio®の技術によるロングリードの RNA-Seq（Iso-Seq®法）は、全長 cDNA のシーケンシングできるためアセンブリが必要なく、多くのアプリケーションでのアイソフォームの新発見を可能にします（図 1）。

*Kinnex™ full-length RNA kit*はトータル RNA をインプットとし、cDNA を連結したライブラリを作成することにより典型的な Iso-Seq ライブラリと比べてスループットを 8 倍に向上させるキットです。SMRT® Link ソフトウェアの Iso-Seq ワークフローを使用することで、他のシーケンシング法を必要としない、費用対効果の高いシーケンシングデータを得ることができます。SMRT Link ソフトウェアは、発現量情報を含む[三次解析ツール](#)で使用できる、アイソフォーム分類レポートを作成します。

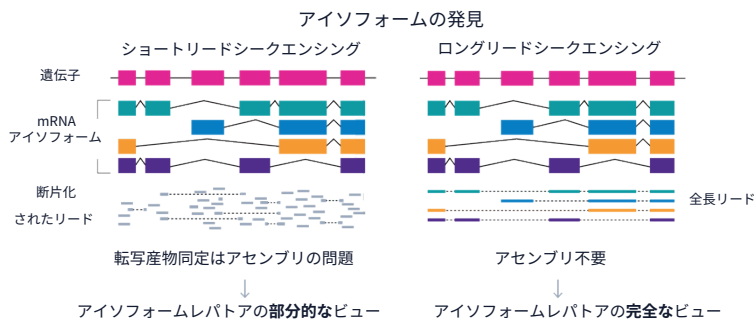


図 1. ロングリードの RNA シークエンシングでは、転写産物のアセンブリが不要でアイソフォーム構造を正確に解析できます。PacBio (Iso-Seq 法) を用いたロングリード RNA-Seq では、全長 cDNA をシークエンシングし、トランスクリプトームの明確な全体像を提供します。

Iso-Seq 法による全長 RNA シークエンシング

従来の RNA-Seq では、ショートリードシークエンシング (100-200 bp) 用に、cDNA を断片化し、その後に元の転写産物アイソフォームを推測するため、コンピュータによる解析が必要となります。しかし、選択的スプライシングの複雑さから、多くのアイソフォーム間で非常に類似した構造を共有しているため、推測された転写産物はしばしば不正確となります。PacBio の HiFi リードは、cDNA の断片化および転写産物のアセンブリが不要で全長 RNA アイソフォームをシークエンシングし (図 1)、全長アイソフォームの明確な検出を可能にします。

全長 RNA シークエンシングにおける HiFi シークエンシングの利点

- 5'末端から 3'末端までの全長アイソフォームをシークエンシング
- スプライシング部位を正確にキャラクタライズ
- 新規遺伝子およびアイソフォームを検出
- アイソフォームリードカウント情報を取得

Iso-Seq 法は、PacBio の HiFi シークエンシングを使用して全長転写産物をシークエンシングし、生物学や疾患における多くの分野で活用されています。ヒトの疾患研究で、Iso-Seq 法は希少疾患、表現型形質、神経疾患に関連するスプライシング異常の特定に使用されています。がん研究では、がんドライバー変異、融合遺伝子や、がんワクチン候補として使用できる可能性のあるネオエピトープの発見に Iso-Seq 法が使用されています (Li et al., 2023)。

植物および動物の研究でも、Iso-Seq 法は高品質なゲノムアノテーションの作成 (Zhang et al., 2023) や親特異的アイソフォーム発現の同定 (Wang et al., 2020) に使用されています。

Kinnex full-length RNA kit

Kinnex full-length RNA kit では、PacBio ロングリードシークエンサーでのスループット向上に MAS-Seq 法を利用しています。MAS-Seq は、cDNA 分子をより長い断片に連結する手法です (Al'Khafaji et al., 2023)。連結した分子のシークエンシングで生成された HiFi リードは、バイオフィォマテイク的に分割され、元の cDNA 配列を復元します。その結果、スループットが向上し、シークエンシングの必要回数が減ることで、費用対効果の高いアイソフォームシークエンシングが可能となります。

