Whole Transcriptome解析への高速シーケンサーの利用

■ はじめに

高速シーケンサーを用いたさまざまなゲノム解析手法が開発されていますが、Whole Transcriptome解析への利用法として、RNA sequencing技術も急ピッチで開発されています。RNA sequencingのうち、メッセンジャーRNA(mRNA)配列を網羅的に解読する手法はmRNA sequencingと呼ばれ、真核生物に対する新規のWhole Transcriptome解析手法として注目を集めています。mRNA sequencingでは転写産物の発現定量化だけではなく、その配列決定から新規転写産物および新規スプライシングジャンクションの探索までを網羅的に行うことができ、マイクロアレイ法やSAGE法などにはない利点を多くもつ魅力的な解析法となっています。

同様に、高速シーケンサーを利用したマイクロRNA(miRNA)の解析手法(miRNA sequencing)も開発が進んでいます。miRNAは18～26塩基程度の短いRNAで、mRNAの分解や翻訳抑制などを行っていると考えられており、癌をはじめとする多くの疾患との関連が注目されています。しかしながら、miRNAは非常に短いこと、比較的低発現のものが多いこと、1塩基異なるだけで別種のmiRNAになることなどから、ハイブリダイゼーションを基本とするマイクロアレイではその検出に限界がありました。一方、1塩基レベルの解像度をもつ高速シーケンサーではこれらは問題とはならず、さらにマイクロアレイよりもダイナミックレンジを広く確保できるため、低発現のmiRNAについてより詳細に解析ができます。

本稿では、タカラバイオで解析受託を承っている高速シーケンサーを用いたmRNA sequencing解析およびmiRNA sequencing解析について具体的な解析例を交えてご紹介いたします。

■ Illumina® Genome Analyzerを使用したヒト脳のmRNA sequencing(mRNA-Seq)解析例

【方法】

Human Brain Total RNA(以下Adult)(製品コード 636530)およびHuman Fetal Brain Total RNA(以下Fetus)(製品コード 636526)の各10 μgを材料に、mRNA-Seq Sample Preparation Kit(イルミナ社)の推奨条件に従ってpolyA$^+$ RNAを精製後、150～250 bpサイズのcDNAライブラリーを作製した。これを用いて、イルミナ社Illumina® Genome Analyzer(以下GAII)にて1レーン分のシングルリードシーケンス解析(50塩基)を行った。今回の解析では、ゲノム配列へのリードのマッピングソフトウェアにBowtie、データ解析ソフトウェアにERANGEを使用した。

ERANGEは、mRNA-SeqおよびChIP-Seq実験から得られる大量のシーケンスデータを解析するために開発されたPythonスクリプトのパッケージである。RNA-Seq解析では、RPKM(Reads per kilo base of exon model per million mapped reads)というノーマライズされたリード数を計算することができる。また、新規スプライシングジャンクションの検出にはTopHatソフトウェアを使用した。いずれもフリー、オープンソースのソフトウェアであり、それぞれ下記アドレスからダウンロードした。

Bowtie：http://bowtie-bio.sourceforge.net/index.shtml
ERANGE：http://woldlab.caltech.edu/html/woldlab
TopHat：http://tophat.cbcb.umd.edu/

【結果】

1. シーケンスデータの概要

GAIIによる1レーン分のシングルリードシーケンス(50塩基)で取得できたデータ量は、Adultで約850万リード、Fetusで約880万リードでした。ERANGEによる解析の結果、ゲノムにマッピングされた使用可能なリード数はAdultで6,842,062、Fetusでは6,837,087となりました。この計算過程で既存の遺伝子モデルに適合しないリードを抽出し、こうしたリードをさらに解析後、「RNAFAR」として定義したうえで、新規転写産物の候補としました。本検討でRNAFARとして定義されたリード数はAdultで19,724、Fetusで10,030あり、合計116種類が新規転写産物の候補として特定されました(表1)。

表1　mRNA-Seqシーケンスデータ解析結果の概要

	Adult		Fetus	
	Num.	%	Num.	%
総リード数	8,459,750	100	8,822,738	100
ゲノムにマッピングされたリード	6,842,062	80.9	6,837,087	77.5
ユニークリード	5,514,554	65.2	5,356,589	60.7
マルチリード	569,505	6.7	547,399	6.2
スプライスリード	710,577	8.4	832,296	9.4
既知遺伝子モデルに適合したリード	4,768,543	56.4	4,675,393	53.0
既知遺伝子モデルに適合しないリード	746,011	8.8	681,196	7.7
RNAFAR	19,724	0.2	10,030	0.1

