Cancer Outlier Analysis Based on Mixture Modeling of Gene Expression Data

盛 啓太

博士論文

総合研究大学院大学
複合科学研究科
統計科学専攻

2014 年 9 月

目次

図目次

表目次

1　要旨

生物学的にがんの発生は分子レベルでの異常が関連していると考えられている．また，あるがん疾患に対して，表現型が同様でも，分子レベルでは全く別の疾患であるという異質性が報告されている．あるがん関連遺伝子に対して，あるがんサンプルでは，遺伝子発現量が正常サンプルとくらべて高発現，または，低発現しているが，他のがんサンプルでは正常サンプルとほぼ同じ発現量であるといったことが起こりうる．このような様々な遺伝的機序によって同一または類似の表現型となる事を genetic heterogeneity という言葉で表す．

近年，分子レベルで多数のがん関連遺伝子候補を同時に調べることができるハイスループット技術が広まりを見せている．これにより遺伝的異質性を検出することでがん関連である可能性が高い遺伝子候補の選抜が可能となってきた．ハイスループット技術の応用としてよく用いられるのはマイクロアレイを使った実験である．このマイクロアレイ実験の結果を用いてがんに関連が強いと考えられる遺伝子の選抜を行うことができる．ただし結果から意味のある情報を引き出すためには統計解析が必要である．

この種のがん関連遺伝子同定のためのスクリーニングには解析手法として差の検定を用いることが多い．さらに遺伝子は多数存在するので，多重検定の枠組みで議論されてきている．この統計解析は単に二群の多重検定を適用すれば解決できるものではない．これは，差の検定手法としてよく用いられる t 検定は群間の一様な差を検出する方法であり，一部のがんサンプルでのみ正常サンプル発現量より高発現，または，低発現が見られる様な遺伝子では検出力が低くなるためである．このように，ある遺伝子において，一様に正常サンプルよりがんサンプルの中で高発現，または，低発現しているのではなく，一部のがんサンプルでのみ正常サンプルよりも高発現，または，低発現している発現量を先行研究では Cancer Outlier と呼んでいる．

先行研究では，遺伝子ごとに標準化を行い，がんサンプルを大きな発現量から並べ替え，あらかじめ決めておいた Cancer Outlier とする閾値となる分位点を適用し，その発現量をその遺伝子の統計量とする手法（COPA 統計量）．また，箱ひげ図の考えを利用し，遺伝子発現量から四分位範囲を算出し，がんサンプル発現量の 75 パーセンタイル点からさらに四分位範囲の大きさより大きな値を取る発現量を足し合わせる手法（OS 統計量）．

OS 統計量と同様の考え方で，四分位範囲を正常サンプルのみから作成する手法（ORT 統計量）．また，恣意性を排除するため，遺伝子毎の標準化後，がんサンプル発現量を大きな値を持つものから並べ替え，一番大きな値のみが Cancer Outlier となるときの統計量，1 番大きな発現量と 2 番目の発現量が Cancer Outlier という形ですべての場合を考える．さらに，この考え方の場合は中央値までは数が増えれば統計量が大きくなるので，順序調整数を導入して比較し最大の値を遺伝子の統計量とする MOST 統計量があった．

提案されてきた統計量は遺伝子レベルで多重検定をするための統計量としての開発が主流であった．しかし，バイオロジストはどの遺伝子ががんに関連が強いかが分かったとき，どのサンプルが Cancer Outlier となる発現量と判断されたのかも知りたいはずである．そのため我々はこの問題に対して，がん関連遺伝子の同定だけでなく，Cancer Outlier と判断された発現量に関しても定量的に比較可能となるような統計量を考えた．まず遺伝子内のがんサンプル発現量を，遺伝子内の正常サンプル発現量のデータを使って標準化する．その後，すべての遺伝子，すべてのがんサンプル発現量を通して，発現量の分布に関する共通のモデルを仮定する方法である．具体的なモデルはパラメトリックに 3 コンポーネントの正規混合モデルを用いることを考える．3 つのコンポーネントはそれぞれ正常発現量，負の Cancer Outlier，正の Cancer Outlier を表している．それぞれコンポーネントの密度関数において，分散は 1 に固定し，平均は 0，δ_1，δ_2 としそれぞれ負の Cancer Outlier コンポーネント，正の Cancer Outlier コンポーネントを表す．正規混合モデルの未知パラメータである δ_1，δ_2，混合割合と一緒に EM アルゴリズムを用いて推定する．EM アルゴリズムとは確率モデルのパラメータを最尤法に基づいて推定する手法の一つである．この推定値を用いて，それぞれのがんサンプル発現量が得られたもとでの Cancer Outlier であることの事後確率を計算できる．この事後確率を用いて，遺伝子レベルの統計量を作成する事を考える．その統計量は発現量が得られたもとでそのサンプルが Cancer Outlier でない確率を計算し，それを遺伝子内のがんサンプル全てで掛け合わせる．それを 1 から引くことで，Cancer Outlier でない発現量であれば小，一つでも入っていれば 1 近づく数値となる．

我々は，従来法と提案法の比較を行うために，モンテカルロ・シミュレーションを行った．シミュレーションにおいて遺伝子数を 1 万，サンプルを正常サンプル，がんサンプルそれぞれ 20，40，100 とした．全遺伝子数に対する関連なし遺伝子，関連あり遺伝子（高発現・低発現）の割合をそれぞれ 0.6，0.2，0.2 した．関連あり遺伝子のがんサンプル中の

Cancer Outlier の割合は 0.1，0.3，0.5 とした．正常サンプル発現量と Cancer Outlier サンプルの平均値の差は-2，2 としてそれぞれ低発現，高発現を示した．このようなシナリオに対して，従来法と提案法の統計量を算出した．比較には横軸に偽発見率，縦軸に検出力をプロットした ROC 曲線を用いた．

提案法は，従来法よりも多くのシナリオにおいて任意の偽発見率のとき，高い検出力を示していた．ただしデータにおける正規性の仮定が崩れ，さらに Cancer Outlier の数が少なかった場合に従来法の方が小さな偽発見率のとき高い検出力を示していた． t 統計量に基づく遺伝子発現の場合はがんサンプル内の Cancer Outlier の割合が小さなとき，小さな検出力であった．しかし，がんサンプル内の Cancer Outlier の割合が大きな値のとき，検出力は大きな値に改善した．これは前述しているように t 統計量が一様な差を検出することを得意とする統計量であるからと考えられた．COPA 統計量や OS 統計量はがんサンプル内の Cancer Outlier の割合が大きなとき悪いパフォーマンスを示していた．ORT 統計量と MOST 統計量はシナリオ全般を通してよい検出力を示していた．しかし，提案法には及ばなかった．シミュレーションデータを t 分布から発生していると考えたときと同様の傾向が観察されていた．

実データに関しては一般に公開されている血液腫瘍のデータを用いた．このデータは骨髄異形成症候群 139 例と白血病でない 69 例のマイクロアレイ実験からのものとなっていた．探索候補は 54675 遺伝子であった．このそれぞれに対して従来法と提案法で統計量を算出した．その後，各統計量のにおいて，がん関連が高いとされる上位 200 遺伝子をピックアップした．さらに，それぞれの手法での上位遺伝子を照合し共通している遺伝子の個数を確認した．

これにより，従来法で上位で検出されている遺伝子のいくつかが同様に検出できていることが確認できた．さらに提案法は従来法のどれかに似た遺伝子を検出する傾向にあるのではなく，検討にあげた統計量すべてと重なる遺伝子が満遍なく一定数存在しているということがわかった．また，提案法では選ばれたが，従来法では選ばれなかった遺伝子もあった．

以上のことから我々は，シミュレーションベースでがん関連遺伝子のスクリーニングにおける従来法より高い検出力の統計量を提案できた．また，実データに適用することで，これまで検出されなかったプロファイルのがん関連遺伝子候補を検出することが出来た．これにより我々はがんの新しい疾患分類の開発や創薬により貢献できると考える．

2 序論

2.1 遺伝的異質性を考慮した新薬開発

生物学的にがんの発生は分子レベルでの異常が関連していると考えられている．また，あるがん疾患に対して，表現型が同様でも，分子レベルでは全く別の疾患であるという異質性が報告されている．あるがん関連遺伝子に対して，あるがんサンプルでは，遺伝子発現量が正常サンプルとくらべて高発現，または，低発現しているが，他のがんサンプルでは正常サンプルとほぼ同じ発現量であるといったことが起こりうる．このような様々な遺伝的機序によって同一または類似の表現型となることを genetic heterogeneity という言葉で表す．遺伝的異質性を扱う研究は数多く存在し，pubmed で検索すれば，1955 年の Dempster の研究ですでにその言葉が表題で用いられていた [4]．異質性 という言葉であれば，1905 年の Torrey JC の赤痢の研究ですでにその言葉が文中で用いられていた [19]．しかし遺伝的異質性に着目した研究はあまり行うことが出来なかった．これは分子レベルで考えるとがん疾患機序に関与する可能性が否定出来ない因子が数千，数万と存在したからである．それをランダムにスクリーニングするよりは，殺細胞性が知られている化合物を候補とするほうがよいと考えられていた．

近年，分子レベルで多数のがん関連遺伝子候補を同時に調べることができるハイスループット技術が広まりを見せている．これにより遺伝的異質性を検出することでがん関連である可能性が高い遺伝子候補の選抜が可能となってきた．ハイスループット技術の応用としてよく用いられるのはマイクロアレイを使った実験である．マイクロアレイは多数の DNA 断片をプラスチックやガラス等の基板上に高密度に配置した分析器具のことである．マイクロアレイ実験では細胞内の遺伝子発現量を測定することができる．このマイクロアレイ実験の結果を用いてがんに関連が強いと考えられる遺伝子の選抜を行うことができる．選抜された候補遺伝子に対しては実際に研究室で実験が行われる．実験の結果によっては，臨床応用に向けての研究に進む．この時，がん関連が同定できた分子をがんの分子標的と呼ぶ．分子標的を同定してからそれをもとに開発された薬を分子標的薬と呼ぶ．

以下に例として血液腫瘍に関連する Abl チロシンキナーゼと肺がんに関連する ALK 融

合遺伝子について述べる．

2.2 Abl チロシンキナーゼ

Abl チロシンキナーゼ阻害剤としてイマチニブ（グリベック）が挙げられる．グリベックは第一世代 Abl チロシンキナーゼ阻害剤と言われている．本国，厚生労働省医薬品医療機器情報ホームページにあるグリベックのインタビューフォームによれば，1992 年にスイスの製薬会社によってチロシンキナーゼ活性を選択的に阻害する候補物質からイマチニブが選択された．イマチニブの臨床試験はインターフェロンアルファ不応，または，不耐の慢性骨髄性白血病患者に対して行われた．さらに KIT(CD117) 陽性の消化管間質腫瘍にも行われ，フィラデルフィア染色体陽性急性リンパ性白血病に対しても行われた．その結果，一定の有効性があると判断し規制当局への承認申請を行い無事認可されている．この期間は 10 年弱とこれまでの治療開発期間からすると短時間で行われている．ただし対象となっている患者でも，IRIS という臨床試験の結果 [11] からイマチニブを投与されてもイマチニブの毒性や有効性が原因で約 20% が離脱しているということがわかった．つまりこの部分集団はイマチニブを投与すべき集団とされていたにもかかわらず投与できない異質な集団であったのでこの対象への治療開発が求められる．そこで第二世代チロシンキナーゼ阻害剤としてニロチニブとダサチニブが開発され日本でも承認されている．同様の理由で第一世代，第二世代が全く効果が無い対象に対しても第三世代チロシンキナーゼ阻害剤の開発が進んでいるところである．

2.3 ALK 融合遺伝子

肺がんにおいては，EGFR(Epidermal Growth Factor Receptor) 遺伝子変異がしられている．これはある研究によれば全肺がん患者の 40% で変異が起こっているという報告がある [6]．最近の例であれば，日本からは JST 課題達成型基礎研究の報告から知ることが出来る [16]．肺がんはこれまでも病理組織などで予後の違いが知られているなど，比較的治療戦略の検討が進んでいるがん種である．これに次いで，間野らは 2007 年に肺腺がんの細胞から肺がんの原因となる EML4-ALK 融合遺伝子を発見した [14]．この融合遺伝子は肺がんの分類の 1 つである非小細胞肺癌の 4% から 6%[15] で見られ，これまでの研究から，臨床情報などと付き合わせた結果，若年性の肺腺がんで多い (約 35%) という

ことが知られている [16]．さらに研究は続けられ ROS1 融合キナーゼ遺伝子を肺がんにおいて発見，さらに KIF5B-RET 融合キナーゼががん化能を持っていることを確認し，RET 阻害剤を用いてがん化の進行を抑えることに成功した [16]．このように上げられた遺伝子は肺がんに罹患したすべての人が同様に持っている遺伝子変異や融合ではなく，肺がん患者の一部にしか現れないものである．

2.4 ゲノムワイドデータの解析

分子標的を同定するためには，ハイスループット技術で得ることが出来たゲノムワイドデータの統計解析が必要となる．このデータ解析では解決しなければならない問題がある．例えば遺伝的異質性によるものである．ある遺伝子において，一様に正常サンプルよりがんサンプルの中で高発現，または，低発現しているのではなく，一部のがんサンプルでのみ正常サンプルよりも高発現，または，低発現している発現量があったとする．先行研究は，このようながん発現量を Cancer Outlier とよんでいる．この Cancer Outlier 型のプロファイルを持つ遺伝子の正常サンプルとがんサンプルにおいて発現量の差の検定をおこなおうとすると，検出力が低くなることが知られている．このため Cancer Outlier 型のプロファイルを検出できる検定統計量の研究が行われてきている．マイクロアレイのような高次元データにおけるスクリーニングでは偽陽性が深刻な問題である．様々な研究では，この数万の遺伝子を同時に検定すると考え，多重検定の枠組みで議論していることが一般的である．マイクロアレイ研究は統計的な観点からは多重検定で用いられる偽発見率 (false discovery rate: FDR) のコントロールが重要と考えられている．また，遺伝子のランキングで結果を出力するようなことも行われている．さらにデータに異常発現と正常発現を仮定し混合分布として扱うこととしたり，その混合分布のパラメータに事前分布を仮定して階層型混合モデルを考えるようなことも行われている [5]．

このような背景から，この論文では，これまで提案されてきている，がんに関連する可能性が他の候補よりも高い遺伝子を選抜するための統計手法よりもさらに検出力の高い統計量を提案したい．

2.5 論文の構成

本論文は次の構成となっている．第 2 章では，序論として研究背景と動機についてまとめた．第 3 章ではこれまでの Cancer Outlier 解析，特に，多重検定に関する既存の研究をレビューする．第 4 章では，提案法である正規混合モデルを用いた Cancer Outlier 解析について述べる．第 5 章では，いくつかのシナリオのもとでのモンテカルロシミュレーションを通して，提案法と従来法の性能比較を行う．第 6 章では骨髄異形成症候群の実データへの適用を行った．第 7 章でまとめと考察を与える．

3 Cancer Outlier の検出法： これまでの研究

3.1 Cancer Outlier の定義

Cancer Outlier は Tomlins ら (2005)[18] により導入された概念であり，その論文では，前立腺がんの予後を規定する遺伝子を同定するための手法として提案された．その中でも例えば組織型や分化度で分類され予後の良さを予測してきた．しかし，これまで行われている分類を使ったとしても，まだ治療効果において異質性を持つ集団であったが，がんが遺伝子に関する疾患であったということがわかってもなお，そのがんに関連する遺伝子を同定するために少なからず可能性があるものを候補として上げると，多数の遺伝子が上がってくる．

このがん関連遺伝子同定問題の難しさは，現在の疾患概念においてがんと診断されている人すべてで遺伝子が特異な発現をしているわけではないというところである．つまり，簡単な模式図を用意して説明すると，図 1 では正常サンプルとがんサンプルのデータが取られており，青色が正常サンプルやがんに関連しない発現量を表している．赤色ががんに関連しないサンプル発現量に比べて高発現，緑色が低発現であると考える．ここで GeneC はがんサンプルの発現量も青色のままなので，がん関連遺伝子ではないことを示している．

