Vol.2012-BIO-28 No.1
2012/3/28

異質倍数体の遺伝子発現量の定量化

山　田　　恵[†1]　清水（稲継）理恵[†2]
清水　健太郎[†2]　瀬々　　潤[†3]

本研究では，異なる種のゲノムを併せ持つ種である異質倍数体の遺伝子発現量をRNA-seq により定量化した．しかし，異質倍数体の親となる種は近縁の種で構成されることが多いため，通常の RNA-seq では親種同士のホモログ遺伝子の配列が非常に近く，いずれの遺伝子が発現したかを見分けることが困難である．本研究では，近縁種からの変異のパターンに着目し，最尤推定法を用いることで，親種を分けた発現量の定量化を可能とした．

Quantified expression levels of genes in allopolyploid species

Megumi Yamada,[†1] Rie Shimizu(Inatsugu),[†2]
Kentaro Shimizu[†2] and Jun Sese [†3]

In this research, We tackled a measurement of the expression levels of genes in alloploidies. For the measurement, we used an RNA-seq. However, high sequence similarity between genes originating from different parental species prevented us to apply the method to the problem directly. To overcome the difficulty, our method used genome of model species, which is closely related to the target alloploid, and aligned RNA-seq sequences to the genome. Based on the mutation distributions of the alignment result, our method determined the originated species of each read using maximum likelihood estimation method. By counting the number of the sequences of each parental species on each gene, our method successfully quantified expression levels of genes in allopolyploid species.

†1 お茶の水女子大学 (Ochanomizu University)
†2 チューリッヒ大学 (University of Zurich)
†3 東京工業大学 (Tokyo Institute of Technology)

1. はじめに

異質倍数体は異なる種のゲノムを併せ持つ種で，ゲノム重複に似た状態を有するため，種分化や環境適応のメカニズム発見に重要と考えられる．これらのゲノムが環境に応じて同調して働くか，個々に働くかの調査に向けて，本研究では，由来するゲノムを区別して遺伝子の発現を同定する手法を開発した．異質倍数体は，植物で多く観測されるため，本論では，*A. thaliana*（シロイヌナズナ）の近縁で起こった異質倍数体の解析を考える．

1.1 ホメオログ

種 A, B を親に持つ異質倍数体 C が存在するとき，A, B がそれぞれ遺伝子 a, b を持つとすると，C は，a, b に由来する c_a, c_b を持つ（図 1）．a, b をホモログとすると，c_a, c_b はホメオログという．A, B は近縁の種である場合が多く，ホメオログは非常に配列相同性が高い．

1.2 ホメオログ遺伝子の発現量定量化

本研究では，遺伝子発現量の定量に RNA-seq を用いる．RNA-seq は mRNA の配列断片を大量に読み（配列をリードと呼ぶ），既知の遺伝子配列やゲノム配列と対応したリードの本数を数えることで遺伝子の発現量を計測する．RNA-seq をホメオログ遺伝子発現量の定量に用いることを考えよう．c_a, c_b 由来のリードは，a, b の配列両方に対応するため，由来する遺伝子を特定できず，発現量の定量を誤る可能性がある．本手法では，ホメオログ遺伝子間の変異を利用して各リードがいずれの親由来であるか分離する．

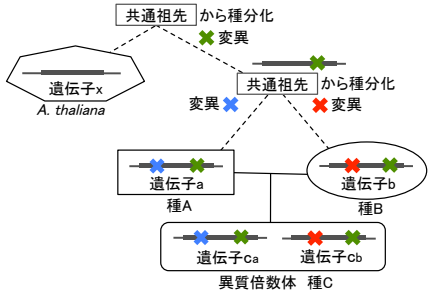

図 1　系統樹，遺伝子の持つ変異の関係
Fig. 1　Relationships between Mutations and Phylogenetic Tree

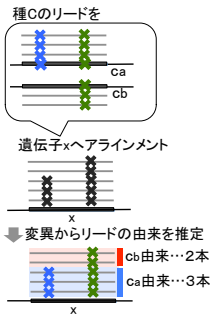

図 2　手法の概要
Fig. 2　Outline of Proposed Method

© 2012 Information Processing Society of Japan

Vol.2012-BIO-28 No.1
2012/3/28

2. 手　　法

RNA-seq では，リードをアラインメントするゲノムが必要だが，異質倍数体と親種のゲノムは未知のため，近縁で，ゲノムや遺伝子配列が既知の *A. thaliana* を利用する．

2.1 変異の種類と定義

a, b を *A. thaliana* の遺伝子 x のオーソログとする．系統樹を考えると，x と c_a, c_b の配列間の差異は，親種特有の変異（図 1 の c_a, c_b 上の×印，赤青），祖先種から親種へ進化する過程で起きた変異（緑）（以下，親由来の変異，祖先由来の変異）の 2 種類に分けられる．c_a, c_b 由来のリードを x にアラインメントした模式図が図 2 である．