2. mRNA-Seq解析の定量性評価

mRNA-Seqで検出された転写産物の定量再現性を評価するために、同一サンプルで行ったマイクロアレイ解析(アフィメトリクス社)のFetus/Adult比の結果と比較しました(図1)。

まず、マイクロアレイデータおよびmRNA-Seqデータの両方で一致して発現していることが確かな7,436遺伝子を特定し、これを用いて両実験間のピアソン相関係数を計算したところ、0.78となりました。このようにmRNA-Seqとマイクロアレイの結果は高い相関を示しました。この例はGAIIでわずか1レーン分(約800万リード)のシーケンス解析の結果ですが、さらに解析量を増やせば、低発現遺伝子を含めたより多数の遺伝子の発現に対して、定量性の高い結果を得られることが期待できます。

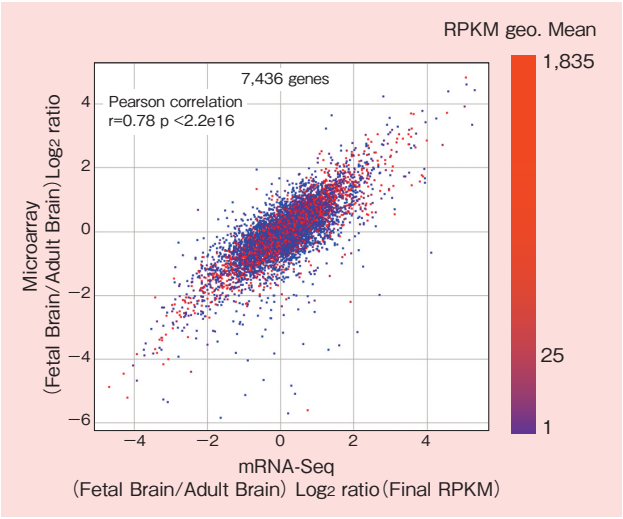

図1　mRNA-Seqとマイクロアレイ解析データとの比較

マイクロアレイおよびmRNA-Seqのデータから、Adult/Fetusの発現比率をそれぞれ計算し、Log_2変換して比較した。散布図上の各点が各転写産物の発現比率を表す。右端の棒グラフはmRNA-Seq結果におけるAdultとFetusのRPKM値の平均値の分布を表しており、25RPKM値を境に高発現の転写産物を赤、低発現の転写産物を青いドットで表している。高発現の転写産物ほどy＝xの直線上にプロットされる傾向がみられ、より高い相関を示すことがわかった。

3. スプライシングジャンクションの検出

TopHatによるスプライシングジャンクションの解析結果はUCSCウェブサイト(http://genome.ucsc.edu/)のブラウザを用いて表示することができ、既存遺伝子モデルには存在しないエキソンにマップされるリード配列を検出し、新規スプライシングジャンクションとして視覚的に把握することができます(図2)。

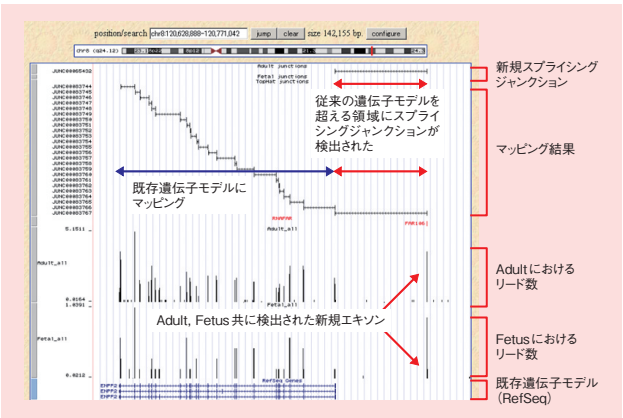

図2　新規スプライシングジャンクションの検出例

TopHatでの解析結果。RefSeqには登録されていない領域にエキソンを検出。既知エキソンと繋げて新規スプライシングジャンクションとして検出されている。

本検討ではAdultおよびFetus両方から合計88,846のスプライシングジャンクション配列を検出し、この中から新規のものとしてAdultでは3,976個、Fetusでは4,989個のスプライシングジャンクション候補が見つかりました。
この解析手法を用いることで、新規スプライシングジャンクション候補の検出だけでなく、発生ステージごとに変化するスプライシングバリアントを検出することなども可能であり、Whole Transcriptome解析の可能性が大いに広がります。

