東京大学
THE UNIVERSITY OF TOKYO

PRESS RELEASE

報道解禁（日本時間）：9 月 14 日 18 時（15 日朝刊）

配付先：大学記者会（東京大学）　文部科学記者会　科学記者会　柏記者クラブ

2023 年 9 月 12 日
東京大学

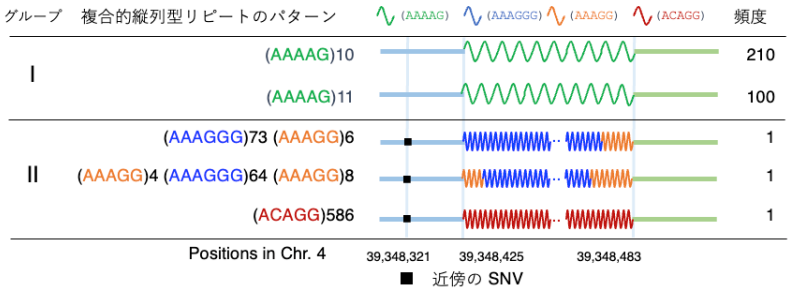
ヒトゲノムの暗黒領域を読み解く

発表のポイント

◆ヒトゲノムの中で暗黒領域と呼ばれてきたうちのひとつである縦列反復配列の組成を、日本人健常者 270 人のゲノムデータを分析し明らかにしました。

◆繰り返し単位が 1 種類の単一型よりは、複数の単位が混在する複合型が多く存在し、複合型は単一型に比べ塩基の変化が大きく、長さは短い傾向にあることがわかりました。

◆約 8900 個の領域では、伸長が顕著な個人ゲノムが観察され、疾患に関連する候補領域として今後重要であるといえます。



複合型縦列反復配列の例

発表概要

東京大学大学院新領域創成科学研究科メディカル情報生命専攻の森下真一教授、市川和樹助教、川原理樹大学院生と同大学理学部生物情報科学科の浅野岳士学部生（研究当時）の研究グループは、ヒトゲノムの暗黒領域である縦列反復配列（注 1）の組成を明らかにしました。

ヒトゲノムの中には暗黒領域(dark matter、注 2)と呼ばれる領域があります。組成を分析することが難しく、その多くが分析されてきませんでした。類似した配列が並び繰り返している配列を（図 1)、縦列反復配列と呼びます。個人差が大きいと見積られ、約 60 カ所の領域は疾患との関連性が報告されています。今回の研究では、日本人健常者 270 人の集団データを解析し、約 200 万カ所の縦列反復配列について、その組成を分析しました。

分析したうち約 322,000 カ所の領域の個人差は大きく、周辺の領域に比べると、多様性が大きいことがわかりました。リピート単位が 1 種類の単一型よりは、複数の単位が混在する複合型である領域の方が多いことがわかりました。複合型は単一型に比べ塩基の変化が大きいが、リピート単位は短く、全長は短い傾向にあります。一方、単一型は塩基の変化が少なく、リピート単位が長くより正確に複製され、全長が長くなります。

過去の研究では、約 60 個の疾患に関連して顕著に長い縦列反復配列配列が報告されています。今回の健常者の集団的調査により約 8900 カ所の領域では、中央値に比べて 100 塩基以上長くなる個人ゲノムを確認することができました。このように多様性の大きな領域は、疾患に関連する可能性があり、今後の疾患研究の基礎的情報として重要です。

