植物科学におけるトランスクリプトーム解析の最前線

市橋 泰範, 福島 敦史
理化学研究所 環境資源科学研究センター
〒230-0045 神奈川県横浜市鶴見区末広町 1-7-22

Yasunori Ichihashi, Atsushi Fukushima
Frontiers of Transcriptomics in Plant Science

Key words: Gene co-expression, Library preparation, Network analysis, RNA-seq, Transcriptome

RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, 230-0045 Japan

1. はじめに

ゲノム配列の解読をきっかけに、多量のデータを体系的に扱う手法であるオミクス解析が始まった（Fukushima and Kusano 2014）。その中でもトランスクリプトーム解析は、ゲノムから最終的な表現型へと情報を橋渡しする転写産物全体を明らかにすることにより、生命現象の理解に大きく貢献する（Ichihashi et al. 2015; Ichihashi and Sinha 2014）。次世代シーケンサーの技術進歩により誕生した RNA-seq は、トランスクリプトーム解析の精度を大幅に引き上げるとともに、原理的にどんな生物種においてもトランスクリプトーム解析を適用可能とした（Wang et al. 2009）。従来のハイプリダイゼーションによるマイクロアレイやサンガー法に基づくシーケンス技術と比べて、RNA-seq は遺伝子発現の検出範囲が広いためトランスクリプトームの複雑性を正確に検出できる（Mader et al. 2011）。さらに新規の転写産物、small RNAs, alternative splicing variants を検出でき、転写産物の塩基配列を直接読むことにより SNPs, fusion transcript, 転写開始点も同定することが可能となる（Ozsolak and Milos 2011）。このように RNA-seq によるトランスクリプトーム解析は、生物学分野全般において知識発見を加速している。

次世代シーケンサーのプラットホームは常に改良され、シーケンスデータの量・質とともに大幅に向上している。そのため、今後さらに大きなスケールで実験が可能となることは想像に難くない。しかしながら、RNA-seq ライブラリー作成はいまだ労力、時間、コストがかかり大規模なプロジェクトへの展開の制限要因となっている。加えて、たとえ大規模のデータを手にいれることができたとしても、多量でかつ複雑なデータの解析手法についてもスタンダードな方法がまだないため、多くのケースで生物学的意味を引き出すことができずデータが埋没する恐れがある。そこで本総説ではこれらの問題を克服できる最新のライブラリー作成技術および大規模データ解析手法を紹介し、植物科学における新しい研究展開について考察する。

Y. Ichihashi & A. Fukushima-1
2. RNA-seq ライブラリー作成技術

ここでは次世代シーケンサーの中でよく使われている Illumina 社のプラットホーム用のライブラリー作成について論じる。

2-1. 今までのライブラリー作成技術

現在までの方法として Illumina 社が 2012 年当時に発売したハイスループット用の TruSeq RNA sample preparation kit をもとに RNA-seq ライブラリー作成の基本的なステップについて説明する（図 1A）。まず生物組織から total RNA を抽出し、mRNA を精製する。得られた mRNA について二価陽イオンを用いて断片化する。断片化した mRNA をもとに 2 本鎖 cDNA 合成を行う。その後、エンドリペアーゼにより 5'末端がリン酸化された平滑末端の cDNA 断片を得る。A テーリングにより 3'末端に A を付加し、TA クローニングを利用してアダプターを cDNA 断片の両端に接続させる。上記の酵素反応ごとに、solid-phase reversible immobilization (SRR1) 磁気性ビーズを使い精製し、併せて cDNA 断片のサイズをある程度で選択していく。得られたアダプター付き cDNA 断片を PCR 増幅して、RNA-seq ライブラリーが完成する。この方法は、生物組織からおそらく 3 日間で 24 ライブラリー作成が可能である。このように RNA-seq ライブラリー作成は、当世の技術でハイスループットを念頭に開発された方法であっても、労力・時間・コストがかかっていた。