■ GAIIを使用したマウス細胞のmiRNA sequencing(miRNA-Seq)解析例

【方法】

異なる状態の4種類のマウス細胞サンプル(SampleA、SampleB、SampleCおよびControl：各10 μg)を材料に、Small RNA Sample Prep Kit ver.1(イルミナ社)の推奨条件に従って18～26 bpサイズのcDNAライブラリーを作製した。これを用いてGAIIによる1レーン分のシングルリードシーケンス(36塩基)解析を行った。得られたリード配列をゲノム配列に対してマッピングし、ゲノム配列に完全一致したリードを抽出して各種データベース(miRNA, piRNA, rRNA, tRNA, snRNA, snoRNA, miscRNA, RefSeq transcripts)を用いてリード配列にアノテーション付けを行った。また、RNA-editingの検出の解析には1塩基ミスマッチを許容してマッピングを行った。

【結果】

1. シーケンスデータの概要

GAIIによる1レーン分のシングルリードシーケンス(36塩基)で取得できたデータ量は、SampleA、SampleB、SampleC、およびControlが、それぞれ560万リード、580万リード、650万リードおよび870万リードでした。リード数の遷移を確認したところ、サンプルの状態によってはゲノムにマッピングされるリードの割合が低くなる場合がみられますが、いずれのサンプルにおいても有効リードのうち8割以上が既知のmiRNAとして検出されています(次ページ表2)。

[isomiRsの検出]

isomiRsはmiRBaseに登録されているリファレンス配列から数塩基5′または3′末端側にずれているmiRNA(miRNA variants)であり、生体内でDroshaもしくはDicerがmiRNA前駆体を異なる箇所で切断することによって生じます[3]。通常、miRBaseに登録されているリファレンスmiRNA配列が最も高く発現していますが、特定の条件下では特定のisomiRsが代表配列になることが報告されています[4]。本解析においては、mmu-miR-7aで、すべてのサンプルにおいてリファレンス配列(青)より3′末端側に1塩基長い配列(赤)が最も多く検出されました(次ページ図3)。

表2　GAIIによる解析リード数の遷移

分類	SampleA		SampleB		SampleC		Control	
総リード数	5,661,483	100.00 %	5,867,095	100.00 %	6,597,873	100.00 %	8,722,703	100.00 %
フィルタリングで除去されたリード数	970,083	17.13 %	1,479,495	25.22 %	2,497,329	37.85 %	2,797,314	32.07 %
フィルタリング通過リード数	4,691,400	82.87 %	4,387,600	74.78 %	4,100,544	62.15 %	5,925,389	67.93 %
ゲノムにマップされたリード数	4,004,393	70.73 %	3,529,935	60.16 %	2,397,277	36.33 %	4,431,681	50.81 %
解析対象リード	4,004,393	100.00 %	3,529,935	100.00 %	2,397,277	100.00 %	4,431,681	100.00 %
マウスの既知 Hairpin miRNA	**3,911,253**	**97.67 %**	**3,132,133**	**88.73 %**	**1,992,502**	**83.12 %**	**3,919,322**	**88.44 %**
他生物種の既知 Hairpin miRNA	1,765	0.04 %	1,196	0.03 %	3,121	0.13 %	1,550	0.03 %
piRNA	17,125	0.43 %	69,104	1.96 %	93,283	3.89 %	103,748	2.34 %
rRNA	3,397	0.08 %	14,027	0.40 %	15,490	0.65 %	12,915	0.29 %
tRNA	888	0.02 %	4,128	0.12 %	2,506	0.10 %	3,147	0.07 %
snRNA	626	0.02 %	7,589	0.21 %	30,961	1.29 %	22,128	0.50 %
snoRNA	14,251	0.36 %	107,748	3.05 %	65,445	2.73 %	101,700	2.29 %
miscRNA	1,833	0.05 %	10,372	0.29 %	12,545	0.52 %	20,260	0.46 %
RefSeq transcripts	14,450	0.36 %	61,641	1.75 %	99,096	4.13 %	93,766	2.12 %
ゲノムにのみ一致したリード数	38,805	0.97 %	121,997	3.46 %	82,328	3.43 %	153,145	3.46 %