SMRT Link の PacBio Iso-Seq ワークフローでは、全長 cDNA 配列を処理し、リファレンスアノテーション (例 GENCODE) に照合して分類し、新規遺伝子とアイソフォームを同定します。解析結果は、アイソフォーム分類結果を含む、三次解析ツールに互換性のあるリードカウントとして出力できます。

Kinnex full-length RNA kit はトータル RNA (300 ng) をインプットとして使用し、シークエンシング可能なライブラリを 2 日間のワークフローで作製します。

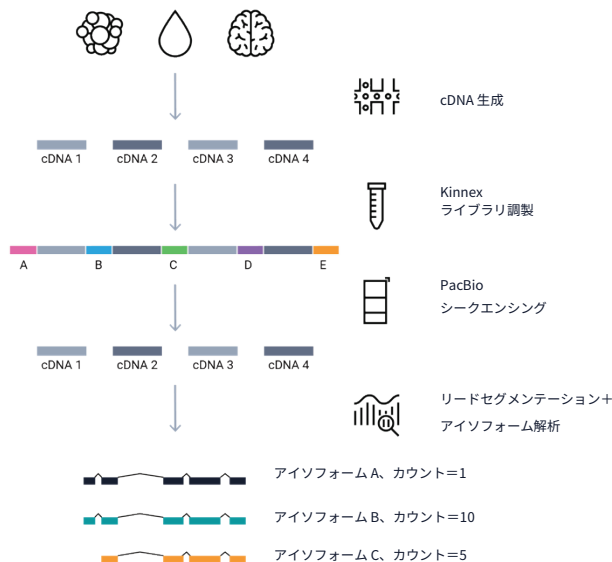


図 2. Kinnex 全長 RNA シークエンシング。全長 cDNA 分子を連結してラージインサートライブラリとし、シークエンシングを行った後に PacBio ソフトウェアで処理する。

Kinnex RNA ライブラリワークフロー

Kinnex 全長 RNA ワークフロー (図 3) では、トータル RNA を起点として、シーケンシング可能なライブラリを作製します。

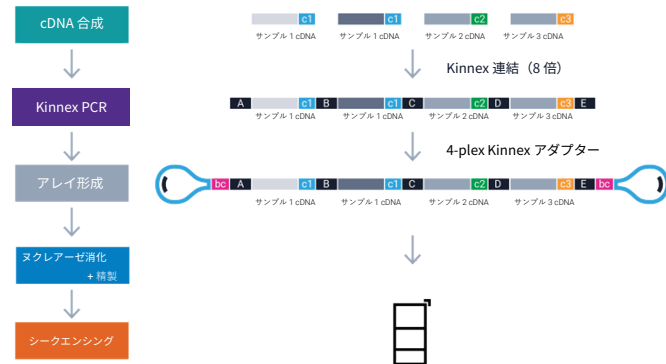


図 3. Kinnex 全長 RNA ライブラリワークフロー

全長 cDNA 分子は、Kinnex アレイ形成に使用できるハンドルと共に合成されます (*Iso-Seq* express 2.0 キットを使用)。cDNA 増幅の一部として、5'末端に cDNA バーコードを付加します。HiFi シーケンシングに先立ち、連結数の少ないアレイはヌクレアーゼ消化中に除去されるため、Kinnex アダプターライゲーションでフルアレイの濃縮を確実に行えます。バーコード付きの cDNA は最大 12-plex をサポートし、Kinnex アダプターはライブラリレベルで 4-plex をサポートしています。

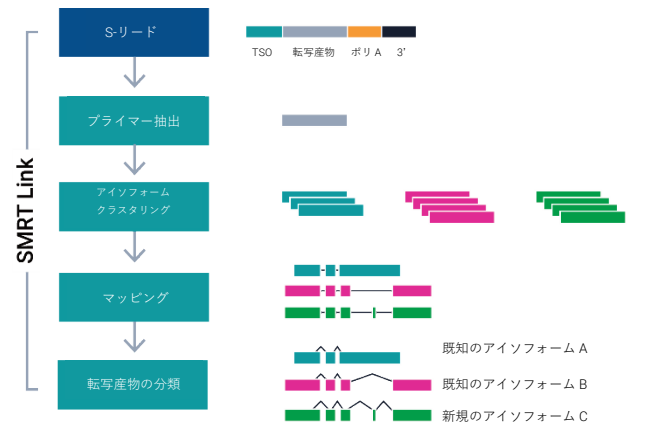
適切なフルアレイ形成と十分なシーケンシングにより、Sequel[®] II/Ile および Revio[™] システム上の 1 つの SMRT[®] Cell でそれぞれ 1,500 万および 4,000 万の cDNA 配列を期待できます (表 1)。

