計量生物学 Vol. 38, No. 2, 127–139 (2017)

オミクス研究における検証的解析と探索的解析：多重検定と P 値を中心に

Confirmatory and Exploratory Analyses in Omics Studies with Particular Focus on Multiple Testing and P-value

松井茂之
Shigeyuki Matsui

名古屋大学大学院医学系研究科生物統計学分野
Department of Biostatistics, Nagoya University Graduate School of Medicine
e-mail:smatsui@med.nagoya-u.ac.jp

In this article, we discuss the role of P values in multiple testing to associate a large number of genetic or molecular features with a phenotypic variable of interest in biomedical omics studies. For multiple tests in such association analyses, we distinguish those conducted for confirmatory purpose, as seen in genome-wide association studies to determine disease-associated variants, from those for exploratory screening of associated features. For the latter, exploratory analysis, we discuss application of the ROC curve analysis used in diagnostic medicine, as an alternative, but more relevant framework, rather than the standard framework based on multiple testing that controls false positives only. Finally, partly based on arguments made in the field of omics studies, we make some comments on future endeavors by statisticians to disseminate discussions given in the ASA's Statement on P-Values (Wasserstein and Lazar, 2016, The American Statistician, 70, 129-133) to improve statistical practice in various scientific fields.

Key words: confirmatory analysis; exploratory analysis; multiple testing; omics studies; P-value; ROC analysis.

1. はじめに

近年，著しいバイオ技術の進展によって生体由来の組織や細胞を分子レベルで捉えることが可能となり，DNA 配列(点変異，挿入・欠失，コピー数，染色体構造等)，転写物(遺伝子発現)，DNA メチル化，タンパク質発現，代謝物質など，様々な分子データの測定が可能となっている．さらに，シークエンスやアレー技術の低コスト化により，ゲノム，転写物，蛋白質などの網羅的な解析が普及し，分子データのビックデータ化が進んでいる．全ゲノム，全遺伝子などの分子全体を表すデータはオミクスデータ(omics data)と総称される．

オミクスデータと生体の表現型(医学の例では，疾患の発生，病態，治療反応性など)との関連を調べることはオミクス研究における基本的な解析の一つである．より具体的には，一つ一つの

オミクス変数(遺伝子多型，プローブセットなど)と表現型変数の関連に対して検定が実施される(多重検定；multiple testing)．オミクス変数は数千〜数百万とあるので，膨大な数の P 値が計算されることになる．

オミクス研究，あるいは，上記の多重検定はどのような目的のもとで行われるのだろうか？ 医学の研究を例にとると，オミクス研究のゴールとして，1) 疾患の発生，進展，病態，治療反応性等の背後にある分子機構(メカニズム)の理解，2) 医療技術(疾患の予防法，診断法，治療法)の開発，が主にあげられる．とりわけ，未知の重要な分子機構の発見や診断に役立つ新しいマーカーの発見といった「新発見」，あるいは，「新たな仮説の生成」においてオミクス研究の大きな意義・価値を見出すことができる．必然的に，オミクス研究からの新発見や仮説は，後続の研究で更に検討されることになる．従って，オミクス研究での多重検定は，探索的解析と位置づけるのが妥当ということになる．しかし，実質的に検証を目的として多重検定が実施されている研究も例外的にみられる．疾患関連遺伝子の同定のためのゲノムワイド関連研究(genome-wide association study; GWAS)がその代表である．

本稿は，2016 年に American Statistical Association(ASA)から出された統計的仮説検定の P 値に関する声明(Wasserstein and Lazar, 2016)を受け，オミクス研究分野の現状を紹介し，その中での P 値の意義や役割について整理することを目的とする．大きく検証的解析と探索的解析に分けて議論する．

2. 検証的解析—ゲノムワイド関連研究

2.1 ゲノムワイド有意水準

ゲノムワイド関連研究 GWAS では，患者と健常者(それぞれ数百〜数十万人)の二つのグループの間で頻度の異なる遺伝子多型(一塩基多型)の同定が試みられる．遺伝子多型の数は，数十万〜数百万，その一つ一つについて頻度の差の検定が実施される．当然，多重性の調整が必須となる．GWAS では，有意水準として，ゲノムワイド有意水準(genome-wide significance level)とよばれる 5.0×10^{-8} を用いることが定着している．この有意水準に対する一つの解釈として，一般ヒト集団(ヨーロッパ集団)の全ゲノム染色体上に約 100 万個の独立断片が存在すると想定し，これにボンフェローニ基準を適用するとこの有意水準が導かれる(Risch and Merikangas, 1996; Hoggart et al., 2007)．いくつかのシミュレーション研究もほぼ同じ有意水準を導いている(Hogaart et al., 2007; Pe'er et al., 2008; Dudbridge and Gusnanto, 2008)．ボンフェローニ法は，検定全体で少なくとも一つの偽陽性が生じる確率である family-wise error rate(FWER)をコントロールする多重調整法の中で最も保守的な方法であり(例えば，Westfall and Young, 1993)，実質，検証的解析として多重検定が実施されているということができる．実際，GWAS において，多重検定の P 値，及び，統計的有意性の有無は研究の成否(≃ パブリケーション)の重要な判断材料の一つになっている．