	Sample			
mmu-miR-7a	A	B	C	Control
TTTGGAAGACTAGTGATTTTGTTG	0	0	1	0
AATGGAAGACTAGTGATTTT	0	1	4	6
TTGGAAGACTAGTGATTTTG	0	0	1	0
GTGGAAGACTAGTGATTTTGTT	0	1	0	0
ATGGAAGACTAGTGATTTT	0	0	0	2
TGGAAGACTAGTGATTTTTT	0	1	8	18
TGGAAGACTAGTGATTTTT	0	2	3	6
TGGAAGACTAGTGATTTTGTTGTTTTT	0	0	1	0
TGGAAGACTAGTGATTTTGTTGTTTT	0	1	24	2
TGGAAGACTAGTGATTTTGTTGTTGT	0	4	17	1
TGGAAGACTAGTGATTTTGTTGTTG	0	7	34	9
TGGAAGACTAGTGATTTTG	0	20	218	53
TGGAAGACGAGAGATTTTGTTG	0	0	1	0
TGGAAGACCAGTGATTTTG	0	3	3	3
TGGAAAACTAGTGATTTTG	0	0	2	2
TGGAAGACTAGTGATTTTGTTGTTT	1	47	274	61
TGGAAGACTAGTGATTTTGTTGTT	142	1,657	6,255	1,950
TGGAAGACTAGTGATTTTGTTGT	45	1,441	4,863	1,650
TGGAAGACTAGTGATTTTGTTG	3	309	668	590
TGGAAGACTAGTGATTTTGTT	4	63	160	186
TGGAAGACTAGTGATTTTGT	1	26	305	97
GGAAGACTAGTGATTTTGTTGTTT	0	1	2	1
GGAAGACTAGTGATTTTGTTG	0	1	6	5
GGAAGACTAGTGATTTTGTT	0	1	2	1
GGAAGACTAGTGATTTTGT	0	0	3	0
GGAAGACTAGTGATTTTGTTGTT	1	4	18	10
GGAAGACTAGTGATTTTGTTGT	1	8	26	4
GAAGACTAGTGATTTTGTTGTTTT	0	0	1	0
GAAGACTAGTGATTTTGTTGTTT	0	0	0	2
GAAGACTAGTGATTTTGTTGTTG	0	2	0	0
GAAGACTAGTGATTTTGTTGTT	0	1	4	7
GAAGACTAGTGATTTTGTTGT	0	2	3	2
ACTAGTGATTTTGTTGTTTT	0	0	1	1

.UGUGGAAGACUAGUGAUUUUGUUGUUUUUAG.

図3　isomiRsの例

mmu-miR-7aでは青色で示したmiRBase Reference配列よりも、赤色で示した3′末端側に1塩基長い配列が最も多く検出されている。

2. RNA-editingの検出

RNA-editing[5]はpri-miRNAsにおいて特定の塩基がアデノシンからイノシンへと変換される現象であり、ターゲットmRNAの翻訳抑制調節に関わっていると考えられています[6]。データ解析上、miRNA-editingの検出はミスマッチを許容したゲノムアライメントを行うことで解析結果に含めることが可能ですが、検出されるmiRNAが18～22塩基という短い配列であることを考えると、ゲノム位置の正確性が幾分低下する可能性があります。Controlサンプルの解析結果について1塩基ミスマッチを許容して解析をした結果、マウスにおいて既知のmiRNA-editingの例として報告されているmmu-miR-376c[5]と相同なリード配列の中に、本解析においても同様のeditingを受けたと考えられるリードを確認することができました(図4)。

