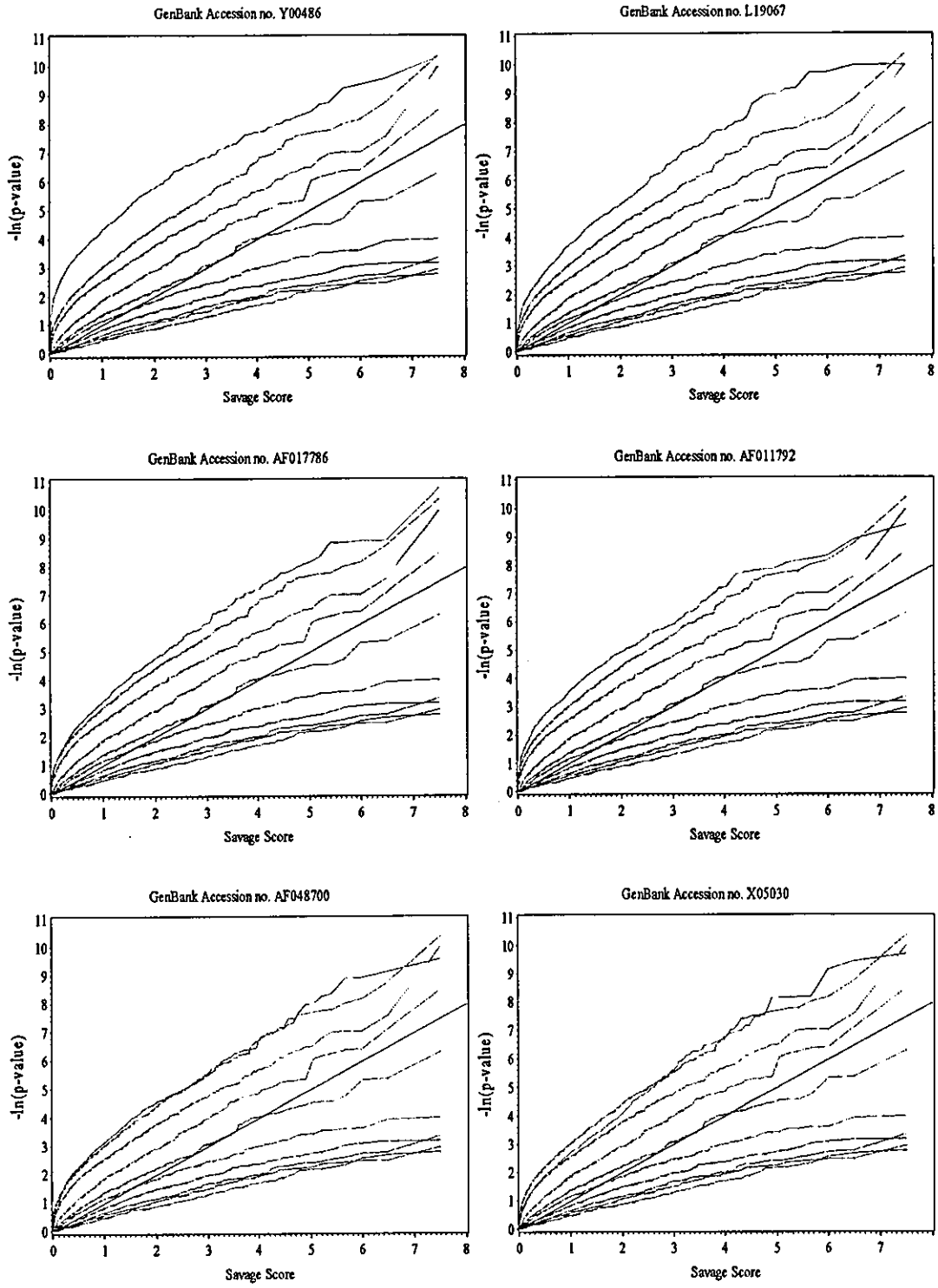
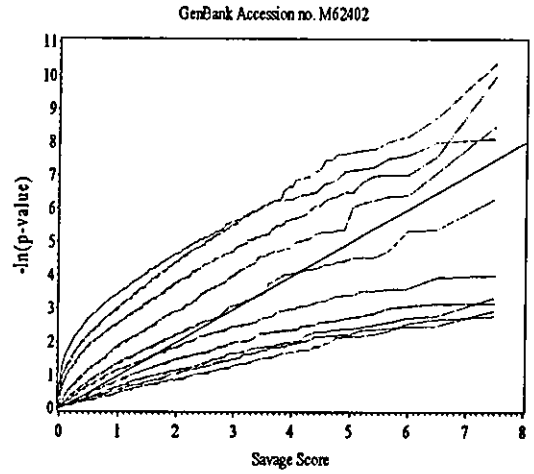
