別紙３

厚生労働科学研究費補助金（エイズ対策研究事業）
（分担）研究報告書

「エイズ関連悪性腫瘍誘発機序の理解と抗体療法の有効性評価」研究
分担課題：小サンプルの In vitro 実験による全ゲノムＤＮＡメチル化データから、変化した遺伝子を客観的／効率的に抽出するための統計学的解析方法の検討

研究分担者　田中　紀子　国立国際医療研究センター医学統計研究室長

研究要旨
一般的に統計学を用いた研究仮説の検証は大標本理論をもとにした要約統計量を求めることにより群間比較や要因分析により行われる。DNA メチル化データの統計学的解析方法においても、広く適用されているものはこの大標本理論に基づいたものがほとんどである。しかしながら、網羅的ゲノム解析は現在もコストがかかることにより大標本を得ることは難しく、効率的研究デザインおよび小サンプルでの仮説探索が不可欠となる。今回の研究のように In vitro で特に制御された環境下においては実験デザインの工夫により環境要因による分散成分を限りなく小さくしていくことにより効率を上げていくことが可能である。さらに、臨床検体を用いた先行研究より、メチル化状態が真に変化していると考えられる個体あるいは遺伝子に関しては、全体よりかなりはずれた値を示すことが示され、より小さいサンプルサイズにより変化をとらえることの可能性が示されたことを踏まえ、a:全ゲノムＤＮＡメチル化測定データの分布に基づく個体間差の検出、b:小サンプルでの複数のバラツキの指標を用いた多群比較法を用いた DNA メチル化感受性遺伝子の検出方法の検討を行うこととする。a については、臨床検体を用いた先行研究データを用いて適切なデータの視覚化および、パラメトリックおよびノンパラメトリックな分布当てはめと形状パラメタの推定のためのプログラムを R を用いて開発する。ｂについては、IQR(四分位差)、MAD（中央値絶対差）など正規性からの逸脱に頑健なスケールパラメタによりなるべくバラツキが小さく、平均群間差の大きいあるいは群間でのメチル化状態のバラツキ具合自体に差のある遺伝子領域を選択できる指標を GEO にて公開済みのデータを用いて検討した。その結果、個体間差の検出のための混合β分布の当てはめは、ノイズに非常に敏感であること、計算時間がかかることなどから、実用性に低い可能性が示された。今後実用性の高い別の方法の開発が必要であることが示唆された。候補領域の絞り込みには、検出したい差に基づいた複数の手法の複合的評価基準を設けることが必要であることが示唆された。

別紙３

A. 研究目的

一般的に統計学を用いた研究仮説の検証は大標本理論をもとにした要約統計量を求めることにより群間比較や要因分析により行われる。DNA メチル化データの統計学的解析方法においても、広く適用されているものはこの大標本理論に基づいたものがほとんどである。しかしながら、網羅的ゲノム解析は現在もコストがかかることにより大標本を得ることは難しく、効率的研究デザインおよび小サンプルでの仮説探索が不可欠となる。今回の研究のように In vitro で特に制御された環境下においては実験デザインの工夫により環境要因による分散成分を限りなく小さくしていくことにより効率を上げていくことが可能である。さらに、臨床検体を用いた先行研究より、メチル化状態が真に変化していると考えられる個体あるいは遺伝子に関しては、全体よりかなりはずれた値を示すことが示され、より小さいサンプルサイズにより変化をとらえることの可能性が示されたことを踏まえ、本研究では、a:全ゲノムＤＮＡメチル化測定データの分布に基づく個体間差の検出、b:小サンプルでの複数のバラツキの指標を用いた多群比較法を用いた DNA メチル化感受性遺伝子の検出方法の検討を行うことを目的とする。

B. 研究方法

目的 a および b について用いた実データはすでに出版済みの HIV とリンパ腫との関連を検討したイルミナ４５０K チップにより測定されノーマライゼーション済みの 28 検体分、375639 プローブのデータである。

目的 a および b について用いたシミュレーションデータは、全ゲノムメチル化データが２～３混合ベータ分布に従うことから以下のようないくつかのパラメタを設定して発生させた。

Scenario 1 (two-peaks model): a1=1, b1=12, a2=13, b2=2, w1=0.4, w2=0.6;
Scenario 2 (three-peaks model): a1=7.0, b1=43.6, a2=2.6, b2=3.9, a3=14.0, b3=1.8,w1=0.26, w2=0.31, w3=0.43;
Scenario 3 (four-peaks model): a1=5, b1=70 a2=8, b2=26, a3=45, b3=25, a4=64, b4=6, w1=0.35, w2=0.15, w3=0.16, w4=0.34.