なぜ，GWAS ではゲノムワイド有意水準という，極めて厳しい有意基準が確立しているのだろうか？ 一般に，疾患の発生には，遺伝のみならず，環境の影響も考えられるが，遺伝は環境曝露よりも時間的に先行するものとみなせるので，関連遺伝子多型を探索することで疾患原因変異

(causal variant)の発見に結びつく可能性がある．疾患原因変異または遺伝子の発見は，それ自体，サイエンスの進歩として(一般人にも)認識しやすく，さらに新しい治療の開発にも結びつく可能性もあるので社会的な意義としてもわかりやすい．また，多くの GWAS は，国家やコンソーシアムの一大プロジェクト研究として実施される．GWAS の結果は Nature やその系列雑誌など，総合科学トップジャーナルで報告されることが多いが，以上のことがその主な背景にあると考えられる．これらのジャーナルは大きな社会的インパクトをもち，そこで発表された結果が思わぬ方向で一人歩きするリスクも大きい．それゆえ，偽陽性に関して常に大きな危惧を抱いていることは容易に想像できる．特に，数十万〜数百万の遺伝子多型を一度に調べる GWAS ではなおさらであろう．ジャーナルの戦略として，ゲノムワイド有意水準(さらに追加の検証も要請)といった，極めて保守的な基準を採用することは理解できる(一方，研究者はとにかくトップジャーナルで研究発表したいので，どんな厳しい基準でもなんとかクリアしようとがんばる！)．しかし，改めて，GWAS を第 1 節で述べたオミクス研究としてみると，ゲノムワイド有意水準の使用は，オミクス研究の本来の価値である「新発見」の可能性を大きく損ねているといわざるを得ない．

2.2 他の指標，基準

もちろん，この問題は関連する学会や研究者コミュニティーにおいても十分認識されている．家系解析から推定される遺伝的疾患要因のほんの一部しか解明されてないという問題，いわゆる，missing heritability(失われた遺伝率)(例えば，Maher, 2008)の一因になっていることも明らかである．そこで，P 値，あるいは，ゲノムワイド有意水準に代わる基準がいくつか提案されている(註：あわせて第 3 節の ROC 曲線解析も提案に加えることができるだろう)．

ベイズ流アプローチからの提案の一つとして，false-positive report probability (FPRP) がある(Wacholder et al., 2004)．任意の遺伝子多型に対して，検定の P 値を P，有意水準を α で表すと，FPRP は，

$$FPRP = \Pr(H_0 \mid P \le \alpha) = \frac{\Pr(P \le \alpha \mid H_0)\Pr(H_0)}{\Pr(P \le \alpha \mid H_0)\Pr(H_0) + \Pr(P \le \alpha \mid H_1)\Pr(H_1)} = \frac{\alpha\pi}{\alpha\pi + (1-\beta)(1-\pi)}$$

と表される．一行目の式はベイズの定理を用いている．二行目にある $1-\beta$ は(想定される効果サイズのもとでの)検出力である．π は，帰無仮説 H_0 が真である事前確率である．FPRP は，ベイズ流の false discovery rate (Efron and Tibshirani, 2002)と解釈することができる(註：3.3 節の FDR(c) に対応)．

ベイズファクター(Bayes factor; BF)の利用も提案されている(Wakefield, 2009)：

$$BF(y) = \frac{\Pr(Y = y \mid H_0)}{\Pr(Y = y \mid H_1)}$$

ここで，Y は検定統計量を表す．効果サイズ(例えば，遺伝子多型と疾患の有無に関する対数オッズ比の真値に相当)を θ で表すと，分母の $\Pr(Y = y \mid H_1)$ は，任意の効果サイズ θ のもとでの確率密度 $f(Y = y \mid \theta)$ を効果サイズの事前分布 $g(\theta)$ について積分をとることで得られる．$BF(y)$ は H_0 に関する事後オッズ比と事前オッズに対して，

$$\frac{\Pr(H_0 \mid Y = y)}{1 - \Pr(H_0 \mid Y = y)} = BF(y) \times \frac{\pi}{1-\pi}$$

の関係にある．ベイズ流意思決定の観点からは，ベイズファクターに基づく判定は，（データ Y 所与のもとでの）H_0 の事後確率に基づく判定に対応し，第一種の過誤に伴う損失に対して第二種の過誤に伴う損失を最小化するものとみることができる（例えば，Robert, 2007, pp. 224–228）．この点は第一種の過誤のコントロールのみを考える検定と大きく異なる．

ベイズ流アプローチでの課題は，事前情報（事前確率 π，効果サイズの事前分布 $g(\theta)$）の設定である．客観性の担保や正確性の向上を考えると，事前情報を研究データから直接推定するアプローチが魅力的である（経験ベイズ）．幸い，GWAS を含め，オミクス研究では膨大なオミクス変数（遺伝子多型，プローブセットなど）のデータが測定されている．つまり，似たような分布・構造をもつ変数のデータが潤沢に存在する．変数間の情報共有を試みる経験ベイズのアプローチはオミクスデータの解析で特に有効と考えられる（Efron, 2008; 2010）．