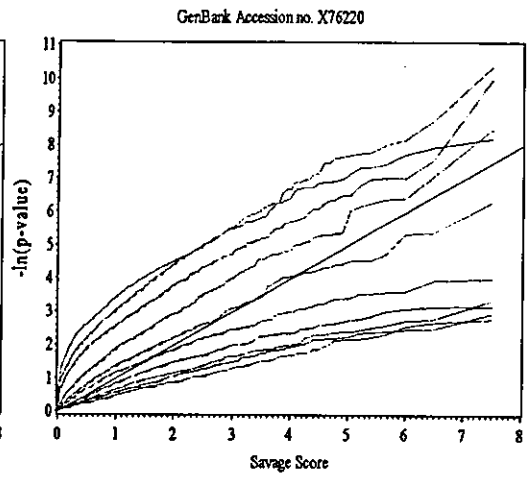
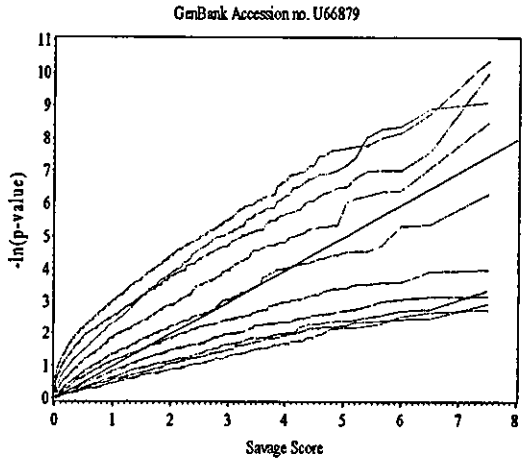
資料18　group 1の遺伝子のQ-Q plot

