Vol.2016-BIO-45 No.7
2016/3/18

乳癌データベースを用いた遺伝子発現プロファイルの数値変換の検討

草田　義昭[1,2]　瀬尾　茂人[1]　竹中　要一[1]　野口　眞三郎[2]　松田　秀雄[1,a)]

概要：遺伝子発現プロファイルの臨床応用は、近年精力的に研究が行われている。しかしマイクロアレイを用いたデータ解析においては、”バッチ効果”を取り除くことが不可欠であり、さらに逐次サンプルが追加される臨床現場では、1サンプル毎に正規化が完結することが求められている。我々は、ノンパラメトリック Z 標準化（NPZ）法を提案し、既存の手法と比較検討を行った。まず、公共のデータベースからエストロゲン受容体（ER）とヒト上皮増殖因子受容体 2（HER2）の免疫組織化学 (IHC) 染色の結果を有する 2,813 症例（24 データセット）のマイクロアレイの発現データを抽出した。続いて、CEL ファイルからバックグランド補正及び、log2 変換のみを行ったもの（Raw）, 既存の 4 つの正規化法 [Microarray Suite 5.0 (MAS5)、 frozen robust multiarray analysis (fRMA)、 radius minimax (RMX)]. に対して、下記の 6 つの数値変換 [無変換、シングルアレイ数値変換 (RANK,、Z、 NPZ,、 YuGene) 、マルチアレイ数値変換 (ComBat)] を加えて、各々の ER と HER2 の IHC 染色の結果と mRNA の発現の一致率を比較した。シングルアレイ数値変換を行うことで IHC 染色と mRNA の発現の一致率は改善した。一方で、マルチアレイ数値変換は、主成分分析ではバッチ効果を他の手法に比して除去しているように図示されたが、実際には IHC 染色との一致率が低下していた。さらに、乳癌の予後と数値変換の検討の結果、MAS5 後に NPZ を加えることで、無変換、マルチアレイ数値変換と比べて 2 群の差が明瞭となった。今回、我々は乳癌のデータセットを用いて数値変換の与える影響について検討を行った。シングルアレイ数値変換を追加することで、臨床における発現データのバッチ効果の除去に有効である可能性が示唆された。

キーワード：遺伝子発現プロファイル, 数値変換, 正規化, 乳癌

1. はじめに

近年、マイクロアレイを用いることで転写産物のハイスループットな情報を比較的簡便に、より安価に入手可能となってきた。さらに、Gene Expression Omnibus (GEO: http://www.ncbi.nlm.nih.gov/geo/) などの公共のデータベースに非常に多くの実験データが日々蓄積されている。これらのデータの中には、新たな生物学的な意味づけを行う手がかりが含まれている可能性がある。しかし、マイクロアレイには、所謂”バッチ効果”と呼ばれるデータのばらつきが存在する [1]。 これらのバッチ効果を改善するために、これまでに様々な正規化法が報告されている [2]。Robust multiarray average (RMA) 法 [3] は、バッチ効果に対してロバストな正規化法の一つであるが、他の複数のマイクロアレイの発現値を用いて正規化を行う方法（マルチアレイ正規化）である。したがってサンプルの増減に伴い個々の発現値が変化するという問題点がある。一方、Microarray Suite 5.0 (MAS5) [4] や frozen robust multiarray analysis (fRMA) [5] や radius minimax (RMX) [6] に代表される正規化法は、各々のアレイごとに正規化が完結する（シングルアレイ正規化)。しかし、一般的にマルチアレイ正規化の方が、シングルアレイ正規化よりもバッチ効果に対しては、頑強であると報告されている　[7]。そのため、様々な付加的な数値変換法（rank transformation 法 [8] や Z scale 法 [9]）がしばしば正規化後のデータに対して用いられている。近年、経験ベイズ法に基づいた ComBat [10] 法による数値変換法が、主成分分析 (PCA) を用いた解析でバッチ効果を効果的に取り除くことができると報告された。しかし、この数値変換法もまた、複数のアレイを用いて行っているために（マルチアレイ数値変換)、1 例単位での不変的な値の算出が困難である。一方で Kim-Anh 等は、累積割合値を応用した YuGene 法 [11] を提案し、シングルアレイ数値変換法においても生物学的な情報を失うことなくバッチ効果を取り除けると報告した。しかし、YuGene 法