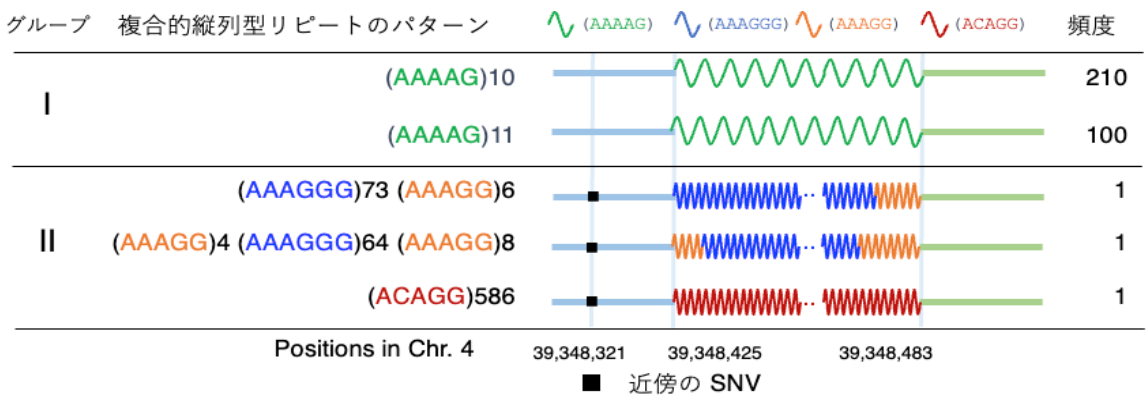


図 1：複合型縦列反復配列の例

脳疾患 CANVAS の原因と考えるリピート伸長が存在する遺伝子 *RFC1* の第 2 イントロンに存在する複合型の縦列反復配列の例。4 色の波形は 4 つのリピート単位を表現しており、最右列はパターンの頻度を示す。下の３パターンの長さは 400 塩基を超えており、一番下のパターンの長さは約 3000 塩基である。従来のショートリードでは見落とされ、ロングリードにより初めて見出された。近傍の SNV を見ただけでは、内部のリピート伸長は推定できず、縦列反復配列配列の多様化の速さを例示している。

発表内容

〈研究の背景〉

2022 年に半数体のヒトゲノムが完全に解読されました。それ以前は「暗黒領域」として黒く塗りつぶされてきたセントロメア、ゲノム重複、縦列反復配列領域を、1 個体の半数体とはいえ完全解読することはできました。ただし、「集団内で暗黒領域にはどのような個人差があり、疾患に関係しているのか」という疑問に答えるのは今後の課題です。本研究では縦列反復配列領域に焦点を当て、日本人 270 人の分析を行いました。

〈研究の内容〉

日本人 270 人から収集されたロングリードデータ（注 3）を分析し、ゲノム中の約 200 万カ所の縦列反復配列を分析しました。分析したうち約 322,000 カ所の領域は、周辺の 1 塩基バリアント（注 4）と比較して多様性が著しく大きいことを観察しました。

縦列反復配列には、複数の種類のリピート単位が存在する複合型領域があります。一部のリピート単位が異常に伸長し脳疾患を引き起こすことが、近年いくつか報告されています。これまで、複合型の縦列反復配列単位の組成を解析することは困難でした。そこで、本研究グループは高精度で分析することが可能なアルゴリズムを開発しました（Masutani *et al. Bioinformatics*, 2023）。この手法を使うことで複合型の検出が容易になり、270 人の日本人データを分析することができました。その結果、複合型は単一型に比べ塩基の変化が大きいが全長は短い傾向にあることがわかりました。その一方で、単一型は塩基の変化が少なく、リピート単位が長く、より正確に複製され長くなる傾向にあることもわかりました（図 2）。

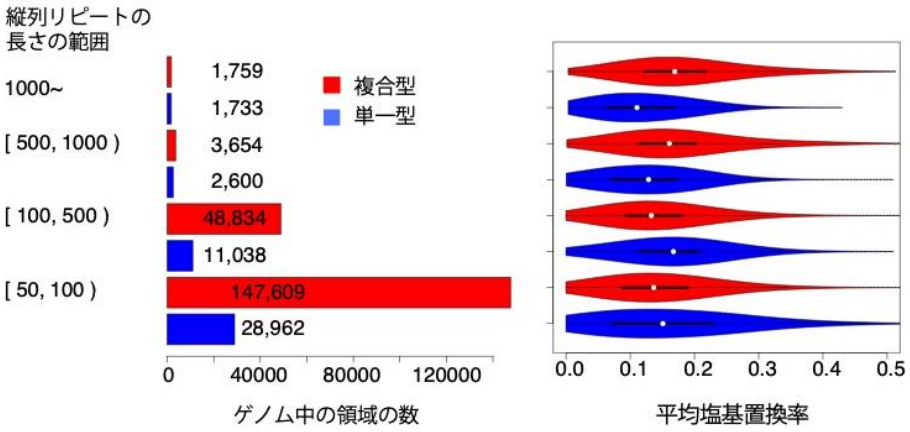


図 2：複合型と単一型で分類した縦列反復配列長の分布
左側の棒グラフは、日本人集団における縦列反復配列長の中央値の分布を示す。長さの範囲を 4 つのグループに分け、さらに複合型と単一型に分類して、各グループのゲノム中の領域数を表示している。
右側は、各グループでの領域の塩基置換率の平均値の分布を示す。