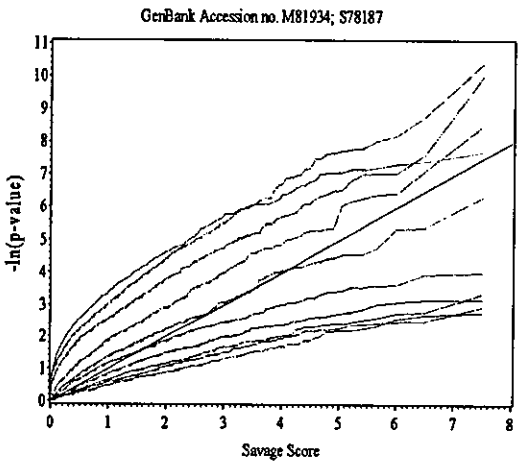


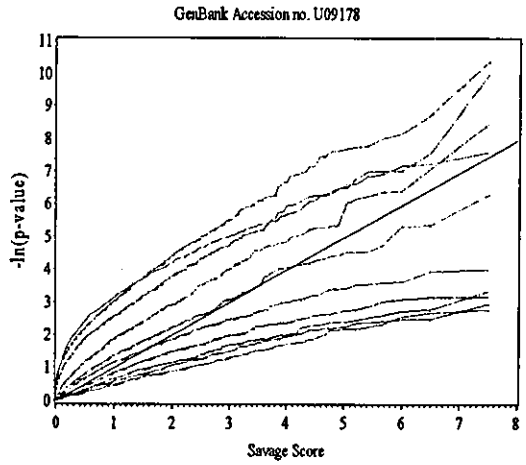
資料19　group 2の遺伝子のQ-Q plot











資料20　2乗差の中央値の大きい遺伝子

Gene Name	Coordinate	Median of Squared Difference
RHO GDP-dissociation inihibitor 1 (RHO-GDI 1); RHO-GDI alpha (GDIA1); ARHGDIA	C3e	7.924
cell division protein kinase 6 (CDK6); serine/threonine protein kinase PLSTIRE	A3k	4.831
Human paxillin mRNA, complete cds	F13d	4.590
neurotrophin 3 precursor (NT3); nerve growth factor 2 (NGF2)	A14l	4.444
rhoHP1	C3l	3.767
tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon polypeptide (YWHAE);14-3-3 protein epsilon; protein kinase C inhibitor protein 1 (KCIP1); mitochondrial import stimulation factor L subunit	C7g	3.477
c-myc proto-oncogene	A9m	2.994
Casein kinase I alpha isoform (CKI-alpha); CK1; CSNK1A	C2f	2.949
cAMP-dependent protein kinase I alpha regulatory subunit (PRKAR1); tissue-specific extinguisher 1 (TSE1)	B9h	2.943
H1 histone family member 0 (H1F0; H10); H1FV	E11b	2.573
TSG101 tumor susceptibility protein	B11g	2.176
Metastasis-associated protein 1 (MTA1)	B10n	2.037
Rho-related GTP-binding protein RhoE; Rho8; ARHE	C3f	1.982
cell cycle progression 2 protein (CPR2)	A4h	1.812
ERBB2 proto-oncogene; NEU proto-oncogene; HER2	B2h	1.770
interleukin 6 precursor (IL6); B-cell stimulatory factor 2 (BSF2); interferon beta 2 (IFNB2); hybridoma growth factor	D11e	1.690
Growth factor receptor-bound protein 2 (GRB2); abundant SRC homology protein (ASH)	D7b	1.687
cyclin-dependent kinase 4 inhibitor 2D (CDKN2D); p19-INK4D	A8j	1.535
Homo sapiens pyruvate dehydrogenase kinase, isoenzyme 2 (PDK2), mRNA	F12g	1.417
Retinoic acid receptor alpha 1 (RAR-alpha 1; RARA); PML-RAR protein	B8j	1.403

資料21　２乗差の分散の大きい遺伝子

Gene name	Coordinate	Variance of Squared Difference
RHO GDP-dissociation inihibitor 1 (RHO-GDI 1); RHO-GDI alpha (GDIA1); ARHGDIA	C3e	18.250
tumor necrosis factor type 1 receptor-associated protein (TRAP1)	D12b	14.453
Tyrosine-protein kinase receptor UFO precursor; axl oncogene	B11f	13.687
Homo sapiens v-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma) (AKT3), mRNA	F12e	9.059
Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon polypeptide (YWHAE);14-3-3 protein epsilon; protein kinase C inhibitor protein 1 (KCIP1); mitochondrial import stimulation factor L subunit	C7g	8.892
neurotrophin 3 precursor (NT3); nerve growth factor 2 (NGF2)	A14l	8.585
TIS11B protein; butyrate response factor 1 (BRF1); EGF response factor 1 (ERF1)	D9c	7.818
type II cytoskeletal 11 keratin (KRT11); cytokeratin 1 (CK1); 67-kDa cytokeratin; hair alpha protein	F8a	7.539
interleukin 1 alpha precursor (IL1-alpha; IL1A); hematopoietin 1	D10d	6.726
Casein kinase I alpha isoform (CKI-alpha); CK1; CSNK1A	C2f	6.611
transforming growth factor beta 3 (TGF-beta3; TGFB3)	D8d	6.250
serine/threonine protein kinase SAK	B11n	6.082
68-kDa tumor protein (TP68); p51B	A6n	5.980
Insulin-like growth factor-binding protein 2 (IGF-binding protein 2; IGFBP2; IBP2)	D7j	5.908
Homo sapiens cytochrome P450, subfamily IIIA (niphedipine oxidase), polypeptide 5 (CYP3A5), mRNA	F13f	5.640
Human paxillin mRNA, complete cds	F13d	5.460
ras-related C3 botulinum toxin substrate 2; p21-rac2; small G protein	B5m	5.459
farnesyl pyrophosphate synthetase (FPS); farnesyl diphosphate synthase (FDPS); dimethylallyltranstransferase; geranyltranstransferase; KIAA0032	C6i	5.314
Type II cytoskeletal 2 epidermal keratin (KRT2E); cytokeratin 2E (CK2E)	F8b	5.199
integrin beta 7 precursor (ITGB7)	C9g	5.152