ある一つの遺伝子上に変異が k 箇所あるとき，変異を有する位置の集合を $\mathbf{M} = \{M_1, M_2, ..., M_k\}$ とする．M_i の位置を含むリードの集合を R_i，要素数を $|R_i|$ と表す．変異の有無で分けたリードの集合を $R1_i, R2_i$ とする．この時，$|R1_i| \leq |R2_i|$ とする．

2.2 由来別の遺伝子発現量の算出

図 2 のように，アラインメントした時点では変異の種類は分からず，変異が，親由来か祖先由来かを見分ける方法が必要になるが，後述する．計算により判定した親由来の変異に注目すると，リードは変異の有無から 2 本と 3 本に分けられ，リードごとにいずれの親由来か（青，赤）が分かる．更にリードの本数から，リードの由来別のホメオログ遺伝子の発現量が求められる．ここでは特定の変異に着目して親種の比率を求めたが 1 点のみの推定ではシークエンスエラーによる影響も多く確度が低い．本研究では，リード毎にいずれの親由来であるかを決定し，その本数を遺伝子全体にわたって数える事で計算の精度を向上させる．

2.3 変異の種類を見分ける方法

変異 M_i に着目した時，この変異は親由来か祖先由来のどちらかである．RNA-seq はランダムに配列を選択する手法であり，M_i のリードの本数が $R1_i, R2_i$ である確率 $P(M_i)$ を，ポアソン分布を用いてモデル化する．λ_p を変異が親由来である時の，変異のあるリードの本数の期待値，同様に λ_a を祖先由来の変異である時の期待値とすると，M_i が親由来である確率は，$P(M_i) = \lambda_p^{|R1_i|} e^{-\lambda_p} / |R1_i|!$ ，祖先由来である確率は，$P(M_i) = \lambda_a^{|R1_i|} e^{-\lambda_a} / |R1_i|!$ と表せる．t を祖先由来の時の変異率とすると，$\lambda_a = t \times |R_i|$ となる．t はあらかじめ固定した値である．λ_p は遺伝子ごとに異なる値であり，データから求める．

それぞれの変異が独立な事象とすると変異 $\mathbf{M}$ の尤度 $P(\mathbf{M})$ は $P(M_1) \times P(M_2) \times \cdots \times P(M_k)$ で表せるので，この尤度が最大になるように，すべての変異を親由来か祖先由来かに分ければ良い．この時，親由来と考えられる変異セットを $\mathbf{M_P} (\subseteq \mathbf{M})$ とする．

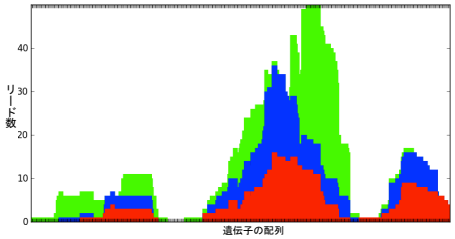

図 3　親種を区別したリード本数の分布
Fig. 3　Distribution of Number of Reads Classified by Originated Parental Species

$\mathbf{M_P}$ を現実的な時間で求めるため，以下の工夫を行った．
（1）$M_i \in \mathbf{M_P}$ に対して λ_p は $\mathbf{M_P}$ に属する変異の平均に近いと考えられるので，$I = \{i \mid M_i \in \mathbf{M_P}\}$ とすると，$\lambda_p = |R_i| \times (\sum_{i \in I} |R1_i| / \sum_{i \in I} |R_i|)$ とする．
（2）$|R1_i|/|R2_i|$ が大きいもの程，λ_p を大きくするので，この比率が 1 に近いものから順に $\mathbf{M_P}$ に入れて，尤度が最大となる組み合わせを計算した．

3. 実データを用いた実験の結果

異質倍数体の *C. flexuosa* に対し Life Technologies 社の SOLiD5 を用いて RNA-seq を行ったリード約 6 千万本に対し，本手法を適用した．アラインメント先に *A. thaliana* (TAIR10)[1] のゲノムを，アラインメントソフトには SHRiMP[2] を用いた．図 3 は，一つの遺伝子に対し本手法を適用した結果を示している．横軸に遺伝子上の位置，縦軸に塩基ごとのリードの本数を表す．色分けは，リードごとに判定した親種に対応し，赤と青が片親，緑は由来が不明のリードである．この遺伝子の両親種ごとの遺伝子発現量は，それぞれ 5.93, 6.56 だった（RPKM[3] 換算）．グラフからは，この遺伝子の場合，親種ごとの発現量を求めるのに使うことの出来たリードがほぼ遺伝子の全体を網羅し，発現の比が遺伝子上を一貫して一定であったことが分かる．

参　考　文　献

1) Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40** (D1): D1202-D1210., 2012.
2) Rumble, S.M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLOS Comput. Biol.* **5**, e1000386, 2009.
3) Mortazavi, A. *et al.* Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature Methods*, **5**(7):621–628, 2008.

© 2012 Information Processing Society of Japan