発見した領域を詳しく分析すると、縦列反復配列領域が従来の 1 塩基置換・挿入・削除だけでなく、リピート単位の重複および縮退が高い頻度で起こっていることもわかりました。このような重複と縮退を考慮した進化系統樹を描くことは、疾患に関連する縦列反復配列の伸長の特徴を理解するのに有用と考えられます（図 3）。

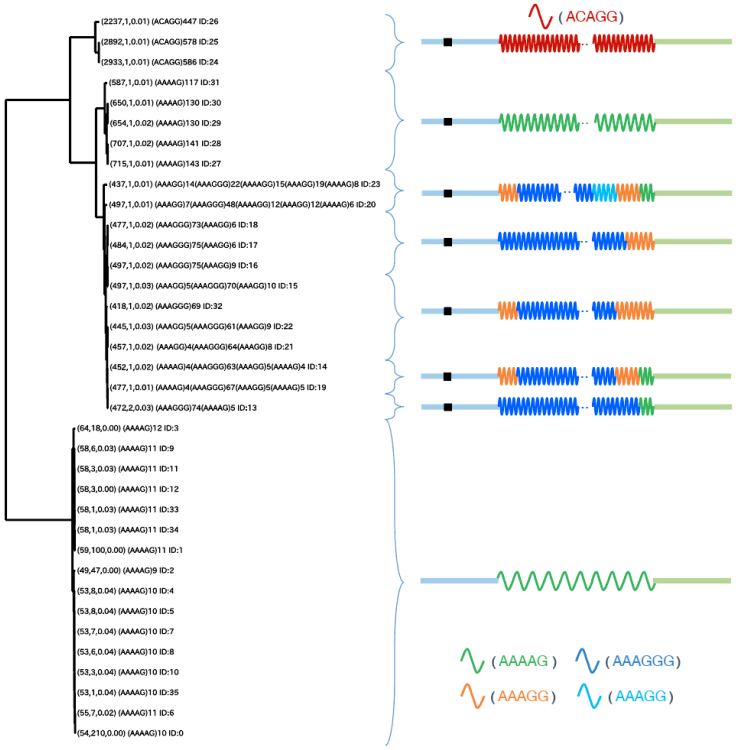


図 3： 脳疾患 CANVAS の原因遺伝子 *RFC1* の第 2 イントロンに存在する複合型の縦列反復配列の進化系統樹
1 塩基置換・挿入・削除だけでなく、replication slippage が生むリピート単位の重複および縮退が高頻度で起こる。それを考慮した精密な進化モデルの作成は今後の課題である。

〈今後の展望〉

過去の研究では、約 60 個の疾患の罹患者で顕著に長くなる縦列反復配列領域が報告されています。今回の健常者の集団的調査では、中央値に比べて 100 塩基以上長くなる個人ゲノムが見つかる領域が約 8900 カ所観察されました。その特徴は複合型より単一型の頻度が高い傾向にあり（図 4a)、リピート単位の長さは単一型が顕著に長く 10 塩基を超える場合がほとんどです（図 4b)。具体例として、図 4c に筋萎縮性側索硬化症（ALS）罹患者に関連するリピート単位 69 塩基の単一型縦列反復配列のコピー数の分布を示します。最長値と中央値の間にはコピー数換算で 20 個以上の差があります。これは日本人健常者での分布ですが、欧州での ALS 罹患者の分布より長くなる傾向にあるため、縦列反復配列分布は民族別に異なる可能性があります。今後は、日本人集団だけでなく、全世界の様々な民族集団での調査し、民族別に疾患に関与する縦列反復配列を分析することが重要になります。