ゲノムワイド有意水準よりも緩い有意基準として，suggestive level とよばれる 1.0×10^{-6} が用いられることがある．また，一つの偽陽性も許したくない FWER 基準ではなく，（有意な検定のセットにおける）偽陽性の割合をある程度以下に抑える false discovery rate（FDR）のコントロール（Benjamini and Hochberg, 1995; Storey, 2002）も一部の研究で用いられている（例えば，Nelson et al., 2017）．

2.3 提　　言

しかし，2.1 節で述べたような GWAS の性格を考えると，「ゲノムワイド有意水準を含む極めて保守的な基準を使用する」というプラクティスが今後覆ることはまずないであろう．このように認識すると，「ゲノムワイド有意水準の是非を問う」，「ゲノムワイド有意水準，FDR のいずれがよいか？」などの発想に留まっていては大きな前進はないと思われる．一つの現実的なアプローチとして，ゲノムワイド有意水準を用いた検証的解析に加えて，上記のベイズ流指標，suggestive level, FDR などを用いた探索的解析を同時に実施することが考えられる．例えば，治療法を評価する臨床試験では，通常，主要な評価項目（プライマリーエンドポイント）の解析とその他の副次的な項目（セカンダリーエンドポイント）の解析が同時に行われる．もちろん，両者は厳密に区別されなければならない．GWAS の論文では，例えば，「Exploratory Analysis for Hypothesis Generation」などと銘打ったセクションあるいはサプリメントを設けるようにしてみるのはどうであろうか？　一方，探索的な基準（suggestive level, FDR など）を用いる際には，別の独立集団を用いての検証的研究（外的妥当性の評価にも対応）と組み合わせるなど，偽陽性の可能性を十分小さくするための工夫が求められる（例えば，Nelson et al., 2017）．このような方向性については未だ系統だった検討はほとんどみられないが，ゲノムワイド有意水準に代わるアプローチとして今後重要になると考えられる．

3. 探索的解析―関連のスクリーニング

3.1 多重検定の枠組みは適切か？

第 1 節で述べたように，オミクス研究におけるオミクスデータと表現型変数の関連解析は，一般に，関連のスクリーニングのための探索的解析として位置づけられる．関連解析のこれまでの議論を振りかえると，そのほとんどは多重検定の枠組みで行われてきた．オミクス解析（あるいは

高次元データの解析)が出現する以前は，比較的限られた数の検定を対象とした「検証的解析のための FWER のコントロール」の議論が主流であった．そこに，近年，高次元データの解析が出現し，それまでの多重検定の方法論・方法をどのように適合(adaptation)すればよいか？ が議論されるようになり，その中で，FWER 基準に代わる FDR 基準などの新しい展開が生じ，今日のオミクス研究における多重検定の枠組みが確立されてきたと著者は理解している．つまり，元々の検証的解析の流れを汲んでいる，あるいは，検証的解析の延長線上にあるものが確立しているといえる．しかし，検証的解析(関連の検証)と探索的解析(関連のスクリーニング)は，本来，根本的に異なるものではないか？ という疑問が生じないだろうか．

改めて，オミクス研究の意義・価値を考えると，第 1 節で述べたように，それは「新発見」にある．実際，オミクス研究を行う研究者は「新発見」に大いに期待しているし，ときめいている(著者の経験)．オミクス研究に対しては，それがどれだけ新発見できる能力をもっているのか，そのことが非常に重要と考えられる．多重検定の枠組みでいうと，真陽性をどれだけ検出できるか，という側面が重要になる．しかしながら，多重検定の枠組みでは偽陽性のみが考慮され，コントロールの対象になる．これでは片手落ちといえるのではないか？

3.2 ROC 曲線解析

診断医学(diagnostic medicine)の一分野に，健常人の集団を対象とした疾患スクリーニングがある．ある疾患の検出に関して，第一段階として，比較的非侵襲で簡便な診断法が実施され，疾患の疑いのある対象者がふるいわけられる(疾患のスクリーニング)．そこで “陽性” となった対象者には，第二段階として，より侵襲を伴う，より確定的な診断が試みられる．第一段階での疾患スクリーニングで用いる(簡便な)診断法の性能は，感度(sensitivity)と特異度(specificity)により評価されるのが基本である．感度は，疾患を有する集団における陽性の確率，特異度は，疾患を有さない集団における陰性の確率である．多くの場合，陽性の判定は，特定のマーカーに対し，それが閾値(カットオフ，cut-off)を超えたかどうかで行われる．カットオフを動かすと，感度，特異度も変化する．この関係を，縦軸に感度，横軸に 1 – 特異度(偽陽性の確率)をとって表現したものが ROC 曲線(receiver operating characteristic curve)である．カットオフの選択は，感度と偽陽性度(あるいは特異度)のバランスを考慮して行われる．著者にとって，この ROC 解析の枠組みは，オミクス研究での関連スクリーニングにもすんなりとフィットする自然な枠組みと考えられる．