2-2. High-throughput RNA-seq (HTR) ライブラリー作成技術

そこで著者らはよりハイスループットな RNA-seq を実現するため、簡便・迅速・安価にできる方法を開発した（HTR, Kumar et al. 2012）。この HTR では、Illumina 社の方法をもとに以下の点について大きく改良した：1）組織から直接 mRNA を抽出する，2）mRNA でなく cDNA を断片化する，3）cDNA 断片の末端修飾反応の全てを SRR1 ビーズ上で行う（Fisher et al. 2011），4）96 パーコードをアダプターに組み込む（図 1B）。これらの改良により、生物組織から 2 日間で 96 ライブラリー作成が可能となり、従来の方法に比べて 6 倍のハイスループット化が実現した。さらにコストも当時の市場価格において 3-11 倍削減することに成功した。加えて、ライブラリーの質も Illumina 社の方法とほぼ同等であり、むしろリポソーム RNA のコンタミネーションが少ない良質のものであった。HTR と Illumina 社のプロトコールを使って同じサンプルセットについて遺伝子発現の差異を解析すると、大半の遺伝子群については両プロトコールによって検出されたが、プロトコール間で異なる挙動を示す遺伝子群も多く見つかった。そのため、異なるプロトコールで得たサンプルを一緒に解析する際はプロトコール間の違いに十分注意すべきである。また今回の解析結果からは、HTR のほうが技術的反復間での誤差が少なかったため、統計的検出力が高いデータを出すこともわかった。

この方法は幅広い植物種に応用でき、現在までに多くの草本植物に加えて、水分を多く含む藻類（Caulerpa taxifolia）やリグニンが蓄積した木本の植物（Gevuina avellana）でも高い精
度のデータを取得している（Ranjan et al. 2015; Ostria-Gallardo et al., unpublished）。またこの方法を用いることで大規模な RNA-seq を行うことが可能となった。例えば、著者らはトマトと近縁種間の Introgression line 76 系統において RNA-seq 解析（生物学的反復を加えて、およそ 300 サンプル）を行い、introgession の境界を一塩基レベルで特定した（Chitwood et al. 2013a）。同時にこのデータを使って、遺伝子発現を量的形質と扱い、それに影響するゲノム領域を特定する expression quantitative trait locus（eQTL）解析を行うことで、遺伝子発現制御をゲノムレベルで明らかにすることができた（Ranjan et al., unpublished）.

2-3. Breath Adapter Directional sequencing (BrADseq) ライブラリー作成技術

著者らは近年、さらに簡便、高速、超低価格を実現した方法を開発した（BrADseq, Townsley et al. 2015）。この BrADseq では、上記の HTR にあるステップを改善しただけでなく、アダプター付加を 2 本鎖 cDNA 合成とともに行う方法を独自に開発し、プロトコールに取り入れた（図 1C）。具体的には 2 本鎖核酸の末端が化学的特性によって開閉する現象を利用して（von Hippel et al. 2013），5'末端特異的にアダプターを組み込むことに成功した。この BrADseq は、組織からたった 6 時間で 96 ライブラリー作成が可能となり、コストも HTR よりもさらに 7 倍、従来の方法よりも 21-77 倍削減できた。加えて strand-specificity のあるライブラリーであるため、情報量がより充実したデータを取得することができる。HTR と BrADseq のプロトコールで遺伝子発現の差異を比較すると、非常に高い相関が得られたため、HTR と同様に BrADseq も統計的検出力が高いデータを出すことができる。

一方で上記の方法以外にも数多くの RNA-seq ライブラリー作成方法が開発されてきている。例えば、dUTP を使った strand-specific ライブラリーや多くのサンプルをプールして 1 反応でライブラリー作成を行う方法があり、これらもまた独自にコストダウンやハイスルーポット化が実現されている（Shishkin et al. 2015; Wang et al. 2011）。そのため RNA-seq による網羅的遺伝子発現解析は、個々の遺伝子を対象とした定量的 PCR のような発現解析に取って代わり、より身近な技術になると著者らは予測している。

Y. Ichihashi & A. Fukushima-3
図1 RNA-seq ライブラリー作成方法と要する日数の比較

（A）Illumina 社のTruSeq RNA sample preparation kit, （B）HTR, （C）BrADseq ライブラリー作成のワークフロー。近年の技術開発により、通常3日で24サンプルのところが、6時間で96サンプルというハイスループット化が実現された。

