第1回 「RNA-Seqによる網羅的トランスクリプトーム解析」

成相直樹（東北大学 東北メディカル・メガバンク機構 ゲノム解析部門）

はじめに

近年、次世代シークエンサ（NGS）によるRNA-Seq技術により、細胞の遺伝子発現を網羅的かつ一塩基レベルの高解像度で解析することが可能となった。これまで一般的に行われてきているマイクロアレイ技術を利用した遺伝子発現解析においては、Agilent社、Affymetrix社等が提供する解析ソフトウェアを使用したり、RのBioconductorパッケージ等を使用したりすることが比較的簡単に出来る様になっている。しかしながらRNA-Seqデータ解析については、新しいシークエンス技術の登場と共に様々な解析手法が日々提案されている。また、ツールごとに必要なマシンの環境、使用方法、出力結果の解釈等が異なることもあり、特にウェットの研究者にとってRNA-Seqデータ解析の敷居が高いと感じてしまうことがあるかもしれない。今回の解説記事ではRNA-Seqデータ解析をなるべく身近に感じてもらえるように、RNA-Seqデータ解析を行う前の事前準備から、リードのアライメント（マッピング）方法、アライメント結果からの遺伝子発現レベルの正規化、定量まで、最も基礎的かつ重要なステップを解説したい。また、具体的にいくつかの既存ツールのアルゴリズム、長所・短所を概観する（**表1**）。本稿の目的は、RNA-Seqデータ解析が実は非常にシンプルであり、かつ応用範囲が広いことを認識していただくことである。

RNA-Seqの特徴・利点

RNA-Seqは細胞から抽出したmRNAをcDNAに逆転写してPCR増幅した後、数百bp程度に断片化し、次世代シークエンサで各フラグメントの片側、もしくは両側から塩基を読む。通常、各フラグメントの両側100-250bp程度を読み取る（PacBio RSII などでの一分子長鎖配列解析については別の機会に触れる）。転写量推定の最初のステップとして、読んだ配列（リード）をリファレンス配列（ゲノム配列もしくはcDNA配列）にアライメントする。次に、そのアライメント結果に基づき、リードがどのリファレンス配列から出ているかを判断し、それぞれの転写産物の遺伝子発現量を推定する。その際、真の遺伝子発現量はシークエンサで読めたcDNA断片の数に比例する、という仮定に基づく。マイクロアレイによる発現量の定量方法が蛍光色素の検出を基にしていることに対して、NGSによるRNA-Seqではリード数を基に遺伝子発現量を定量することから、リード数を十分読めば、マイクロアレイに比べてダイナミックレンジが広いことが報告されている（Wang et al., 2009）。また、cDNAプローブがデザイン済みであるマイクロアレイ解析とは異なり、転写産物の配列を直接シークエンシングできること、そしてサンプルのゲノム多型（cDNA配列から間接的に、ではあるが）、及びアリル特異的な遺伝子発現が検出できることもRNA-Seqの大きな利点である。本稿では特に、遺伝子発現量の定量手法について解説する。

表1 各ツールの比較

ツール	長所	短　所	ウェブサイト
Cufflinks	新規転写産物を推定できる	モデル生物の解析のみ対応	http://cufflinks.cbcb.umd.edu/
RSEM	高速。非モデル生物の解析も可能	新規転写産物を推定できない	http://deweylab.biostat.wisc.edu/rsem/
TIGAR	高精度な遺伝子発現量推定。非モデル生物の解析も可能	新規転写産物を推定できない	https://github.com/nariai/tigar/

RNA-Seqにおける遺伝子発現量の正規化

RNA-Seqデータから転写産物ごとの遺伝子発現量を定量するためには、リファレンスcDNA配列にアライメントした総リード数、及び転写産物（cDNA）の長さで正規化した値、FPKM（Fragments Per Kilobase of exon per million fragments mapped）が良く使われる（Trapnell et al., 2012）。例えば、ある転写産物にアライメントされたリード数が3,000リード、cDNA全長が5,000 bp、リード総数が4,000万リードだとする。そうするとこの転写産物のFPKMは

