



Tokyo Tech



2019年4月22日

報道機関各位

東京工業大学
情報・システム研究機構 国立遺伝学研究所

両親由来のゲノム配列を個別に決定する新手法

ーゲノム多様化領域に起因した生命現象の解明へー

【要点】

- 両親由来のゲノム配列を高精度にかつ個別に決定する情報解析手法
- 哺乳類、無脊椎動物、植物などを対象にしたテストで性能を確認
- 従来は解析が困難だった両親間のゲノムが多様化した領域を解析

【概要】

東京工業大学 生命理工学院 生命理工学系の梶谷嶺助教、吉村大大学院生（博士後期課程3年・研究当時）、奥野未来研究員、伊藤武彦教授らの研究チームは、国立遺伝学研究所の豊田敦特任教授、小原雄治特任教授、東京大学の窪川かおる特任教授らと共同で、真核生物のゲノム配列決定において、両親由来の配列を区別し、高精度にそれぞれを決定する、新しい情報解析手法の開発に成功した。

ヒトなど真核生物のゲノム情報は、両親から受け継いだ情報を持ち合わせているが、今までは両親由来ゲノムの差異を無視して配列決定を行うことが一般的だった。しかしながら、この差異の大きな領域は、種々の昆虫の表現型（紋様）との関連や、ヒトでの免疫型の決定、さらには疾患との関連も報告されるようになっている。そのため、簡便に両親由来の配列を区別して解析できる手法が求められていた。

研究チームが開発した「Platanus-allee（プラタナス アリー）」と呼ばれる新しいプログラムは、特殊な装置や前処理を必要とせず、現在の主流になっている次世代シーケンサー（注1）の大規模な断片配列データのみから、両親由来の配列を高精度に再構築できる画期的なものだ。

本成果は、2019年4月12日付けの「*Nature Communications*」に掲載された。

●背景

2001年にヒトの全ゲノム配列が決定されてから、わずか20年足らずの間に、次世代シーケンサーの登場と解読システムの性能向上により、ゲノム配列を決定するコストは数万分の1になり、読み取り時間も劇的に短くなっている。

ヒトに代表される一般的な高等動植物のゲノムは、母親と父親の両親から受け継いだ両方の情報を持ち合わせている。しかしながら現状では、この差異を無視して、両親由来のゲノム情報全体をモザイク状につなぎ合わせることで、ゲノム配列を解読する手法が一般的に用いられている。

近年、両親由来のゲノム配列（相同染色体）間で差異が大きい領域が存在し、これが様々な表現型（例えば、昆虫の体の紋様や性決定、ヒトの免疫型決定や疾患など）とリンクしている事例が報告されている。そのため、両親由来のゲノム配列を“分けて”解析することの重要性が認識されるようになってきたが、その実現は技術・コストの面から多くの問題が存在していた。

●研究の内容

本研究で開発された解析手法は、**ショートリード**（注2）と呼ばれる次世代シーケンサーが産出するデータの精度の高さを活かして、まず、大量の断片配列内に存在する一塩基の違いをも区別できる**グラフ構造**（注3）に変換する（図a）。次に、そのグラフ構造をショートリード間のペア情報を用いて単純化することで、より長く繋がったゲノム配列を再構築する（図b）。相同染色体との対応付けを配列の類似情報から行い（図c）、エラー修正などを行った上で最終的な配列を導き出す。

研究チームでは、この手法を「Platanus-alley」というプログラムに実装し、ホームページで公開した（<http://platanus.bio.titech.ac.jp/platanus2>）。さらにこの手法を実際に、線虫、シロオビアゲハ、ナメクジウオ、サクラ、ヒトなどの各種生物に適用したところ、その実効性を証明できた。