3. データ解析

ここではRNA-seqのデータ解析のうち、上流の配列解析についてではなく、大規模なデータセットにより統計的な問題が生じやすい、発現量を算出し正規化した後のデータ解析手法について論じる（図2A）。

3-1. 単変量・多変量解析

前述したライブラリー作成技術の革新により、従来よりもはるかに大規模な実験が可能となる一方で、データの内容が複雑化してくることが容易に想像できる。典型的オミクスデータの解析アプローチとしては、2群間での比較が通常よく行われている。全発現遺伝子について2群間の発現量の平均の差を検出し、t検定（正規分布からデータが得られていることが前提）などの仮説検定をする有意水準で行うことにより、対象とする2群間で発現様式が異なる一団の遺伝子群（differentially expressed genes, DEGs）を同定できる（図2B）。DEGs解析にはリード数が負の二項分布に従うことを仮定した手法であるedgeR（Robinson et al. 2010）やDEseq（Anders et al. 2013）などの計算ソフトRで利用できるパッケージが現在よく

Y. Ichihashi & A. Fukushima
利用されている。さらに下流解析として、これら DEGs について Gene ontology（GO）解析や Gene set enrichment（GSE）解析を行うことで、特定の機能や代謝経路に関与する遺伝子群が DEGs 内に統計的に有意に多く存在しているか検定することができ（Hung et al. 2012）、2 群間での遺伝子発現の違いを特徴づけることができる。

時系列データなどの複雑なデータについては多変量解析を行うことで見通しがよくなる（図 2C）。例えば、発現パターンを抽出する目的で、k-means clustering や hierarchical clustering といったクラスタリング解析がよく使われている（Andreopoulos et al. 2009）。また発現パターンの全体的な特徴を把握する目的で、主成分分析や多次元尺度構成法がよく使われる。これらは多変量データの変数に重みをつけて少数の合成変数を作ることにより、多次元データをより低次元にすることでデータの解釈を助ける。さらに大きなスケールのデータでは、複数の要因が入れ子状になった多変量データになるケースが考えられる。最近、そういった多次元データを解析する手法として後述する ΔPC を用いた superSOM クラスタリング解析が有効であることが示された（Chitwood et al. 2013b）。SOM は、自己組織化マップ（Self-organizing maps）という人工神経回路ネットワークの一種で、教師なし学習によって入力データを任意の次元へ写像することができる（Wehrens and Buydens 2007）。特に superSOM は、変数ごとに重み付けができるため、多次元データのクラスタリング解析に用いることができる。そこで多次元データ（例えば、異なる処理条件における時系列の遺伝子発現を比較する）から要因間の相互作用（例えば、処理 × 時間の相互作用）を抽出するため、まず主成分分析により全データを低次元空間（PC space）上にマップして、異なる要因の関で変化する遺伝子を PC space 上での変化量（ΔPC）として PC 軸ごとに計算する。全 ΔPC について、変数ごとに重み付けができるクラスタリング方法である superSOM で解析することにより、異なる要因間での相互作用（例えば、異なる処理条件下で時系列の遺伝子発現パターンが変化する特徴）をシステムレベルで記述し理解することができる。このような解析によって、従来のペアワイズ DEGs 解析やクラスタリング解析では検出できない遺伝子発現のダイナミクスが明らかになる。例えば Chitwood et al. (2013b) では、トマトと近縁種の異なる組織における RNA-seq データを使って、ΔPC を用いた superSOM クラスタリング解析を応用することにより、種間での組織別の遺伝子発現パターンの変化の特徴を明らかにした。その中で、トマトの遺伝子は近縁種の相同遺伝子に比べて、メタシステムを含む組織で高い発現、分化した組織では低い発現を示す傾向にあり、これは種間のメタシステムの大きさの違いを反映していた。

3-2. ネットワーク解析

生物種が示す様々な現象の多くは、数千から数万の異なる遺伝子群の複雑な制御ネットワークの振る舞いに依存する。このような複雑なシステムを理解するために、遺伝子制御ネットワークの推定とそのモジュール（サブネットワーク）の効率的な同定は、現代のシステム生物学における中心的課題の一つである。サンプル数が増加したトランサクリプトームデータ