$$FPKM = 3{,}000 / (5{,}000 / 1{,}000) / (40{,}000{,}000 / 1{,}000{,}000) = 15.0$$

と計算される。以上の定義から明らかである様に、FPKMの正確な算出のためには、そもそもリードが正しく転写産物（cDNA）のリファレンス配列にアライメント出来ていることが前提である。しかしながら通常、リードの長さは100-250 bpであり、シークエンス精度は向上しつつあるものの0.1% -1.0%程度の置換・挿入・欠失エラーを含むため、必ずしもリードが正しい転写産物のリファレンス配列にアライメントされるとは限らない。更に、選択的スプライシングにより生成される同じ遺伝子由来の転写産物は当然ながら塩基配列がお互い非常に良く似ており、あるリードがどの転写産物から生成されたかを正しく判断することは必ずしも容易ではない。また、ある転写産物についてパラログ、偽遺伝子の関係にある遺伝子から得られる産物とはリファレンス配列が似ているため、リードをリファレンス配列にアライメントする際には上記と同様、マルチマップの問題が発生する。このような曖昧なアライメント候補が複数存在する場合、一つの解決方法として、マルチマップリードを無視することにして、ユニークにアライメントされたリードのみを基にして遺伝子発現レベルを定量する、という方法が考えられる。しかしながらこの方法では本来ならば考慮に入れるべきリードを捨ててしまっているため定量性が落ちることが知られており（Mortazavi et al., 2008）、更にはトリプレットリピート等のlow complexity sequenceを含むような転写産物の定量が難しくなると予想される。従って現在、標準的に使用されているRNA-Seqデータ解析手法の多くはマルチマップリードを捨てるのではなく、有効利用するアルゴリズムを実装している。以降、各ツールの特徴について概説する。

既存手法（TopHat/Cufflinks）

最もポピュラーに使用されているRNA-Seqデータ解析手法の一つに、TopHat/Cufflinksが挙げられる（Trapnell et al., 2009, Trapnell et al., 2012）。TopHatはリードをリファレンスゲノムにアライメントするツールであり、Cufflinksはそのアライメント結果から各転写産物の発現量を推定するツールである。当然、リードはリファレンスゲノムのexon領域にしかアライメント出来ないため、アライメントの際にはexon/intron 領域のアノテーション情報が必要となるが、その際に使用するgtfファイルなどもまとめてパッケージとして提供されている（http://ccb.jhu.edu/software/tophat/）。

第一のステップとして、TopHatではリードをそれぞれ25 bp程度に断片化した後、マルチマップを許した上でリファレンスゲノムにアライメントする（その際、bowtieというアライメントツールを使用する）。これによりリード断片がリファレンスゲノムにおけるexon領域にマッピングされたり、されなかったりする。マッピングされなかったリード断片の前後にexon領域にマップしたリード断片があれば、そのunmappedリードはexonとexonのジャンクションに由来するリードであると推定される。このアルゴリズムにより、新規の転写産物を推定することも可能となっている。

次のステップとして、CufflinksはTopHatのアライメント結果を入力として、転写産物ごとのFPKMを計算する。リードが複数の転写産物にマルチマップしている場合、デフォルトオプションではマルチマップしている転写産物に等分にリードを分配する（たとえば、1本のリードが2つの転写産物にアライメントしている場合、それぞれの転写産物から0.5本ずつ出力された、と考える）。最新のバージョンでは“-u”オプションを付けることにより、周辺のユニークマップしたリードの数に基づいて、マルチマップのリードを比例配分する“rescue method”（Mortazavi et al., 2008）を実行することが出来る。しかしながらTopHatにおいてはリード

を25 bpの長さに断片化してリファレンスゲノムにアライメントするアルゴリズムを採用しているため、アライメント自体の偽陽性を含むことは避けられず、結果として定量性が損なわれる、あるいは間違って新規の選択的スプライシングを推定してしまう、といった可能性を常に留意しておく必要がある（この点については、後述する）。

TopHat/Cufflinksの長所として、Cuffdiffなどいくつかの下流解析を行うツールがパッケージとして提供されているため、とりあえず一連のRNA-Seq解析を行うことが出来ることが挙げられる。また、多くの研究者が既に使用しているため、問題が発生したときに助けてくれる仲間が多いことが期待できる。逆に欠点として、TopHat/Cufflinksはモデル生物（ヒト、マウス、ラット等）の解析をメインターゲットにしているため、非モデル生物のトランスクリプトーム解析には向いていない。また、デフォルトオプションでは新規の転写産物を予測してくれるものの、この予測精度については現時点では十分に検証されているとは言い難い（短いリード断片から新規の選択的スプライシングを予測することが難しいことは容易に想像できる）。新規の転写産物を予測したくない場合は、TopHat/Cufflinksのオプションで“-G”オプションを付けることにより、既知の転写産物のみの発現レベル推定を行うことも出来る。