[1] 大阪大学大学院情報科学研究科
[2] 大阪大学大学院医学系研究科
[a)] matsuda@ist.osaka-u.ac.jp

は、全体の和から逐次、発現値の高い順に値を減算し、累積割合値を算出するために大きな外れ値を伴うデータの場合には、少なからず影響を受ける可能性がある。この点において rank transformation は、より頑強であるが、発現データが一様分布へと変換され、実際のデータ分布と大きく異なる点が問題である。

乳癌において、OncotypeDx (real-time PCR 法を活用; Genomic Health, Redwood City, CA, USA) [12]) や mammaprint (マイクロアレイを活用; Agendia, Amsterdam, The Netherlands) [13] などの mRNA をターゲットとした多重遺伝子診断は、現在急速に臨床応用化がなされており、その有用性に期待が集まっている。この背景には、治療方針決定のために重要なエストロゲン受容体 (ER) とヒト EGFR 関連物質受容体 2（HER2）の発現が、従来の免疫組織化学染色法（IHC 法）と mRNA をターゲットとしたリアルタイム-PCR 法やマイクロアレイ法と非常によく相関することが関与しているものと思われる [14]。
臨床応用に向けたマイクロアレイを用いた研究は積極的に行われているが、克服すべき課題も山積しているのが現状である。特に、個々の症例において独立して、随時、値を算出することは不可欠であり、施設、日時や電圧などの様々な避けられないバッチ効果が想定される。今回、我々はノンパラメトリック-Z-スケーリング (NPZ) 法をを提案した。この方法は、中央値を用いてセンタリングしてるため分布の形状に左右されず、四分位範囲を用いているため外れ値の影響を受けにくい特徴がある。今回の検討では、臨床応用と同様な状況を想定し、乳癌のマイクロアレイデータをシングルアレイ毎に正規化した後に、各々の数値変換によるバッチ効果の除去やそれに伴う臨床アウトカムを評価し比較検討を行った。

2. 方法

2.1 正規化

前処理は、CEL ファイルからバックグランド補正及び、log2 変換のみ (Raw) を行った。今回の検討では、下記のシングルアレイ正規化である MAS5、 RMX や fRMA.RWA (fRMA-robust weighted average method) を用いて行った (**表 1**)。

2.2 数値変換

正規化終了後に、それぞれのデータに下記の数値変換法を付加し、比較検討を行った (表 1)。

(1) Untransformed
(2) RANK: rank transformation（R コアパッケージ）
(3) Z: z-score transformation （R コアパッケージ）
(4) NPZ: nonparametric z-score $NPZ = (X_i - X_m)/NIQR$ where X_i is the value of the sample, X_m is the median of all probes, and $NIQR$ is the normalized interquartile range;
(5) YuGene: semi-rank-based transformation [15]
(6) ComBat: multi-array-based approach based on an empirical Bayes method [16]

2.3 データセット

患者は、公共のデータベース GEO から下記の適格基準を満たす症例を選択した。適格基準：(1) 乳癌である; (2) Affymetrix HG-U133 (GPL96) または、Affymetrix HG-U133 plus 2.0 (GPL570) を用いて得られた発現データである; (3)ER と HER2 の情報が利用可能であり, 下記に示す同様な基準で判定されている (ER: 10 fmol/mg cytosol protein または 10 percent tumor; HER2: fluorescence in situ hybridization (FISH) もしくは IHC テストでスコアが 3+); (4) 本検討の集積データ内で新たに登録される症例である（全てのプローブで同じ発現パターンを示す症例は除外)。上記の基準で 24 データセットから合計 2,813 症例が選択された。(**表 2**).

2.4 mRNA の閾値の決定

乳癌において、ER および HER2 のステータスは、治療のターゲットとなるため日常臨床で測定されている。IHC 染色とマイクロアレイの一致率を評価するために、下記の既に ER と HER の蛋白質とそれぞれ発現が一致することが知られている [14] プローブ [ER; ESR1 (estrogen receptor 1, Affymetrix probe ID 205225_at), HER2; ERBB2 Erb-B2 receptor tyrosine kinase 2 (216836_s_at)] を選択した。続いて、IHC 染色と mRNA の発現の閾値を決定するため、1000 回のブートストラッピング抽出に対して、随時 ROC 曲線を用いて算出し、それらの中央値を最終的な閾値と決定した。

