

項目応答理論による加齢関連ゲノム領域のプロファイリング

中村 みゆき¹ 五島 杏奈¹ 安田 圭志¹ 持橋 大地²

概要：哺乳類ゲノムでは加齢に伴い DNA メチル化状態が変化するが、その分子機構は不明である。本研究では項目反応理論の連続応答モデルをヒト血液のメチル化データに適用し、各領域の変化時期と速度を評価した。年齢と正相関する領域ではエンハンサーのクロマチン状態と関連する早期変化サブセットが、負相関する領域では全体的により早期の変化が見られるなど、加齢メチル化変化の分子機構解明に資する知見を提供した。

キーワード：DNA メチル化、加齢、連続応答モデル

1. 背景

ゲノム中の DNA は CG 配列がメチル化修飾を受けることで、周辺遺伝子の発現に影響を与える。この DNA メチル化は細胞分裂を経てもある程度維持されるが、哺乳類においては、このパターンが加齢に伴い変化することが知られている。近年では、DNA メチル化パターンから年齢を推定する研究も行われている[1]。ゲノム中の CG 配列において、メチル化の傾向の変化は様ではなく、加齢に伴いメチル化率が高くなるゲノム領域も減るゲノム領域も存在するが、こうした DNA メチル化の変化の分子機構は十分に解明されていない。DNA メチル化パターンからの年齢予測の従来研究においては、Elastic Net が用いられることが多いが、この手法では各領域の変化の「時期」と「速度」を統合的に評価することが困難であり、変化パターンの多様性から分子機構を推定する手がかりが限られていた[2]。本研究では、項目応答理論(IRT)の一種である連続応答モデル(CRM)を、連続値として測定される DNA メチル化率に対して適用し、加齢に伴い変化する CG 配列のプロファイリングをおこなった[3]。

2. 研究手法

2.1 関連研究

DNA メチル化から年齢を推定する手法の多くは Elastic Net 回帰を用いている。この手法は特徴量選択とパラメータ推定を同時におこない高い精度での年齢推定を実現している。一方、Elastic Net 回帰を用いて構築されたモデルは多数あるがそれらで共通に選択された CG 配列は非常に限られている[4]。このため、ゲノム上で加齢に伴い変化する CG 配列は多数存在し、その全容を捉えるには至っていない。

2.2 提案手法

IRTはテスト分野において利用される技術であるが、我々は DNA メチル化分析の研究において用いている[5], [6]。我々が実施した従来研究においては、最もシンプルな 2 parameters logistic (2PL) model を用いている。通常、メチル化率は連続値として計測されるが、2PL model を用いて分析するためには、計測されたメチル化率を前処理で二値化する必要がある。二値化により、2PL model のパラメータ推定が容易になるというメリットがあるが、連続値で測定されたメチル化率の情報が劣化するというデメリットもある。

本研究では、IRT 関連技術の中でも連続値を扱うことができる CRM を用いた分析を実施する。IRT をテスト分野で用いる場合は、各テスト受験者の能力を潜在特性としている。また、各テスト項目は、困難度パラメータと識別力パラメータによりモデル化している。IRT をメチル化分析に用いる場合においては、潜在特性を被験者の年齢とし、テスト項目を各 CG 配列領域として扱う。また、困難度パラメータはメチル化変化開始年齢、識別力パラメータは加齢に伴いメチル化変化が進む速度として次式によりモデル化している。

$$p(X_{nj} \geq x_{nj} | \theta_n, a_j, b_j, \alpha_j) = \int_{-\infty}^{a_j(\theta_n - x_{nj})} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (1)$$

$$\text{where } x_{nj} = b_j + \frac{1}{\alpha_j} \log \frac{z_{nj}}{1 - z_{nj}} \quad (2)$$

ここで、 z_{nj} は被験者 n , CG 配列 j におけるメチル化率を表し、 a_j は CG 配列 j における識別力で、加齢に伴う CG 配列の変化の傾きを表し、ここでは「加齢の速度」を意味する。 b_j は CG 配列 j における困難度でメチル化変化開始のタイミングを表し、ここでは「加齢の段階」を意味する。 θ_n は被験者 n の潜在特性を表し、ここでは年齢を示す。 α_j はスケーリング係数である。

1 株式会社 KDDI 総合研究所
KDDI Research, Inc.
2 統計数理研究所
Institute of Statistical Mathematics