それでは，オミクス研究において，どのように ROC 曲線を構成すればよいのだろうか？

疾患スクリーニングで用いられる診断マーカーに相当するものは，個々のオミクス変数と表現型変数の関連を捉える適当な統計量である(検定統計量など)．陽性判定のカットオフに相当するものはこの統計量上でのカットオフとなる．検定統計量(P 値)を用いる場合は，棄却限界値(有意水準)がカットオフに相当する．

縦軸の感度に対応するものとしては，H_0 が成立しない，すなわち，non-null が正しいときの検出力，あるいは，真に表現型と関連するオミクス変数の検定における検出力を用いることが考えられる．具体的には，全ての non-null のオミクス変数における平均検出力である．これはオミクス研究の検出力評価で用いられている(Tsai et al., 2005; Shao and Tseng, 2007; Tong and Zhao,

2008 など)．(註：これらはあくまで多重検定の検出力評価，または，サンプルサイズ設計を目的としており，以下に述べるようなデータから効果サイズ分布や ROC 曲線を推定し，カットオフを選択するという発想はない)．以上の平均検出力は，non-null のオミクス変数 “全体” での平均検出力であるので，全体検出力(overall power)とよぶことにする(Matsui and Noma, 2011a)．しかし，non-null のオミクス変数の中には，効果サイズが小さいものがかなり多く存在するため，全体検出力を計算すると，一般にがっかりするほど小さな値(例えば，< 20%)になることが多い．そもそも，微小な効果サイズをもつものも含め，non-null のオミクス変数全体の検出を試みることは到底無理なことであろう．そこで，一定以上の効果サイズをもつ non-null のオミクス変数に限定，あるいは，ターゲットとしたもとでの平均検出力である “部分” 検出力(partial power)(Matsui and Noma, 2011a)の使用が考えられる(具体的な設定法については以下の 3.3 節を参照)．大きな効果サイズをもつオミクス変数(遺伝子多型，遺伝子など)は，生物学的にもより重要と考えることができるのであれば，部分検出力の使用は妥当であろう．

一方，横軸の偽陽性度に相当する指標としては FDR を用いるのがよいだろう．その理由として，スクリーニングされたオミクス変数セットにおける偽陽性の割合として解釈しやすいということ，また，その使用は多くのオミクス研究ですでに十分確立していることなどがあげられる．

ROC 曲線を描くことにより，平均検出力(真陽性度)，FDR(偽陽性度)の関係を視覚的に捉えることができ，双方を考慮したカットオフの選択に役立つ有用な情報を得ることができる．オミクス研究でスクリーニングされたオミクス変数(遺伝子など)は，生物学的機序を調べるための機能解析，パスウェイ解析，あるいは，より簡便なプラットフォーム(例えば，臨床現場で普及している RT-PCR など)での検証研究の対象となる．これらの後続研究での研究リソース(資金，時間，マンパワーなど)の制約を考えると，あまりに多くのオミクス変数をスクリーニングすることは現実的でないだろう．その意味で，FDR は，多少厳しい水準も含んで，1–20%程度が一般に受け入れられる範囲だろう．一方，縦軸の平均検出力については，改めて，(全ての non-null オミクス変数に対する)全体検出力はかなり低い水準になることが多く，あまりおすすめできない．一定以上の効果サイズの制限をおいた部分検出力の使用がより現実的であろう．なお，効果サイズの制限の入れ方にはいろいろ考えられるので，いくつか部分検出力を定義できる．以下の 3.3 節でその一例を示す．なお，3.3 節は，統計モデル，推定法に関する少しテクニカルな内容を含んでいるが，ROC 曲線(全体検出力，部分検出力)のイメージは図 2 をみることでつかめるだろう．最終的には，FDR も踏まえ，共同研究者(生物学・医学の研究者など)と協議の上，カットオフを選択するとよい．

3.3 ROC 曲線の推定

上記の ROC 解析は，データから ROC 曲線を推定することで初めて可能になる．オミクス研究において ROC 曲線の推定問題はこれまでほとんど議論されてこなかった．その理由として，これまでの多重検定偏重の流れももちろんあると考えられるが，ROC 曲線の推定に役立つオミクスデータ全体に対するモデリングの枠組みが十分認識，整理されてこなかったこともあげられるだろう．

以下では，まず，ROC 曲線を推定するためのオミクスデータ全体のモデリングについて触れ，

その上で ROC 曲線の推定を考える．なお，事例を用いた方がより具体的で理解しやすいと思われるので，ここでは，Setlur et al. (2008) による前立腺がんの遺伝子発現マイクロアレー研究を考える．前立腺がんの悪性度や予後に関係することが知られているがん融合遺伝子（TMPRSS2-ERG）を有している $n_1 = 103$ 名の前立腺がん患者とこれを有していない $n_0 = 352$ 名の前立腺がん患者に対して，がん細胞の cDNA アレー解析（6,144 個のプローブセット）が行われた．ここでは，二つの患者グループ間で発現量の異なるプローブセット（以下，便宜的に遺伝子と呼ぶ）のスクリーニングを考えよう．