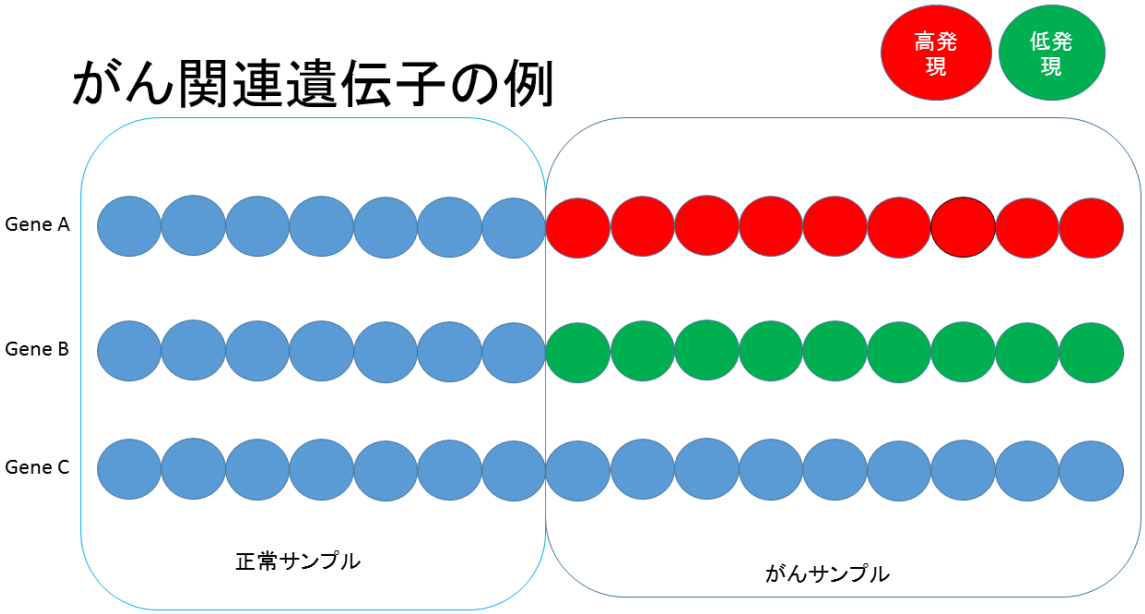

図 1　がんサンプル全てで正常サンプルに対して高発現，または，低発現している模式図

そして，GeneA や GeneB のようにがんサンプル全体で正常サンプルよりも高発現，または，低発現しているような状況も考えられるが，実は少ないということが上でも書いたとおりであり，候補の中にあったとしても，遺伝子によってはがんサンプルで，細胞増殖能に影響を与える，または，増殖能を抑制しているというような性質を持たず，正常サンプルと同じ振る舞いをしているものが存在するということが考えられる．このように，がんサンプルの一部の発現量で，正常サンプル発現量よりも高発現，または，低発現をしている発現量のことを Cancer Outlier と定義している．この模式図を図 2 に示す．

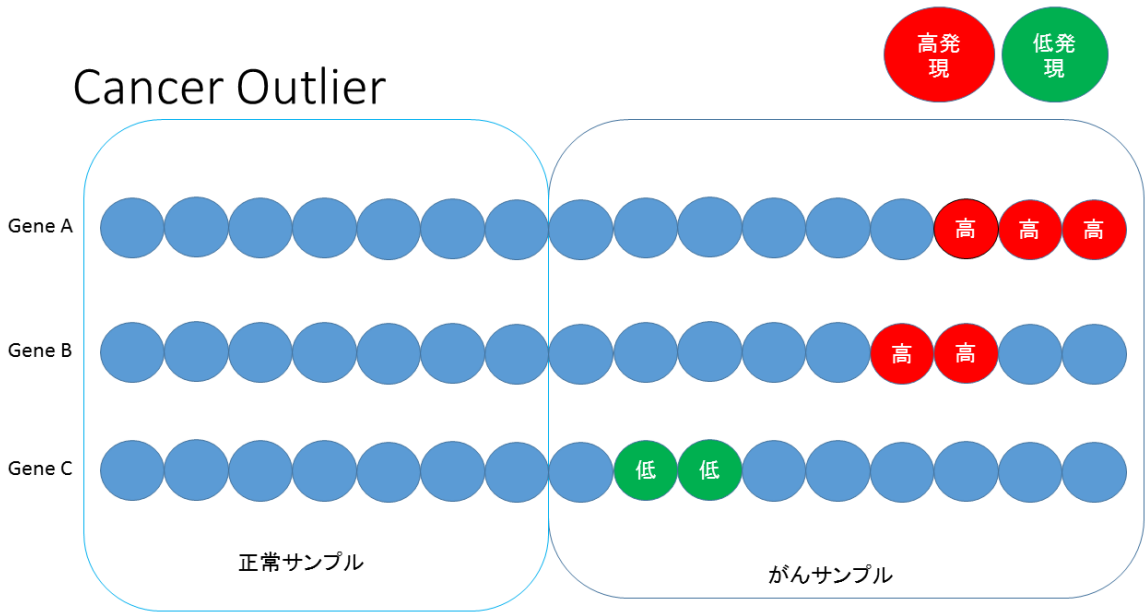

図 2　Cancer Outlier を含むがん関連遺伝子の模式図

つまり，同じ遺伝子のなかでそれぞれのサンプルから取られている発現量があり，正常サンプルの発現量をコントロール群とみたとき，がんサンプルの発現量の中には正常サンプルの発現量と同様の振る舞いをしているものや，高発現，低発現しているものがあり，その高発現や低発現している発現量を Cancer Outlier と呼ぶ．Cancer Outlier 検出問題は統計学的には，数万の遺伝子に対して同じ統計量を考え，そのときの多重検定問題であると帰着できる．先行研究では，n 個のサンプルの遺伝子発現量のデータからなり，それぞれのサンプルに対して莫大な G 個 (数千から数万) の遺伝子からがん関連遺伝子を同定するためのマイクロアレイ研究を考えている．このとき，n 個のデータは n_0 人分が正常なサンプルであり，n_1 人分はがんサンプルであるとしている．ここでは，実験を行って出来た生データに対して，対数の比を正規化した two-color cDNA アレイからのデータや，オリゴヌクレオチドアレイからのシグナルに対数をとったものを正規化したようなデータが遺伝子発現データとして想定されている．遺伝子 $g(g = 1, \cdots, G)$, において，x_{gi} はサンプル $i(i = 1, \cdots, n_0)$ の正常サンプルの発現量とし，y_{gj} はサンプル $j(j = 1, \cdots, n_1)$ としてがんサンプルの発現量をあらわすこととする．このとき，それぞれの遺伝子発現量において正常サンプル，がんサンプルで差があるかどうかを見るためには，単純に伝統的

な二群の差の検定を行うことを考えるのは自然である．これに対して，数万の遺伝子において，同様に検定を行うような状況であるので，その検定を何度も独立に行うという多重検定の枠組みを考えるというのが先行研究において用いられている枠組みである．実際には高発現や，または，低発現であることが考えられるが，先行研究では高発現，または，低発現のどちらか一方のみにしか Cancer Outlier がないという状況での議論となっている．つまり二群の検定では片側検定を考えていると考えることができるが，高発現，または，低発現である場合は，片側検定を二回考えることで対応できる．つまり，片側のときの性質がわかれば，一般性を欠くことなしに，我々は過剰発現や発現抑制が同じ遺伝子の一部で同時に見られる Cancer Outlier を含むものを選抜することも出来る．

3.2 t 統計量

従来法を提案するそれぞれの論文の書き出しでも触れられているが，一般的にある二群の平均値に差があるかどうかを検定するためには，伝統的な二標本 t 検定が用いられる．今回の問題に適用することを考えれば，g 個のそれぞれの遺伝子に対してがんサンプル，正常サンプルを用いて統計量を計算することになるので，定式化すれば，

$$t_g = \frac{\bar{y}_g - \bar{x}_g}{\bar{s}_g}, \tag{1}$$

となる．ここで，$\bar{y}_g$ は遺伝子 $g(g = 1, \cdots, G)$ に関してのがんサンプルの平均発現量，$\bar{x}_g$ は遺伝子 g に関しての正常サンプルの平均発現量である．そして，$\bar{s}_g$ は遺伝子 $g(g = 1, \cdots, G)$ においてのがんサンプル，正常サンプルの二群をプールしたときの標準偏差である．ただ，ここで t 検定は一方の群のサンプル発現量が，他方の群のサンプル発現量よりも「一様」に高発現，または，低発現するときに検出力の高い方法であり，Cancer Outlier のように一部でしか他方の群の発現量より，高発現，または，低発現しないような遺伝子の場合は検出力が低いということが報告されている [18]．さらに今回のように多重検定を考えているときは，深刻な検出力の低下が起こることが知られているので，これを改良できないかということが先行研究のモチベーションとなっている．

3.3 COPA 統計量

Tomlins ら (2005)[18] は COPA 統計量を考えた．COPA は Cancer Outlier Profile Analysis の略である．彼らが提案した方法ではまず遺伝子内発現量の標準化に正常サンプルがんサンプルすべてをプールしたときの中央値と絶対中央偏差を用いていた．そもそも Cancer Outlier を含むことを前提としているので，標本平均よりもロバストである代表値を用いたと考えられる．また，標準化を行った後，がんサンプルの発現量を大きな方から並べ替え，あらかじめ決めておいた Cancer Outlier とする閾値となる分位点を適用し，その発現量をその遺伝子の統計量とする．例えば，標準化を行った後，90% 点を閾値とするというような形であれば，がんサンプル発現量が 10 個あれば一番大きな方から 2 番目を統計量とする．これは高発現のときの方法だが，同様に考えることで低発現に対しての Cancer Outlier も考えることが出来る．

$$Copa_g = \frac{q_r(y_{gj} : 1 \leq j \leq n_1) - med_g}{mad_g}. \tag{2}$$

ここで $q_r(\cdot)$ は発現量の $r\%$ 点であり， med_g は遺伝子 g においてすべてのサンプルからの発現量の中央値，そして， mad_g は同様に遺伝子 g において遺伝子内すべてのサンプルの絶対中央偏差となっている．

$$med_g = \text{median}(x_{gi}, y_{gj}; i = 1, \cdots, n_0, j = 1, \cdots, n_1).$$
$$mad_g = 1.4826 \times \text{median}(|x_{gi} - med_g|, |y_{gj} - med_g|; i = 1, \cdots, n_0, j = 1, \cdots, n_1).$$

なお，ここで 1.4826 という数字が出てくるが，これは mad を用いて標準偏差を推定するときに用いられるスケールパラメータであり，分布に依存している [13]．mad を MAD と表すとして X を正規分布に従う確率変数，μ を X の平均とする．この場合

$$\frac{1}{2} = Pr(|X - \mu| \leq MAD) \tag{3}$$
$$= Pr(|\frac{X - \mu}{\sigma}| \leq \frac{MAD}{\sigma}) \tag{4}$$
$$= Pr(|Z| \leq \frac{MAD}{\sigma}) \tag{5}$$

これより

$$\frac{1}{2} = Pr(|Z| \leq \frac{MAD}{\sigma}) \tag{6}$$

から

$$\Phi(\frac{MAD}{\sigma}) - \Phi(\frac{-MAD}{\sigma}) = \frac{1}{2} \tag{7}$$

これより

$$\Phi(\frac{-MAD}{\sigma}) = 1 - \Phi(\frac{MAD}{\sigma}) \tag{8}$$

なので

$$\frac{MAD}{\sigma} = \Phi^{-1}(\frac{3}{4}) \tag{9}$$

つまり

$$\sigma = \frac{1}{\Phi^{-1}(\frac{3}{4})} * MAD \tag{10}$$

より

$$K = \frac{1}{\Phi^{-1}(\frac{3}{4})} = 1.4826 \tag{11}$$

$q_r(\cdot)$ の r の値は Cancer Outlier と判断する閾値であり，これは研究者で決めることとしている．例えば $r = 75, 90,$ や 95 という値を用いることが多い．

この COPA 統計量では，標準化されたサンプルにおける $r\%$ 点の値を用いており恣意的である．また，カットオフ値を固定することで，すべての遺伝子で Cancer Outlier の個数が一定であるという仮定を暗においていることになっている．それを改善するために，次の統計量が考えられた．

3.4 OS 統計量

Tibshirani ら (2007)[17] では，新たに OS 統計量が提案された．OS は Outlier Sum の略である．ここでは，遺伝子内のサンプル全体の中央値と絶対中央偏差によって標準化

を行い，以下のように定義された．

$$OS_g = \frac{\Sigma_{j \in R_g}(y_{gj} - med_g)}{mad_g}. \tag{12}$$

ここで遺伝子 g における Cancer Outlier の集合を R_g とし，

$$\begin{aligned} R_g = \{j : y_{gj} > q_{75}(x_{gi}, y_{gj} : i = 1, \cdots, n_0; j = 1, \cdots, n_1) \\ + IQR(x_{gi}, y_{gj} : i = 1, \cdots, n_0; j = 1, \cdots, n_1)\} \end{aligned} \tag{13}$$

と定義した．ここで IQR はデータの四分位範囲であり， $IQR = q_{75} - q_{25}$ とかける．q_{25}, q_{75} はそれぞれ 25% 点 75% 点である．OS では，このように統計量を定義することで，全体から相対的にはずれた発現量がなければ，0 になる．つまり，COPA のようにパーセンタイル点のみを関連あり遺伝子としての情報とするよりも，更に情報を有効活用していると考えられ，それにより Cancer Outlier を検出しやすい統計量となると考えられる．模式図を図 3 で示す．

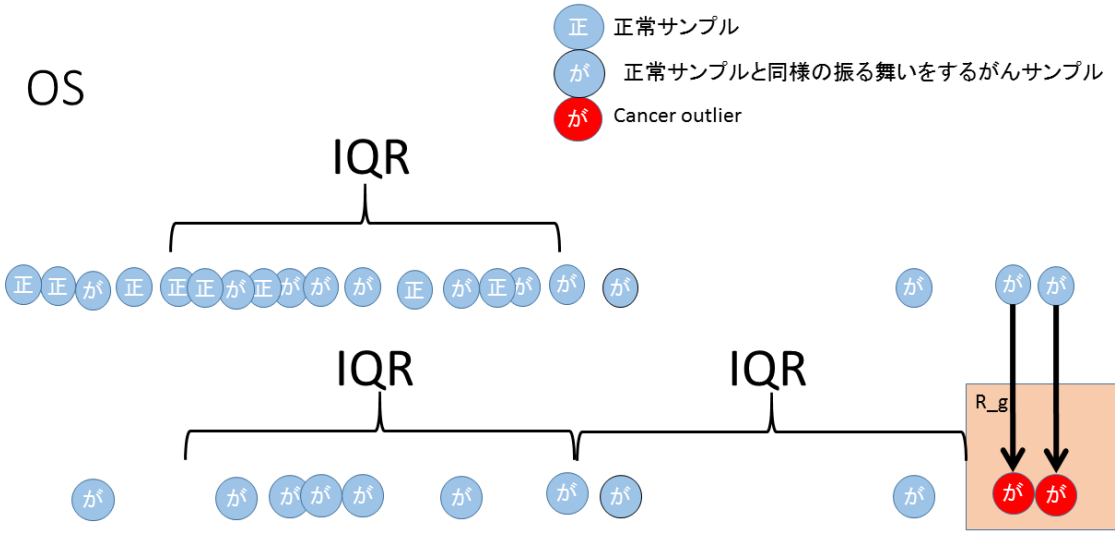

図 3　OS 統計量作成を説明する模式図

この図 3 では青で正常サンプル，または，がんサンプルであってもその遺伝子ではがん

関連を示さない様な発現量を示すサンプルを示している．上部で IQR の作成部分を表しており，右になるほど大きな発現量をもつサンプルが並んでいると考える．さらに下部では正常サンプル，がんサンプルすべてを下部で R_g に入る遺伝子を考えるために上部から正常サンプルを取り除いたがんサンプルのみを並べている．

3.5 ORT 統計量

Wu(2007)[22] は先の OS 統計量の提案を改良する形で，ORT 統計量を提案した．ORT は Outlier Robust T-statistics の略である．ここで ORT という用語自体に統計量の意味が入っているが，他の手法と合わせるため，ORT 統計量という用語を用いることとする．この提案では，正常サンプル発現量，がんサンプル発現量すべてプールしたところから Cancer Outlier を定義するのではなく，あくまでも正常サンプルの振る舞いから乖離しているがんサンプルを Cancer Outlier とする方法を考えた．このようにすることで，Cancer Outlier が正常サンプル発現量からの乖離を指標としていることがより明確になり，がんサンプルの中で正常サンプルと同様の振る舞いをするサンプルからの影響を受けにくくなるという利点がある．

$$ORT_g = \frac{\Sigma_{j \in O_g}(y_{gj} - med_{g,x})}{mad_g}. \tag{14}$$

ここで，

$$O_g = \{j : y_{gj} > q_{75}(x_{gi} : i = 1, \cdots, n_0) + IQR(x_{gj} : i = 1, \cdots, n_0)\}$$

$$med_{g,x} = \text{median}(x_{gi}; i = 1, \cdots, n_0)$$

$$med_{g,y} = \text{median}(y_{gj}; j = 1, \cdots, n_1)$$

であり，

$$mad_g = 1.4826 \times \text{median}(|x_{gi} - med_{g,x}|, |y_{gj} - med_{g,y}|; i = 1, \cdots, n_0, j = 1, \cdots, n_1)$$