Kinnex RNA バイオフィォーマティクスワークフロー

SMRT Link リードセグメンテーションと *Iso-Seq* ワークフロー (図 4) は、Kinnex 全長 RNA ライブラリから生成された HiFi リードを処理して、[三次解析ツール](#) に互換性のある、リードカウントを含む分類済のアイソフォームを生成します。

アイソフォームのクラスタリング

FLNC リードは配列の類似性でクラスター化され、アイソフォームコンセンサス配列を生成します。ゲノムが提供されない場合、このステップが *Iso-Seq* 解析最後のステップとなります。



ID	遺伝子	転写産物	サンプル 1 の TPM	サンプル 2 の TPM
PB.1.1	ARGN	Novel	100	200
PB.1.2	ARGN	ENST00000379370.7	10	1.4
PB.10.1	ACTB	ENST00000646664.1	8132	8000

図 4. リードセグメンテーションと *Iso-Seq* ワークフローを用いた、Kinnex 全長 RNA 解析

マッピング

ゲノムが提供される場合、前ステップ由来のアイソフォームコンセンサス配列がマッピングされます。また可視化用にアイソフォームを GFF ファイルとして生成するため、エクソン構造に従い分割します。

転写産物の分類

アノテーション (例 GENCODE) が提供される場合、既知および新規遺伝子/アイソフォームを同定するため、アイソフォームを [pigeon](#) (SQANTI3 の PacBio インプリメンテーション) を用いて分類します。*Iso-Seq* ワークフローは、プールされたサンプルリードを一緒に解析し、統一されたアイソフォームアノテーションを、サンプルごとに Raw カウントおよび 100 万あたりで正規化された (CPM) カウント値付きで作成することができます。

指標	性能
サンプル調製時間	2 日間
期待されるライブラリサイズ	11,000–18,000 bp
目標の P1 ローディング	60–80%
期待される HiFi 収量	150 万–250 万 HiFi リード (Sequel II/Ile) 400 万–600 万 HiFi リード (Revio)
期待されるフルアレイ%	80–90%
期待されるリード収量	~1,500 万リード (Sequel II/Ile) ~4,000 万リード (Revio)

表 1. 想定される Kinnex 全長 RNA ライブラリの性能

SMRT Link は現在、ヒトおよびマウスサンプル用の転写産物分類のみをサポートしています。ヒト/マウス以外のサンプルは、カスタムアノテーション GTF ファイルを用いて、[コマンドラインで解析する](#)必要があります。

SMRT Link の留意事項

以下は、Iso-Seq ワークフローを実行する上での、いくつかの一般的な留意事項および推奨事項です。

リードカウントを含む分類済アイソフォームの生成に、SMRT Link リードセグメンテーションと Iso-Seq ワークフローは現在、ヒトおよびマウスのリファレンスゲノムとアノテーションをサポートしています。他の生物をご利用の場合、推奨解析方法については表 2 をご覧ください。

リファレンス/アノテーション	推奨解析方法
ヒトまたはマウス	分類とリードカウント情報を含む、マッピングされたユニークアイソフォームの取得 (FASTA, GFF, TXT) には、ヒト/マウスのアノテーションをプレロードした Iso-Seq ワークフローを使用してください。
良好なアノテーションを持つモデル生物	マッピングされたユニークアイソフォームの取得には、アップロードしたリファレンスゲノム (FASTA, GFF) で Iso-Seq を実行してください。 pigeon 準拠アノテーション を生成し、コマンドラインを使用して、リードカウント情報 (TXT) 付きのアイソフォームを分類してください。
ゲノム情報がある非モデル生物	マッピングされたユニークアイソフォームの取得には、アップロードしたリファレンスゲノム (FASTA, GFF) で Iso-Seq を実行してください。
ゲノム情報なし	ユニークアイソフォームの取得には、リファレンスゲノム (FASTA) を使用せずに Iso-Seq を実行してください。