統計的モデリングによるマルチマップリードの解決と転写量推定

RNA-Seqデータ解析において、リファレンス配列（ゲノム配列、もしくはcDNA 配列）にアライメントされたリード数に基づいて転写量推定を行うこと、従ってリードが複数箇所のリファレンス配列にアライメントする場合（マルチマップ）の扱い方が問題になることを説明した。前述の通り、TopHat/Cufflinksにおいては“-u”オプションを付けることにより、rescue methodと呼ばれるアルゴリズム（周辺領域のユニークマップリードの情報に基づく重み付け）によって、ある程度マルチマップの問題を解決している。しかしながらこのアルゴリズムでは、マルチマップの問題を完全に解決しているとは言い難い。

近年、マルチマップリードのアライメントを隠れ変数として扱い、転写量をパラメータとして推定する統計的な手法が提案されている（Li et al., 2010, Nariai et al., 2013）。説明のため、シンプルな例を挙げて説明する。今、isoform A, B, Cという転写産物が発現しているとする。それぞれのリードがどの転写産物から生成されたものであるか、リファレンスcDNA配列にリードをアライメントすることにより同定する。まず簡単のため、アライメントは一意に決まるとする。それぞれの転写産物について、アライメントされたリード数に基づいて、転写量（ここではリードの割合）を推定することができる（**図1**）。アライメントが一意に決まらない場合（マルチマップ）、統計的枠組みの下、アライメントを隠れ変数として推定することが出来る（ここでは詳細な説明は省く）。この様な統計的手法の利点として、シークエンスデータ（リード）自体のエラーをモデリングできること、そしてデータの生成モデルを仮定した上で統計的に尤もらしいアライメントと転写量を同時に推定できることが挙げられる。このような統計的手法を適用することで、シークエンシングエラーやアライメントエラーなどに強い、ロバストな転写量推定が可能となる。パラメータ推定はEMアルゴリズム（もしくは変分EMアルゴリズム）により、隠れ変数の推定とパラメータのアップデートを交互に繰り返すことで、解析的に（局所）最適解が得られる。以下、著者らが開発した統計的手法であるTIGAR（Transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference）の利点について解説する。

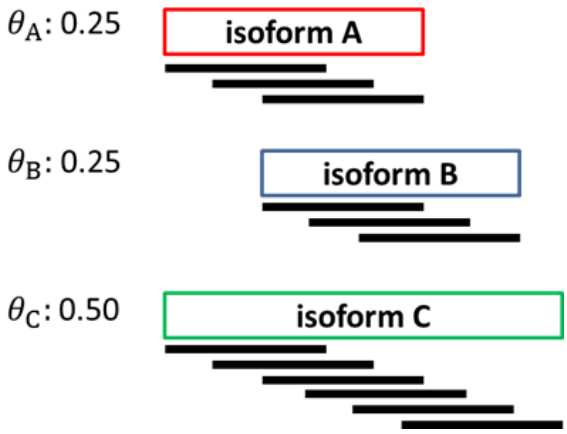

図1 リードをリファレンス cDNA 配列にアライメントすることにより、それぞれの転写量を推定することが出来る

RSEM（Li et al., 2010）においては、リードデータは挿入・欠失を含まないことを前提としてモデリング

されている。この仮定は、HiSeqやMiSeqなど、Illumina社のシークエンサ機器から得られる100bp程度のショートリードに関しては挿入・欠失エラーが比較的少ないため、大きな問題とはならないかもしれない。しかしながらLife Technologies社のIonPGM、Pacific Biosciences社のRS IIなどの比較的、挿入・欠失エラーを多く含むシークエンスデータに対しては、必ずしも適切では無い。筆者らが開発したTIGARにおいては、挿入・欠失エラー、すなわちリファレンスcDNA配列に対するギャップ付きアライメントの状態をPair HMMという確率モデルを利用することで表現し、データの尤度を計算する。この様にギャップ付きアライメントを受け付けるモデルを採用することにより、bowtie2、bwa-mem、novoalignなどの高感度なアライメントツールが利用できる（これに対して、RSEMはギャップ付きアライメントを行わないbowtieを使用する）。