である．模式図を図 4 で示す．

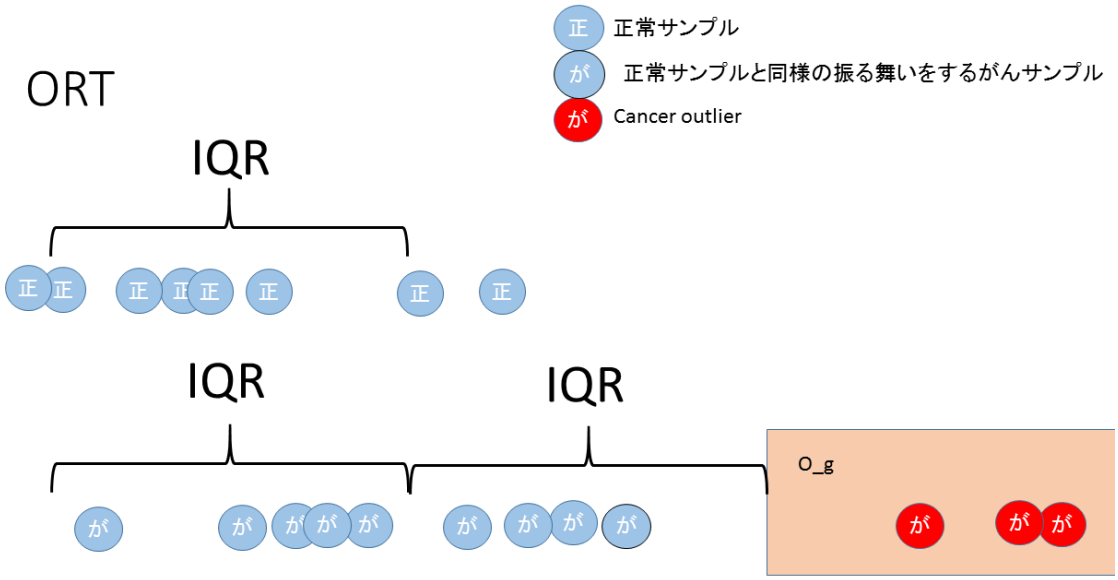

図 4　ORT 統計量作成を説明する模式図

この図 4 は図 3 も同じ発現量を持つサンプルのときで作成しているが，ORT の場合は OS と違い，IQR を正常サンプルのみから作ることとしていたので，下部 O_g には OS では選ばれなかったサンプル発現量が Cancer Outlier に指定されることになっている．

3.6　MOST 統計量

ORT 統計量は，Cancer Outlier であるとする領域を，分位点を用いて定義するため恣意的であるという問題が残っていた．この問題を解決する一つの方法として，Lian(2008)[7] は Cancer Outlier であると判断する閾値を，可能性がある部分をすべて検討してから考える統計量を考案した．この統計量を MOST とよび提案している．MOST は Maximum Ordered Subset T-statistics の略である．これも MOST 統計量という言い方を使う．MOST 統計量の作成においては，まず，遺伝子毎にがんサンプルを遺伝子発現量の大きさで並べ替える．g 番目の遺伝子においてのがんサンプルで一番大きな値をとっているものを $y_{g.(1)}$，2 番目に大きな発現量を $y_{g.(2)}$ というように一番小さな発現量を $y_{g.(n_1)}$ と

する．

$$y_{g.(1)} \geq y_{g.(2)} \geq \cdots \geq y_{g.(n_1)} \tag{15}$$

このとき，統計量の候補として，以下の様な式を提案している．

$$M_{gk} = \frac{\sum_{1\leq j\leq k}(y_{g.(j)} - med_{g,x})}{med(\{x_{gi} - med_{g,x}\}_{1\leq i\leq n_0}, \{y_{gl} - med_{g,y}\}_{1\leq l\leq n_1})} \tag{16}$$

このとき，k は Cancer Outlier の個数であり，この真の値を知ることは出来無い．そこで，遺伝子の統計量として以下を定義する．

$$M_g = \max_{1\leq k\leq n_1} M_{gk} \tag{17}$$

しかし，k の値の違いによる M_{gk} はがんサンプル，正常サンプルがそれぞれ標準正規分布に従っているという帰無仮説のもとでは直接比較することが出来無い．ここで，新しくがんサンプルの数だけ標準正規乱数を発生させ，これもがんサンプル $\{y_{(j)}\}$ と同様に大きな物から順に並べることとする．すなわち，帰無分布である標準正規分布に従う z に対して順序統計量

$$z_{(1)} > z_{(2)} > \cdots > z_{(n_1)} \tag{18}$$

としたとき，それぞれの k に対して，

$$\mu_k = E[\sum_{1\leq j\leq k} z_{(j)}] \tag{19}$$

$$\sigma_k^2 = Var(\sum_{1\leq j\leq k} z_{(j)}) \tag{20}$$

を定義する．これによってがんサンプルが帰無仮説に従うとすれば，

$$M_{gk} = \left\{ \frac{\sum_{1\leq j\leq k}(\tilde{y}_{gj} - med_{g,x})}{1.4826 \times med(\{x_{gj} - med_{g,x}\}_{1\leq j\leq n_0}, \{\tilde{y}_{gj} - med_{g,y}\}_{1\leq j\leq n_1})} - \mu_k \right\} \Big/ \sigma_k \tag{21}$$

となり，これは近似的にそれぞれ平均 0，標準偏差 1 に従うと考えられる．例えばこのそれぞれの Cancer Outlier をどこまでにするのが良いかを判断するための統計量候補が一

番大きくなるところをそれぞれの遺伝子統計量として採用することとし，改めて，

$$M_g = \max_{1 \leq k \leq n_1} M_{gk} \tag{22}$$

を考えることとする。このように考えることで閾値に関しての恣意性を排除することが可能となっている。

4 提案法

4.1 遺伝子発現データの混合モデル

我々は盛ら (2013)[10] において遺伝子レベルの情報だけでなく，がんサンプルの情報も共有するために，我々は，がんサンプルの遺伝子発現データにおいて単純なパラメトリック正規混合モデルを考えることを提案した．まず遺伝子内での正常サンプルを対照として，がんサンプルの標準化を考える．これは従来法での ORT や MOST でも考えられてきた方法である．つまり，この段階では，従来法でよく用いられる多重検定の枠組みで考えられる統計量と同じ手順である．式で表せば,

$$u_{gj} = \frac{y_{gj} - \bar{x}_g}{s_{g,x}}. \tag{23}$$

となる．ここで，$s_{g,x}$ は遺伝子 $g(g = 1, \cdots, G; j = 1, \cdots, n_0)$ の中の正常サンプルで推定される．我々は正規混合モデルにおいて 3 つのコンポネントを仮定する．

$$f(u_{gj}) = \pi_0 f_0(u_{gj}) + \pi_1 f_1(u_{gj}) + \pi_2 f_2(u_{gj}). \tag{24}$$

密度関数 f_0 は正常サンプル発現量やがんサンプルでありながら Cancer Outlier でないと判断される発現量の密度関数として定義される．この発現量を null 発現量と呼ぶこととする．f_1 と f_2 の密度はそれぞれ正常サンプル発現量よりも，低発現，高発現の Cancer Outlier を代表とする non-null 発現量のコンポネントに対応する．我々は f_0, f_1, と f_2 の正規分布としてそれぞれ,$N(0, 1^2)$,$N(\delta_1, 1^2)$, そして $N(\delta_2, 1^2)$, と仮定する．$\pi_q(q = 0,$ 1, 2) はそれぞれの null, 負の non-null, 正の non-null コンポネントの混合割合であり，$\pi_0 + \pi_1 + \pi_2 = 1$ である．

分布の混合化の際には，背景となる部分母集団が特定できないことが問題を複雑にしていると考えられる．このため，これは Mclhalan ら (2000)[8] のように，我々は観測できないランダムな指示変数 $Z_{gj,h}$ を考えた．遺伝子 g において j 番目のサンプルが h 番目のコンポネントに入るときに $Z_{gj,h} = 1$ とした．それ以外だったときは $Z_{gj,h} = 0$ として与えた $(g = 1, \cdots, G; j = 1, \cdots, n_1)$．これにより，我々は，それぞれの観測が得られたとき，$Z_{gj,h} = 1$ の下での条件付き密度を考えることが可能になり，この $\mathbf{Z}$ の分布は部分母

集団の割合を反映した観測総和 1 の多項分布になると考えることができる．このようにして，この分布に含まれる種々のパラメータ δ_1,δ_2,π_1,π_2 の値を EM アルゴリズムを用いて推定する．Mclhalan ら (2000)[8] の方法を用いて，我々の提案する方法に必要なパラメータ推定を行うための更新式を以下に示す.

まず混合割合の推定であるが，一般に確率変数 V,W の同時密度を $f(v,w), W=w$ が与えられたという条件のもとで V の条件付き密度を $f(v|w)$，W の密度関数を $f(w)$ とすると同時密度 $f(v,w)$ は,

$$f(v,w) = f(v|w)f(w) \tag{25}$$

とかける．したがって，このとき，単一観測 $\mathbf{X}^{*T} = (y, \mathbf{Z}^T)$ に関する同時分布 $f(x^*|\theta)$ は

$$f(x^*|\theta) = \prod_{j=1}^{g} f_j^{z_j}(y|\theta_j) \prod_{j=1}^{g} \pi_j^{z_j} \tag{26}$$

である．母集団全体に対する割合である．また，このことから観測 Y の密度関数は,

$$f(y|\theta) = \sum_{\mathbf{Z}} f(x^*|\theta) = \sum_{j=1}^{g} \pi_j f_j(y|\xi_q) \tag{27}$$

である．θ は未知パラメータ全体を表し，未知パラメータの数を n_g とすると，ξ_j は $a \leq q \leq n_g$ となるときの j 番目の 未知パラメータである．ここで $\Sigma_{\mathbf{Z}}$ は起こりうるすべての $\mathbf{z}$ に関する和を示している．このことから，Y が与えられたという条件のもとでの，$\mathbf{Z}$ の確率関数は $Z_j = 1$ である場合,

$$f(z|\mathbf{y},\theta) = \frac{f(\mathbf{y},\mathbf{z}|\theta)}{f(\mathbf{y}|\theta)} = \frac{\pi_j f_j(y|\xi_q)}{\sum_{j=1}^{g} \pi_j f_j(y|\xi_q)} \tag{28}$$

と表現できる.

このとき，n 個の標本からの観測 $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ と対応する指示確率変数 $\mathbf{Z} = (Z_1^T, \cdots, Z_n^T)^T$ から完全観測 $\mathbf{X} = (\mathbf{Y}^T, \mathbf{Z}^T)$ を構成すると $\mathbf{X}$ と $\mathbf{Y}$ の密度関数はそれぞれ,

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} (\prod_{j=1}^{g} f_j^{zij}(y_i|\xi_q) \prod_{j=1}^{g} \pi_j^{z_{ij}}) \tag{29}$$

$$f(\mathbf{y}|\theta) = \sum_{(\mathbf{Z_1},\cdots,\mathbf{Z_n})} f(\mathbf{x}|\theta) = \sum_{(\mathbf{Z_1},\cdots,\mathbf{Z_n})} \prod_{i=n}^{n} [\prod_{j=1}^{g} \{\pi_j f_j(y_i|\xi_q)\}^{z_{ij}}] \tag{30}$$

となる．ただし，ここで z_{ij} は観測 y_i がどの部分母集団へ属するかを示す指示変数ベクトル z_i の j 番目の要素を示し，$\sum_{(\mathbf{Z_1},\cdots,\mathbf{Z_n})}$ はすべて可能な，$(\mathbf{Z_1},\cdots,\mathbf{Z_n})$ に関する和を示している．さらに，例えば，$\mathbf{Y}$ と $\mathbf{Z_n}$ の同時密度は，$Z_{nl} = 1$ の場合，

$$f(\mathbf{y},\mathbf{z}_n|\theta) = \sum_{(\mathbf{Z_1},\cdots,\mathbf{Z_{n-1}})} f(\mathbf{x}|\theta) \tag{31}$$

$$= \sum_{(\mathbf{Z_1},\cdots,\mathbf{Z_{n-1}})} f(\mathbf{y}, z_1, \cdots, z_n|\theta) \tag{32}$$

$$= \sum_{(\mathbf{Z_1},\cdots,\mathbf{Z_{n-1}})} \prod_{i=1}^{n-1} [\prod_{j=1}^{g} \{\pi_j f_j(y_i|\xi_q)\}^{z_{ij}}]\} \prod_{i=1}^{g} \{\pi_j f_j(y_n|\xi_q)^{z_{nj}}\} \tag{33}$$

これは結局，l 番目だけの項が残るため，

$$= (\sum_{j_1=1}^{g} \pi_{j_1} f_{j_1}(y_1|\xi_{q_1})) \cdots (\sum_{j_{n-1}=1}^{g} \pi_{j_{n-1}} f_{j_{n-1}}(y_{n-1}|\xi_{q_{n-1}})) \pi_l f_l(y_n|\xi_l) \tag{34}$$

となるので，一般に $\mathbf{Y}$=$\mathbf{y}$ が与えられたという条件のもとでの $\mathbf{Z}_k$ の条件付き密度は z_{kl}=1 の場合，

$$f(z_k|\mathbf{y},\theta) = \frac{f(\mathbf{y},\mathbf{z_k}|\theta)}{f(\mathbf{y}|\theta)} \tag{35}$$

$$= \frac{\pi_l f_l(y_k|\xi_l)}{\sum_{j=1}^{g} \pi_j f_j(y_k|\xi_q)} \tag{36}$$

となる．ここで，$f(\mathbf{x}|\theta)$ の式から，完全観測 $\mathbf{X}$ に基づく，対数尤度 $l^C(\theta,\mathbf{x})$ は

$$l^C(\theta,\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\theta) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij} \log f_j(y_i|\xi_q) + \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij} \log \pi_j \tag{37}$$

となる．EM アルゴリズムの E ステップでは，現時点でのパラメータ値 $\theta^{(k)}$ が得られ，観測 $\mathbf{Y} = \mathbf{y}$ が与えられたという条件のもとでの $l^C(\theta,\mathbf{x})$ に関する条件付き期待値 $Q(\theta,\theta^{(k)})$ を計算する．この場合，対数尤度 $l^C(\theta,\mathbf{x})$ のなかで，$\mathbf{Z}$ はその成分が，線形に

取り込まれているので，条件付期待値の計算では，単純に Z_{ij} をその条件付き期待値

$$E_{\theta^{(\mathbf{k})}}[Z_{ij}|\mathbf{Y}=\mathbf{y}] = Pr_{\theta^{(\mathbf{k})}}\{Z_{ij}=0|\mathbf{Y}=\mathbf{y}\} + Pr_{\theta^{(\mathbf{k})}}\{Z_{ij}=1|\mathbf{Y}=\mathbf{y}\} \tag{38}$$

とかける．ここで右辺第 1 項は 0 になるので,

$$= Pr_{\theta^{(\mathbf{k})}}\{Z_{ij}=1|\mathbf{Y}=\mathbf{y}\} \tag{39}$$

$$= \frac{\pi_j^{(k)} f_j(y_i|\xi_q^{(k)})}{\sum_{j=1}^{g} \pi_j^{(k)} f_j(y_i|\xi_q^{(k)})} \tag{40}$$

$$= z_{ij}^{(k)} \tag{41}$$

で置き換えればよい．したがって $Q(\theta,\theta^{(k)})$ は

$$Q(\theta,\theta^{(k)}) = E_{\theta^{(\mathbf{k})}}[l^C(\theta,\mathbf{x})|\mathbf{Y}=\mathbf{y}] \tag{42}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k)} \log f_j(y_i|\xi_q) + \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k)} \log \pi_j \tag{43}$$

となる．

M ステップでは E ステップで得られた $Q(\theta,\theta^{(k)})$ をそれぞれの θ に対して最大化すればよい．まずそれぞれの母集団に対する各母集団の割合 $\pi_j(j=1,2,3)$ に対して考える．

$$\frac{\partial}{\partial \pi_j} Q(\theta,\theta^{(k)}) = \sum_{i=1}^{n}\left(\frac{z_{ij}^{(k)}}{\pi_j} - \frac{z_{ig}^{(k)}}{\pi_g}\right) = 0 (j=1,2) \tag{44}$$

という方程式を解けば良いということになる．π_3 は混合割合の和が 1 であるので，解かなくても計算できる．ここで，方程式に戻る．

$$\sum_{j=1}^{g} z_{ij}^{(k)} = 1 \tag{45}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(k)} = n \tag{46}$$