2.5 Leave-one-out 交差検証及び評価

IHC 染色と mRNA の発現の一致率を評価するため、Leave-on-out 交差検証（LOOCV) 法を採用した。つまり、我々は検証用として用いるデータセットを除いた残り全てのデータセットで閾値を決定し、その閾値を用いて検証用データに適応し一致率を評価した。それをすべてのデータセットが 1 回ずつ評価されるまで繰り返した。最終的に、それぞれのコホートの一致率の単純平均を結果として採用した。

2.6 統計解析

主成分分析 (PCA) は、R コアパッケージを用いて、全てのプローブを対象として、第 1 主成分と第 2 主成分に対して行った。生存分析は、内分泌療法症例における無再発生存期間 (RFS) は、Kaplan-Meier 曲線を用いて算出し、検定はログランクテストにおける P 値を採用した。

表 1　正規化と数値変換

	Method	Detail
Normalization	Raw	Single array-based
	MAS5	Single array-based
	RMX	Single array-based
	fRMA.RWA	Single array-based
Transformation	Untransformed	
	Rank	Single array, does not assume a distribution
	YuGene	Single array, does not assume a distribution
	Non-parametric-Z-score	Single array, does not assume a distribution
	Z-score	Single array, does not assume a distribution
	ComBat	Multiple array, Gaussian distribution

表 2　乳癌データセットの臨床病理学的特徴

GEO		Total	ER status by IHC			HER2 status by IHC /FISH		
Datasets	Platform	n	Positive	Negative	Unknown	Positive	Negative	Unknown
GSE42822	GPL96	90(91) †	0	0	90	34	54	2
GSE33658	GPL570	11	11	0	0	0	0	11
GSE32646	GPL570	115	71	44	0	34	81	0
GSE32518	GPL96	50 (74) †	25	25	0	11	39	0
GSE29431	GPL570	54	0	0	54	25	25	4
GSE26971	GPL96	276 (277) †	0	0	276	0	276	0
GSE25066	GPL96	433 (508) †	261	172	0	4	418	11
GSE23593	GPL570	50	36	14	0	0	50	0
GSE23177	GPL570	116	116	0	0	0	116	0
GSE20271	GPL96	72 (178) †	32	39	1	22	47	3
GSE20194	GPL96	67 (278) †	33	34	0	36	31	0
GSE20181	GPL96	59	59	0	0	0	0	59
GSE18864	GPL570	84	0	0	84	18	64	2
GSE18728	GPL570	21	0	0	21	5	16	0
GSE17705	GPL96	298	298	0	0	0	0	298
GSE16446	GPL570	120	0	120	0	31	62	27
GSE16391	GPL570	55	55	0	0	3	42	10
GSE12093	GPL96	136	136	0	0	0	0	136
GSE10810	GPL570	31	19	12	0	0	0	31
GSE9195	GPL570	77	77	0	0	0	0	77
GSE6532	GPL96	206 (327) †	161	45	0	0	0	206
GSE6532	GPL570	87	87	0	0	0	0	87
GSE5460	GPL570	23 (127) †	12	11	0	23	0	0
GSE2034	GPL96	286	209	77	0	0	0	286

GEO : Gene Expression Omnibus　n : number of patients　ER : estrogen receptor
IHC : Immunohistochemistry staining　HER2 : human epidermal growth factor receptor 2
GPL96 : Affymetrix Human Genome U133A Array　GPL570 : Affymetrix Human Genome U133 Plus 2.0 Array
† : total number of patients before exclusion

3. 結果

3.1 同一プラットフォーム内での比較検討

この検討では、GPL96 のプラットフォームから 1973 症例（11 施設）と GPL570 のプラットフォームより 844 症例（13 施設）の合計 2,817 例が選択された。まず、それぞれのプラットフォーム内での数値変換の影響を調べるために、ER と HER2 のステータスを mRNA の発現と IHC 染色との両方で LOOCV 法を用いて検討を行った。GPL96 内では、正規化に加えてシングルアレイ数値変換法を行うことで一致率の改善を認めた（ESR1; シングルアレイ数値変換 90.0% vs. 無変換 88.3%, HER2; シングルアレイ数値変換 90.5% vs. 無変換 85.0%)。一方で、マルチアレイ数値変換法では、一致率が低下する結果 (ESR1; 82.1%, HER2; 82.3%) となった (**表 3**)。同様な結果が、GPL570 においても認められた（シングルアレイ数値変