資料２２　フィルターアレイの信頼性および抗癌剤感受性遺伝子

フィルターアレイを使用して、代表的な乳癌細胞株に対してエストロジェンおよびタモキシキフェン接触を行ったときの遺伝子発現変動を測定した。今回使用したアトラスフィルターアレイは825種類の遺伝子に関する発現情報を一度に測定することが可能である。本実験ではBT474、MCF7、MDA-MB-231、SK-BR-3という4種類の代表的な乳癌細胞株に対して、それぞれ4種類のエストロゲン接触処理を行った。4種類の処理はそれぞれ、エストロゲンのみ接触、エストロゲンとタモキシフェン低用量に接触、エストロゲンとタモキシフェン高用量に接触、およびエストロゲンフリー（エストロゲン枯渇状態）である。従って、細胞株と処理の組み合わせは計16通りである。本実験ではアレイデータの信頼性を評価するために、この16通りの処理を2回繰り返して実施した。通常、アレイデータ解析における繰り返しは推定値の精度を高めるために行われるため、同一実験日に同一条件の実験を繰り返すが、今回の目的は信頼性の評価であることから、2回の実験はあえて実験日を変えて行った。アレイデータの信頼性はこれら2回の繰り返し実験における結果間のピアソン相関係数を計算することによって評価する。フィルターアレイはRIによる単独標識を用いるため、本実験で使用したアレイの数は32枚である。実験の順序に伴う系統的変動を除去する目的で、本実験では各16種類の組み合わせに対するハイブリダイゼーション処理はランダム化した。従って、統計学的には、本実験は繰り返し数2の4細胞株×4処理の要因実験とみなすことができる。表１は細胞株と処理の組み合わせを示したものである。なお、実験は単独の実験者によって行われ、RI使用に関する制約から実験は8回に分けて行われた。

図１は細胞株と処理の16種類の組み合わせにおける遺伝子発現強度の散布図である。横軸は1回目の実験における遺伝子発現強度、縦軸は2回目の実験における発現強度である。この図より、実験間で発現強度の位置と尺度が大きく変化していること、および発現強度の分布が右上に大きく裾を引くことがわかる。このようなデータに対してはピアソンの相関係数を用いても信頼性を妥当に評価することはできず、また、後述する薬剤感受性遺伝子の抽出においても、統計解析上困難が生じる。従って、統計解析のためには位置・尺度の調整だけではなく、適当な変数変換が必要である。

図2は遺伝子発現強度の遺伝子毎の平均と標準偏差について、散布図をとったものである。大きな点はデータ系列に対する移動平均である。図より、平均強度の増加に伴って標準偏差が著しく増加することがわかる。