であることに着目すると，π_j に関するパラメータの更新式として

$$\pi_j^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(k)} \tag{47}$$

$$(j = 1, 2) \tag{48}$$

を得る．さらに，各母集団，今回の研究では分散 1 の正規分布を仮定するが，このときの平均パラメータは，対数尤度の式の第二項が平均パラメータに依存しないことから，式はさらに簡単となり，混合割合の更新式も利用して，以下のように与えられる．正規分布であることも考慮して書き下すと，解くべき方程式は

$$0 = \frac{\partial}{\partial \mu_l}\sum_{i=1}^{n}\sum_{j=1}^{3} z_{ij}^{(k)}\{\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y_i - \mu_l)^2\} \tag{49}$$

$$= \sum_{i=1}^{n} z_{il}^{(k)}\{\frac{1}{\sigma^2}(y_i - \mu_l)\} \tag{50}$$

となり，パラメータの更新式は

$$\mu_l^{(k+1)} = \frac{\sum_{i=1}^{n} z_{il}^{(k)} y_i}{\sum_{i=1}^{n} z_{il}^{(k)}} \tag{51}$$

$$(l = 1, 2) \tag{52}$$

である．

4.2 遺伝子選抜のための統計量

がんサンプル y_{gj} が与えられたもとで，事後確率 $w_{gj,k}$ を考える．このとき，がんサンプ標準化した場合の発現量の u_{gj} が k 番目のコンポネントであるとき下記の式で与えられる．

$$w_{gj,k} = \frac{\hat{\pi}_k \hat{f}_k(u_{gj})}{\hat{f}(u_{gi})}. \tag{53}$$

このとき，$\hat{\pi}_k$ は EM アルゴリズムによるパラメータの推定値であり，$\hat{f}_k$,$\hat{f}$ はそれぞれパラメータの推定値を使用した k 番目のコンポネントの密度関数，パラメータの推定値を使用した混合分布の密度関数となっている．

片側検定による Cancer Outlier のプロフィールをもつ高発現の遺伝子を探索するために我々は，遺伝子に基づく統計量を使用する方法を提案する．式としては，

$$S_g = 1 - \prod_{j=1}^{n_1}(1 - w_{gj,2}) \tag{54}$$

である．この統計量は過剰発現をもつ Cancer Outlier ががないことを示すものとなっており，(1-事後確率) としてあたえられる．このとき，一番大きな S_g を選抜する．がん関連遺伝子により発現抑制されている遺伝子を同定するための遺伝子に基づく統計量は同様に開発できる．この枠組みでは，両側検定を考えることで，過剰発現，発現抑制の両方を同時に探すこともできる．これは，

$$T_g = 1 - \prod_{j=1}^{n_1}\{1 - (w_{gj,1} + w_{gj,2})\} \tag{55}$$

という統計量を用いると可能である．

事後確率 $w_{gj,k}$ を考えるときに注意する重要なことは，それ自身はがん関連遺伝子を同定するだけでなく，その遺伝子の中のどのサンプル発現量が Cancer Outlier と考えることができるかまで情報が与えられることである．一方，従来の Cancer Outlier を同定する方法では，Cancer Outlier 発現量の同定のために，発現量レベルでの統計量は提案されていなかった．

次の章ではモンテカルロシミュレーションにより様々なシナリオを用いて，それぞれの手法を ROC 曲線を用いて比較する．

5　シミュレーション

5.1　従来法提案論文でシミュレーション内容のレビュー

我々が考えた評価方法を紹介する前に従来法の提案時に行われていたそれぞれの手法評価のためのシミュレーションをレビューする．

5.1.1　OS 統計量

OS 統計量が提案された Tibshirani と Hastie(2007)[17] では t 統計量と COPA 統計量との比較が行われている．すべての遺伝子発現量は，それぞれの定義式の様に標準化された．COPA 統計量の閾値は 0.90 とされた．シミュレーションには 1000 遺伝子と 30 サンプルを用意した．半分の 15 サンプルが正常サンプル群，残りの 15 サンプルががんサンプル群とされている．すべてのデータはまず標準正規分布から発生させている．データの中で一つ (gene1) のみ Cancer Outlier を含む遺伝子であるとし，その場合に 15 サンプルの内 15 個すべてが正常サンプルに比べて高発現を表すために発現量に 2 を加えた．また，同様に Cancer Outlier の数を 8 サンプル，6 サンプル，4 サンプル，2 サンプルと変化させ，シミュレーションを行っている．それぞれの統計量でランキングし，gene1 よりも上位の割合を計算して p 値を計算している．つまり，遺伝子が 1000 個という設定で gene1 がランキング 1 位になっている場合はそれより上位の遺伝子は 0 個のため p 値は 0/1000=0 となる．また，100 番の場合は 99 遺伝子が上位に位置しているため 99/1000=9.9% となる．論文内では，50 回シミュレーションを行ったときの gene1 に対する p 値の分布を示している．t 統計量での p 値は 1 に近い値があり，Cancer Outlier の検出が上手くできていないことがわかる．対して OS 統計量は比較的小さな値に半分ほどのデータがあつまり，関連ありとセッティングした遺伝子が検出されているのがわかる．また，論文内では，シミュレーションを 50 回行い，p 値の中央値，平均値，標準偏差を算出している．

5.1.2　ORT 統計量

Wu(2007)[22] の論文においては, OS 統計量, COPA 統計量, T 統計量と ORT 統計量に対しての検出力を評価している．また, 検出力と同時に, Benjamini と Hochberg(1995)[1]

で提案された False Discovery Rate(偽発見率) を用いている．シナリオとしては，正常サンプルとがんサンプルはそれぞれ 25 サンプルずつあり，遺伝子数は 1000 としている．発現量データは標準正規分布から発生しているとし，Tibshirani と Hastie(2007)[17] のシミュレーションと同様に Cancer Outlier としてセッティングする gene1 のがんサンプルにだけ発現量に 2 を加えることにしている．がんサンプル内での Cancer Outlier の数は，1,5,10,15,20,25 というシナリオを用意して考察する．シミュレーションは 1000 回行い偽陽性の割合をカットオフ値として真陽性の割合を計算し ROC 曲線としてプロットしている．

また，遺伝子の中でがん関連遺伝子が 100 個，200 個，300 個のときの検討も同様に行っている．この場合は全体に対する関連ありと判断した遺伝子の割合と偽発見率をプロットしている．

5.1.3 MOST 統計量

Lian(2008)[7] では MOST 統計量が提案された．シミュレーションシナリオは，正常サンプル，がんサンプル，それぞれ 20 サンプルずつとした．それぞれの値は標準正規乱数で生成された．関連あり遺伝子数は 1000 とされた．関連なし遺伝子数は 1000 とし全部で 2000 遺伝子で考えるとされていた．また，関連あり遺伝子の中で，がんサンプルのいくつで cancer outllier とするかを変更している．Cancer Outlier とされた発現量には 1，2 または，4 を加えることで異常な発現量を表現する．Cancer Outlier とするサンプルの数は 10,15,20 と変化させている．

5.1.4 まとめ

以上のようにがんサンプルと正常サンプルの数は同じ数としており，遺伝子数は 1000, または，2000 であった．また，遺伝子を一つだけ Cancer Outlier 型の遺伝子としてセッティングしたときの検出力や偽発見率を議論し，他には複数の遺伝子でがん関連ありとしてセッティングしたときの ROC 曲線を確認している研究も存在した．マイクロアレイデータとしては数万の遺伝子情報が一挙に手に入り解析しなければならない状況も多いと考えられるので更に多くの遺伝子としても良いのではないかと考える．高発現とするときに Cancer Outlier としてセッティングするために足す（引く）値 (効果サイズ) であるが，これは，1,2,4 と言う値が取られていた．我々の研究では 2 を用いることとした．が

ん関連あり遺伝子としてセッティングされたときのがんサンプルの数は 1,2,5,10,15,20,25 と言う値での検討がなされていた．我々の検討では標本の大きさを変更して検討をしているのでがんサンプルの割合でセッティングすることとした．

5.2 シミュレーションシナリオと評価方法

提案法と従来法の評価を行うためにモンテカルロシミュレーションを実施した．マイクロアレイ研究で得ることが出来たデータを想定し，$G = 10000$ 遺伝子 ， $n = 40, 80,$ または，200 サンプルとし，$n_0 = n_1 = n/2$ と考え，それぞれのシナリオにおいてサンプルの半分を正常サンプル (n_0 個)，半分を cancer サンプル (n_1 個) とした．まずデータは標準サンプルはそれぞれの発現量ごとに標準正規分布 $N(0, 1^2)$ から発生させた．または，データに正規性がなかったときを仮定して，自由度 20 の t 分布から発生させた乱数を用いることとした．自由度のであるがこれは後で説明する実データに対してそれぞれの自由度のときの遺伝子毎の尤度を算出しそれぞれの自由度での尤度を確認し決定した．遺伝子間の交互作用はないと仮定した．遺伝子 1 万 (G) 個に対して，関連なし遺伝子の混合割合は $0.6(= \tau_0)$，高発現，低発現の混合割合はそれぞれ $0.2(= \tau_1 = \tau_2)$ とする．混合割合が違う場合も考えることができるが，高発現している遺伝子が多いデータセットなのか低発現している遺伝子が多いデータセットなのかを確かめる方法はないため，まずは混合割合を同じくするシナリオでのそれぞれの手法の性能を評価することを考えた．また，それぞれの関連あり遺伝子の中で，がんサンプルにおいて高発現，または，低発現している割合は $\phi = 0.1$，0.3，または，0.5 として各シナリオでシミュレーションする．Cancer Outlier は高発現，低発現をそれぞれ $N(2, 1^2)$，$N(-2, 1^2)$ から発生するデータとして表現する．つまり，$\delta_1 = -2.0$ であり， $\delta_2 = 2.0$ である．

この内容を盛り込んだ模式図を図 5 に示す．

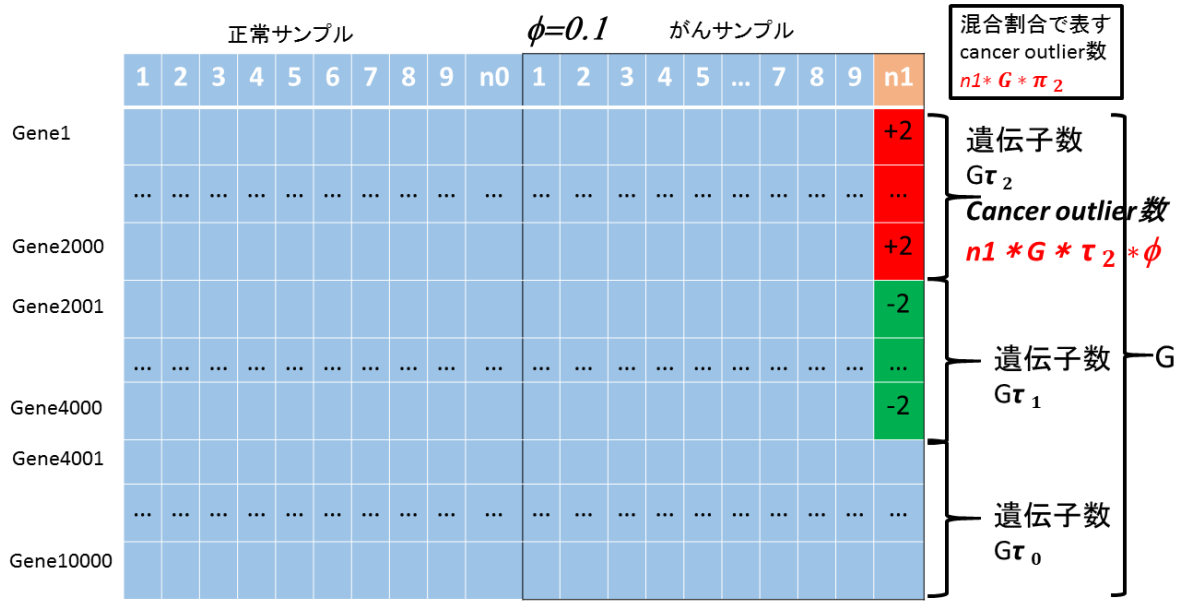

図 5　$\phi = 0.1$ のときのシミュレーションデータセットの模式図

このそれぞれのシナリオにおいて，評価をするために偽発見率と検出力を計算する．偽発見率の定義は有意であるとした遺伝子の中での偽陽性の割合，TPR の定義は関連あり遺伝子として定義した遺伝子の中で真に陽性とされたものの割合である．この偽発見率を横軸，検出力を縦軸にとって ROC 曲線を書くことにする．一般に ROC 曲線は偽発見率ではなく偽陽性率を用いることが多いが，近年偽発見率を横軸に取るものであっても雑誌によっては ROC 曲線と呼んでおり，今回の論文でもそれに習って ROC 曲線と呼ぶこととした [3]．また，偽発見率は偽陽性率と一対一に対応している．

我々はそれぞれのシナリオに対して，200 回のシミュレーションを行い，偽発見率が与えられたときの検出力を 200 個作成し，その平均値をその手法の偽発見率が与えられときの検出力とした．これを従来法，提案法に適用した．

5.3　シミュレーション結果

図 6 ，図 7 はそれぞれ発現量データを正規分布から生成の場合のシナリオの結果，t 分布から生成した場合の結果として，前項で考えた内容で作成している．それぞれの曲線

は，提案法統計量，MOST 統計量，ORT 統計量，OS 統計量，COPA 統計量， t 統計量とそれぞれ順番に赤色，青色，緑色，黒色，水色，黒色 (破線) となっている．9 つのグラフはそれぞれのシナリオの結果となっている．上段中段下段の順で全体のサンプルの大きさがそれぞれ 40,80,200 としている．また，左からそれぞれ $\phi = 0.1, 0.3, 0.5$ となっている．