表 3　leave-one-out 交差検証における m RNA と IHC 染色と一致率

		GPL96		GPL570	
		ESR1 / ER	ERBB2 / HER2	ESR1 / ER	ERBB2 / HER2
Normalization	Transformation	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)
Raw	Untransformed	86.5	80.7	94.8	76.3
MAS5	Untransformed	89.6	90.6	94.9	79.6
RMX	Untransformed	86.9	82.3	94.8	76.3
fRMA.RWA	Untransformed	90.2	86.3	94.8	76.3
Raw	RANK	89.6	90.7	94.8	76.8
MAS5	RANK	90.1	89.9	95.0	82.2
RMX	RANK	90.0	89.3	94.8	76.8
fRMA.RWA	RANK	90.3	90.8	94.8	76.8
Raw	YuGene	88.6	91.2	94.8	77.0
MAS5	YuGene	90.1	89.3	95.2	82.0
RMX	YuGene	89.6	89.9	94.8	77.0
fRMA.RWA	YuGene	90.5	90.8	94.8	77.0
Raw	NPZ	89.7	88.8	95.0	81.6
MAS5	NPZ	90.0	92.2	94.8	82.3
RMX	NPZ	90.5	89.7	95.0	81.6
fRMA.RWA	NPZ	90.1	91.0	95.0	81.6
Raw	Z	90.0	90.5	94.8	77.6
MAS5	Z	89.9	91.4	94.9	81.0
RMX	Z	90.4	90.4	94.8	77.6
fRMA.RWA	Z	90.0	91.3	94.8	77.6
Raw	ComBat	81.9	80.3	68.2	78.1
MAS5	ComBat	82.4	84.2	67.2	76.4
RMX	ComBat	80.5	80.7	68.2	78.1
fRMA.RWA	ComBat	83.9	84.1	68.2	78.1

GPL96 : Affymetrix Human Genome U133A Array　GPL570 : Affymetrix Human Genome U133 Plus 2.0 Array
ESR1 : estrogen receptor 1 (205225_at)　ER : estrogen receptor　HER2 : human epidermal growth factor receptor 2
ERBB2 : Erb-B2 receptor tyrosine kinase 2 (216836_s_at)

換： ESR1; 94.9%, HER2; 79.2%, マルチアレイ数値変換: ESR1; 94.8%, HER2; 77.1%)。

3.2　臨床アウトカムに対する影響

我々は、数値変換が生物学的特徴を失うことなくバッチ効果を取り除くことができるかどうかについて検討を行った。この問題を解決するため、臨床アウトカムに着目して数値変換の効果を評価した。一般的に、ER 陽性乳癌は、内分泌療法のターゲットであり、その治療により ER 陰性乳癌とは異なり予後が改善することが知られている [17]。そのため、我々は術後補助療法として内分泌療法のみを行った症例を対象として、無再発生存期間（RFS) を数値変換ごとに比較した。我々は、この検討を行うことでバッチ効果が生物学的差異を縮小させて、それを取り除くことで臨床的な差異がより明確になるのではないかと想定した。**図 1** は MAS5 後の NPZ を行うことで (P= 2.79E-5, log-rank test) ESR1 陽性及び陰性乳癌の RFS を無変換 (P= 1.08E-4, log-rank test) やマルチアレイ数値変換 (P= 0.322, log-rank test) に 2 群間の差が大きくなる事が示された。同様な結果が、fRMA や RMX の場合にも観測された。

4.　考察

マイクロアレイを用いた研究においてバッチ効果を取り除くことは不可欠であり、これまでに様々な正規化や数値変換が提案されている。しかしながら、本検討のような数千を超えるサンプルを対象として、数値変換の効果を実際の蛋白質発現と比較検討した研究はあまり報告されていない。近年、マルチアレイ数値変換が PCA グラフにおいてバッチ効果を改善することが報告されているが [18]、PCA グラフにおける改善が、必ずしも生物学的特性を失わずに効果的にバッチ効果を改善しているかどうかは不明瞭である。今回の検討では、当初我々は、マルチアレイ数値変換が最も頑強な方法であり、シングルアレイ数値変換の非劣勢を証明する精度を目標として検討を開始した。驚くべきことに、マルチアレイ数値変換は、IHC 染色との一致率においては最も低い結果となった。我々も同様に PCA を行った結果（ **図 2**)、PCA グラフは、マルチアレイ数値変換によって一見個々のバッチ効果が取り除かれているように推察される。しかし、**表 ??**の結果から、数値変換によっ