各組み合わせの発現強度が比較可能になるように尺度を調整し、かつ平均強度とその標準偏差が無関係になるようにデータを変換することにより、エストロゲン接触関連遺伝子を統計学的に抽出することが比較的容易になる。通常のマイクロアレイ解析では、データに対して対数変換を施し、平均値あるいは中央値を用いて正規化を行なうのが一般的である。しかし、本データにおいて上記の正規化を試みたところ、以前として平均強度とその標準偏差に強い相関が認められるため、従来の方法とは異なる正規化の手法が必要になる。

本解析では、Huber(2002)による正規化および分散安定化変換の手法をフィルターアレイのデータに対して適用した。アレイi(=1、…、m)における遺伝子j(=1、…n)の発現強度をx_{ij}とする。アレイiにおける位置および尺度の調整用パラメータをそれぞれo_iとs_iとする。このときデータの変換を、

で定義する。一般性を失わずo_1=0とs_1=1とすることができるので、推定すべきパラメータ数は2m-2個である。次に分散を安定化させるために、発現強度の平均と標準偏差にモデルを仮定す　$x_{ij} \alpha z_{ij} = o_i + s_i x_{ij}$　る。図2より、平均と標準偏差には線形関係が仮定できるので、この場合の分散安定化変換は、

$$h(y_{ij}) = \operatorname{arcsinh}(a_i + b_i y_{ij})$$

として得ることができる。ここでa_i=a+bo_i、b_i=bs_iであり、aとbは分散安定化に用いるパラメータである。a_i、b_iの推定はプロファイル尤度法で行なった。ここで、当フィルターアレイデータの解析にはオープンソースの統計解析ソフトウェアであるR、およびその遺伝子データ解析パッケージであるBioconductorを使用した。

2．薬剤感受性遺伝子の抽出

次に、エストロゲン接触によって発現強度に違いが生じる遺伝子を統計的に抽出した。解析はScholtens and Gentleman のアプローチに基づいて行なった。当実験における発現強度データをy_{ijkl}とする。ここでiは遺伝子（i=1、…、825）、jは細胞（j=1、…、4）、kは処理（k=1、…、4）、lは繰り返し（l=1、2）とする。当実験の発現強度データに対して、下記の分散分析モデルを遺伝子毎に適用した。

$$y_{ijkl} = \mu_i + C_{ij} + T_{ik} + CT_{ijk} + e_{ijkl}$$

ここでμ_iは主効果、C_{ij}は細胞の効果、T_{ik}は処理の効果、CT_{ijk}は細胞と処理の交互作用である。次に、処理効果T_{ik}の検定を行なうために次の帰無モデルを用いた。

$$y_{ijkl} = \mu_i + C_{ij} + e_{ijkl}$$

処理効果のｐ値はF検定で導出した。また、細胞と処理の交互作用CT_{ijk}の検定を行なうために、次の帰無モデルを当てはめた。

$$y_{ijkl} = \mu_i + C_{ij} + T_{ik} + e_{ijkl}$$

発現強度データには依然として実験間で系統的な影響が生じていたことが探索的な検討から判明したため、実際の解析では上記の全てのモデルに実験の効果E_m(m=1、…、8)を加えて解析を行なった。従って、解析に用いたモデルはそれぞれ下記の通りである。

$y_{ijkl} = \mu_i + E_m + C_{ij} + T_{ik} + CT_{ijk} + e_{ijkl}$

$y_{ijkl} = \mu_i + E_m + C_{ij} + e_{ijkl}$

$y_{ijkl} = \mu_i + E_m + C_{ij} + T_{ik} + e_{ijkl}$

上記の解析では一度に825回のF検定を行なうため、検定の多重性の問題が当然生じる。多重性の問題を考慮するために、本解析ではBenjamini and Yekutieli (2001)の多重比較