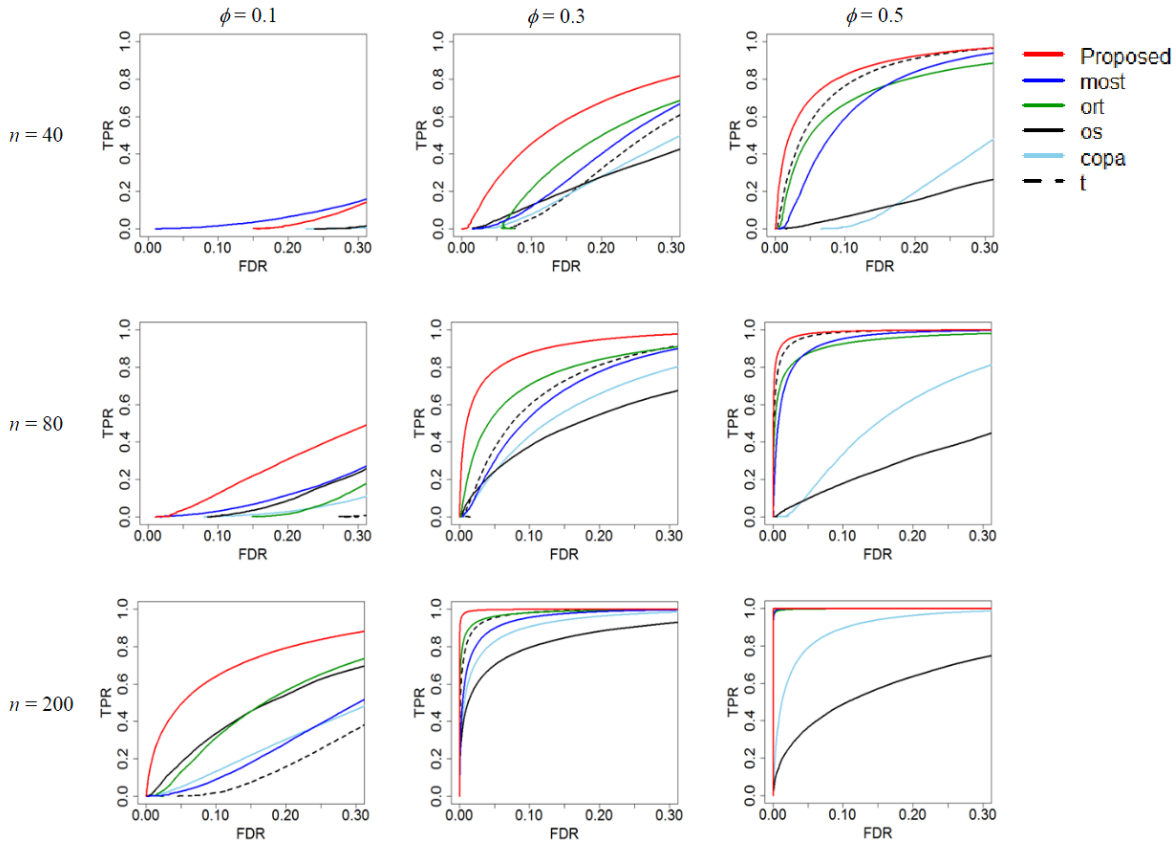

図 6　正規分布からの乱数を用いたシナリオの ROC 曲線

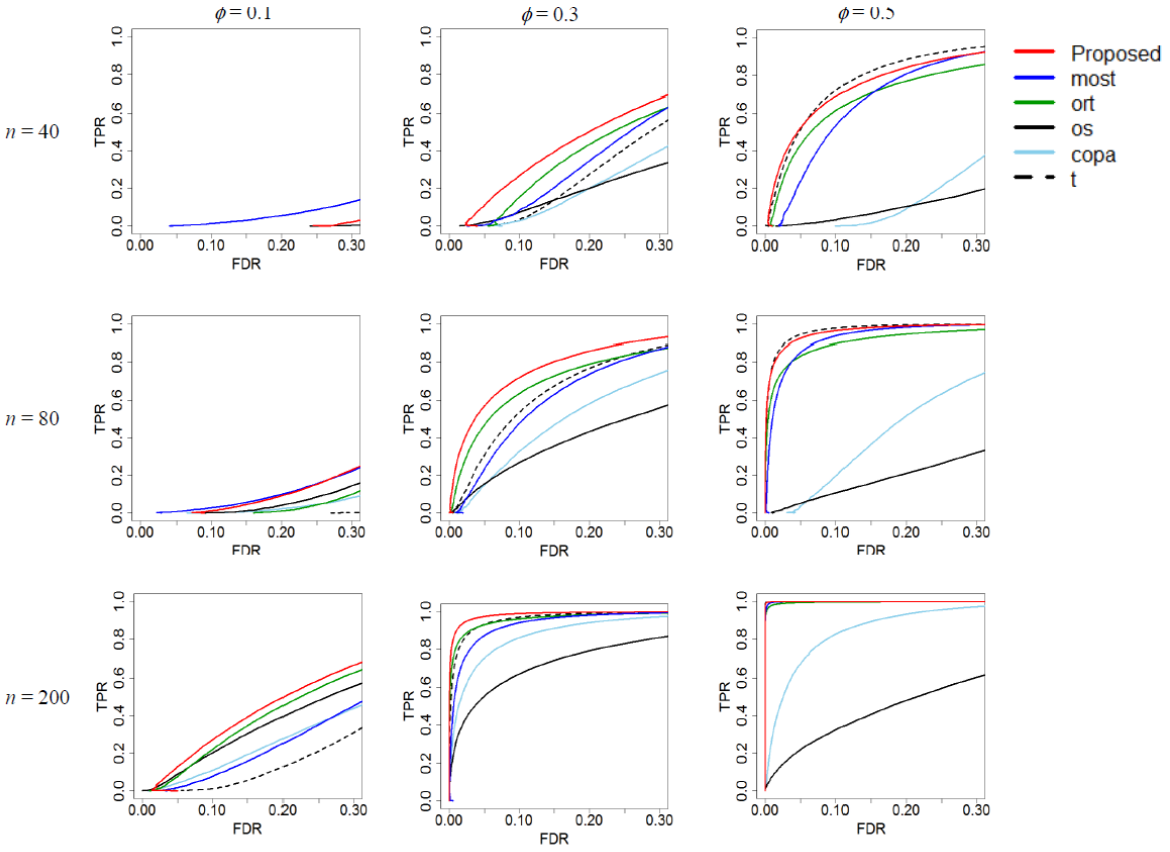

図 7　t 分布からの乱数を用いたシナリオの ROC 曲線

この曲線は算出されたそれぞれの統計量を用いて，いくつの遺伝子を関連ありとするかに関してのカットオフ値を動かしそれぞれの偽発見率に対して得られた検出力を用いて作成されている．まずは図 6 と図 7 を比べる．同じシナリオ同士で比べれば，正規乱数から生成したデータの結果のほうが任意の偽発見率で高い検出力を示している．ただし，ほとんどの場合で性能の順序は変わることはない．ここでの性能とは任意の偽発見率を考えたときに大きな検出力をとっているかどうかである．提案した統計量 S_g に基づく遺伝子が，従来法よりも，任意の与えられた偽発見率に対して一番大きな検出力値をとっている．また，この結果から，提案の統計量 S_g に基づいての遺伝子選抜は ϕ が大きなとき，大きな検出力を提供することが期待される． t 統計量に基づく遺伝子発現の場合は ϕ=0.1 のようなとき，つまり，がんサンプル内の Cancer Outlier の割合が小さいときに一番小さな検出力を提供する．しかし，ϕ=0.5 のときのようながんサンプル内での

Cancer Outlier の割合が大きな値のとき，検出力は大きな値に改善する．これは前述しているように t 統計量が一様な差を検出することを得意とする統計量であるからと考えられる．COPA 統計量や OS 統計量は ϕ=0.5 のときのような特に ϕ の値が大きな傾向にあるとき，検出力が低い．特に，$\phi = 0.5$ のときを見ると任意の偽発見率で，OS 統計量の検出力が悪いことから，徐々に悪化していく傾向にあることが示唆されている．これは統計量の構成に分位点や IQR という恣意的な値を用いてたり，標準化の際にがんサンプルを用いていることにより，本来検出したいの正常サンプルからの乖離がマスクされているためと考えられる．ORT 統計量と MOST 統計量はシナリオ全般を通してよい検出力を示すことが確認できる．しかし，提案法の S_g に基づく方法には及ばない．t 分布から遺伝子発現量が発生していると考えたとき (図 7) も同様の傾向が観察されている．ただし特記すべきこととして，S_g に基づいた提案法は $n = 40$，ϕ=0.1 のときを除いていつも任意の偽発見率おいて検出力が一番大きな値を示している．$n = 40$，ϕ=0.1 のときは ORT 統計量や MOST 統計量の方がよい傾向を示している．

6　実データへの適用

6.1　実データ解析結果のレビュー

シミュレーションでは真陽性の遺伝子が明らかであるが，実問題としてはどれが真にがん関連遺伝子かどうかは分からない．そこでこれまでの論文でもいろいろな方法で実データへの適応評価が行われてきているので我々の実データの話をする前にレビューしておく．

6.1.1　COPA 統計量

COPA 統計量を初めてデータ解析に用いたのは Tomlins ら (2001)[18] の論文である．この論文では，Oncomine データベースに登録させているすべてのデータセットを対象にして考えた．データセットの中でなんらかの疾患の原因となる変異がすでに知られている可能性の高い方から 10 遺伝子に対して，COPA 統計量を計算している．その遺伝子でのエビデンスレベルを書いてその解析が妥当である場合があることを示している．

6.1.2　OS 統計量

OS においては Rieger ら (2004)[12] のデータを用いている．この中で扱われているデータは 12625 の遺伝子，サンプルは全例がん患者から取られておりで 58 人分存在している．58 人の中で 14 人は放射線に対して感受性があり，のこりの 44 人では感受性がなかった．この感受性ありなしの二群に対してデータ解析を行っている．このデータに対して Tusher ら (2001)[20] で考案された Significance analysis of microarray(SAM 法) のアプローチの中で OS を用いて outlier を判定している．SAM を用いて放射線に対して感受性ありとなしのラベルに対しての並べ替え検定の考え方から FDR を推定している．考察には t 統計量と OS 統計量をそれぞれ横軸縦軸に取ったものを用いている．また，関連ありとする遺伝を数を変化させたときの FDR の推定値をプロットすることによって評価を行っていた．OS 統計量の大きさを比較したとき，上位 12 の遺伝子に対しては横軸に発現量，縦軸で群をしめしていた．

6.1.3 ORT 統計量

ORT 統計量の実データへの適用では West ら (2001)[21] の乳がんのデータによって行われている．乳がんサンプルが 49 例分，7129 遺伝子であり，リンパ節転移のありなしによってサンプルラベルの差としている．リンパ節転移ありの乳がんサンプルは 25 例，リンパ節転移なしの乳がんサンプルは 24 例であった．データは Bolstad ら (2003)[2] により提案された quantile nomalization により正規化され，さらに対数変換され統計解析が行われた．リンパ節転移無しのグループをノーマルグループと考えたときに，t 統計量，COPA，OS さらに提案した ORT の高発現の Cancer Outlier 検出のために適用した．それぞれの結果を統計量でランキングして，それぞれの統計量で上位 25 遺伝子を選んだ．その遺伝子が乳がんの分野ですでに PubMed 等で議論されているかを確かめた．論文中ではそれぞれの統計量上位で議論されている遺伝子の名前を示し，他の統計量はその遺伝子が何番にランキングされているか確かめていた．

また，ORT 統計量で上位 25 番までに入りかつ乳がんで議論の行われている遺伝子に対してリンパ節転移あり，なしに分けてそれぞれの発現量のプロット図を作成していた．それぞれがどの統計量で選ばれたものかが一緒に記載されている．

6.1.4 MOST 統計量

MOST 統計量が提案された論文 [7] では ORT 統計量が提案された論文 [22] と同様に West ら (2001)[21] のデータを用いていた．MOST 統計量の提案部分では比較的高発現による Cancer Outlier を検出する方法で説明していたが，そのまま逆に小さな値から順序統計量を考えることで，比較的低発現による Cancer Outlier を検出できると書かれている．そのため提案と同様に高発現の統計量として M_g 低発現の統計量として m_g を用意し，絶対値が大きな方をその遺伝子 g の統計量とすることとした．West ら (2001)[21] のデータにおいて，乳がんに関連があると言われている 908 遺伝子の 655 のプローブに対してこれまで提案されている伝統的な T 検定を含めた 5 種類 (残り COPA 統計量，OS 統計量，ORT 統計量，MOST 統計量) それぞれで計算を行った．計算の結果は遺伝子ごとに MOST とその他の方法の差をとることとした．遺伝子は関連ありなはずであり，これらの統計量で大きな値を示すはずなので，(従来法統計量)-(MOST 統計量) が負であり，さらに絶対値が大きければ MOST 統計量が他の手法に対して優越しているということがで

きる．また，2 つ目の適用例として急性リンパ性白血病のデータに関しても解析を行っている．79 サンプルでの解析であり．内 37 サンプルが転座による BCR か，または，ABL 融合遺伝子であり，残り 42 サンプルは正常サンプルとなっている．

6.2 実例への適用

我々は提案法が異質性を持つがんサンプルの中からがん関連遺伝子を同定できるかどうかを確かめるために，骨髄異形成症候群：myelodysplastic syndromes (MDSs) の検討 [9] から得られたマイクロアレイ実験による遺伝子発現データへの適用を試みた．MDS はさまざまな染色体異常を持ち臨床病期的兆候に遺伝的異質性がある複雑な血液腫瘍である．MDS の遺伝的異質性をもつ臨床病期的兆候の発見のために，139 例の MDS サンプルと，69 例の白血病でないデータに対して提案する混合分布を用いた方法を適用した．白血病でないサンプルは，そのような人の骨髄の単核細胞からのデータとなっている．それぞれのサンプル群はがんサンプルと正常サンプルとみなすこととする．Mills らは MDS サンプルから染色骨髄単球性白血病は除外している．われわれは Bolstad ら [2] の RMA 正規化を生データに適用した．生データのファイルは Gene Expression Omnibus database(GEO,http://www.ncbi.nlm.nih.gov/geo/, accession number GSE15061) からダウンロードできる．我々はそれぞれの遺伝子の発現強度のスケールに対数をとり，それぞれの手法を適用することを考えた．その際の正常サンプル，がんサンプルのヒストグラムをそれぞれ，図 8 と図 9 で示す．

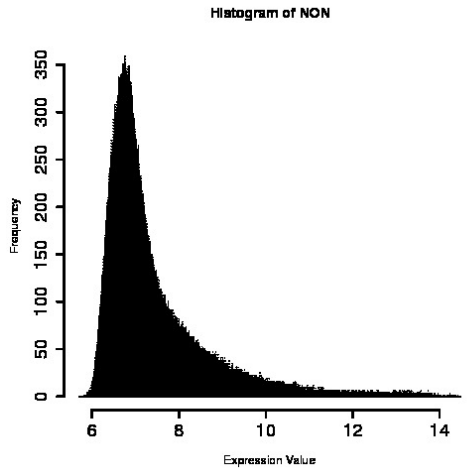

図 8　正常サンプルのヒストグラム

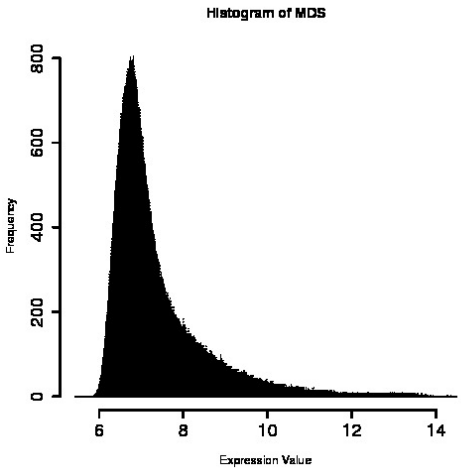

図 9　がんサンプルのヒストグラム

候補遺伝子は G=54675 個であり，ここから Cancer Outlier をもつ遺伝子のスクリーニングを行うために従来法と提案法に適用し，それぞれの統計量を計算した．それぞれの方法において，統計量の大きなものから 200 遺伝子を選び出した．実際の計算では，EM アルゴリズムのもとで，混合分布 (24) のパラメータの推定を行った．EM アルゴリズムで用いる収束条件は 10^{-4} とし，パラメータ推定値としては $\hat{\pi}_1 = 0.0018$,$\hat{\pi}_2$, = 0.0018, $\hat{\delta}_1$ = -1.22 ,$\hat{\delta}_2$ = 3.54 が得られた．表 1 ではそれぞれの遺伝子選択の間で上位 200 個の中で共通の遺伝子が選ばれている数をまとめている．たとえば T 統計量と COPA 統計量がクロスするセルでは，13 という数字が入っているが，これは T 統計量で上位 200 遺伝子までに入ったものを選抜した後，COPA 統計量でも同様に調べたとき，上位 200 番までに共通して入っていた遺伝子の数をあらわしている．

表 1　各手法で上位 200 遺伝子が共通したものの個数

	T	COPA	OS	ORT	MOST	PROPOSED
T		13	14	50	56	56
COPA	13		150	0	99	51
OS	14	150		139	108	86
ORT	50	0	139		151	89
MOST	56	99	108	151		75
PROPOSED	56	51	86	89	75	

OS 統計量，ORT 統計量，MOST は選ばれた遺伝子の一部が重なっていた．重なりの度合いは標準化の方法や outlier とするカットオフ値のことなど手法の説明の部分にて書いたことで説明することができる．一方で，遺伝子に基づいた統計量 S_g に，基づいた提案法は興味深く，それにおいては，従来法のすべての方法に対して重なっている遺伝子がある．これは，提案法がさまざまなプロフィールの Cancer Outlier 由来のがん関連遺伝子を候補としてあげることが出来るということを示している．

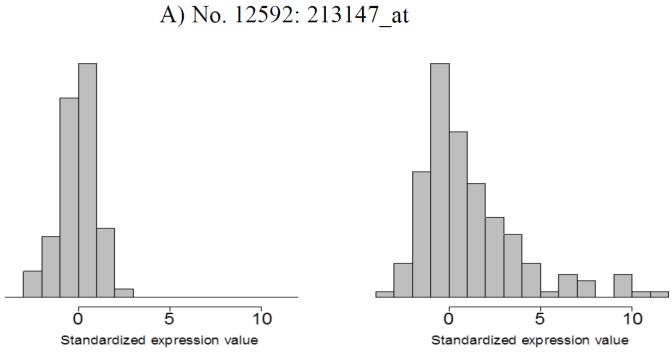

図 10　提案法の統計量で上位にも関わらず, 他の手法では上位とならなかった遺伝子 A

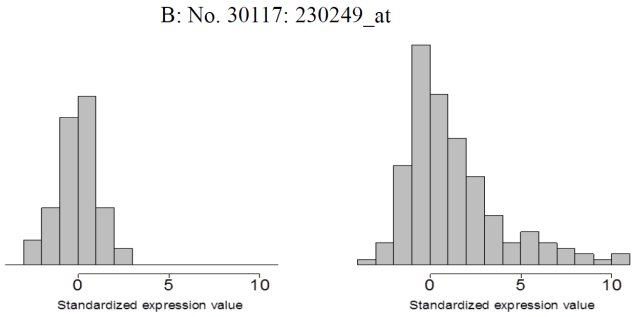

図 11　提案法の統計量で上位にも関わらず, 他の手法では上位とならなかった遺伝子 B

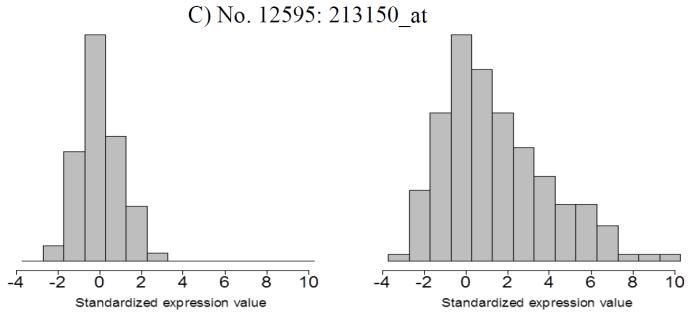

図 12　提案法の統計量で上位にも関わらず, 他の手法では上位とならなかった遺伝子 C

図 10, 図 11, 図 12 は統計量でランキングを行い，上位 200 個の遺伝子の中で，標準化した発現量を見たとき，Cancer Outlier の形，つまり一部のがんサンプルのみで高発現