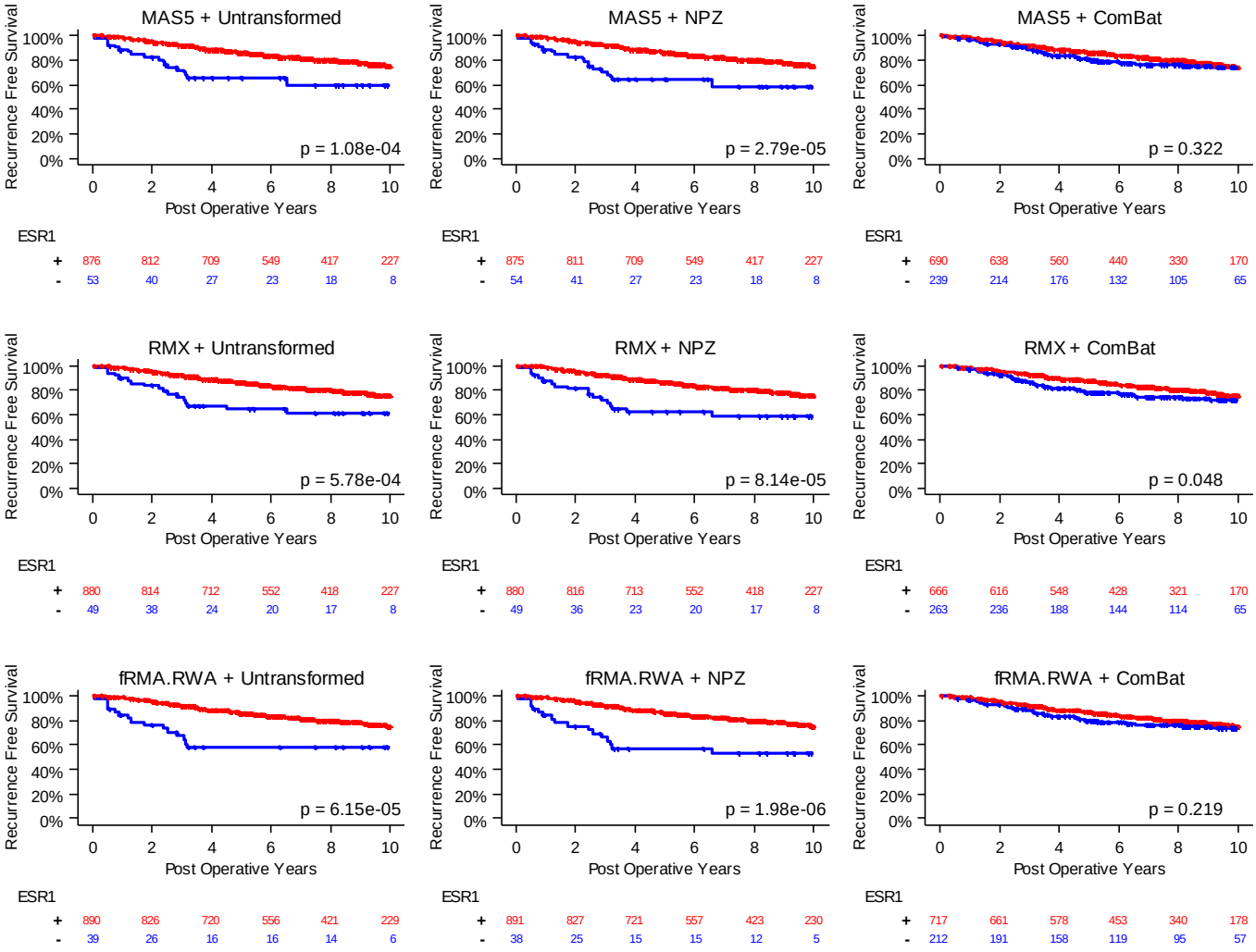

図 1　内分泌療法患者における再発曲線と ESR1　左: Untrasformed, 中: NPZ, 右: ComBat

て生物学的特徴も失われている可能性が示唆された。さらに、1 症例ずつ結果を得る必要がある臨床サンプルにおいては、マルチアレイに基づいた手法は、不利であるだけでなく不便である。我々の結果は、正規化に NPZ のような数値変換を加えることで、同一プラットフォーム内において効果的にバッチ効果を取り除けることを示している。さらに興味深いことに、乳癌の IHC 染色は国際的なガイドラインによって厳格に既定されているが、今回の検討でバッチ効果を取り除かれたマイクロアレイによる ESR1 の発現が、IHC 染色に比べて内分泌療法の適応をより明確に判別できる可能性が示唆された。

5. 結論

今回、我々は乳癌のデータセットを用いて m RNA と対応する蛋白質の発現の一致率を評価することで、数値変換のバッチ効果の除去に与える影響を検討した。臨床では、個別に値を算出できるシングルアレイ数値変換にアドバンテージがあり、更にバッチ効果の除去に伴い生物学的特徴がより強調されうる可能性が示唆された。今後、マイクロアレイの臨床応用のため更なる検討及び研究が期待される。

参考文献

[1] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A.: Tackling the widespread and critical impact of batch effects in high-throughput data, *Nat Rev Genet*, Vol. 11, No. 10, pp. 733–9 (2010).

[2] Allison, D. B., Cui, X., Page, G. P. and Sabripour, M.: Microarray data analysis: from disarray to consolidation and consensus, *Nat Rev Genet*, Vol. 7, No. 1, pp. 55–65.

[3] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, Vol. 4, No. 2, pp. 249–64 (2003).

[4] Hubbell, E., Liu, W. M. and Mei, R.: Robust estimators for expression analysis, *Bioinformatics*, Vol. 18, No. 12, pp. 1585–92 (2002).

[5] McCall, M. N., Bolstad, B. M. and Irizarry, R. A.: Frozen robust multiarray analysis (fRMA), *Biostatistics*, Vol. 11, No. 2, pp. 242–53 (2010).