表 2. リファレンスゲノムおよびアノテーションに基づく、Iso-Seq データの推奨解析法

研究の目的	中程度から希少な転写産物のアイソフォームの発見と定量	高発現転写産物のアイソフォームの発見	種の包括的な転写産物のアノテーション
例	複数のレプリケートを含む疾患組織対正常組織	>20+サンプルの疾患コホート	複数種類の組織を含む植物または動物
ターゲットデプス	1 サンプルあたり 1,000 万リード	1 サンプルあたり 500 万リード	1 サンプルあたり 500 万リード
ライブラリ	1 つの Revio SMRT Cell に対して 4-plex の cDNA、または 1 つの SMRT Cell 8M に対して 2-plex の cDNA		1 つの Revio SMRT Cell に対して 8-plex の cDNA、または 1 つの SMRT Cell 8M に対して 3-plex の cDNA
解析	「Pool reads and cluster together」オプションで、リードセグメンテーションと Iso-Seq ワークフローを実施し、サンプル毎の全長リードカウントを持つマスターアイソフォーム分類を取得		

表 3. 様々な研究目的別のシーケンシングと推奨分析例

シーケンシングデプスは、実験目的とサンプルによって変わりますが、一般的な推奨事項は表 3 をご覧ください。

Kinnex 公開データセットリリース

公開 Kinnex RNA データセットは、主に HG002 細胞株とユニバーサルヒトリファレンス RNA (UHRR) から構成されています。比較用として WTC-11 細胞株、ヒト脳、ソルガムおよびマウス由来の追加サンプルが公開予定です。リードセグメンテーションと Iso-Seq ワークフローの後、各サンプルで >20,000 個のユニーク遺伝子を取得しました (表 4)。転写産物長は 100 bp から 11,000 bp の範囲で、サンプルおよび種依存的にわずかに違いました (図 5)。