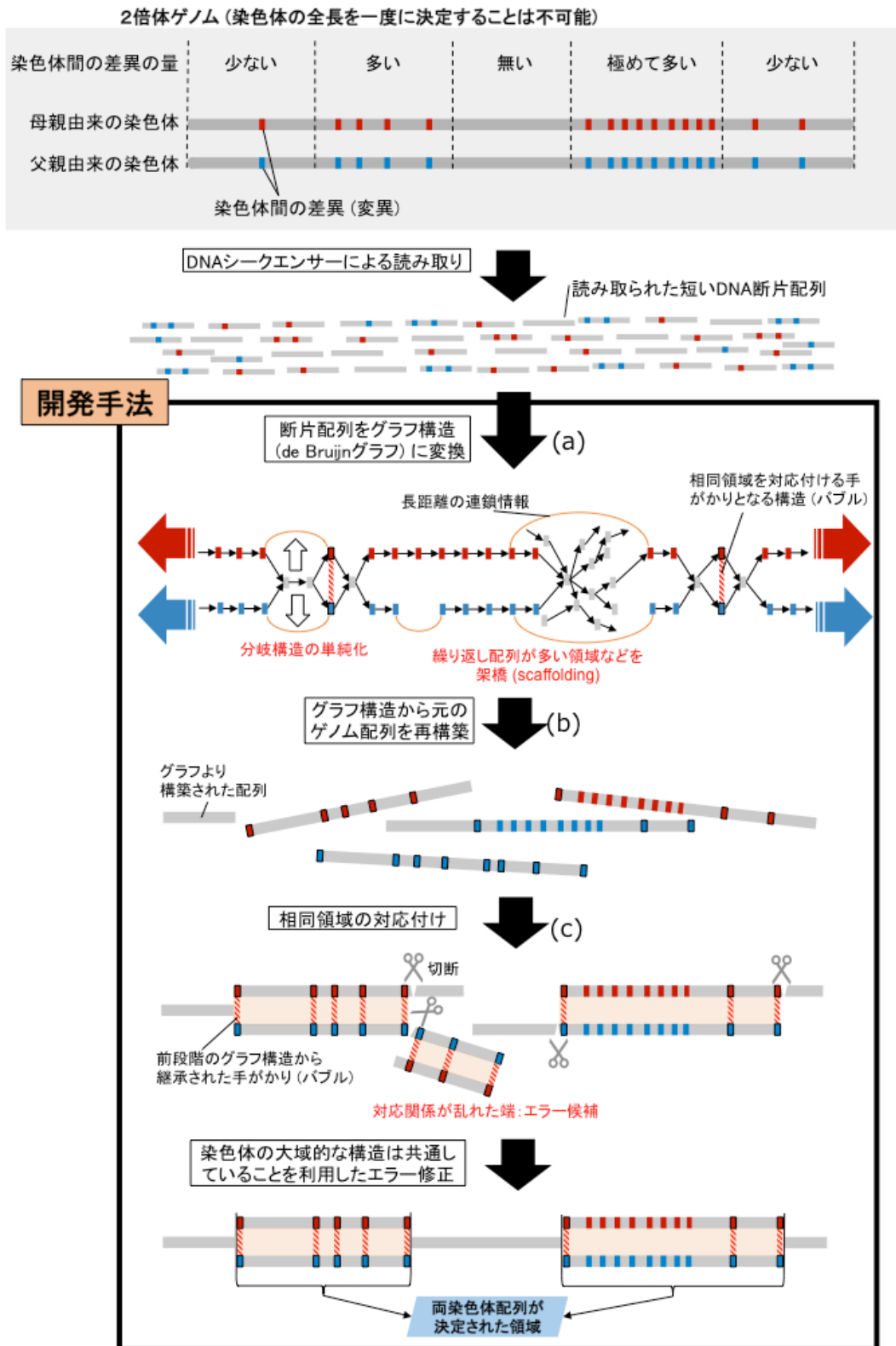


図: 新たな情報解析プログラム「Platanus-allee」のアルゴリズムの模式図

●今後の展開

研究開発された解析手法で、相同染色体間の複雑な変異情報が網羅的に収集可能となる。これにより変異が蓄積したゲノム領域との関連が疑われている種分化、多様性維持、免疫など重要な生命現象の解明が進むと考えられる。また、究極的には、本研究成果を用いて、我々ヒトを含む“2倍体生物が2セットの少し異なるゲノムを維持することで何を得たのか？”という根源的な問いに対する理解が深化すると期待される。

【用語説明】

(注1) 次世代シーケンサー：

2004年頃から登場した新しいタイプの塩基配列解読装置（シーケンサー）。最大の特徴は、それまでの機器（第一世代）と比べて圧倒的に産出するデータ量が多い。その後も次々と新しいタイプのシーケンサーが登場している。

(注2) ショートリード：

次世代シーケンサーの中で主流な機種が出力するタイプのデータのこと。具体的には、ある程度の長さのゲノム断片の両端が100~250文字程度ずつの断片ゲノム配列（ショートリード）としてペアで読まれる。最新の次世代シーケンサーからは、ショートリードが一度に数億本（ペア）のレベルで産出される。

(注3) グラフ構造：

ノード（節）とエッジ（辺）で表現されるデータの集合体。ゲノムデータでは、あるゲノム部分配列（ノード）が他のゲノム部分配列（ノード）に繋がっているか（エッジ）の関係を記載したデータ集合体。

本研究は、文部科学省科研費「新学術領域研究『学術研究支援基盤形成』」先進ゲノム解析研究推進プラットフォーム(16H06279)および、16H04719, 15H0597などの支援を受けて行われました。

【論文情報】

掲載誌：*Nature Communications*

論文タイトル：Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions

著者：Rei Kajitani, Dai Yoshimura, Miki Okuno, Yohei Minakuchi, Hiroshi

Kagoshima, Asao Fujiyama, Kaoru Kubokawa, Yuji Kohara, Atsushi Toyoda,
and Takehiko Itoh

DOI : 10.1038/s41467-019-09575-2

【問い合わせ先】

東京工業大学 生命理工学院 生命理工学系 教授
伊藤 武彦(いとう たけひこ)

【取材申し込み先】

東京工業大学 広報・社会連携本部 広報・地域連携部門

情報・システム研究機構 国立遺伝学研究所
リサーチ・アドミニストレーター室 報道担当