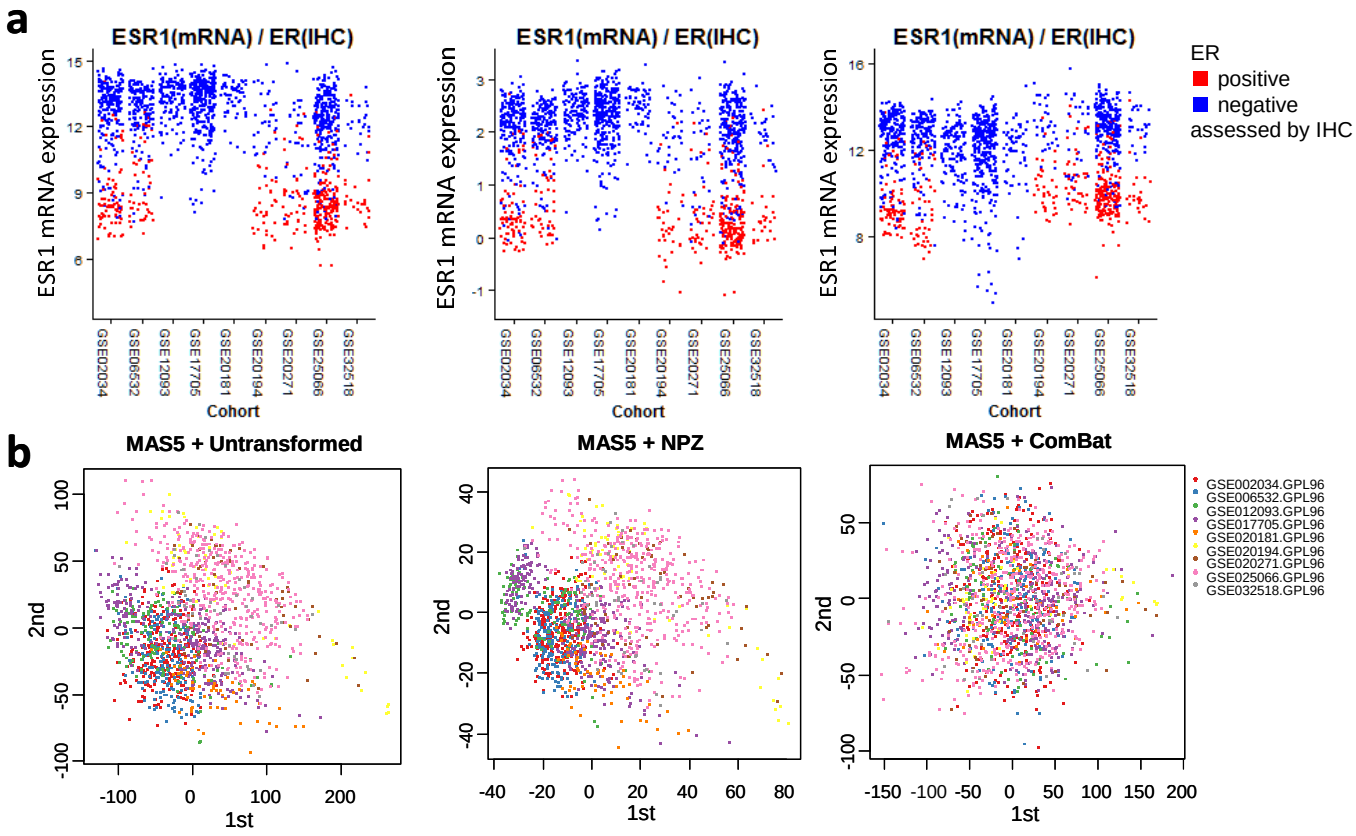

図 2　ESR1 mRNA 発現の散布図と主成分分析

[6] Kohl, M. and Deigner, H. P.: Preprocessing of gene expression data by optimally robust estimators, *BMC Bioinformatics*, Vol. 11, p. 583 (2010).
[7] Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, Vol. 19, No. 2, pp. 185–93 (2003).
[8] Koh, C. H. and Wong, L.: Embracing noise to improve cross-batch prediction accuracy, *BMC Syst Biol*, Vol. 6 Suppl 2, p. S3 (2012).
[9] Cheadle, C., Vawter, M. P., Freed, W. J. and Becker, K. G.: Analysis of microarray data using Z score transformation, *J Mol Diagn*, Vol. 5, No. 2, pp. 73–81 (2003).
[10] Johnson, W. E., Li, C. and Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics*, Vol. 8, No. 1, pp. 118–27 (2007).
[11] Le Cao, K. A., Rohart, F., McHugh, L., Korn, O. and Wells, C. A.: YuGene: a simple approach to scale gene expression data derived from different platforms for integrated analyses, *Genomics*, Vol. 103, No. 4, pp. 239–51 (2014).
[12] Partin, J. F. and Mamounas, E. P.: Impact of the 21-gene recurrence score assay compared with standard clinicopathologic guidelines in adjuvant therapy selection for node-negative, estrogen receptor-positive breast cancer, *Ann Surg Oncol*, Vol. 18, No. 12, pp. 3399–406 (2011).