mRname	sequences	Control
mmu-miR-376c	AACATAGAGGAAATTTCACGT	388
mmu-miR-376c	AACATAGAGGAAATTTCACGTT	23
mmu-miR-376c	AACATAGAGGAAATTTCACG	22
mmu-miR-376c	AACATGGAGGAAATTTCACGT	13
mmu-miR-376c	AACATAGAGGAAATTTCACGTA	10
mmu-miR-376c	ACATAGAGGAAATTTCACGTTT	7
mmu-miR-376c	AACATAGAGGAAATTTCACGG	6
mmu-miR-376c	AACATAGAGAAAATTTCACGT	4
mmu-miR-376c	ACATAGAGGAAATTTCACGTT	3
mmu-miR-376c	AACATAGACGAAATTTCACGT	3
mmu-miR-376c	AACATAGAGGAAATTTCA	3
mmu-miR-376c	AACATAGAGGAAATTTCACGTG	3
mmu-miR-376c	AACATGGAGGAAATTTCACG	3
mmu-miR-376c	ACATAGAGGAAATTTCACGTA	2
mmu-miR-376c	AACATAGAGGAAATTTAACGT	2
mmu-miR-376c	AACATAGAGGAAATTTCACGTTT	2
mmu-miR-376c	CATAGAGGAAATTTCACGTTT	1
mmu-miR-376c	CATAGAGGAAATTTCACGTTTT	1
mmu-miR-376c	ACAAAGAGGAAATTTCACGT	1
mmu-miR-376c	ACATAGAGGAAATTTCACGT	1
mmu-miR-376c	ACATAGAGGAAATTTCACGTTA	1
mmu-miR-376c	AACACAGAGGAAATTTCACGT	1
mmu-miR-376c	AACAGAGAGGAAATTTCACGT	1
mmu-miR-376c	AACATAAAGGAAATTTCACG	1
mmu-miR-376c	AACATAGAGGAAATTTCACGA	1
mmu-miR-376c	AACATAGAGGAAATTTCACGCT	1
mmu-miR-376c	AACATAGAGGAAATTTCACGTTA	1
mmu-miR-376c	AACATAGAGGAAATTTCACT	1
mmu-miR-376c	AACATAGAGGAAATTTTACGTTT	1
mmu-miR-376c	AACATAGGGGAAATTTCACGT	1
mmu-miR-376c	AACATATAGGAAATTTCACGTT	1
mmu-miR-376c	AGCATAGAGGAAATTTCACGT	1
mmu-miR-376c	TACATAGAGGAAATTTCACGT	1
mmu-miR-376c	CAACATAGAGGAAATTTCACGT	1
mmu-miR-376c	GGAACATAGAGGAAATTTCA	1
mmu-miR-376c	GTAACATAGAGGAAATTTCA	1

図4　miRNA-editingの例

マウスにおいて既知のmiRNA-editingとして報告されているmmu-miR-376cに1塩基ミスマッチでアラインメントされるリードを示す。青色で示したmiRBase Reference配列の6塩基目のAがIにeditingされG(赤色)として検出されている。緑色で示したAは3′nucleotide additionの可能性が示唆されるものである。黄色で示した塩基はゲノムと不一致であった塩基である。

3. miRNA-Seq解析の定量性評価

同一サンプルでmiRNAマイクロアレイ解析(アジレント社)とmiRNA-Seqを行い、共通して検出されたmiRNAに対してSampleAとControlの発現比率を計算し、両実験間の比較を行いました(図5)。

ピアソン相関係数で0.86と高い相関を示しており、miRNA-Seqでもマイクロアレイ解析の結果と矛盾しないデータを

得ることができています。なお、miRNA-Seqのデータについては、得られたmiRNAのリード数に対して、各サンプルのマッピングできた総リード数をそろえるようにglobal normalizationを行っています。