さて、上述の通り、TIGARではリードのアライメントの際、ギャップ付きアライメントを行うことで感度を向上することが出来る。しかしながら、同時にアライメントの偽陽性についても注意を払わなければならないことは前項のTopHatのアルゴリズム解説の中でも述べた。RSEMにおいては、EMアルゴリズムによって隠れ変数（アライメント）の推定とパラメータ（各転写産物の転写量）の最尤推定を行う。これに対して、TIGARは、変分ベイズ推定によってパラメータを事後分布として推定する。ベイズ推定ではパラメータを点推定ではなく事後分布として推定するため、ノイズに強いロバストな推定を行うことが出来る。また、事前分布のハイパーパラメータの設定により、モデルの複雑さ（この場合は推定されるべき転写産物の数）をコントロールできるという特徴がある。

ここで、なぜ“推定されるべき転写産物の数”を気にする必要があるかについて述べる。極端な例を挙げると、ある遺伝子座位について、exonの数をnとすれば、選択的スプライシングにより生成可能な転写産物の組み合わせの数は理論的には2^n種類が可能であり、これを全ての遺伝子座位について考えると途方も無い転写産物候補を仮想的には考えることが可能である。また、融合遺伝子の可能性についても考えると、更に膨大な候補数となる。しかしながら、通常、ある細胞のある時点においては、これら膨大な候補のうち、せいぜい数千〜数万程度が実際には遺伝子発現していると考えるのが自然である。TIGARは、ベイズ推定の枠組みの下、より少ないパラメータ数（転写産物の候補の数）でデータを説明するモデルを選択するという“オッカムの剃刀”の精神に従うようなパラメータ推定を行うことが出来る（**図2**）。より詳細なアルゴリズムについては（Nariai et al., 2013）を参照していただきたい。

TIGARの使用方法

TIGARを使用する一連の解析の流れをまとめてパイプラインと呼ぶことにする。パイプラインの構成要素

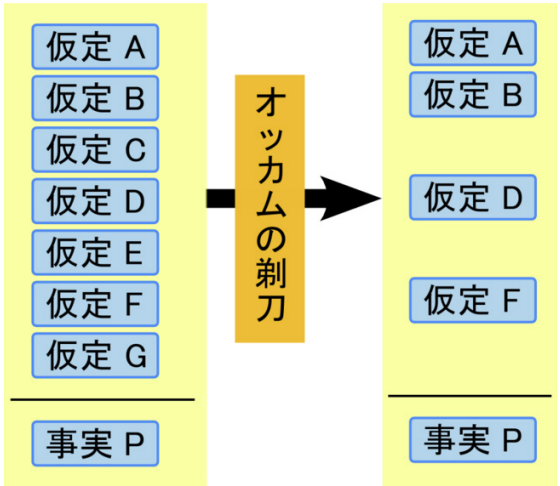

図2 オッカムの剃刀の概念図。データをより少ない仮定で説明できれば、それに越したことは無い、とする精神（http://ja.wikipedia.org/wiki/オッカムの剃刀）

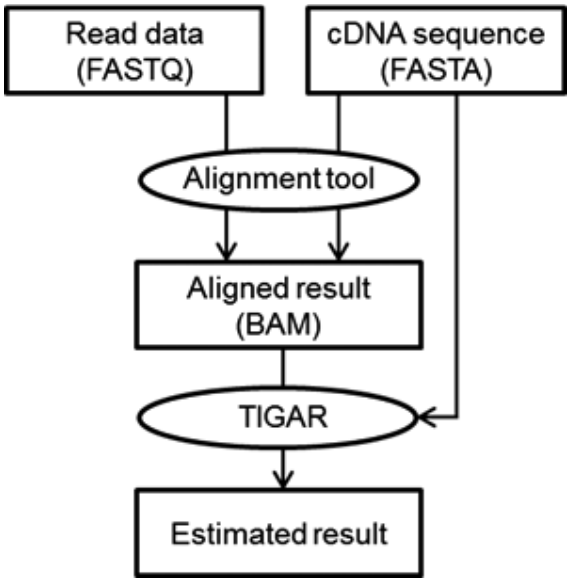

図3 TIGARのパイプライン

は、1)RNA-Seqリードのリファレンスc DNA 配列へのアライメント、2) アライメント結果からの転写量推定、の2つに大きく分けられる（**図3**参照）。