法を用いた。Benjamini and Yekutieli法（以下BY法と略す）は任意の従属関係を持つ検定統計量から得られたp値に対して、試験全体の第一種の過誤ではなく、False Discovery Rate（以下FDR）を調整する多重比較法である。Vを誤って棄却した真の仮説の数、Rを棄却した仮説の数としたとき、確率変数Qを次のように定義する。

$$Q=\begin{cases} V/R & R>0\text{のとき} \\ 0 & \text{その他} \end{cases}$$

このとき、FDRはQの期待値E(Q)である。つまり、棄却された仮説に真の仮説が含まれる割合の期待値である。

任意の従属関係をもつn個の検定統計量から得られたn個のp値をp_1、 p_2、 …、 p_nとする。これらのp値を小さい順に並べ変えたものを$p_{(1)}$、 $p_{(2)}$、 …、 $p_{(n)}$ としたとき、

$$k=\max\left\{i:p(i)\le\frac{i}{n}\frac{1}{\sum_1^n(1/i)}q\right\}$$

を計算し、$p_{(1)}$、 $p_{(2)}$、 …、 $p_{(k)}$ に対応する仮説を棄却する。この検定方式によって、FDRはq以下に調整される（Benjamini and Yekutieli、2001）。なお本研究ではFDRの基準値として0.05を選択した。

次に探索的な目的（仮説発見）のために、タモキシフェン介入に伴って系統的に変動する遺伝子発現プロファイルを各細胞株毎に解析する。解析においては平均値に関する線形傾向性検定を用いた。線形対立仮説を表現する線形対比、

$$L=\sum_k^K c_k\overline{Y}_k$$

に対する検定を考える。ここでkは処理を示す因子の水準数であり、線形対比はそれぞれ$c_1=-3$、$c_2=-1$、$c_3=1$、$c_4=3$である。このとき検定統計量は、

$$T=\frac{\sum_{k=1}^4 c_k\overline{Y}_k}{\sqrt{10\hat{\sigma}^2}}$$

であり、帰無仮説のもとで自由度4のt分布に従う。この検定を細胞株毎、遺伝子毎に適用する。なお本解析では仮説発見の目的のため、多重性の調整は行っていない。

上記の解析結果を視覚的に明確化するために、ボルケーノプロットを描画した。ボルケーノプロットは横軸に検定統計量の分子、縦軸にp値の対数の負値をプロットしたものである。本解析においては、図の右側に位置する遺伝子は介入に伴って発現が増加する傾向を、左側は減少する傾向を示す。

図3は分散安定化変換後の各遺伝子の平均強度と標準偏差の関係を示す図である。図3より、分散安定化変換を行うことで、標準偏差は平均強度によらずほぼ一定になっていることがわかる。これより、本データにHuberの分散安定化手法を用いることで、通常の分散

分析など、当分散性を仮定する解析手法を適用することが妥当であることが分かった。

図4は分散安定化変換後の実験間の散布図である。図4より、データの重心は各散布図の中心にほぼ位置していることがわかり、変数変換前の右上への裾の広さが解消されていることがわかる。これにより、分布の非対称性に影響されやすい、正規分布を仮定した統計解析手法を本データに適用することが妥当であることが示された。

2．フィルターアレイの信頼性検討

本実験で使用したフィルターアレイの信頼性を検討するために、上記の分散安定化変換後のデータに対して、ふたつの実験間における発現強度のピアソン相関係数を各組み合わせ毎に計算した（表2）。なお、分布の非対称性はピアソン相関係数に大きく影響することから、分散安定化前のデータに対してピアソン相関係数を計算することは望ましくないことを注記する。解析の結果、いずれの細胞と処理の組み合わせにおいても、ピアソン相関係数は0.8以上の値を示した。今回の実験のように、マイクロアレイ間で別々のハイブリダイゼーションを行なった場合、対応する発現強度間の相関は60%から80%であるという報告があり（Churchill、2002）、それを考慮すると、この結果はフィルターアレイの信頼性が他のアレイに比べて高水準であることを示唆している。