[13] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H.: Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, Vol. 415, No. 6871, pp. 530–6 (2002).
[14] Roepman, P., Horlings, H. M., Krijgsman, O., Kok, M., Bueno-de Mesquita, J. M., Bender, R., Linn, S. C., Glas, A. M. and van de Vijver, M. J.: Microarray-based determination of estrogen receptor, progesterone receptor, and HER2 receptor status in breast cancer, *Clin Cancer Res*, Vol. 15, No. 22, pp. 7003–11 (2009).
[15] Cao, K.-A. L., Rohart, F., McHugh, L., Korn, O. and Wells, C. A.: YuGene: A Simple Approach to Scale Gene Expression Data Derived from Different Platforms for Integrated Analyses, (online), available from ⟨http://CRAN.R-project.org/package=YuGene⟩ (2015).
[16] Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. and Storey, J. D.: sva: Surrogate Variable Analysis.
[17] Howell, A., Cuzick, J., Baum, M., Buzdar, A., Dowsett, M., Forbes, J. F., Hoctin-Boes, G., Houghton, J., Locker, G. Y., Tobias, J. S. and Group, A. T.: Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer, *Lancet*, Vol. 365, No. 9453, pp. 60–2 (2005).
[18] Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C. M. and Marron, J. S.: Adjustment of systematic microarray data biases, *Bioinformatics*, Vol. 20, No. 1, pp. 105–14 (2004).