2019 年 6 月 21 日

オミクス関連研究における選択マーカー数を固定した二段階スクリーニング法

医学分野におけるオミクス関連解析研究の主な役割の一つは，ゲノム，トランスクリプトーム，プロテオームなどのオミクスデータから，興味のある表現型変数と関連のあるマーカーをスクリーニングすることである．典型的には，false discovery rate（FDR）を制御した多重検定によるマーカー選抜が実施され（第一段階），遺伝子機能解析やパスウェイ解析などの生物学的な検討を通してマーカーがさらに絞られる（第二段階）．最終的に選抜されたマーカーセットは，バリデーション目的で別の分析プラットフォームでの検討対象となることが多い．表現型変数が治療後の反応や予後の場合には，臨床現場で普及しているプラットフォーム（例えば，PCR 法）にて選抜マーカーを用いた表現型変数の判別・予測システムの開発が試みられることもある．多くの場合，以上のような後続研究で対象にできるマーカー数には上限がある（数十程度）．第一段階の多重検定で有意となるマーカー数は確率変数であるので，第二段階目では，この上限に収まるよう，マーカー数の帳尻合わせが行われることになる．

第二段階目での選抜は統計学的視点と生物学的視点が反映されたものとなるが，その基準は研究によってまちまちであり，現時点で明確なものは存在しない．このことは，第二段階目で選抜される K 個のマーカーの FDR レベルは明らかでなく，制御されてないことを意味する．そこで本研究は後続研究に進む K 個のマーカーに対して FDR 制御を可能とする二段階スクリーニング法を提案する（図１） [1].

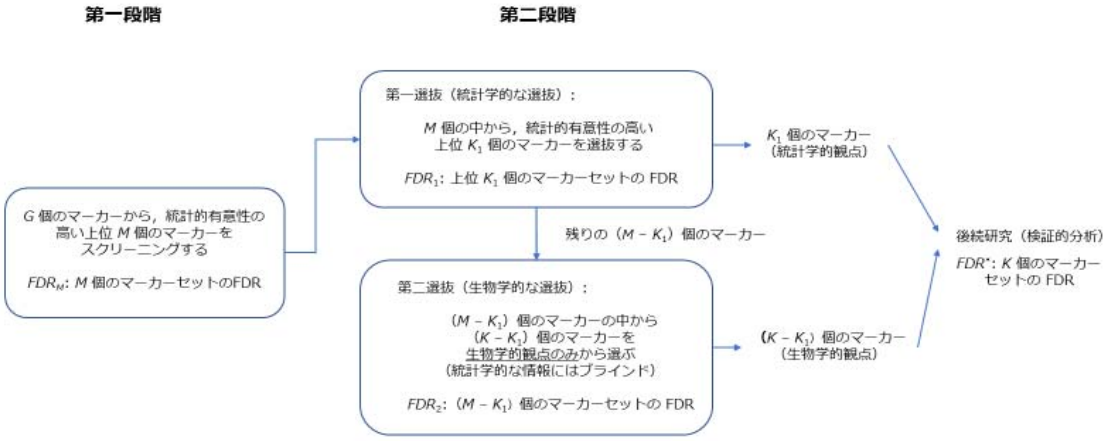

図 1．提案する二段階スクリーニング法

第一段階目では従来のように統計的有意性が最も高い $M(>K)$ 個のマーカーを選抜する．第二段階目では，統計学的視点と生物学的視点による選抜を明確に分ける．すなわち，統計学的な視点を最大限に尊重して統計的有意性が最も高い K_1（$<K$）個のマーカーを先に選抜する．残りの（$M-K_1$）個のマーカーに対しては生物学者が「生物学的な視点のみに基づいて」（$K-K_1$）個のマーカーを選抜する．その際，（$M-K_1$）個のマーカーに関する P 値や効果サイズなどの統計学的な情報が伏せられる（ブラインドがかけられる）．併せて，既に選んでいる K_1 個のマーカーについてはそのマーカー名のみが生物学者に提示される．これより，K_1 個のマーカーとは別の生物学的側面をもつマーカーが選抜されるであろう．

第二段階目において，統計的有意性が最も高い K_1 個のマーカーに対する FDR レベル（FDR_1 とおく）は標準的な FDR 推定法（Storey 法など）によって推定可能である．一方，生物学的視点から選んだ残りの（$K-K_1$）個のマーカーに対する FDR レベルは，「（$K-K_1$）個のマーカーは（$M-K_1$）個のマーカーからのランダムサンプリング」とみなすことで，（$M-K_1$）個のマーカーに対する FDR レベル（FDR_2 とおく）に等しいと考えることができる．すなわち，最終的に選抜される K 個のマーカーの FDR レベル（FDR^*）は以下で表される：

$$FDR^* = \frac{K_1 FDR_1 + (K - K_1) FDR_2}{K}$$

ところで，第一段階目で選抜する M 個のマーカーに対する FDR_M は以下で表現できる：

$$FDR_M = \frac{K_1 FDR_1 + (M - K_1) FDR_2}{M}$$

このとき，$FDR_1 < FDR_2$，$(K-K_1)/K < (M-K_1)/M$ より，$FDR^* \leq FDR_M$ が成立する．これより，提案法によって FDR^* を α 以下（5%や 10%以下）に制御できるようになることで，FDR_M に対して α を上回る FDR レベルを許容できることになる．これは，第一段階目において，（従来の $FDR_M \leq \alpha$ とする基準よりも）より多くのマーカーをスクリーニングできることを意味する．これより，第二段階目での生物学的視点に基づく選抜でより多くのマーカーが対象となり，生物学的に重要なマーカーをより多く発見できることが期待される．

数値実験の結果，FDR^* の制御によって，従来の基準と比較して，FDR_M の値で 25%程度，M の値に関しては 5 倍程度，許容できることがわかった（注：真陽性の数はほとんど同じ）．その意味で提案法はオミクス研究でのマーカースクリーニングの有効な方法であると考えられる．

文献：

1. Kawabata T, Emoto R, Nishino J, Takahashi K, Matsui S. Two-stage analysis for selecting fixed numbers of features in omics association studies. Statistics in Medicine 2019 (In press).