している遺伝子を拾い上げることができているということを示している．このように，さまざまなプロフィールの Cancer Outlier を含むがん関連遺伝子を我々の方法では突き止めることができるということが再度示された．さらにこの 3 つの遺伝子に関して確認すると，図 10 図 12 は HOXA10 に関連しており, 図 11 は HG-U133B に関連していた．どちらも血液腫瘍に関連する細胞に関連しており，さらに HOXA10 は MDS に関連することが基礎実験のレベルでは確かめられている．これは我々の手法が，生物学的に意義のある遺伝子を関連遺伝子として同定できている可能性が高いことを示していると考えられる．

7　考察

この研究では，遺伝子の情報とがんサンプルの情報を共有を通して，Cancer Outlier の解析を効果的に行うことができるように改良した．これを示すためのシミュレーションにおいて，提案した遺伝子発現量のデータのパラメトリックな正規混合モデルを基にした遺伝子選抜法はいくつかの場合で有効なことが示された．

シミュレーションに関しては限られた状況でしか試していない．例えば，サンプルを等分とした場合のみでしかシミュレーションをしておらず，症例数がアンバランスな場合も考えるべきであるという意見も考えられる．しかし，正常サンプルが，がんサンプルに比べて多くなるときは，今回比較するこれまでに紹介してきた従来法，提案法それぞれにおいて，がんサンプル発現量データの遺伝子内標準化を考える際，真のパラメータに近づくと考えることができるため，サンプルが同程度のときはこの意味で一番検出力が低い状況でのシミュレーションを行っているということができる．また，逆にがんサンプルが多い状況は理論的には考えることができるが，実務的には，倫理的にも許容されないため，やはり，がんサンプルと正常サンプルの数が同数であるときの検討のみでよいと考えた．

また，標準正規分布の場合だけでなく，t 分布からの乱数によるシミュレーションを行うのは，後で考える実データの分布を確認したときや，裾を引く分布であることが確認できたためである．提案法では暗にデータが正規分布から発生しているため，その仮定が崩れたときにどのような振る舞いになるかも確認することが必要であると考えた．

また，シミュレーションデータを共通のモデルから生成すると，遺伝子毎の違いが反映されないため，提案法に有利に働いている可能性がある．そのためすべての遺伝子に対して共通のモデルを仮定していいかどうかは将来検討が必要である．

提案した統計量は，がん関連遺伝子の選択に効果的となるであろう．そしてがん関連遺伝子として同定された遺伝子のがんサンプルの中で主にアクティベートしているサンプルを含んでいると考えられる．今回のシミュレーションで提案法においてパフォーマンスが悪いとされた ϕ が小さいケースに関しては，将来の研究で解決していかなければならない．

また，特記すべきこととして，今回の検討は遺伝子レベルでの混合分布の構造を付け加えるということも考えることができるであろう．つまり，遺伝子はがんに関連している遺

伝子と関係していない遺伝子に分けることができるが，これにおいても遺伝子選択において，真陽性と偽陽性の評価を提供できる．我々は 混合構造において，3 つのコンポネント，f_0, f_1, f_2, において遺伝子間で共通であると仮定した．しかしいくつかのケースで，Cancer Outlier のコンポネントがさらにたくさんある場合などもがんサンプルの遺伝的異質性が大きなときも考えることができる．我々の手法は，さらに多数のコンポネントに拡張可能であり，AIC や BIC [8] などのモデル選択規準に基づいた Cancer Outlier コンポネントの数の決定なども可能であると考える．

我々のモデルでは遺伝子間での交互作用などは考えることができていない．遺伝子レベルでの混合分布で記述された調査によると (例えば [8])，相関の影響は小さいとされているが更なる研究が必要である．

従来の Cancer Outlier 解析の方法と提案法の性能評価を行ったとき，我々のシミュレーションによれば，多くの場合で我々の提案法の性能がよいことが確認できた．そのひとつの理由に，正常サンプルでのデータを対照として標準化し，がんサンプルと正常サンプルがプールされたデータは用いていないことが影響しているのではないかと考えられる．しかし，プールして計算をすればよいという簡単な話でないことが従来法を見てわかる．$\phi = 0.5$ などの Cancer Outlier の割合が大きなとき，OS 統計量のパフォーマンスは悪い．これは，正常サンプルがんサンプルをプールしたデータを下に IQR を計算しているからである．このような Cancer Outlier の数が比較的多い状況では，IQR がいくつかの Cancer Outlier を含む形として規準が作られてしまうため，統計量の性能が落ちてしまう．対照的に ORT 統計量の性能はこれを改善するために正常サンプルのみで IQR を作成しているため，純粋に正常サンプルとの乖離を考えることとなり，ϕ が大きくても，よい振る舞いをすることがわかった．

Cancer Outlier としてのプロフィールを持つがん関連遺伝子を選抜した後は，研究者は，さらに同時にコントロールされている遺伝子を同定するために遺伝子のクラスタリングなどを行うだろう．そして，同じ生物学的疾患や活性に関連するような分子を同定しようとするだろう．同時に，遺伝子クラスタ同定に基づくがんサンプル発現量のクラスタリングは Cancer Outlier のプロフィールを持った遺伝子発現に基づくがんの新しい分類の助けになることができるであろう．そして，予後や治療効果などが明確になり臨床を変える可能性があるだろう．Cancer Outlier 分析を用いての遺伝子とサンプルの two-way クラスタリングの手法は，この研究で行われたモデルベースの手法の拡張である．それは重要な

トピックであり，クラスタリングを使うひとつの手法としてこれから報告していく予定である．

8 結論

この研究ではマイクロアレー実験から生成される発現量データを用いて，がん関連遺伝子を同定するための手法について検討された．提案した統計量はすべての遺伝子，すべてのがんサンプル発現量を通して，発現量の分布に関する共通のモデルを仮定する方法であった．モデルには 3 コンポーネントの正規混合モデルを用いた．未知パラメータは EM アルゴリズムにより推定した．その推定値を用いて，それぞれのがんサンプル発現量が得られたもとでの Cancer Outlier である事後確率を計算した．この事後確率から遺伝子毎に統計量を構成し従来法と比較した．比較のためにモンテカルロ・シミュレーションを行った．シナリオとしてサンプルサイズを小中大で 3 段階，Cancer Outlier の数で 3 段階，それぞれを組み合わせて 9 つのシナリオで検討した．さらに発現量データが正規分布に従っていると考えられる場合と誤特定してしまっている場合の一つとして t 分布に従っているとかんがえられる場合の 2 つを考えた．それぞれのシナリオでそれぞれの手法の ROC 曲線を描き，比較した．結果，多くの場合において提案法が従来法よりも任意の偽発見率の時に検出力が高いことが示された．ただしデータにおける正規性の仮定が崩れ，さらに Cancer Outlier の数が少なかった場合に従来法の方が小さな偽発見率のとき高い検出力を示していた．また，実データへの適用を行った．データは Mills(2009)[9] により一般に公開されている血液腫瘍のデータを用いた．データは骨髄異形成白血病かそうでないかの二群に分かれていた．これに対して，提案法と従来法それぞれの統計量を作成した．さらにそれぞれの手法において統計量の意味で，上位 200 遺伝子をピックアップした．それぞれの手法での上位遺伝子を照合し共通している遺伝子の個数を確認した．これにより，従来法で上位で検出されている遺伝子のいくつかが同様に検出できていることが確認できた．以上のように，盛ら (2013) [10] では，Cancer Outlier を含むがん関連遺伝子の同定のための新たな統計量を提案した．

9　謝辞

本論文は，筆者が総合研究大学院大学複合科学研究科統計科学専攻在学中に松井茂之先生（初めの２年間），ならびに江口真透先生（その後の１年）のもとで研究した結果をまとめることにより構成されています．両先生には本研究に関して終始ご指導ご鞭撻を頂き，心より感謝申し上げます．また，論文を審査して頂いた先生方におかれましては大変有意義なコメントを頂くことが出来ました．松浦正明先生には，長きに渡り臨床現場で活躍されている経験からのコメントを頂けました．間野修平先生には主査になっていただき，論文に関して細かい部分までコメントしていただきました．野間久史先生におかれましては，研究を始める際から，プログラミング等基礎的なことを含めてご指導いただけました．本当にありがとうございました．

加えて実データの解析のことなど論文化の際にコメントいただけた共著者の大浦智紀先輩，さらに，お互いが同期入学，フルタイムの学生ということで何かと話すことが多かった野津昭文さんをはじめ統計数理研究所の先生方，また学生の皆様に大変感謝申し上げます．

さらに職場の上司でありました山本信之先生，現在の上司であります安井博史先生には学位取得を目指す過程で，格別の配慮をいただけましたことを御礼申しあげます．また論文の実例部分に対してコメントしてくださった今井久雄先生を初め職場の医師，その他スタッフの皆様に厚く御礼申し上げます．最後になりますが，多くの人より少しだけ長く学生でいることを認めてくれた私の両親である芳彦氏，幸子氏にも感謝申し上げます．

参考文献

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[2] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, Vol. 19, No. 2, pp. 185–193, 2003.

[3] Sung E Choe, Michael Boutros, Alan M Michelson, George M Church, and Marc S Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome biology*, Vol. 6, No. 2, p. R16, 2005.

[4] Everett R Dempster. Maintenance of genetic heterogeneity. In *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 20, pp. 25–32. Cold Spring Harbor Laboratory Press, 1955.

[5] Frank Emmert-Streib and Matthias Dehmer. *Statistical Diagnostics For Cancer: Analyzing High-Dimensional Data*. John Wiley & Sons, 2012.

[6] Takayuki Kosaka, Yasushi Yatabe, Hideki Endoh, Hiroyuki Kuwano, Takashi Takahashi, and Tetsuya Mitsudomi. Mutations of the epidermal growth factor receptor gene in lung cancer biological and clinical implications. *Cancer research*, Vol. 64, No. 24, pp. 8919–8923, 2004.

[7] Heng Lian. Most: detecting cancer differential gene expression. *Biostatistics*, Vol. 9, No. 3, pp. 411–418, 2008.

[8] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[9] Ken I Mills, Alexander Kohlmann, P Mickey Williams, Lothar Wieczorek, Weimin Liu, Rachel Li, Wen Wei, David T Bowen, Helmut Loeffler, Jesus M Hernandez, et al. Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of aml transformation of myelodys-

plastic syndrome. *Blood*, Vol. 114, No. 5, pp. 1063–1072, 2009.

[10] Keita Mori, Tomonori Oura, Hisashi Noma, and Shigeyuki Matsui. Cancer outlier analysis based on mixture modeling of gene expression data. *Computational and mathematical methods in medicine*, Vol. 2013, , 2013.

[11] Stephen G O'Brien, Francois Guilhot, John M Goldman, Andreas Hochhaus, Timothy P Hughes, Jerald P Radich, Marc Rudoltz, Jeiry Filian, Insa Gathmann, Brian J Druker, et al. International randomized study of interferon versus sti571 (iris) 7-year follow-up: sustained survival, low rate of transformation and increased rate of major molecular response (mmr) in patients (pts) with newly diagnosed chronic myeloid leukemia in chronic phase (cmlcp) treated with imatinib (im). In *ASH Annual Meeting Abstracts*, Vol. 112, p. 186, 2008.

[12] Kerri E Rieger, Wan-Jen Hong, Virginia Goss Tusher, Jean Tang, Robert Tibshirani, and Gilbert Chu. Toxicity from radiation therapy associated with abnormal transcriptional responses to dna damage. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 17, pp. 6635–6640, 2004.

[13] David Ruppert. *Statistics and data analysis for financial engineering.* Springer Texts in Statistics., 2011.

[14] Manabu Soda, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina, Hisashi Hatanaka, et al. Identification of the transforming eml4–alk fusion gene in non-small-cell lung cancer. *Nature*, Vol. 448, No. 7153, pp. 561–566, 2007.

[15] Kengo Takeuchi, Young Lim Choi, Manabu Soda, Kentaro Inamura, Yuki Togashi, Satoko Hatano, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Yukitoshi Satoh, et al. Multiplex reverse transcription-pcr screening for eml4-alk fusion transcripts. *Clinical Cancer Research*, Vol. 14, No. 20, pp. 6618–6624, 2008.

[16] Kengo Takeuchi, Manabu Soda, Yuki Togashi, Ritsuro Suzuki, Seiji Sakata, Satoko Hatano, Reimi Asaka, Wakako Hamanaka, Hironori Ninomiya, Hirofumi

Uehara, et al. Ret, ros1 and alk fusions in lung cancer. *Nature medicine*, Vol. 18, No. 3, pp. 378–381, 2012.

[17] Robert Tibshirani and Trevor Hastie. Outlier sums for differential gene expression analysis. *Biostatistics*, Vol. 8, No. 1, pp. 2–8, 2007.

[18] Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *Science*, Vol. 310, No. 5748, pp. 644–648, 2005.

[19] John C Torrey. A comparative study of dysentery and dysentery-like organisms. *The Journal of experimental medicine*, Vol. 7, No. 4, pp. 365–384, 1905.

[20] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, Vol. 98, No. 9, pp. 5116–5121, 2001.

[21] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A Olson, Jeffrey R Marks, and Joseph R Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, Vol. 98, No. 20, pp. 11462–11467, 2001.

[22] Baolin Wu. Cancer outlier differential gene expression detection. *Biostatistics*, Vol. 8, No. 3, pp. 566–575, 2007.

付録

モンテカルロシミュレーションに使用した R のプログラムソースを以下に示す.