ここでリファレンス配列はcDNA配列であり、例えばマウスのcDNA配列であれば以下の様にUCSCのウェブサイトからダウンロードできる：

```
wget http://hgdownload.soe.ucsc.edu/goldenPath/mm9/bigZips/refMrna.fa.gz
```

リードデータ（FASTQ）を上記で得られたリファレンスcDNA配列にアライメントする際には、ギャップ付きアライメントを行うことができる bowtie2、あるいはbwa-memなどを推奨している。いずれのツールを使う場合でも、最初に一度だけ、リードの高速なアライメントのためリファレンス配列のFMインデックスを作成する必要がある（FMインデックスの詳細はLi and Durbin, 2009を参照）：

```
mkdir ref
bowtie2-build index refMrna.fa ./ref/refMrna
```

FMインデックスを作成後、マルチマップを許したアライメントを行う：

```
シングルエンドデータの場合：
bowtie2 -k 100 --very-sensitive ./ref/refMrna sample.fastq > sample.sam

ペアエンドデータの場合：
bowtie2 -k 100 --very-sensitive ./ref/refMrna -1 sample_1.fastq -2 sample_2.fastq
> sample.sam
```

最後に、TIGAR本体のプログラムを実行する：

```
シングルエンドデータの場合：
java -jar Tigar.jar refMrna.fa sample.sam --alpha_zero 0.1 sample_out.txt

ペアエンドデータの場合：
java -jar Tigar.jar refMrna.fa sample.sam --is_paired 1 --alpha_zero 0.1 sample_out.txt
```

上記の様に、cDNAリファレンス配列さえ準備できれば、TIGARは非常にシンプルなパイプラインで実行することが出来る。これはすなわち、非モデル生物のRNA-Seqデータ解析も同様のパイプラインで実行可能であることを意味する。尚、プログラム、実行マニュアルはプロジェクトのウェブサイトから入手できる（https://github.com/nariai/tigar）。

TIGAR と既存ツールのパフォーマンス比較

ここではシミュレーションデータを使用することで、TIGARと既存ツール（TopHat/Cufflinks、RSEM）とのパフォーマンス比較を行う。RefSeqのデータベースからヒトのcDNA配列をランダムに10,000個選び、それぞれの遺伝子発現量が対数正規分布に従っていると仮定して“正解の”FPKMを設定する。ここでは入力データとして、100 bp x 2（ペアエンド）の人工的なRNA-Seqデータを100万リード生成した。シークエンサのエラーとして、置換・欠失・挿入エラーを1%導入する。それぞれのソフトウェアで転写量推定を行って、真の転写量との相関を見た（**図4**）。転写量が大きい転写産物（たとえばFPKM＞50）はどのツールでも上手く推定できるが、転写量が小さい転写産物は転写量の定量がRSEM、TopHat/Cufflinksでは難しくなることが分かる。この理由として、既存ツールがリードのアライメントを間違えている、あるいはマルチマップリードを適正に配分できていないため、結果として正確な転写量を推定することが出来ていないことが考えられ