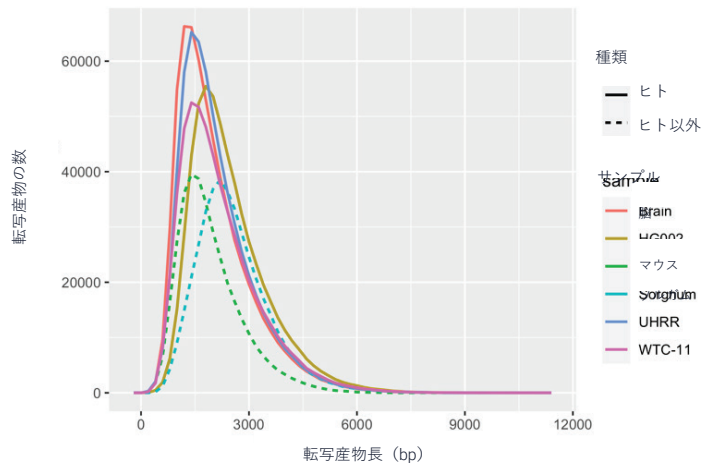


図 5. 様々な Kinnex ライブラリの転写産物長は、サンプルおよび種特異的に違いがあるが、大部分は同程度となる。

サンプル	ライブラリ	HiFi リード	S-リード	平均 S-リード長	既知遺伝子	新規遺伝子	既知アイソフォーム	新規アイソフォーム
UHRR	Non-Kinnex – Sequel II/IIe	3,194,311	n/a	n/a	12,921	121	27,821	16,323
	Kinnex – Sequel II/IIe	2,720,033	20,453,853	1,918	18,903	1369	53,623	102,059
	Kinnex-Revio	6,546,645	47,250,258	1,914	22,365	4,223	68,087	231,467
HG002	HG002	5,984,046	38,740,671	2,227	18,230	8,448	55,689	281,460
WTC-11	Day 0 – rep 1	6,920,750	54,110,504	1,856	19,905	3,636	58,651	212,523
	Day 0 – rep 2	8,611,025	67,547,611	1,764	21,170	4,932	63,553	277,189
	Day 0 – rep 3	8,124,744	63,251,235	1,864	20,744	4,822	62,349	257,665
	Day 1 – rep 1	6,430,958	49,897,067	1,743	20,204	4,629	60,451	213,429
	Day 2 – rep 1	7,353,759	58,217,895	1,201	21,169	6,570	64,066	165,496
	Day 3 – rep 1	5,483,994	42,173,159	1,844	19,436	2,692	56,533	185,650
	Day 3 – rep 2	6,687,580	52,317,384	1,705	21,270	4,482	63,430	241,726
	Day 4 – rep 1	7,295,962	57,061,795	1,727	21,751	3,594	63,466	225,636
	Day 5 – rep 1	6,645,009	51,741,094	1,751	21,754	3,195	62,217	185,807
	Day 5 – rep 2	7,542,604	59,092,202	1,792	21,721	3,613	65,369	228,584
	Day 5 – rep 3	6,358,300	49,466,302	1,803	21,389	3,652	59,638	187,394

表 4. UHRR、HG002 および他の共同研究者の WTC-11 細胞株の Kinnex 全長 RNA データ。HiFi リードは SMRT Link v13.0 のリードセグメンテーションと *Iso-Seq* ワークフローを使用して解析した。非 Kinnex、および Sequel II/IIe システムでシーケンシングされた Kinnex を除き、全サンプルは Kinnex ライブラリで、1 つの Revio SMRT Cell でシーケンシングした。新規遺伝子とアイソフォームは pigeon を使用し、GENCODE v39 に照合して決定した。

Kinnex と非 Kinnex のライブラリを比較すると、Kinnex 連結または他のシーケンシングプラットフォームを使用した場合でも、転写産物長はシフトしないことが示されました (図 6)。さらに、大部分のアイソフォームの存在量は一致していました (図 7)。

Kinnex ライブラリはテクニカルレプリケートの再現性が高く (表 5)、イルミナの同じデータで示された再現性の内容と一致しました (未掲載)。

飽和曲線が示すように、既知遺伝子とアイソフォームの多くが 1,000 万リードで検出できました (図 9)。シミュレーションされた飽和曲線の裏付けとして、WTC-11 サンプルを 500 万および 1,000 万リードでそれぞれサブサンプリングし、Sequel II/IIe および Revio で取得可能な総収量のシミュレートのために異なる plex 数でプールしました。予想通り、3-plex 5M (Sequel II/IIe では SMRT Cell 8M あたりトータル 1,500 万リード) では、8-plex 5M (Revio SMRT Cell あたりトータル 4,000 万リード) や 4-plex 10M ライブラリよりも検出アイソフォーム数が少なくなりました (図 10)。

飽和およびサブサンプリングしたデータを合わせると、1 サンプルあたり 1,000 万リードで、~80%の既知アイソフォームを検出できることが示されました。シーケンシングデプスを増やすことで、検出される新規アイソフォームの数は増えますが、新たに検出されたアイソフォームのほとんどは、存在量が少なくなりました。