3．薬剤感受性遺伝子の抽出

Huberの分散安定化変換によってアレイデータの非対称性と不等分散性がほぼ解消されたので、先に紹介した分散分析法と対比を用いた方法を本データに対して適用した。分散分析に関しては、BY法によって検定の多重性を調整した場合、主効果および交互作用の検定のいずれにおいてもk=0であり、残念ながら、処理および処理と細胞の交互作用に関して、統計的な発現差をもつ遺伝子は検出されなかった。参考までに未調整p値が0.01未満の遺伝子を表3および表4に列挙する。また、これら遺伝子において、特徴的な変動を示したp21-activated kinase alphaの発現プロファイルを図5に示す。

図5において、物理的あるいはタモキシフェンによるエストロゲン接触低下によって、MCF7とBT474でp21-activated kinase alphaの発現が増加していることがわかる。p21の発現は細胞周期の進行を遅延させ、またアポトーシスを誘発させることが知られている。一方で、MDA-MB-231とSK-BR-3ではこのような傾向性をみることができなかった。MCF7とBT474はエストロゲン陽性細胞株であり、一方でMDA-MB-231とSK-BR-3はエストロゲン陰性細胞株であることから、タモキシフェンの介入によってエストロゲン陽性細胞株のみのp21の発現が促進され、細胞周期の進行遅延およびアポトーシス誘導が生じる可能性が示唆された。

次に探索的解析の結果を示す。乳癌細胞株の各種類に対して、線形対比を用いた解析を適用し、ボルケーノプロットを用いて顕著な変動がみられた遺伝子群を列挙した。本研究では探索的に、p値の対数の負値が2以上の遺伝子を選択した。ここで、細胞株の各種類に

ついて個別に線形傾向性の検定を行ったため、各検定でのサンプルサイズは8である。残念ながら有意水準5%で統計的に有意な遺伝子は抽出されなかったが、これはサンプルサイズの小ささに起因するものである可能性が高い。もちろん、本解析で探索的に抽出した遺伝子には第一種の過誤の危険性があり、あくまでも仮説発見のための情報であることを注記する。

図６はMCF７細胞株におけるボルケーノプロットであり、表5、表6はその結果から抽出した遺伝子群である。表５はタモキシフェンの用量増加に伴って遺伝子発現に増加傾向がみられる遺伝子群であり、表６は減少傾向がみられる遺伝子群である。Multidrug resistance-associated proteinやcytokeratinに関連した遺伝子が比較的多く抽出されていることがわかる。これらの傾向性は図7と図8で確認できる。

次にBT474細胞株について検討を行った。図9はBT474細胞株のボルケーノプロットであり、表7、表8は抽出された遺伝子群である。Metalloproteinaseやプロテインキナーゼに関連する遺伝子群が多く抽出されていることがわかる。これらの傾向性は図10および図11で確認できる。

続いて、図12のプロットよりSK-BR-2細胞株に対して遺伝子を抽出した。表9、表10はそのリストである。Cadherinやras関連遺伝子に発現変動が生じているのがわかる。これらの傾向性は図13と図14に図示した。

最後に、MDA-MB-231細胞株におけるボルケーノプロットを図15に示した。表11、表12は抽出された遺伝子群である。Rasに関連した遺伝子群が多く抽出されていることがわかる。これらの遺伝子の傾向性は図16と図17を参照。

参考文献

Benjamini Y and Yekutieli D (2001), The control of the FDR multiple testing under dependency, Annals of Statistics, 29, 1165-1188.

Churchill GA (2002), Fundamentals of experimental design for cDNA microarrays, Nature Genetics, 32 Suppl.、 490-495.

Huber W, von Heydebreck A, Sueltmann H, et al. (2002), Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics, 18 Suppl., S96-S104.

Scholtens D and Gentleman R, Estrogen 2×2 factorial design, a Vignette of Bioconductor, 1- 12.

