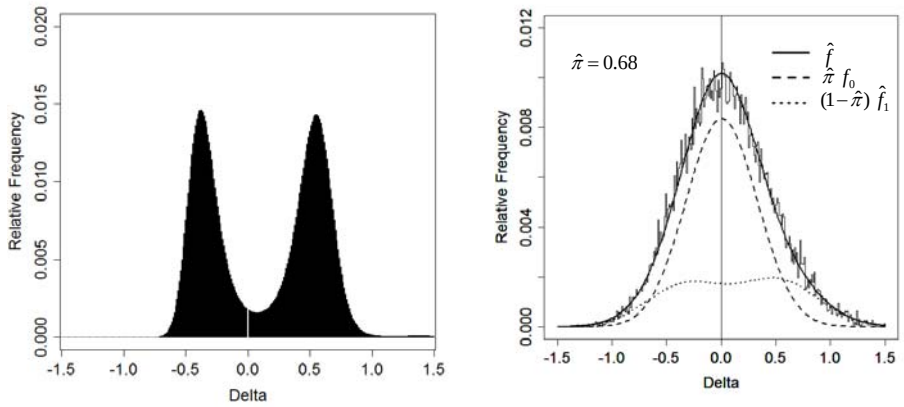
階層混合モデルと経験ベイズによるゲノムデータの解析

多くのゲノム解析研究では関連遺伝子のスクリーニングが行われます．例えば，マイクロアレーを用いた遺伝子発現解析では，一つ一つのプローブセットまたは遺伝子について表現型（病型，臨床アウトカムなど）と発現量との関連性が検定され，そこで有意となったものが関連遺伝子の候補となります．しかし，数万の検定を繰り返すことで深刻な偽陽性と偽陰性（低検出力）の問題に直面します．この問題を改善するための有効なアプローチは，一つ一つの遺伝子を別々に解析するのでなく，全遺伝子の背後にある共通の構造をモデリングし，これを抽出することです．

具体的には，階層混合モデル（hierarchical mixture models）を用いて，全遺伝子をごく一部の関連遺伝子と残りの（関連なしの）遺伝子に分け，関連遺伝子では関連の大きさがある分布に従っていると仮定します．ここで，関連の大きさの分布は特に指定しません（ノンパラメトリックな指定）．このモデルは EM アルゴリズムを用いてデータに基づいて推定できます（経験ベイズ）．この解析そのものは推定の方法であり，解析対象である遺伝子の数が増えるほど推定精度が改善します．つまり，遺伝子数が多いほど好都合といえます．これは深刻な偽陽性や過適合（p>>n）の問題に直面する多重検定や統計的機械学習（予測解析）とは対照的です．



小児白血病の予後関連遺伝子の検出(Kirschner-Schwabe et al. Clin Cancer Res 2006;12:4553-61)．22,283 遺伝子の内の約 32%(=100 – 68%)が予後関連遺伝子と推定．左図は関連遺伝子での関連パラメータ（関連の大きさ）の推定分布．右図は全遺伝子での頻度分布と推定分布（実線）．さらに，推定分布を関連ありとなしの遺伝子の分布に分解．

いったん全遺伝子のモデルを精度よく推定できれば，このモデルから導かれる解析ツールは最も有効なものと考えられます．関連遺伝子のスクリーニング[1]のみならず，遺伝子のランキング[2]，さらには，判別解析での判別精度推定[3]において，従来法よりも性能の良い方法が得られます．遺伝子スクリーニングでの偽陰性（低検出力）の問題に対する解決策としてのサンプルサイズ推定においても同様です[1]．現在は，がんの単群第 II 相試験での奏功率の推定，第 III 相ランダム化試験での治療効果関連遺伝子と予後関連遺伝子の同時検出などの解析法を開発中です．Lasso や boosting などでの stage-wise 変数選択での応用など，統計的機械学習との融合も興味深いテーマです．

以上のように，階層混合モデルと経験ベイズ解析は高次元ゲノムデータに対する有効な解析の枠組みを与えるものであり[4]，今後この枠組みでの多くの研究が期待されます.

文献：

[1] Matsui S, Noma H. Estimating effect sizes of differentially expressed genes for power and sample size assessments in microarray experiments. Biometrics 2011; 67: 1225-1235.

[2] Noma H, Matsui S. Empirical Bayes ranking and selection methods via semiparametric hierarchical mixture models in microarray studies. Statistics in Medicine 2013; 32: 1904-1916.

[3] Matsui S, Noma H. Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. Biostatistics 2011; 12: 223-233.

[4] Noma H, Matsui S. Gene Screening in the development of genomic signatures: beyond multiple testing. In Design and Analysis of Clinical Trials for Predictive Medicine (eds. Matsui S, Buyse M, Simon R), CRC Press (To appear in 2014).