まず，遺伝子 j $(j = 1, \cdots, 6{,}144)$ に対して，効果サイズパラメータをグループ間の平均発現量の差 $\delta_j = (\mu_j^{(1)} - \mu_j^{(0)})/\sigma_j$ として定義する．ここで，$\mu_j^{(1)}$, $\mu_j^{(0)}$ はがん融合遺伝子保有グループ，非保有グループの平均発現量であり，グループ内での（共通の）標準偏差 σ_j で除して標準化している．これに対する統計量として，$Y_j = (\hat{\mu}_j^{(1)} - \hat{\mu}_j^{(0)})/\hat{\sigma}_j$ を用いる．なお，二標本 t 検定統計量は，$t_j = Y_j/\tau_n, \tau_n = \sqrt{1/n_1 + 1/n_0}$ と表されるので，実質，t 検定を用いることと同じである．

以下の混合モデルを仮定する．すなわち，Y の確率密度関数 $f(y)$ に対して，

$$f(y) = \pi f_0(y) + (1 - \pi) f_1(y)$$

を仮定する．ここで，π は null（関連なし）の事前確率である．f_0 は null のときの Y の分布（確率密度関数），f_1 は non-null（関連あり）のときの Y の事前分布である．f_0 に対しては，理論的に導かれる null 分布，例えば，漸近分布として $N(0, \tau_n^2)$，を指定するのが一般的であるが（theoretical null），$N(\gamma_0, \phi_0^2)$ とおいて，γ_0, ϕ_0^2 をデータから推定すること（empirical null）も考えられる（Efron 2004a）．事後確率 π と事前分布 f_1 は解析者が適当に指定してもよいが，2.2 節で述べたように，データから推定する経験ベイズのアプローチが有効である．なお，最尤推定では，ほとんどの場合，便宜的にオミクス変数間での独立性が仮定されるが，相関に対する一定のロバストネスが示されている（McLachlan, Bean, and Jones, 2006）．もし，生物学的な観点から相関を指定できるのであれば，これを取り入れた推定も可能である．以上の混合モデルは，FDR の評価を目的とした多重検定の議論で多く用いられてきたが，以下に与える全体検出力（Ψ）の評価にも使えることに注意したい．

いま，null/non-null の混合構造に加えて，non-null 分布 f_1 に階層を入れ，効果サイズ分布を導入する．

$$Y_j \mid \delta_j \sim N(\delta_j, \tau_n^2), \qquad \delta_j \sim g$$

ここで，g は効果サイズ分布に対応する．g に対して適当なパラメトリック分布を仮定してもよいが，（遺伝子数が十分大きいことから）ノンパラメトリック推定も現実的である．一般に，効果サイズ分布についての事前情報はほとんどないので，分布を仮定しないで済む後者のアプローチは魅力的である．周辺分布 f を適当なスプライン関数で推定し，逆変換により求める方法 (Efron, 2004b)，g に離散分布を想定した EM アルゴリズムである smoothing-by-roughening 法（Shen and Louis, 1999）を適用する方法（Matsui and Noma, 2011a, b）が提案されている．

効果サイズ分布 g にノンパラメトリック分布を仮定し，smoothing-by-roughening 法により推定した結果を図 1 に示す．図 1 (a) は，全 6,144 遺伝子の統計量 Y の相対頻度のヒストグラムに

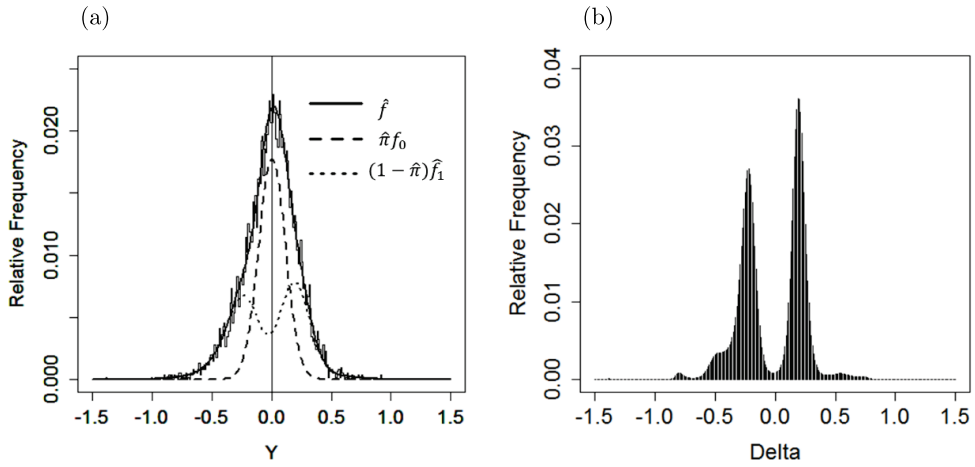

図 1. Setlur et al. (2008) の前立腺がん研究でのマイクロアレー遺伝子発現データを用いた融合遺伝子保有と非保有の比較．パネル(a)は，全 6,144 遺伝子の統計量 Y のヒストグラム，周辺分布 f, null コンポーネント πf_0，non-null コンポーネント $(1-\pi)f_1$ の推定(Y のグリッド：0.01)．パネル(b)は，推定されたノンパラメトリック効果サイズ分布 g.