データの発生には R vers.2.15 を用いた。

目的 a については、臨床検体を用いた先行研究データを用いて適切なデータの視覚化および、パラメトリックおよびノンパラメトリックな分布当てはめと形状パラメタの推定のためのプログラムを R を用いて開発した。シミュレーションデータについては真の分布の混合数がわかっているが、実データについては真の混合数が分からないため、データを視覚化した際に観測されるピークの数を数えることで、真の混合数を設定した。尚、観測者によるバイアスを除去するために、ピーク数の数え上げは独立した場所で 3 人によって行われ、多数決によって真の混合数が決定された。3 人とも回答が異なった場合は、判定不能とした。

目的ｂについては、まず本年度は、2 群比較のみを行うこととした。一般的に適用されている t 検定（群ごとの分散が等しいという仮定のもとで平均値の比較)、ウィルコクソン順位和検定（群ごとの分布形が等しいという仮定のもとでノンパラメトリックな中央値の比較）に加え、ノンパラメトリックな F 検定に対応する Ansari-Bradley 検定（群ごとの分布型が等しいという仮定の下でノンパラメトリックな分散の比較)、および、分布型も分布の代表値も違うことを検出するためのコルモゴロフスミルノフ検定を実データに適用し、選択される候補領域にどのくらい差があるのかを観察した。

C. 研究結果

a 前年度において、多次元β分布を各サンプルのメチル化測定データに当てはめて、個体間での分布比較が可能かどうかについて実データおよびシミュレーションデータで検討を行った結果、感度が 70~90%と比較的良好であったが、特異度が 0~50%と低い値であった。そこで、測定の生データはβ分布に従うことが理論的に示されているが、変数変換することで、漸近的に多次元正規分布に従うと考えられるため、サンプ

ルごとに変数変換後多次元正規分布を当てはめ、再度感度・特異度についてプレリミナリーな検討を行った結果、感度・特異度ともにβ分布で検討するよりも高い結果となった。

ｂ　４つの検定手法を NCGM データに適用した結果、たとえば、ｐ値の小さい順 20 プローブを選択するとした場合、375639 プローブ中 66 プローブが選択された。このうち 12 プローブが二つの手法で選択されていたが、そのうち９プローブはウィルコクソン順位和検定とコルモゴロフスミルノフ検定で選択されていたものであった。ｔ検定やウィルコクソン順位和検定で検出されず、コルモゴロフ検定で検出されたものに関して分布型を検討した結果、コントロール群で比較的標準的な単法性の分布型であるのに、ケース群（今回は HIV 感染群）で 2 峰性をとっているものなど、分子生物学的には重要そうな差を検出している可能性が示唆された。

D. 考察

a について

混合β分布の当てはめは、ノイズに非常に敏感であること、計算時間がかかることなどから、実用性に低い可能性が示された。多次元正規分布を当てはめることで、個人のメチル化分布の差の検出が可能であることが示されたが、さらに、多次元正規分布に従わないようなノイズの多い分布が実データでは発生することも考えられるため、ノンパラメトリックに核関数で分布を推定し、分布のピークを数え上げることで分布の個人間差を検討する方法についても同様に性能評価を行うこととした。

ｂについて

今後は、どのくらいのサンプルサイズで漸近性が保たれるのか、あるいはある程度分布型が異なる場合においても t 検定でも十分検出されるようになるのかなどをシミュレーションデータで検討することも必要と考える。また、検定手法によって検出したい差が異なるので、複合的評価基準を設けることで候補領域を見逃さない手法の提案を行う必要性があると考えられた。

A. 結論

混合β分布の当てはめは、ノイズに非常に敏感であること、計算時間がかかることなどから、実用性に低い可能性が示された。今後実用性の高い別の方法の開発が必要であることが示唆された。

候補領域の絞り込みには、検出したい差に基づいた複数の手法の複合的評価基準を設けることが必要であることが示唆された。

B. 研究発表

(学会発表)

Tanaka N, Kurosawa T, Inaba Y, Toyo-oka L, Yoshida L, Kawasaki Y. Filtering samples based on Beta-Micture model for DNA methylation data Quantified by Bisulphite microarrays. International Biometric Conference 2014. Florence. Italy. July. 2014.

C. 知的財産権の出願・登録状況

無し.