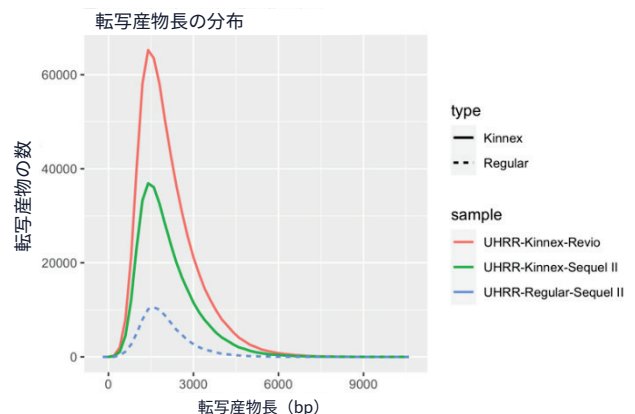


図 6. 同一 UHRR サンプルでは、Kinnex (連結) および非連結ライブラリ間で転写産物長の分布は同一傾向にある。

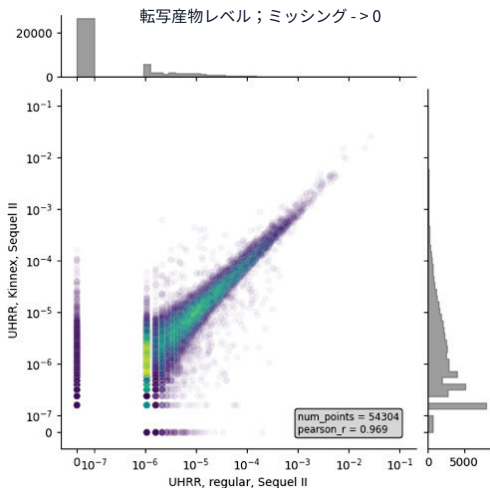


図 7. Kinnex 連結でアイソフォーム存在量を歪めることはなかった。Sequel II/e システムにおいて Kinnex と Kinnex 以外の UHRR データ間では、アイソフォーム存在量の高い相関が見られた。既知の GENCODE アイソフォームのみを比較した。Kinnex-Revio のデータとも、同様の相関 (0.964) が見られた (未掲載)。

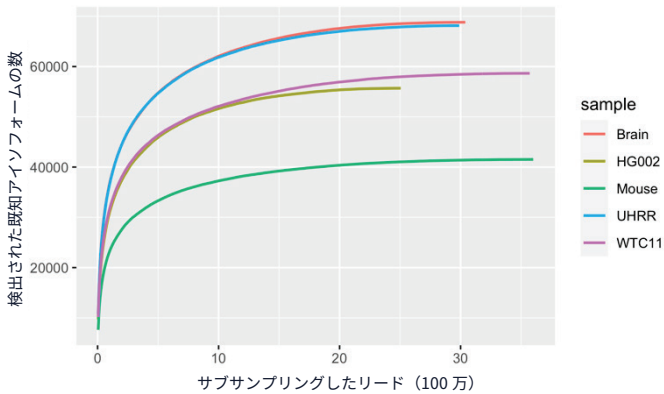


図 9. アイソフォームレベルでの WTC-11 Kinnex サンプルの飽和曲線。1,000 万リードで、既知アイソフォームの大部分が検出された。

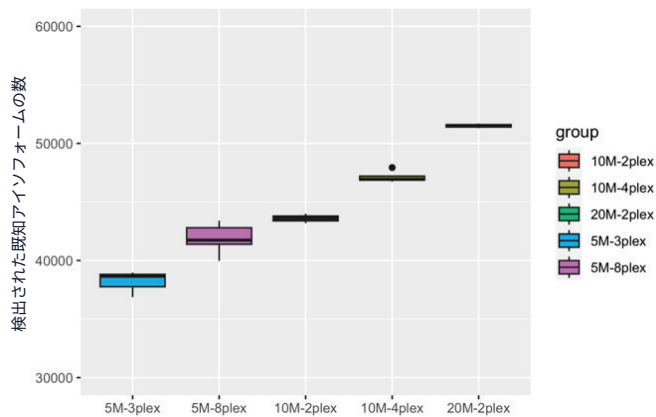


図 10. 1 サンプルあたり 500 万または 1,000 万リードデプスのシミュレーションで検出された、既知アイソフォーム数。プールのシーケンシングデプスの総数、サンプルリードあたりの数が増えるほど、より多数の既知アイソフォームが検出される。新規アイソフォームの数は、30,000-70,000 (5M-3-plex、10M-2-plex、5M-8-plex) から 90,000-130,000 (10M-4-plex、20M-2-plex) の範囲で検出された。

WTC-11 day 0				WTC-11 day 5			
	Rep 1	Rep 2	Rep 3		Rep 1	Rep 2	Rep 3
Rep 1	1.00	0.80	0.79	Rep 1	1.00	0.80	0.80
Rep 2	0.80	1.00	0.81	Rep 2	0.80	1.00	0.79
Rep 3	0.79	0.81	1.00	Rep 3	0.80	0.79	1.00