対して，周辺分布 f, null のコンポーネント πf_0，non-null のコンポーネント $(1-\pi)f_1$ の推定値を示している．図 1 (b) は，効果サイズ分布 g の推定結果である．なお，null の事前確率 π の推定値は $\hat{\pi} = 0.6$ であった．

いま，遺伝子スクリーニングとして，統計量 Y に対して $|Y| \geq c\,(>0)$ のときに陽性と判定するルールを考えよう(両側検定に相当)．このとき，FDR，全体検出力 Ψ は，カットオフ c の関数として以下のように表される：

$$FDR(c) = \frac{\pi_0\{F_0(-c) + 1 - F_0(c)\}}{F(-c) + 1 - F(c)}$$

$$\Psi(c) = F_1(-c) + 1 - F_1(c)$$

ここで，F, F_0, F_1 は f, f_0, f_1 に対応する分布関数である．

次に，一定以上の効果サイズに限定，あるいは，ターゲットとした平均検出力である部分検出力を求めてみよう．明らかに，部分検出力の評価は，non-null 分布 f_1 に階層構造を考え，効果サイズ分布 g を導入することで可能になる．前立腺がんの事例において，効果サイズ δ に関して，マイナスの方向で $\delta \leq \eta_1\,(<0)$，プラスの方向で $\delta \geq \eta_2\,(>0)$ を満たすものを意味のある効果サイズ，または，検出したい効果サイズと定義する(つまり，この条件を満たす効果サイズ δ_j をもつオミクス変数 j の集まりが検出したいターゲットとなる)．定数 η_1，η_1 の指定については，例えば，効果サイズ分布 g の推定値の 10%点，90%点などとしてもよい．前立腺がんのデータでは，10%点は $\eta_1 = -0.34$，90%点は $\eta_2 = 0.26$ である．あるいは，ノンパラメトリック推定された効果サイズ分布の形状をみて，興味のあるピークを検出するように指定してもよいであろう．例えば，図 1 (b) の効果サイズ分布の推定結果をみると，マイナス，プラスそれぞれの方向で大きな効果サイズのピークがいくつかみられる．これらを検出するために，例えば，$\eta_1 = -0.60$, $\eta_2 = 0.42$

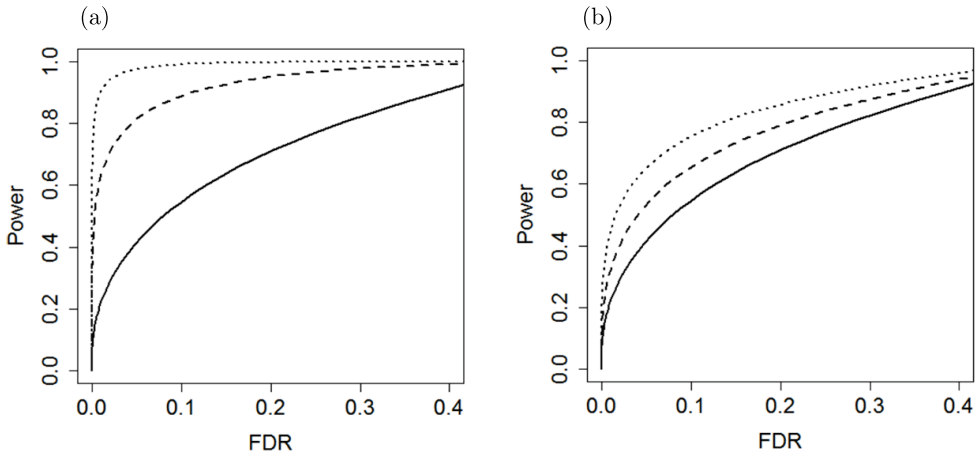

図 2. 前立腺がん研究データ．パネル(a)は，推定された ROC 曲線．実線は全体検出力，破線・点線は部分検出力(破線：$\eta_1 = 0.34$, $\eta_2 = 0.26$，点線：$\eta_1 = -0.40$, $\eta_2 = 0.40$)．パネル(b)は，サンプルサイズを変えたときの全体検出力の見積もり(実線：$n = 455$，破線：$n = 600$，点線：$n = 800$)．

などと指定してもよいだろう．部分検出力は，

$$\Psi_P(c;\eta_1,\eta_2) = \frac{\int_{|u|\geq c}\int_{\delta\leq\eta_1,\delta\geq\eta_2} g(\delta)\varphi_{\delta,\tau_n^2}(u)d\delta du}{\int_{\delta\leq\eta_1,\delta\geq\eta_2} g(\delta)d\delta}$$

と表現できる．ここで，$\varphi_{\delta,\tau_n^2}(\cdot)$ は，平均 δ，分散 τ_n^2 をもつ正規分布の密度関数である．図 2 (a) は，全体検出力，または，部分検出力を用いたときの ROC 曲線の推定結果である．例えば，FDR = 10% に指定すると，全体検出力は 50%程度に達する(註：サンプルサイズが比較的大きいこともあり珍しく高い水準といえる)．一方，部分検出力は，全体検出力よりもずっと高い水準にある．効果サイズが特に大きい上位遺伝子群を捉えたいのであれば，FDR = 5% 程度に厳しくしても 80%を超える部分検出力を達成できるだろう．