従来法関数

従来法の統計量をデータから作成するための統計量は以下のとおり.

```
#従来法
cal.6.statistics<-function(
sample
,n1
,n2
,copa.quantile=0.9
){
rowNames <- rownames(sample)
#common
n<-n1+n2;samplex<-sample[,1:n1]; sampley<-sample[,(n1+1):n]
medx<-apply(samplex,1,median)
medy<-apply(sampley,1,median)
med<-apply(sample,1,median)
diffxmedx<-abs(samplex-medx)
diffymedy<-abs(sampley-medy)
stdforeachrow<-apply(cbind(diffxmedx,diffymedy),1,median)+0.01
#most
{
result<-array(0,dim=c(dim(sample)[1],n2))
temp<-generateorder(n2)
a<-temp[,1]; b<-temp[,2];
for(ktest in 1:n2)
{
if (ktest>1){
```

```
meanoutlier<-rowMeans(sampley[,1:ktest])
} else {
meanoutlier<-sampley[,1]
}
result[,ktest]<-((meanoutlier-medx)*ktest/stdforeachrow/1.4826-a[ktest])/b[ktest]
}
most<-apply(result,1,max)
}
#ort
{
threshold1<-apply(samplex,MARGIN=1,FUN=quantile,probs=0.75)+apply(samplex,1,IQR)
outlierflag<-(sampley>threshold1)
ortstatistics<-(rowSums(sampley*outlierflag)-
rowSums(outlierflag)*medx)/stdforeachrow
}
#os
{
threshold2 <- apply(sample,MARGIN=1,FUN=quantile,probs=0.75)+apply(sample,1,IQR)
outlierflag <- (sampley > threshold2)
osstdforeachrow <- apply(abs(sample-apply(sample,1,median)),1,median)+0.01
osstatistics <- (rowSums(sampley*outlierflag)-
rowSums(outlierflag)*med)/osstdforeachrow
}
#copa
{
q90<-apply(sampley,1,quantile,probs=copa.quantile)
copastdforeachrow<-apply(abs(sample-med),1,median)+0.01
copastatistics<-(q90-med)/copastdforeachrow
comment(copastatistics) <- paste("q=",copa.quantile,sep="")
}
#t
{
meanx<-rowMeans(samplex); meany<-rowMeans(sampley);
```

```
tvar<-rowSums((samplex-meanx)^2)+rowSums((sampley-meany)^2)
tstatistics<-(meany-meanx)/(sqrt(tvar)+0.01)
}
names(most) <- rowNames
names(ortstatistics) <- rowNames
names(osstatistics) <- rowNames
names(copastatistics) <- rowNames
names(tstatistics) <- rowNames
return(list(
most=most
,ort=ortstatistics
,os=osstatistics
,copa=copastatistics
,t=tstatistics
))
}#cal.6.statistics
generateorder<-function(n){
#used by calculation of most
r.sample <- rnorm(1000*n)
msample <- matrix(r.sample,nrow=1000)
ordered <- apply(msample,MARGIN=1,FUN=sort,decreasing=T)
cumsumordered <- apply(ordered,2,cumsum)
# a <- apply(cumsumordered,1,mean)
mean <- rowMeans(cumsumordered)
cumsumordered <- cumsumordered - mean
b <- apply(cumsumordered,1,sd)
cbind(mean,b)
}#generateorder
```

シミュレーションソース

#moment 法分散推定

```
 fn.var <- function(x, mo1,mo2,mo3,mo4,mo5,mo6) {
 p1<-x[1]
 m1<-x[2]
 p2<-x[3]
 m2<-x[4]
 v 1<-x[5]
 v2<-x[6]
 c(mo1-(p1*m1+p2*m2)
,mo2-(p1*(m1^2+v1^2)+p2*(m2^2+v2^2)+(1-p1-p2))
,mo3-(p1*(m1^3+3*m1*v1^2)+p2*(m2^3+3*m2*v2^2))
,mo4-(p1*(m1^4 + 6*(m1^2)*(v1^2) + 3*v1^4 )
+p2*( m2^4 + 6*(m2^2) *(v2^4) +3*v2^4)+3*(1-p1-p2))
,mo5-(p1*(m1^5+10*(m1^3)*(v1^2)+15*m1*v1^4)
+p2*( m2^5 +10*(m2^3) *(v2^2)+15*m2*v2^4))
,mo6-(p1*(m1^6+15*(m1^4)*(v1^2)+45*(m1^2)*(v1^2)+15*v1^6)
+p2*(m2^6+15*(m2^4)*(v2^2)+45*(m2^2)*(v2^2)+15*v2^6)+15*(1-p1-p2)))
 }

#moment 法分散推定無
 fn.non.var <- function(x, mo1,mo2,mo3,mo4) {
 p1<-x[1]
 m1<-x[2]
 p2<-x[3]
 m2<-x[4]
 c(mo1-(p1*m1+p2*m2)
,mo2-(p1*(m1^2+1)+p2*(m2^2+1)+(1-p1-p2))
,mo3-(p1*(m1^3+3*m1)+p2*(m2^3+3*m2))
,mo4-(p1*(m1^4 + 6*(m1^2)+ 3 )+p2*( m2^4 + 6*(m2^2) +3)+3*(1-p1-p2))
)
 }
Time0<-Sys.time()
set.seed(8101)
############シナリオ変数######################
```

```
library(nleqslv)
#関連あり遺伝子の数
PI0 <- 0.6
#遺伝指数
m <-10000
#繰り返し回数
re<-1
#関連なし遺伝子の割合
nonnull.prop <- 1-PI0
#データセット内の関連なし遺伝子の個数
m1 <- m*nonnull.prop
################################################
#提案法の false discovery rate および true positive rate の計算結果を入れて
おく変数
fp.mix.var<- matrix(1,m,re)
tp.mix.var<- matrix(1,m,re)
Fp.mix.var<- Tp.mix <- numeric(m)
#var.fp.mix.max<- matrix(1,m,re)
#var.tp.mix.max<- matrix(1,m,re)
#var.Fp.mix.max<- Tp.mix <- numeric(m)
fp.mix.non.var<- matrix(1,m,re)
tp.mix.non.var<- matrix(1,m,re)
Fp.mix.non.var<- Tp.mix <- numeric(m)
#non.var.fp.mix.max.non.var<- matrix(1,m,re)
#non.var.tp.mix.max<- matrix(1,m,re)
#non.var.Fp.mix.max<- Tp.mix <- numeric(m)
var.ERR1<-NULL
non.var.ERR1<-NULL
################################################
#従来法結果をいれておく変数
fp.t <- matrix(1,m,re)
fp.copa <- matrix(1,m,re)
fp.os <- matrix(1,m,re)
```

```
fp.ort <- matrix(1,m,re)
fp.most <- matrix(1,m,re)
tp.t <- matrix(1,m,re)
tp.copa <- matrix(1,m,re)
tp.os <- matrix(1,m,re)
tp.ort <- matrix(1,m,re)
tp.most <- matrix(1,m,re)
Fp.t <- Tp.t <- numeric(m)
Fp.copa<- Tp.copa <- numeric(m)
Fp.os <- Tp.os <- numeric(m)
Fp.ort <- Tp.ort <- numeric(m)
Fp.most<- Tp.most <- numeric(m)
################################################
non.var.pi0<-NULL
non.var.pi1<-NULL
non.var.mu1<-NULL
non.var.pi2<-NULL
non.var.mu2<-NULL
non.var.par1<-matrix(0,1000,5)#変数モニタリングのため
var.pi0<-NULL
var.pi1<-NULL
var.mu1<-NULL
var.pi2<-NULL
var.mu2<-NULL
var.sd1<-NULL
var.sd2<-NULL
var.par1<-matrix(0,1000,7)#変数モニタリングのため
#data_set#########################################################
#サンプルサイズ
n0 <- n1<-20 ; n <-n0+n1
#n0 <- n1<-30 ; n <-n0+n1
n0 <- n1<-40 ; n <-n0+n1
#n0 <- n1<-50 ; n <-n0+n1
```

```
n0 <- n1<-60 ; n <-n0+n1
#n0 <- n1<-70 ; n <-n0+n1
n0 <- n1<-80 ; n <-n0+n1
#n0 <- n1<-90 ; n <-n0+n1
n0 <- n1<-100 ; n <-n0+n1
#効果サイズ
delta <-mu<-2
#がんサンプル内の Cancer Outlier の割合
#qq<-0.1
#qq<-0.2
#qq<-0.3
#qq<-0.4
qq<-0.5
#qq<-0.6
#qq<-0.7
#qq<-0.8
##########detaset 作成#########################
for(q in 1:re){
#標準正規乱数
DATA<- matrix(rnorm(m*n),m,n)
#t 分布からの乱数
#DATA<- matrix(rt(m*n,20),m,n)
#Cancer Outlier の部分のセッティング
DATA[1:((m*(1-PI0))/2),(n1+1):(n0+n0*qq)]
<-(DATA[1:((m*(1-PI0))/2),(n1+1):(n0+n0*qq)]+mu)
DATA[(((m*(1-PI0))/2)+1):(m*(1-PI0)),(n1+1):(n0+n0*qq)]
<-(DATA[(((m*(1-PI0))/2)+1):(m*(1-PI0)),(n1+1):(n0+n0*qq)]-mu)
{
#data の標準化
datastand
<-(DATA-apply(DATA[,1:n0],1,mean))/(apply(DATA[,1:n0],1,sd))
datastandy<-datastand[,(n0+1):n]
y <- matrix(datastandy,1,m*n1)
```

```
##############################################
fn1.var <- function(x) fn.var(x, mean(y),
mean(y^2),mean(y^3),mean(y^4),mean(y^5),mean(y^6))
ans.var<-nleqslv(c(.33, mean( sort(y)[1:round(m*n1/3)])
,.33,mean(sort(y)[(round(m*n1*(2/3))+1):(m*n1)])
 ,sd(sort(y)[1:round(m*n1/3)]) ,sd(sort(y)[(round(m*n1*(2/3))+1):(m*n1)])),fn1.var)
fn1.non.var <- function(x) fn.non.var(x, mean(y), mean(y^2),mean(y^3),mean(y^4))
ans.non.var<-nleqslv( c(.33, mean( sort(y)[1:round(m*n1/3)]) ,.33
,mean(sort(y)[(round(m*n1*(2/3))+1):(m*n1)])) ,fn1.non.var)
pi0 <- 1-ans.var$x[1]-ans.var$x[3]
pi1 <- ans.var$x[1]
mu1 <- ans.var$x[2]
pi2 <- ans.var$x[3]
mu2 <- ans.var$x[4]
sd1<- ans.var$x[5]
sd2<- ans.var$x[6]
var.par0 <- c(pi0, pi1, mu1,pi2,mu2,sd1,sd2 )
time0 <- Sys.time()
#EMアルゴリズム#####################################################################
for(i in 1:1000){
mlike0 <- pi0 * dnorm(y)
mlike1 <- pi1 * dnorm(y, mean=mu1, sd=sd1)#負のCancer Outlier component
mlike2 <- pi2 * dnorm(y, mean=mu2, sd=sd2)#正のCancer Outlier component
tau0 <- mlike0 / (mlike0 + mlike1 + mlike2)+0.000000001
tau1 <- mlike1 / (mlike0 + mlike1 + mlike2)+0.000000001
tau2 <- 1-tau0 - tau1
pi0 <- mean(tau0)
pi1 <- mean(tau1)
pi2 <- 1 - pi0 - pi1
mu1 <- sum(tau1 * y) / sum(tau1)
sd1 <- sqrt(sum(tau1 * (y - mu1)^2) / sum(tau1))
mu2 <- sum(tau2 * y) / sum(tau2)
sd2 <- sqrt(sum(tau2 * (y - mu2)^2) / sum(tau2))
```

```
var.par1[i,] <- c(pi0, pi1, mu1,pi2,mu2,sd1,sd2)
err1 <- min(abs(var.par1[i,] - var.par0)/abs(var.par1[i,]))
var.ERR1<-c(var.ERR1,err1)
if(err1 < 10^-8) break
var.par0 <- var.par1[i,]
time1 <- Sys.time() - time0
cat("var",q,i, var.par1[i,], err1, time1, "\n")
}
pi0 <- 1-ans.non.var$x[1]-ans.non.var$x[3]
pi1 <- ans.non.var$x[1]
mu1 <- ans.non.var$x[2]
pi2 <- ans.non.var$x[3]
mu2 <- ans.non.var$x[4]
non.var.par0 <- c(pi0, pi1, mu1,pi2,mu2 )
time0 <- Sys.time()
#EMアルゴリズム#####################################################################
for(i in 1:1000){
mlike0 <- pi0 * dnorm(y)
mlike1 <- pi1 * dnorm(y, mean=mu1, sd=1)#負の Cancer Outlier component
mlike2 <- pi2 * dnorm(y, mean=mu2, sd=1)#正の Cancer Outlier component
tau0 <- mlike0 / ((mlike0 + mlike1 + mlike2)+0.000000001)
tau1 <- mlike1 / ((mlike0 + mlike1 + mlike2)+0.000000001)
tau2 <- 1-tau0 - tau1
pi0 <- mean(tau0)
pi1 <- mean(tau1)
pi2 <- 1 - pi0 - pi1
mu1 <- sum(tau1 * y) / sum(tau1)
mu2 <- sum(tau2 * y) / sum(tau2)
non.var.par1[i,] <- c(pi0, pi1, mu1,pi2,mu2)
err1 <- min(abs(non.var.par1[i,] - non.var.par0)/abs(non.var.par1[i,]))
non.var.ERR1<-c(non.var.ERR1,err1)
if(err1 < 10^-8) break
non.var.par0 <- non.var.par1[i,]
```

```
time1 <- Sys.time() - time0
cat("non.var",i, non.var.par1[i,], err1, time1, "\n")
}
#収束モニタリング
var.pi0<-c(var.pi0,var.par0[1])
var.pi1<-c(var.pi1,var.par0[2])
var.mu1<-c(var.mu1,var.par0[3])
var.pi2<-c(var.pi2,var.par0[4])
var.mu2<-c(var.mu2,var.par0[5])
var.sd1<-c(var.sd1,var.par0[6])
var.sd2<-c(var.sd2,var.par0[7])
#収束モニタリング
non.var.pi0<-c(non.var.pi0,non.var.par0[1])
non.var.pi1<-c(non.var.pi1,non.var.par0[2])
non.var.mu1<-c(non.var.mu1,non.var.par0[3])
non.var.pi2<-c(non.var.pi2,non.var.par0[4])
non.var.mu2<-c(non.var.mu2,non.var.par0[5])
#3comp
var.w<- ( mean(var.pi0) * dnorm(datastandy,0,1)
+mean(var.pi1) *dnorm(datastandy,mean(var.mu1),mean(var.sd1) ) )
  / ( mean(var.pi0) * dnorm(datastandy,0,1) +
mean(var.pi1) * dnorm(datastandy,mean(var.mu1),mean(var.sd1))
+mean(var.pi2)*dnorm(datastandy,mean(var.mu2),mean(var.sd2))+0.0000001)#var.w_gi
non.var.w<- ( mean(non.var.pi0) * dnorm(datastandy,0,1)
+mean(non.var.pi1) *dnorm(datastandy,mean(non.var.mu1),1 ) )
/ ( mean(non.var.pi0) * dnorm(datastandy,0,1)
+mean(non.var.pi1) * dnorm(datastandy,mean(non.var.mu1),1)
+mean(non.var.pi2)*dnorm(datastandy,mean(non.var.mu2),1)+0.0000001)#non.var.w_gi
var.mix<-var.w
u<-apply(var.mix,1,prod)
#遺伝内の一番 Cancer Outlier である事後確率が高いものを統計量として採用する手法
(MAX)
#u.max<-apply(mix,1,min)
```

```
var.MIX<-u
#MAX<-u.max
var.order.mix<-order(var.MIX)
#order.mix.max<-order(MAX)
f<-NULL
f<-(var.order.mix<=(m1/2))
#f.max<-NULL
#f.max<-(order.mix.max<=(m1/2))
g<-NULL
g<-cumsum(f)
g.max<-NULL
#g.max<-cumsum(f.max)
j<-NULL
j<-1:m
fp.mix.var[,q] <- (j-g)/j
tp.mix.var[,q] <- g/(m1/2)
#fp.mix.max[,q] <- (j-g.max)/j
#tp.mix.max[,q] <- g.max/(m1/2)
non.var.mix<-non.var.w
u<-apply(non.var.mix,1,prod)
#遺伝内の一番 Cancer Outlier である事後確率が高いものを統計量として採用する手法
(MAX)
#u.max<-apply(mix,1,min)
non.var.MIX<-u
#MAX<-u.max
non.var.order.mix<-order(non.var.MIX)
#order.mix.max<-order(MAX)
f<-NULL
f<-(var.order.mix<=(m1/2))
#f.max<-NULL
#f.max<-(order.mix.max<=(m1/2))
g<-NULL
g<-cumsum(f)
```

```
g.max<-NULL
#g.max<-cumsum(f.max)
j<-NULL
j<-1:m
fp.mix.non.var[,q] <- (j-g)/j
tp.mix.non.var[,q] <- g/(m1/2)
#fp.mix.max[,q] <- (j-g.max)/j
#tp.mix.max[,q] <- g.max/(m1/2)
}
#従来法計算して a にいれる
a<-cal.6.statistics(DATA,n0,n1)
t.stat<-a$t
copa.stat<-a$copa
os.stat<-a$os
ort.stat<-a$ort
most.stat<-a$most
t.order<-order(t.stat,decreasing=T)
copa.order<-order(copa.stat,decreasing=T)
os.order<-order(os.stat,decreasing=T)
ort.order<-order(ort.stat,decreasing=T)
most.order<-order(most.stat,decreasing=T)
t.signal<- t.order<=(m1/2)
copa.signal<- copa.order<=(m1/2)
os.signal<- os.order<=(m1/2)
ort.signal<- ort.order<=(m1/2)
most.signal<- most.order<=(m1/2)
t.g<-cumsum(t.signal)
copa.g<-cumsum(copa.signal)
os.g<-cumsum(os.signal)
ort.g<-cumsum(ort.signal)
most.g<-cumsum(most.signal)
j<-1:m
fp.t[,q]<- (j-t.g)/j
```

```
tp.t[,q]<- t.g / (m1/2)
fp.copa[,q]<- (j-copa.g)/j
tp.copa[,q]<- copa.g / (m1/2)
fp.os[,q]<- (j-os.g)/j
tp.os[,q]<- os.g /(m1/2)
fp.ort[,q]<- (j-ort.g)/j
tp.ort[,q]<- ort.g / (m1/2)
fp.most[,q]<- (j-most.g)/j
tp.most[,q]<- most.g / (m1/2)
Sys.time()-Time0
}
Fp.t<-apply( fp.t ,1,mean)
Tp.t<-apply( tp.t,1,mean)
Fp.copa<-apply(fp.copa,1,mean)
Tp.copa<-apply(tp.copa,1,mean)
Fp.os<-apply(fp.os,1,mean)
Tp.os <-apply(tp.os,1,mean)
Fp.ort<-apply(fp.ort,1,mean)
Tp.ort <-apply(tp.ort,1,mean)
Fp.most<-apply(fp.most,1,mean)
Tp.most<-apply(tp.most,1,mean)
Fp.mix.var<-apply(fp.mix.var,1,mean)
Tp.mix.var<-apply(tp.mix.var,1,mean)
Fp.mix.non.var<-apply(fp.mix.non.var,1,mean)
Tp.mix.non.var<-apply(tp.mix.non.var,1,mean)
#Fp.mix.max<-apply(fp.mix.max,1,mean)
#Tp.mix.max<-apply(tp.mix.max,1,mean)
Sys.time()-Time0
```