2.3 データセットと前処理

年齢と全血由来ゲノムワイド DNA メチル化データを EWAS Datahub から取得した [7]. DNA メチル化の測定には Illumina Infinium® HumanMethylation450K BeadChip (45 万サイト以上) が使われており、データは Gaussian Mixture Quantile Normalization によりバッチ効果を補正済みのものを用いている. 2 歳未満のデータを除外し範囲は 2-112 歳. 年齢分布は 2-85 歳の間で概ね均等であった. さらに 10 万 CG サイト以上で欠測のあるサンプルを除外し, 約 1,600 サンプル, およそ 2.3 万ヶ所 CG のデータセットを対象に解析をおこなった. ここからさらに各 CG において標準偏差が 0.05 未満のものについては除き, 年齢との相関係数が 0.5 以上のものと -0.5 以下のものに分けて CRM を適用した. また, ゲノム上の性質データとして ChromHMM を用いた [8].

2.4 CRM の実行

項目パラメータ (識別度, 困難度) と個人パラメータ (潜在特性) は R パッケージ EstCRM [9] で推定した. β 値 (DNA メチル化率 0-1) は各 CpG の最小・最大値で標準化し, 極値回避のため 0 または 1 の場合は疑似値 0.001 で調整した. 推定パラメータはブートストラップ法 (最大 EM サイクル 500, 収束基準 0.01, Shojima タイプ [10], BFGS アルゴリズム, サンプル数 50) で評価した. CRM 推定 θ パラメータを Z-score 標準化後, 実年齢からの統計量による線形変換を適用して潜在特性値を年齢尺度に換算した.

3. 結果

DNA メチル化データを用い, 正・負相関の CG 群を対象にそれぞれ「DNA メチル化が変化する年齢の段階」と「加齢に伴い変化する速度」. 相関閾値 ± 0.5 で選択した正相関 3,125 サイト, 負相関 3,299 サイトについて, ブートストラップ解析により信頼性の低い CG を除外し, 最終的に正相関 3,009 サイト, 負相関 3,241 サイトを得た. ここから推定した「DNA メチル化が変化する年齢の段階」と「加齢に伴い変化する速度」は正相関 CG 群では年齢に沿って非対称分布を示し (図 1), 早期・後期位相の CG は中年期より少なく, いずれも緩やかな加齢ペースを示したが早期位相がより緩慢であった. 一方, 負相関 CG 群からのパラメータ分布は比較的対称的であり (図 2), 高加齢ペースは 40 歳頃に集中した (正相関では 60-80 歳). 正相関 CG には早期位相で緩慢ペースの特異的サブセットが存在した. また推定したパラメータを用いて個々人について年齢変換をおこなったところ, 元の年齢と強い相関がみられた (図 3). 加えて, ゲノム上のエンハンサーやプロモーターといった特徴毎に分布のバイアスが見られるかを確認したところ, エン

ハンサー領域にある正相関 CG の多くが早期かつ緩やかな加齢ペースを示した (図 4).

正の相関CG群 ($r \geq 0.5$)

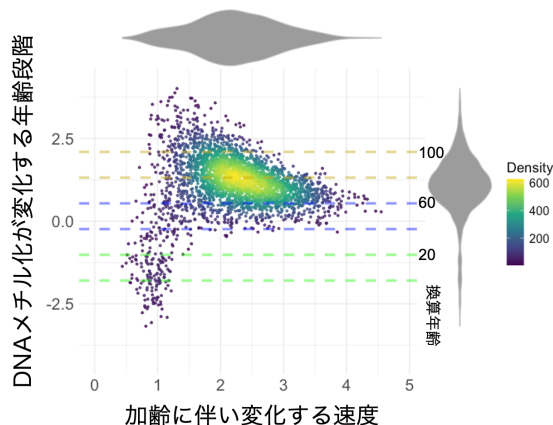


図 1 正の相関を見せる CG 群の年齢段階と速度を示すパラメータ分布

負の相関CG群 ($r \leq -0.5$)