Y. Ichihashi & A. Fukushima*5

BSJ-Review 7:114 (2016)
6.9-BD2>A857CF1
そこで発現差異（DEGs）と遺伝子共発現とを組み合わせた自然な拡張として，共発現差異を考えることができる（図 2E）。これでは，対照群および実験群の 2 群間で異なる共発現関係を同定する，すなわち異なるネットワーク間で変化した相互作用をシステムレベルで明らかにできる（Fukushima 2013; Kayano et al. 2014）。例えば著者らは，トマトの葉と果実から得られたマイクロアレイデータを用いた大規模な共発現差異ネットワークにより，代謝経路におけるキーステップを明らかにできた（Fukushima et al. 2012）。


3-4. オミクスデータ解析における注意事項

上記のデータ解析結果に基づき遺伝子機能実験の計画をする際，注意すべき点を挙げる：

1. 散布図をよく見る — 遺伝子発現パターン間の関係性構造を把握する意味でわめて重要である。例えば，外れ値があった場合のピアソン相関係数にさしたる情報はない。

2. 相関関係は因果関係ではない — よく勘違いされるケースであるが，発現パターンが似ていても，必ずしも同じ機能とは限らず，偶然似たパターンを示した場合もありうる。ましてや，直接的な因果関係を必ずしも示すわけでは無いので，データの解釈には十分な注意が必要である。

3. データベースや解析手法のポリシーを理解する — 利用するデータベースや解析手法がどのようなものであるか，どのようなデータから，どのようなアプローチで，遺伝子発現パターン間の類似性を定量しているか，その背後にある構想や注意点を十分に理解して実験計画を立てることが重要である。

Y. Ichihashi & A. Fukushima
図2: RNA-seqの遺伝子発現量データからの解析方法

(A) シーケンシングから上流解析の配列解析。リファレンスにマップされたリードをカウントすることにより遺伝子発現量を算出する（通常、遺伝子数 p >> サンプル数 n）。(B-E) 発現量を算出した後のデータの解析方法。

(B) DEGs解析。ある遺伝子 A の発現量に差があるか t 検定などで調べる。

(C) 多変量解析のクラスタリング解析と主成分分析。遺伝子発現のパターンを抽出する。

(D) ネットワーク解析。複数の発現データから遺伝子を点、遺伝子間関連性を線で表すネットワークを構築できる。

(E) 共発現差異解析。2つの条件間で各々測定された遺伝子群間の共発現パターンは、2条件間で異なる場合がある。これら共発現差異遺伝子には、酵素系遺伝子や転写調節因子などが含まれる可能性が高い。
### 表1 種間比較が可能な共発現遺伝子データベース

<table>
<thead>
<tr>
<th>名称</th>
<th>種属</th>
<th>データソース</th>
<th>主な類似性尺度</th>
<th>相関遺伝子の決定方法</th>
<th>引用文献</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTED-II</td>
<td>Arabidopsis, field mustard, soybean, medick, poplar, tomato, grape, rice, maize</td>
<td>microarray and RNA-seq</td>
<td>weighted PCC and mutual rank</td>
<td>Reciprocal Best BLAST Hit</td>
<td>Aoki et al. PCP, 2015</td>
<td><a href="http://atted.jp/">url</a></td>
</tr>
<tr>
<td>CoP</td>
<td>Arabidopsis, barley, poplar, rice, soybean, wheat, grape, maize</td>
<td>microarray</td>
<td>cosine correlation</td>
<td>OrthoMCL</td>
<td>Ohyasui et al. PCP, 2015</td>
<td><a href="http://bioinf.mind.meiji.ac.jp/podc/">url</a></td>
</tr>
<tr>
<td>PODC</td>
<td>Arabidopsis, rice, Sorghum, tomato, grape, medicago, potato, soybean</td>
<td>RNA-seq</td>
<td>correspondence analysis and Pearson's correlation coefficient</td>
<td>OrthoMCL</td>
<td>Ohyasui et al. PCP, 2015</td>
<td><a href="http://bioinf.mind.meiji.ac.jp/podc/">url</a></td>
</tr>
<tr>
<td>Maize-rice</td>
<td>maize, rice</td>
<td>microarray</td>
<td>PCC</td>
<td>OrthoMCL</td>
<td>Neutea et al. BMC Genomics, 2014</td>
<td><a href="http://complex.plantgenie.org/">url</a></td>
</tr>
<tr>
<td>ComPEX</td>
<td>Arabidopsis, poplar, rice</td>
<td>microarray</td>
<td>CLR</td>
<td>OrthoMCL</td>
<td>Neutea et al. BMC Genomics, 2014</td>
<td><a href="http://complex.plantgenie.org/">url</a></td>
</tr>
<tr>
<td>PLANEX</td>
<td>Arabidopsis, soybean, barley, rice, tomato, wheat, grape, maize</td>
<td>microarray</td>
<td>PCC</td>
<td>orthom.iivv.cnr.it</td>
<td>Miozzi et al. Plant Mol Biol, 2010</td>
<td><a href="http://ortom.iivv.cnr.it/">url</a></td>
</tr>
<tr>
<td>ORTom</td>
<td>tomato, potato, tobacco, pepper</td>
<td>EST profile/absence profile conservation</td>
<td>binary asymmetric distance</td>
<td>the Eukaryotic Gene Orthologies database</td>
<td>Moroz et al. Plant Mol Biol, 2012</td>
<td><a href="http://ortom.iivv.cnr.it/">url</a></td>
</tr>
</tbody>
</table>