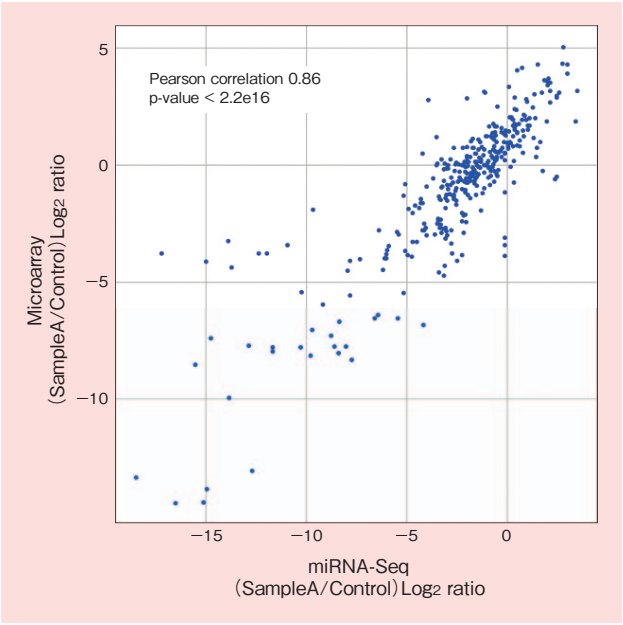

図5　miRNA-Seqマイクロアレイ解析データとの比較

同一サンプルでmiRNAアレイ解析とmiRNA-Seqを行い、共通して検出されたmiRNAに対してSampleAとControlとの対数発現比をとった散布図

■ まとめ

現在、転写産物の大規模発現解析法としてはマイクロアレイを用いた手法が一般的ですが、「既知の転写産物しか検出できない」「クロスハイブリダイゼーションによる高いバックグラウンドノイズ」「シグナルダイナミックレンジの制限」など欠点がみられます。高速シーケンスを用いた解析はハイブリダイゼーションを行わないためこれらの欠点とは無縁であり、さらに、mRNA sequencingには「転写産物の配列を決定できる」「新規転写産物を特定できる」「新規スプライシングジャンクションを特定できる」などの利点があります。また、miRNA sequencingでは前2点に加え、「isomiRsの検出ができる」「RNA-editingとadditionの検出ができる」などの利点が挙げられます。加えて本稿で紹介した解析例のように、mRNA sequencing、miRNA sequencing解析から得られた発現データはそれぞれマイクロアレイとの相関も高く、定量性にも優れています。このように高速シーケンサーを利用したWhole Transcriptome解析は、複合的なデータを1回の実験で得ることができるというアドバンテージを持つ優れた解析手法であると言えるでしょう。

GAIIで解析できるリード長は、現在GAIIxの導入により75塩基に達しており、今後のイルミナ社のロードマップでは、2009年末までには150塩基（ペアエンド法では150塩基×2）まで伸長し、一回の解析で得られるデータ量は95 Gb以上に達する予定になっています。また、Applied Biosystems SOLiD™ 3（アプライドバイオシステムズ社）でもWhole Transcriptome Analysis Kitが発売され、注目を集めています。これらによりデータあたりのコストは確実に下がりますが、今まで以上に大量のデータが得られることになり、データ処理がボトルネックになる可能性が懸念されています。タカラバイオでは情報処理サービスの充実に努め、これまで多くの実績を積み重ねながら最新の技術に対応してまいりました。今後も、加速する高速シーケンスの技術革新にいち早く対応し、専門スタッフが皆様のニーズに合わせた柔軟なサポートを提供いたします。

[謝辞]

miRNA-Seq解析で用いたサンプルとそのアレイ解析データは、京都大学iPS細胞研究センター 山中伸弥先生の研究室からご提供いただきました。

【参考文献】

1) Mortazavi A, *et al.* : Mapping and quantifying mammalian transcriptomes by RNA-seq. (2008) *Nature Methods* **5** (7), 585-7.
2) Trapnell C, *et al.* : TopHat: discovering splice junctions with RNA-Seq. (2009) *Bioinformatics* **25** (9), 1105-11.
3) Morin RD, *et al.* : Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. (2008) *Genome Res.* **18** (4), 610-21.
4) Kuchenbauer F, *et al.* : In-depth characterization of the microRNA transcriptome in a leukemia progression model. (2008) *Genome Res.* **18** (11), 1787-97.
5) Kawahara Y, *et al.* : Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. (2007) *Science* **315** (5815), 1137-40.
6) Reid JG, *et al.* : Mouse let-7 miRNA populations exhibit RNA editing that is constrained in the 5'-seed/ cleavage/anchor regions and stabilize predicted mmu-let-7a:mRNA duplexes. (2008) *Genome Res.* **18** (10), 1571-81.
7) Landgraf P, *et al.* : A mammalian microRNA expression atlas based on small RNA library sequencing. (2007) *Cell* **129** (7), 1401-14.