効果サイズ分布 g を推定することで，サンプルサイズ設計が可能になる．図 2 (b) は，π と g をそれらの推定値に固定したもとで，任意のサンプルサイズに対して，サンプルサイズ項である τ_n^2 を更新し，平均検出力を見積もったものである．具体的には，Setlur et al. (2008) において，TMPRSS2-ERG 融合遺伝子保有，非保有患者の割合($= n_1/n_0 = 103/352$)を一定としたもとで，全体のサンプルサイズ n を 455 ($= n_1 + n_0$) から 600，800 例と変化させたときの全体検出力の推定値を示している．このような解析は，例えば，Setlur et al. (2008) を参考に類似のオミクス研究を新たに計画し，そのサンプルサイズ設計を行う際に有用である．あるいは，Setlur et al. (2008) の研究が継続されているものとして，この種の検討を「中間解析」として実施し，望ましい検出力を達成するための「サンプルサイズ再設計」を行ってもよいであろう．

3.4 階層混合モデル・経験ベイズ推定：関連スクリーニング解析における更なる可能性

原論文 Setlur et al. (2008) では，選抜された遺伝子セットを用いた TMPRSS2-ERG 融合遺伝子の保有の判別が試みられている．オミクス研究後に選抜した遺伝子セットを用いて判別・予測システム(genomic signature)の開発を行う場合，選抜遺伝子セットに対してどの程度の判別能が

期待できるかを事前に評価できれば，後続の判別システム開発研究の go/no-go の判断や研究デザインの検討に役立つ．上記の階層混合モデルを用いた経験ベイズ解析では，選抜した個々の遺伝子に対して効果サイズを推定でき(縮小推定)，任意の遺伝子サブセットに対して遺伝子間の相関を考慮した判別精度の推定が行える．詳細については，Efron (2009), Matsui and Noma (2011b) を参照されたい．なお，よくみかける誤解であるが，統計的有意性と判別・予測能は全く別物である．意味のある予測能を得ることは，統計的有意性を得ることよりも格段にハードルが高い(例えば，Pepe 2005)．従って，genomic signature の開発を考えるのであれば，統計的有意性の解析だけでは全く不十分であり，上記のモデルベースの判別精度評価，あるいは，標準的な判別・予測解析による精度評価を実施する必要がある．

より複雑な状況への応用も興味深い．その一つは，患者・サンプルについて複数のサブグループが存在する場合である．例えば，がん患者の生存時間を新治療群と対照治療群で比較するランダム化臨床試験において，治療前(ベースライン)のがん組織に対してマイクロアレー遺伝子発現解析が実施されているとする(Matsui et al., 2012)．いま，治療群別に生存時間上での遺伝子発現量の効果を調べるとすると(遺伝子発現量を一つの共変量とした単変量 Cox 解析など)，発現量の効果が治療群で共通である遺伝子は予後マーカー(prognostic marker)，そうでないもの(つまり，治療-遺伝子の交互作用)は，治療効果予測マーカー(predictive marker)の候補とみなすことができる．上記の階層混合モデルを二次元に拡張し，新治療，対照治療の二群で求めた発現量の効果に対して同時モデリングを行い，効果サイズ分布をノンパラメトリックに推定する．そうすると，一言に，治療-遺伝子の交互作用といってもさまざまな効果プロファイルをもつものが存在することがわかるだろう．その中には，標準的な治療-交互作用の検定では十分検出できないものも少なからず含まれるだろう．推定された(ノンパラメトリック)効果サイズ分布に基づいて，尤度比検定に相当する optimal discovery procedure(Storey, 2007; Noma and Matsui, 2012)を構成することで，様々な効果プロファイルをもつ遺伝子(予後マーカー，治療効果予測マーカーの候補)を効率的に検出できる(Matsui et al., 2018)．以上の例は，柔軟な階層混合モデルを用いてオミクスデータに含まれるシグナル成分をありのままに推定することが関連のスクリーニングにおいて有効であることを示している．

3.5 P 値の意味・役割

改めて，個々のオミクス変数に対して実施される検定の P 値の役割・意味について考えてみよう．多重検定の枠組みでは，FWER を代表とする検定全体での偽陽性の指標を一定以下に抑えるために P 値(あるいは，有意水準)が調整される．そもそも P 値は単一の検定(任意の一つのオミクス変数の検定)の統計的有意性に関する指標である．そこに検定全体(オミクス変数全体)の偽陽性基準が反映され，個々の検定(一つ一つのオミクス変数の検定)の有意性を判定するための調整 P 値が作られる．すなわち，もし，関連スクリーニングを偽陽性のみの側面で行うならば，(検定全体の偽陽性度を反映した)調整 P 値を通してスクリーニングのカットオフを決めることは自然である．

一方，ROC 曲線解析の枠組みにおいては，スクリーニングのカットオフを決めるという意味での P 値の役割は半減する．なぜなら，真陽性も同時に考慮される枠組みだからである．ただし，

(P 値の隠れた役割としての)オミクス変数のランキングとしての役割は残している(ただし，ROC 曲線解析の枠組みでは必ずしも統計的有意性に基づいてランキングする必然性はないが)．