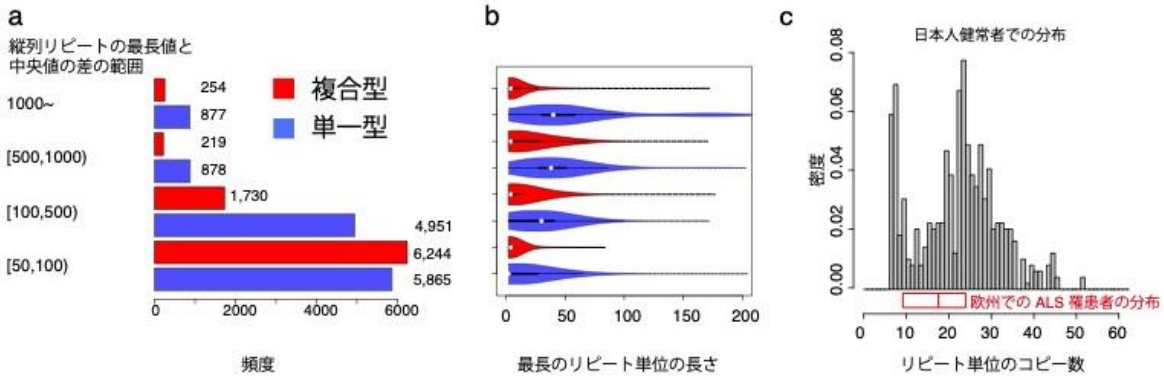


図 4：伸長が顕著な個人が存在する縦列反復配列の分析

a. 集団内における縦列反復配列が、中央値に比べて最長値が 100 塩基以上長くなる領域の頻度を、長さの範囲および複合型と単一型で分類したヒストグラム。

b. 各グループの領域におけるリピート単位の長さの分布。複合型の場合は最長単位の長さの分布。単一型はより長いリピート単位が顕著に多い。

c. 18 番染色体の領域（chr18:57024495-57024955）で ALS 罹患者において伸長することが報告されている 69 塩基を単位とする縦列反復配列が、日本人健常者集団での分布。

発表者

東京大学

大学院新領域創成科学研究科メディカル情報生命専攻

森下 真一（教授）

市川 和樹（助教）

川原 理樹（修士課程）

理学部生物情報科学科

浅野 岳士（学部生：研究当時）

論文情報

〈雑誌〉 Nature Communications
〈題名〉 A landscape of complex tandem repeats within individual human genomes.
〈著者〉 Kazuki Ichikawa, Riki Kawahara, Takeshi Asano, and Shinichi Morishita*
〈DOI〉 10.1038/s41467-023-41262-1
〈URL〉 https://www.nature.com/articles/s41467-023-41262-1

注意事項

日本時間 9 月 14 日 18 時（英国夏時間：14 日午前 10 時）以前の公表は禁じられています。

研究助成

本研究は、国立研究開発法人日本医療研究開発機構「ゲノム医療実現バイオバンク利活用プログラム、ゲノム医療実現推進プラットフォーム・先端ゲノム研究開発、研究課題名ヒトゲノム De Novo 情報解析テクノロジーの創出（課題番号：23tm0424219h0003)」の支援により実施されました。

用語解説

（注 1）縦列反復配列（tandem repeats）：
リピート単位が隣り合って縦列式に重複している繰り返し配列。CACACACA のように単位 CA が重複している配列は典型的な例。

（注 2）暗黒領域（dark matter）：
ヒトゲノムで解読が困難な領域。多くは縦列反復配列から構成される。ロングリードが普及した 2018 年ごろから解読への関心が進み、この言葉も使われるようになる(参考論文:Sedlazeck, F.J., *et al*. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* 19, 329-346 (2018). DOI：10.1038/s41576-018-0003-4)。現在、多様な人種の個人ゲノム内での分析が進み始めている。

（注 3）ロングリード（long reads）：
長さが 1 万塩基以上の DNA 断片を解読した配列。本研究では Pacific Biosciences 社のシーケンサーを利用し、塩基精度が平均 99.9%のロングリードを使用した。長さが 100 塩基以上の縦列反復配列領域は数多く、既存の技術であるショートリードで被覆できないため、その多くは見落とされてきた。ロングリードにより見落としは少なくなった。

（注 4）1 塩基バリアント（single nucleotide variant）：
個人のヒトゲノム内で観察される 1 塩基の置換。例えばシトシンがチミンへ置き換えられるのが 1 塩基バリアントである。集団内で一部の個人だけが持つ稀なバリアントも存在する一方で、多数の個人で共有されるバリアントもある。

問合せ先

〈研究に関する問合せ〉
東京大学大学院新領域創成科学研究科　メディカル情報生命専攻
教授 森下 真一（もりした しんいち）
E-mail：moris@edu.k.u-tokyo.ac.jp

〈報道に関する問合せ〉
東京大学大学院新領域創成科学研究科 広報室
Tel：04-7136-5450
E-mail：press@k.u-tokyo.ac.jp