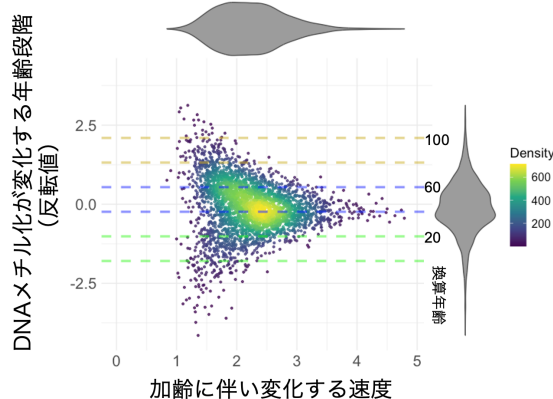


図 2 負の相関を見せる CG 群の年齢段階と速度を示すパラメータ分布

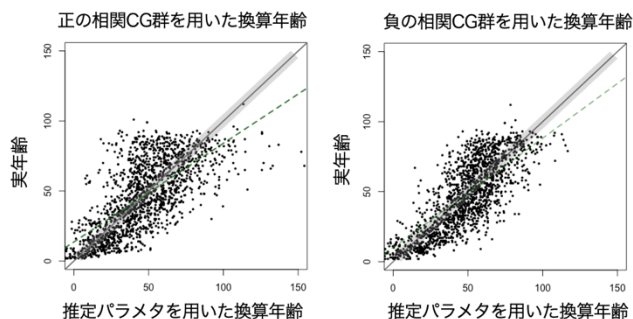


図 3 推定したパラメータから換算した年齢と実年齢の相関

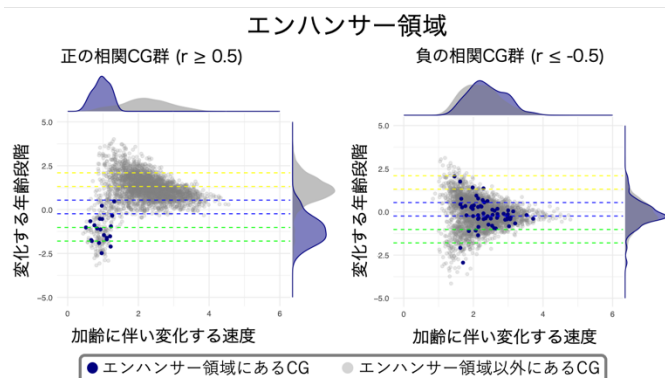


図 4 エンハンサー領域にある CG のパラメタ分布
(左: 正の相関 CG 群, 右: 負の相関 CG 群)

4. 考察と結論

本研究により、加齢に伴う DNA メチル化変化において正相関および負相関 CG 群のパラメタ分布に明確な傾向の違いが認められた。特に正相関 CG 群では若年期に変化する特異的サブセットの存在が確認され、これは従来の手法では捉えることが困難であった加齢メチル化変化の多様性を示している。本手法で得られた各 CG 配列の加齢位相と加齢ペースの定量的評価は、年齢推定に留まらず他の生物学的形質の解析にも応用可能と考えられる。さらに、各 CG 配列の応答特性の詳細な解析結果は、効率的なバイオマーカー用プローブサブセットの設計において、より精度の高いマーカー選択の指針となることが期待される。これらの知見は加齢に関連するエピジェネティック変化の分子機構解明に向けた有用な基盤を提供する。

参考文献

- [1] Warner, B. et al.,: A systematic review of phenotypic and epigenetic clocks used for aging and mortality quantification in humans, *Aging*, Vol.16, No.17, pp.12414-12427 (2024).
- [2] Simpson, D. J. and Chandra, T.: Epigenetic age prediction, *Aging Cell*, Vol.20, No.9, p.e13452 (2021).
- [3] Samejima, F.: Homogeneous Case of the Continuous Response Model, *Psychometrika*, Vol.38, No.2, pp.203-219 (1973).
- [4] Liu, Z. et al.: Underlying features of epigenetic aging clocks in vivo and in vitro, *Aging Cell*, Vol.19, No.10, p.e13229 (2020).
- [5] Nagata, M. et al.: Pilot Study of Item Response Theory-based DNA Methylation Analysis for Aging Association, *Proc. 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.3109-3113 (2022).
- [6] Yasuda, K. et al.: Building a DNA Methylation Aging Clock Model on Less Labelled Data Using Item Response Theory, *Proc. 16th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp.139-145 (2023).
- [7] Xiong, Z. et al.: EWAS Data Hub: a resource of DNA methylation array data and metadata, *Nucleic Acids Research*, Vol.48, No.D1, pp.D890-D895 (2020).
- [8] Vu, H. and Ernst, J.: Universal annotation of the human genome through integration of over a thousand epigenomic datasets, *Genome Biol*, Vol.23, No.1, p.9 (2022).
- [9] Zopluoglu, C.: EstCRM: An R Package for Samejima's Continuous

IRT Model, *Applied Psychological Measurement*, Vol.36, No.2, pp.149-150 (2012).

- [10] Shojima, K.: A noniterative item parameter solution in each EM cycle of the continuous response model, *Japan Society for Educational Technology*, pp.1-2 (2005).

【 この位置に改ページを入れ,以降のページを印刷対象外とする 】