4. おわりに

本稿では，オミクス研究での関連スクリーニング解析に焦点をあて，その現状と統計的方法について整理した．GWAS を代表とする一部の研究での検証的解析とそれ以外の探索的解析に分けて議論したが，オミクス研究での関連解析は基本的に後続研究に向けての探索的解析と位置づけるのが妥当である．第 1 節で述べたように，本稿執筆の動機は ASA による P 値に関する声明(ASA 声明)と一部関係するが，第 3.5 節で議論したように，オミクス研究での関連スクリーニング解析(探索的解析)における P 値の役割・意義は限定的と考えられる．あるいは，関連スクリーニングを P 値に関係するものと捉えることは，かえって関連スクリーニングを誤って捉えてしまう危険性すらある．著者は，関連のスクリーニングは，「仮説検定の問題」というよりは，統計モデルを用いたオミクス全体の関連構造，または，ROC 曲線の「推定の問題」，そして，推定結果に基づく「選択の問題」と捉えるべきだと考えている．

最後に，(オミクス研究に限らず) P 値に関する ASA 声明(Wasserstein and Lazar, 2016)がもつ意味について少し私見を述べる．この声明の主な目的は，実質科学の多くの研究分野でみられる P 値についての誤解を解き，P 値のより適切な使用，あるいは，別の適切なアプローチ・方法の使用を促すことであることは明らかである．多くの研究分野において，この声明の内容が，統計家はもちろんのこと，統計を専門としない研究者・実務者にも十分認知され，統計的方法の実践が改善されることがゴールといえる．そこをめざすには，それぞれの研究分野別に，どのような P 値の誤解や誤用が起こりうるのか，具体的な事例とともに，可能な解決策も併せて提示することが第一歩と考えられる(オミクス研究の分野においては本稿で著者なりにこのことを試みたつもりであるが)．一般論のレベルで解決策を示すだけでは実効性は乏しいだろう．それぞれの研究分野において，統計家がその分野の研究者と一緒になって固有の問題に正面から深く向き合うことが肝要であり，今後統計家にはその努力が一層求められる ── ASA 声明にはそのような大きな意味が込められているように思われる．

参考文献

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology* **32**, 227–234.

Efron, B. (2004a). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99**, 96–104.

Efron, B. (2004b). Selection and estimation for large-scale simultaneous inference. *Technical Report* No. 2005-18B/232, Division of Biostatistics, Stanford University.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science* **23**, 1–22.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association* **104**, 1015–1028.

Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. New York: Cambridge University Press.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., and Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology* **32**, 179–185.

Maher B. (2008). Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21.

Matsui, S. and Noma, H. (2011a). Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. *Biometrics* **67**, 1225–1235.

Matsui, S. and Noma, H. (2011b). Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. *Biostatistics* **12**, 223–233.

Matsui, S., Noma, H., Qu, P., Sakai, Y., Matsui, K., Heuck, C., and Crowley, J. (2018). Multi-subgroup gene screening using semi-parametric hierarchical mixture models and the optimal discovery procedure: Application to a randomized clinical trial in multiple myeloma. *Biometrics* (In Press).

Matsui, S., Simon, R., Qu, P., Shaughnessy, J. D. Jr, Barlogie, B., and Crowley, J. (2012). Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clinical Cancer Research* **18**, 6065–6073.

McLachlan, G. J., Bean, R. W., and Jones, L. B. T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615.

Nelson, C. P., Goel, A., Butterworth, A. S., Kanoni, S., Webb, T. R., Marouli, E. et al. (2017). Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics* **49**, 1385–1391.

Noma, H. and Matsui, S. (2012). The opitimal discovery procedure in multiple significance testing: an empirical Bayes approach. *Statistics in Medicine* **31**, 165–176.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology* **32**, 381–385.

Pepe MS. (2005). Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine* **24**, 3687–3696.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Robert, C. P. (2007). The Bayesian Choice, Second Edition. New York: Springer.

Setlur, S. R., Mertz, K. D., Hoshida, Y., Demichelis, F., Lupien, M., Perner, S. et al. (2008). Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *Journal of the National Cancer Institute* **100**, 815–825.

Shao, Y. and Tseng, C. -H. (2007). Sample size calculation with dependence adjustment for FDR-control in microarray studies. *Statistics in Medicine* **26**, 4219–4237.

Shen, W. and Louis, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics* **8**, 800–823.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B* **69**, 347–368.

Tong, T. and Zhao, H. (2008). Practical guidelines for assessing power and false discovery rate for a fixed sample size in microarray experiments. *Statistics in Medicine* **27**, 1960–1972.

Tsai, C.-A., Wang, S. J., Chen, D. T., and Chan, J. J. (2005). Sample size for gene expression microarray experiments. *Bioinformatics* **21**, 1502–1508.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *Journal of the National Cancer Institute* **96**, 434–442.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genetic epidemiology* **33**, 79–86.

Wasserstein R. L. and Lazar N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician* **70**, 129–133

Westfall, P. H. and Young, S. S. (1993). Resampling-Based Multiple Testing. New York: Wiley.