病原細菌の種内多数のゲノムデータから
組換えのホット領域を推定する手法の開発

久留米大学	矢原　耕史
Imperial College London	Xavier Didelot
University of Oxford	M. Azim Ansari
Swansea University	Samuel Sheppard
Max Planck Institute	Daniel Falush

1. はじめに

突然変異と組換えは、ゲノムに変異を生み出し、遺伝的な個体差・多様性を生み出す源である。突然変異率はゲノム内の特定の領域で高いことが知られ、その疾患との関係が注目されている（Michaelson, 2012, *Cell*）。一方、組換えは、突然変異よりも検出が難しく、その回数をゲノムに沿って推定すること自体が未解決の難問である。今世紀になって、種内多数の全ゲノム配列を比較するというアプローチが可能になり、それによってゲノム全域について、組換え率と「ホット領域」（組換えが繰り返し生じた結果、個体間を高頻度で移動したように見える領域）を推定することが、重要な課題になっている。

2. 内容

本研究では、病原細菌のゲノムに沿って、1 塩基レベルでその推定を可能にする、新たな手法を開発した。この手法の土台となっているのは、N 個体のゲノム全域の 1 塩基多型とそのポジションを入力データとし、ある個体（レシピエント）のゲノムを、残り N-1 個体（ドナー）の DNA 断片の組み合わせ・モザイクとして再構築する、近年開発された隠れマルコフモデル「染色体ペインティング法」である（Lawson 2012, *PLoS Genetics*）。これを病原細菌に応用すると、組換えの痕跡（ある個体の DNA 断片が別個体のゲノムに入り込んだモザイク構造）をゲノム全域に渡って明らかにできる（Yahara 2013, *Mol. Biol. Evol*）。ただし、このモデルでは最近 1 回の組換えしか検出できず、過去の履歴を考慮しておらず、組換えの生じた回数や組換えのホット領域を推定できない。また、病原細菌がクローン増殖する点を考慮していない。本研究では、これらの問題点を解決するアルゴリズムと、各塩基におけるドナーからレシピエントへのコピー確率の行列表現に基づいて塩基あたりの組換え率を推定する、新たな統計量を開発した。また、組換えのホット領域推定に関する bootstrap support value を開発した。さらに、入力データの中に欠損値が含まれる場合への対処法も開発した。

この手法を、まずシミュレーションデータに適用し、その感度・特異度を評価した。次に、大腸菌の 27 本の完全ゲノムデータに適用し、既知の組換えのホット領域が推定できることを確認した。さらに、人獣共通感染性細菌種の 200 本の不完全ゲノムデータに適用し、新たな組換えのホット領域を明らかにした。そして、並列計算によって高速な計算を可能とし、メモリ使用率を低く抑え、シンプルに使用可能なソフトウェア（https://github.com/bioprojects/orderedPainting）として実装し、一般公開した（Yahara, 2014, *Mol. Biol. Evol*）。