表 5. Kinnex ライブラリにおける良好なテクニカルレプリケート。WTC-11 サンプル由来の 3 つの各テクニカルレプリケートは、0 日目と 5 日目の両方でアイソフォーム存在量の高い相関を示し、イルミナで同じテクニカルレプリケートを取った際の相関値と同様であった (0.78-0.82、データ未掲載)。

WTC-11 および UHRR サンプルで生成された Kinnex RNA データセットの解析によって、以下が実証されました：

- 通常の Iso-Seq ライブラリと比較して、Kinnex ライブラリでも検出転写産物のサイズあるいは存在量に変化は無かった。
- Kinnex ライブラリにおいて、生物種やサンプルの種類によって検出される転写産物サイズの差はあるものの、全体的な傾向は変わらなかった。
- ライブラリレプリケートおよび PacBio ロングリードシーケンシングプラットフォームを通して、技術的再現性が高い。
- 1,000 万リードで、大部分の既知遺伝子とアイソフォームが検出された。検出された新規アイソフォーム数は、シーケンシングデプスと共に増加するが、次第に稀になる。

Kinnex full-length RNA kit と SMRT Link 解析の組み合わせで、疾患と生物学へ非常に優れた洞察の提供が期待できる、高品質な全長を提供します。

結論

PacBio の Iso-Seq 法は全長転写産物を高精度にシーケンシングし、アイソフォームの明確なキャラクタライゼーション、アレル特異的アイソフォーム発現、発現差異解析などを可能にします。

Kinnex full-length RNA kit は MAS-Seq 連結技術により、スループットを 8 倍に向上します。フレキシブルなマルチプレックス戦略と SMRT Link バイオインフォマティクスワークフローの組み合わせで、ユーザーは費用対効果の高い方法で包括的なアイソフォームシーケンシングを実現できます。

資料集と参考文献

資料集

[アプリケーション概要 – A more complete cancer transcriptome with the Iso-Seq method – single-cell and bulk RNA sequencing.](#)

[ホワイトペーパー – Bulk and single-cell isoform sequencing for human disease research.](#)

[アプリケーションノート – Bioinformatics tools for full-length isoform sequencing.](#)

Kinnex 全長 RNA データセット：
<https://pacb.com/datasets>

Iso-Seq の資料：<https://isoseq.how/>

pigeon の資料：<https://isoseq.how/classification/>

参考文献

Al'Khafaji, A. M., et al. (2023). High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nature Biotechnology*, 1-5.
<https://doi.org/10.1038/s41587-023-01815-7>

Li, Z., et al. (2023). An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing. *bioRxiv*, 2023-04.
<https://doi.org/10.1101/2023.04.21.536771>

Pardo-Palacios, F., et al., (2023). SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *bioRxiv*, 2023-05.
<https://doi.org/10.1101/2023.05.17.541248>

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11), 631-656.
<https://doi.org/10.1038/s41576-019-0150-2>

Wang, B., et al. (2020). Variant phasing and haplotypic expression from long-read sequencing in maize. *Communications Biology*, 3(1), 78.
<https://doi.org/10.1038/s42003-020-0805-8>

Zhang, R., et al. (2022). A high-resolution single-molecule sequencing-based Arabidopsis transcriptome using novel methods of Iso-Seq analysis. *Genome Biology*, 23(1), 149.
<https://doi.org/10.1186/s13059-022-02711-0>

研究用のみに使用できます。診断目的およびその手続き上の使用はできません。

Research use only. Not for use in diagnostic procedures. © 2023 Pacific Biosciences of California, Inc. ("PacBio"). All rights reserved. Information in this document is subject to change without notice. PacBio assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of PacBio products and/or third-party products. Refer to the applicable PacBio terms and conditions of sale and to the applicable license terms at pacb.com/license. Pacific Biosciences, the PacBio logo, PacBio, Circulomics, Omniome, SMRT, SMRTbell, Iso-Seq, Sequel, Nanobind, SBB, Revio, Onso, Apton, and Kinnex are trademarks of PacBio.

© 2023 PacBio. All rights reserved. Research use only. Not for use in diagnostic procedures.

102-326-591 REV02 NOV2023