性能評価のための ROC 曲線をプロットして jpeg 画像にするために必要なコード

```
jpeg("comp.t,200,0.5.jpeg")
```

```
plot (c(100,100),xlim=c(0,1),ylim=c(0,1),xlab="FDR",ylab="TPR")
lines(Tp.t~Fp.t, lty=2,col="purple")
lines(Tp.copa~Fp.copa, lty=3,col="#006600")
lines(Tp.os~Fp.os, lty=4,col="#ff0099")
lines(Tp.ort~Fp.ort, lty=5,col="black")
lines(Tp.most~Fp.most, lty=6,col="orange")
lines(Tp.mix ~Fp.mix, lty=1,col="red")
legend("topright",lty=c(1,2,3,4,5,6),
col=c("red","purple","#006600","#ff0099","black","orange"),
legend=c("mix","t","copa","os","ort","most"))
legend("bottomright",legend=c("delta",mu,"n",n,"pi",1-nonnull.prop,"q",qq))
dev.off()
```

実データ解析ソース

研究で用いた実データを解析するためのソース

```
 fn.var <- function(x, mo1,mo2,mo3,mo4,mo5,mo6) {
 p1<-x[1]
 m1<-x[2]
 p2<-x[3]
 m2<-x[4]
 v 1<-x[5]
 v2<-x[6]
 c(mo1-(p1*m1+p2*m2)
,mo2-(p1*(m1^2+v1^2)+p2*(m2^2+v2^2)+(1-p1-p2))
,mo3-(p1*(m1^3+3*m1*v1^2)+p2*(m2^3+3*m2*v2^2))
,mo4-(p1*(m1^4 + 6*(m1^2)*(v1^2) + 3*v1^4 );
+p2*( m2^4 + 6*(m2^2) *(v2^4) +3*v2^4)+3*(1-p1-p2))
,mo5-(p1*(m1^5+10*(m1^3)*(v1^2)+15*m1*v1^4);
+p2*( m2^5 +10*(m2^3) *(v2^2)+15*m2*v2^4))
,mo6-(p1*(m1^6+15*(m1^4)*(v1^2);
+45*(m1^2)*(v1^2)+15*v1^6);
```

```
+p2*(m2^6+15*(m2^4)*(v2^2);
+45*(m2^2)*(v2^2)+15*v2^6)+15*(1-p1-p2)))
 }

#moment 法分散推定無
 fn.non.var <- function(x, mo1,mo2,mo3,mo4) {
 p1<-x[1]
 m1<-x[2]
 p2<-x[3]
 m2<-x[4]
 c(mo1-(p1*m1+p2*m2)
,mo2-(p1*(m1^2+1)+p2*(m2^2+1)+(1-p1-p2))
,mo3-(p1*(m1^3+3*m1)+p2*(m2^3+3*m2))
,mo4-(p1*(m1^4 + 6*(m1^2)+ 3 )+p2*( m2^4 + 6*(m2^2) +3)+3*(1-p1-p2))
)
AML<-gexp[,c(1:199,433:435)]
MDS<-gexp[,200:363]
NON<-gexp[,364:432]
############変数群#######################
n0 <- dim(NON)[2]
n1 <- dim(MDS)[2]
n <-n0 + n1
m <-dim(NON)[1]
DATA<-cbind(NON,MDS)
################################################
var.ERR1<-NULL
non.var.ERR1<-NULL
###############################################
non.var.pi0<-NULL
non.var.pi1<-NULL
non.var.mu1<-NULL
non.var.var1<-NULL
non.var.par1<-matrix(0,1000,5)
```

```
var.pi0<-NULL
var.pi1<-NULL
var.mu1<-NULL
var.pi2<-NULL
var.mu2<-NULL
var.sd1<-NULL
var.sd2<-NULL
var.par1<-matrix(0,1000,7)#変数モニタリングのため
###########mix#####################################
{
#data の標準化
datastand<-(DATA-apply(DATA[,1:n0],1,mean))/(apply(DATA[,1:n0],1,sd))
datastandy<-datastand[,(n0+1):n]
y <- matrix(datastandy,1,m*n1)
fn1.non.var <- function(x) fn.non.var(x, mean(y), mean(y^2),mean(y^3),mean(y^4))
ans.non.var<-nleqslv( c(.33, mean( sort(y)[1:round(m*n1/3)]) ,;
.33,mean(sort(y)[(round(m*n1*(2/3))+1):(m*n1)])) ,fn1.non.var)
pi0 <- 1-ans.var$x[1]-ans.var$x[3]
pi1 <- ans.var$x[1]
mu1 <- ans.var$x[2]
pi2 <- ans.var$x[3]
mu2 <- ans.var$x[4]

sd1<- ans.var$x[5]

sd2<- ans.var$x[6]
var.par0 <- c(pi0, pi1, mu1,pi2,mu2,sd1,sd2 )

time0 <- Sys.time()
#EM アルゴリズム#################################################################
for(i in 1:1000){
mlike0 <- pi0 * dnorm(y)
mlike1 <- pi1 * dnorm(y, mean=mu1, sd=sd1)#負の Cancer Outlier component
```

```
mlike2 <- pi2 * dnorm(y, mean=mu2, sd=sd2)#正のCancer Outlier component
tau0 <- mlike0 / (mlike0 + mlike1 + mlike2)+0.000000001
tau1 <- mlike1 / (mlike0 + mlike1 + mlike2)+0.000000001
tau2 <- 1-tau0 - tau1

pi0 <- mean(tau0)
pi1 <- mean(tau1)
pi2 <- 1 - pi0 - pi1
mu1 <- sum(tau1 * y) / sum(tau1)
sd1 <- sqrt(sum(tau1 * (y - mu1)^2) / sum(tau1))
mu2 <- sum(tau2 * y) / sum(tau2)
sd2 <- sqrt(sum(tau2 * (y - mu2)^2) / sum(tau2))
var.par1[i,] <- c(pi0, pi1, mu1,pi2,mu2,sd1,sd2)
err1 <- min(abs(var.par1[i,] - var.par0)/abs(var.par1[i,]))
var.ERR1<-c(var.ERR1,err1)
if(err1 < 10^-8) break
var.par0 <- var.par1[i,]
time1 <- Sys.time() - time0
cat("var",i, var.par1[i,], err1, time1, "\n")
}

pi0 <- 1-ans.non.var$x[1]-ans.non.var$x[3]
pi1 <- ans.non.var$x[1]
mu1 <- ans.non.var$x[2]
pi2 <- ans.non.var$x[3]
mu2 <- ans.non.var$x[4]



non.var.par0 <- c(pi0, pi1, mu1,pi2,mu2 )
time0 <- Sys.time()
#EMアルゴリズム#####################################################################
for(i in 1:1000){
mlike0 <- pi0 * dnorm(y)
```

```
mlike1 <- pi1 * dnorm(y, mean=mu1, sd=1)#負の Cancer Outlier component
mlike2 <- pi2 * dnorm(y, mean=mu2, sd=1)#正の Cancer Outlier component
tau0 <- mlike0 / ((mlike0 + mlike1 + mlike2)+0.000000001)
tau1 <- mlike1 / ((mlike0 + mlike1 + mlike2)+0.000000001)
tau2 <- 1-tau0 - tau1

pi0 <- mean(tau0)
pi1 <- mean(tau1)
pi2 <- 1 - pi0 - pi1
mu1 <- sum(tau1 * y) / sum(tau1)
mu2 <- sum(tau2 * y) / sum(tau2)
non.var.par1[i,] <- c(pi0, pi1, mu1,pi2,mu2)
err1 <- min(abs(non.var.par1[i,] - non.var.par0)/abs(non.var.par1[i,]))
non.var.ERR1<-c(non.var.ERR1,err1)
if(err1 < 10^-8) break
non.var.par0 <- non.var.par1[i,]
time1 <- Sys.time() - time0
cat("non.var",i, non.var.par1[i,], err1, time1, "\n")
}
non.var.pi0<-c(non.var.pi0,non.var.par0[1])
non.var.pi1<-c(non.var.pi1,non.var.par0[2])
non.var.mu1<-c(non.var.mu1,non.var.par0[3])
non.var.pi2<-c(non.var.pi2,non.var.par0[4])
non.var.mu2<-c(non.var.mu2,non.var.par0[5])
pi0.500<-non.var.par1[500,1]
pi1.500<-non.var.par1[500,2]
mu1.500<-non.var.par1[500,3]
pi2.500<-non.var.par1[500,4]
mu2.500<-non.var.par1[500,5]
#3comp
w<- ( pi0 * dnorm(datastandy,0,1)+pi2 * dnorm(datastandy,mu2.500,1))
 / ( pi0 * dnorm(datastandy,0,1) + pi1 * dnorm(datastandy,mu1,1)
 +pi2*dnorm(datastandy,mu2,1)+0.0000001)#w_gi
```

```
mix<-w
u<-apply(mix,1,prod)
MIX<-u
order.mix<-order(MIX)
}
a<-cal.6.statistics(DATA,n0,n1)
t.stat<-a$t
copa.stat<-a$copa
os.stat<-a$os
ort.stat<-a$ort
most.stat<-a$most
t.order<-order(t.stat,decreasing=T)
copa.order<-order(copa.stat,decreasing=T)
os.order<-order(os.stat,decreasing=T)
ort.order<-order(ort.stat,decreasing=T)
most.order<-order(most.stat,decreasing=T)
```

各シナリオでのパラメータ推定値

今回本文で提案する方法，つまり，正常サンプルでがんサンプルの標準化した標本平均と標本分散を用いて行う方法を考えたが，標準化の方法で他を使うものを考えていた．まず遺伝子内の発現量の標準化を遺伝子全体で行う方法と正常サンプル発現量のみを用いて標準化を行う方法である．更に，平均と分散を使う方法とは別に中央値と絶対中央偏差を用いる方法である．それぞれの 2 種類ずつの組み合わせにより，4 種類の方法を考えることが出来る．単に遺伝子ごとに正常サンプル，がんサンプルそれぞれの発現量をプールして，標本平均と標本分散で標準化を行う手法は，明らかに Cancer Outlier と指定することになるサンプルを標準化に含んでしまっているので，その検討ははずした．ただし中央値と絶対標準偏差で標準化を行う場合は，ロバストな代表値を用いているため outlier サンプルの割合によってはどうなるかが不明であったため，この検討には残した．そのため sd1 は正常サンプル発現量のみで標準化，mad1 は正常サンプルのみの中央値，絶対中央偏差で標準化，mad2 は正常サンプル，がんサンプルすべてを遺伝子ごとにプールして標準化を行っている．この場合にモンテカルロシミュレーションを行い，どの手法が過大推定，過小推定しやすいかを確かめた．そのシミュレーションの結果を以下に示す．シナリオは本文と同様に正常サンプルの発現量が標準正規分布に従っている場合，裾の重い分布になっている場合として t 分布からのデータ発生の場合も検討した．それぞれのがんサンプル内での Cancer Outlier 発現量の割合も 0.1,0.3,0.5 とうごかし，それぞれの場合に対してパラメータの推定値がどのようになるか確認した．真の値はそれぞれのシナリオでのパラメータ真値である．

表 2　n=40 のシナリオで様々な標準化法によるパラメータの推定値のまとめ

遺伝子内の outlier サンプルの割合=0.1		π_0	π_1	μ_1	π_2	μ_2
	真の値	96%	2%	2	2%	-2
正規分布	sd1	93%	4%	2.2	4%	-2.1
	mad1	90%	5%	2.4	5%	-2.4
	mad2	95%	3%	2.6	2%	-2.7
自由度 20	sd1	94%	3%	2.4	3%	-2.4
の t 分布	mad1	90%	5%	2.6	5%	-2.6
	mad2	94%	3%	2.8	3%	-2.8

遺伝子内の outlier サンプルの割合=0.3		π_0	π_1	μ_1	π_2	μ_2
	真の値	88%	6%	2	6%	-2
正規分布	sd1	87%	7%	2.3	7%	-2.3
	mad1	85%	7%	2.5	7%	-2.6
	mad2	92%	4%	2.5	4%	-2.6
自由度 20	sd1	89%	6%	2.4	6%	-2.4
の t 分布	mad1	86%	7%	2.6	7%	-2.6
	mad2	92%	4%	2.7	4%	-2.7

遺伝子内の outlier サンプルの割合=0.5		π_0	π_1	μ_1	π_2	μ_2
	真の値	80%	10%	2	10%	-2
正規分布	sd1	80%	10%	2.3	10%	-2.3
	mad1	80%	10%	2.6	10%	-2.6
	mad2	91%	5%	2.4	5%	-2.4
自由度 20	sd1	83%	9%	2.3	9%	-2.4
の t 分布	mad1	80%	10%	2.6	10%	-2.7
	mad2	91%	4%	2.5	5%	-2.5

表 3　n=80 のシナリオで様々な標準化法によるパラメータの推定値のまとめ

遺伝子内の outlier サンプルの割合=0.1		π_0	π_1	μ_1	π_2	μ_2
	真の値	96%	2%	2	2%	-2
正規分布	sd1	94%	3%	2	3%	-2.1
	mad1	93%	4%	2.5	4%	-2.5
	mad2	95%	2%	2.7	2%	-2.6
自由度 20	sd1	95%	2%	2.4	2%	-2.3
の t 分布	mad1	92%	4%	2.6	4%	-2.6
	mad2	95%	3%	2.7	3%	-2.7

遺伝子内の outlier サンプルの割合=0.3		π_0	π_1	μ_1	π_2	μ_2
	真の値	88%	6%	2	6%	-2
正規分布	sd1	88%	6%	2.1	6%	-2.2
	mad1	88%	6%	2.6	6%	-2.6
	mad2	93%	4%	2.5	4%	-2.6
自由度 20	sd1	89%	6%	2.2	6%	-2.2
の t 分布	mad1	87%	6%	2.6	6%	-2.6
	mad2	92%	4%	2.6	4%	-2.6

遺伝子内の outlier サンプルの割合=0.5		π_0	π_1	μ_1	π_2	μ_2
	真の値	80%	10%	2	10%	-2
正規分布	sd1	81%	10%	2.2	10%	-2.2
	mad1	82%	9%	2.6	9%	-2.6
	mad2	91%	4%	2.4	4%	-2.4
自由度 20	sd1	82%	9%	2.2	9%	-2.2
の t 分布	mad1	82%	9%	2.6	9%	-2.6
	mad2	91%	4%	2.4	4%	-2.4

表 4　n=200 のシナリオで様々な標準化法によるパラメータの推定値のまとめ

遺伝子内の outlier サンプルの割合=0.1		π_0	π_1	μ_1	π_2	μ_2
	真の値	96%	2%	2	2%	-2
正規分布	sd1	95%	2%	2	2%	-2
	mad1	94%	3%	2.6	3%	-2.6
	mad2	96%	2%	2.7	2%	-2.7
自由度 20	sd1	96%	2%	2.3	2%	-2.3
の t 分布	mad1	94%	3%	2.7	3%	-2.7
	mad2	95%	2%	2.8	2%	-2.7

遺伝子内の outlier サンプルの割合=0.3		π_0	π_1	μ_1	π_2	μ_2
	真の値	88%	6%	2	6%	-2
正規分布	sd1	88%	6%	2.1	6%	-2.1
	mad1	89%	5%	2.6	5%	-2.6
	mad2	93%	3%	2.5	3%	-2.6
自由度 20	sd1	90%	5%	2.1	5%	-2.1
の t 分布	mad1	89%	6%	2.6	6%	-2.6
	mad2	93%	4%	2.6	4%	-2.6

遺伝子内の outlier サンプルの割合=0.5		π_0	π_1	μ_1	π_2	μ_2
	真の値	80%	10%	2	10%	-2
正規分布	sd1	80%	10%	2.1	10%	-2.1
	mad1	83%	8%	2.6	8%	-2.6
	mad2	92%	4%	2.4	4%	-2.3
自由度 20	sd1	82%	9%	2.1	9%	-2.1
の t 分布	mad1	84%	8%	2.6	8%	-2.6
	mad2	92%	4%	2.4	4%	-2.4