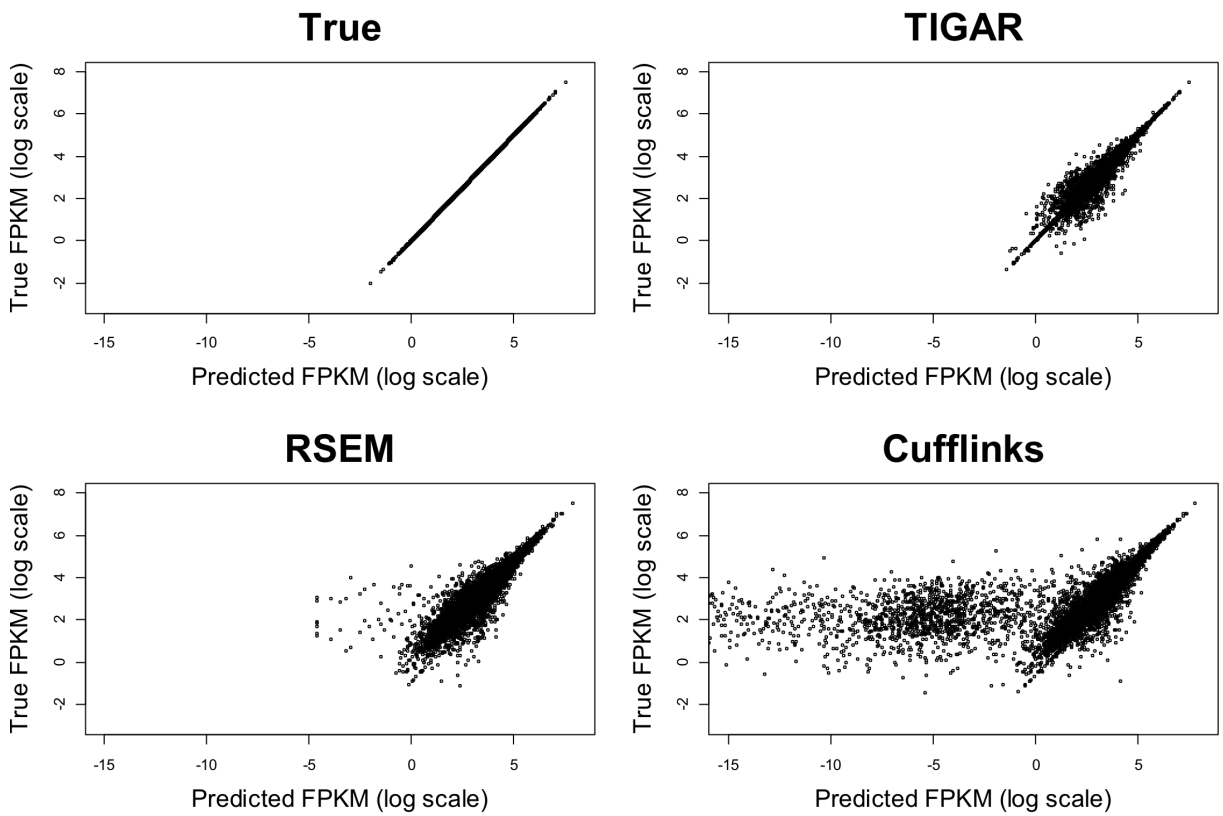

図3 各ツールによる転写量の推定結果と、真の転写量との比較

る。一方、TIGARでは実際の転写量と良く相関した転写量推定が出来ていることが分かる。今後、single-cell sequencingデータへの適用等が考えられる。

まとめ

本稿ではRNA-Seqデータから転写量推定を行う一般的な流れと、既存ツールのアルゴリズム紹介、筆者らが開発した統計的手法TIGARの具体的な使用方法、及びシミュレーションデータによる実行結果の例を紹介した。RNA-Seqデータを解析する際の基本的な考え方、注意点などを理解していただくことが出来たのなら、幸いである。スペースの関係上、新規の選択的スプライシングの検出方法については詳細な解説を行わなかったが、興味のある読者は参考文献（Mezlini et al., 2013）等を参照していただきたい。また、通常、転写量推定を行った後、発現差解析、パスウェイ解析などの下流解析に進むのが一般的である。これらのトピックについても今回は省略する。興味ある読者にとって東京大学・門田幸二先生のウェブサイト（http://www.iu.a.u-tokyo.ac.jp/~kadota/r_seq.html）が参考になると思われる。

参考文献

[1] Wang et al., Nat Rev Genet. (2009) 10:57-63
[2] Mortazavi et al., Nat Methods. (2008) 5(7):621-628.
[3] Trapnell et al., Bioinformatics. (2009) 25(9):1105-1111.
[4] Trapnell et al., Nat Protoc. (2012) 7(3):562-578.
[5] Li et al., Bioinformatics. (2010) 26(4):493-500.
[6] Nariai et al., Bioinformatics. (2013) 29(18):2292-2299.
[7] Li and Durbin, Bioinformatics. (2009) 25(14): 1754-60.
[8] Mezlini et al., Genome Res. (2013) 23(3):519-29