* PCC, Pearson's correlation coefficient; CLR, Context Likelihood of Relatedness
4. おわりに
今後の植物科学においてトランスクリプトーム解析はどのように利用されていくだろうか？本総説で示したように，RNA-seqが高度にハイルーブット化することで，より大規模なプロジェクトが可能となる。これにより統計的に高い精度のデータ取得はもちろん，フィールドを対象とする生態学への展開や，一人の研究者で一つの植物種のトランスクリプトームマップを作成できるなど，新規遺伝子の発見が加速するだろう。加えて，より詳細に組織別の発現解析を行うために組織別プロモーターやレーザーマイクロダイセクションを用いた発現解析や，近年の流体力学の進展により可能となった1細胞のトランスクリプトーム解析も身近な技術となるだろう（Picelli et al. 2014）。また第三世代シークエンサーである1分子シークエンサーはより長い配列を読むことができるため，転写産物全長を一度にシークエンスできる（Tilgner et al. 2014）。この技術がトランスクリプトーム解析に利用されることになれば，より正しく発現量やalternative splicing variantsを定量することができる。

本総説では詳細しなかったが，シークエンシング後の配列データ解析ではより高速で正確なリファレンスへのマッピングや遺伝子発現量推定手法の洗練が求められている。加えて，トランスクリプトームデータのみならず，幅広くメチローム・プロテオーム・メタボローム・フェノームといった様々なオミクデータを統合し，いかに新規の生物学的意義を引き出すかが計算生物学・バイオインフォマティクス分野の中心課題となっている（Cavill et al. 2015）。特にこれらデータを用いた統合的なネットワーク解析は，重要形質に関与する鍵遺伝子の予測やフィールドでの表現型予測を行う効率的なモデルをもたらす可能性をもち，遺伝型—表現型関連性の解明に寄与する。このような技術進歩によりトランスクリプトームを含めたオミク解析の対象がマクロに至るまで拡大し，我々が見ることができる世界が広がる。今後，その新しい世界を見ることによって，いかに世界観すなわち新しい生物学的知識を理解できるかが次世代の研究者に求められる課題であろう。

5. 謝辞
本稿で紹介した著者らの研究の一部は，理化学研究所・基礎科学特別研究員制度および科学研究補助金・若手研究 B（15K18589）（市橋 泰範），及び科学研究補助金・若手研究 B（26850024）（福島 敦史）の支援を得て遂行した。また本研究を進めるにあたり，数々のサポートを頂いた，カリフォルニア大学デービス校・Prof. Neelima Sinha, Dr. Ravi Kumar, Dr. Brad Townsley, Dr. Jie Peng, ドナルドダンフォース研究所・Dr. Daniel Chitwoodに，この場を借りてお礼申し上げます。

6. 引用文献


Y. Ichihashi & A. Fukushima·11

BSJ-Review 7:120 (2016)


Y. Ichihashi & A. Fukushima•14