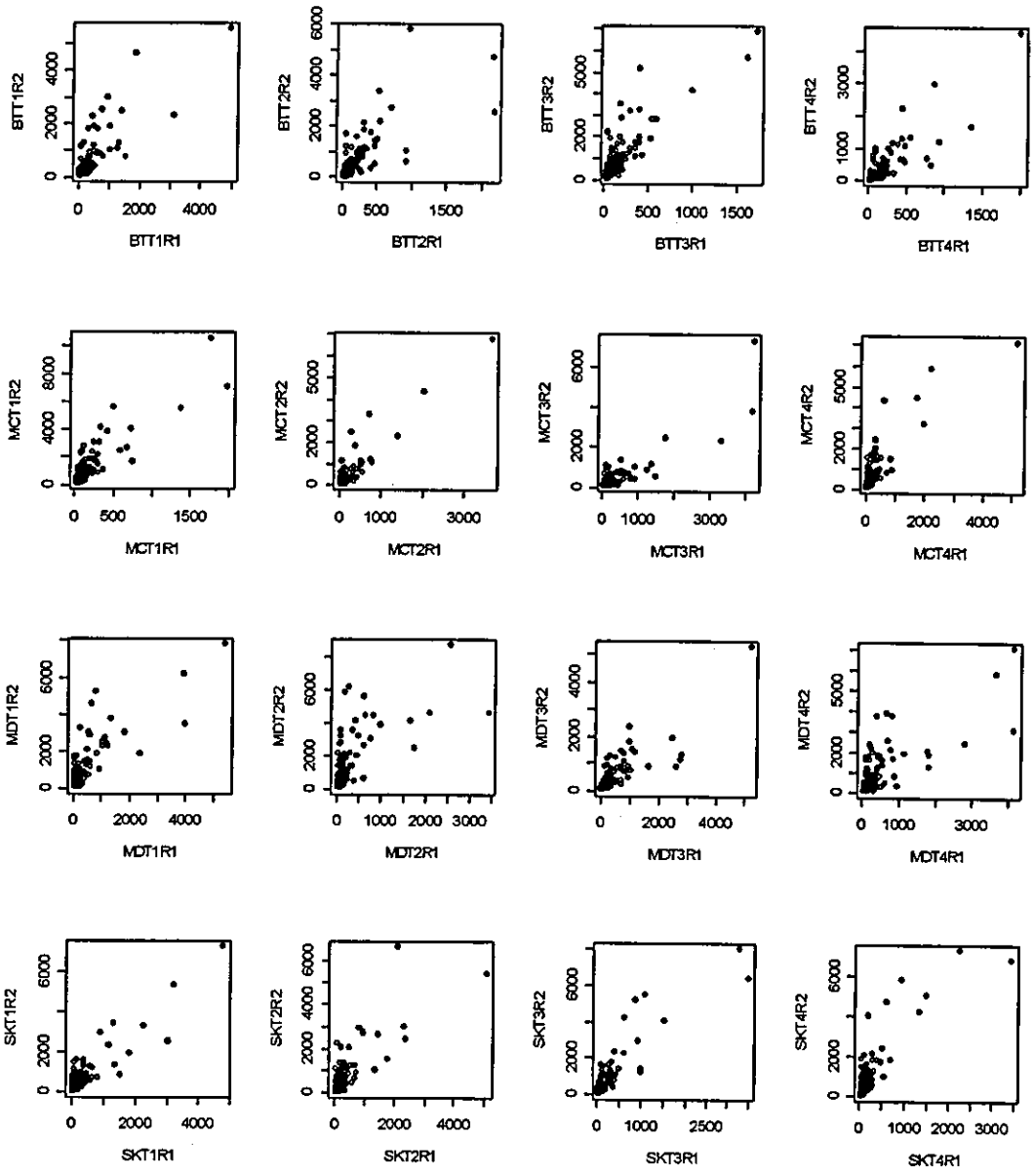


図1　16種類の実験組み合わせにおける遺伝子発現強度の散布図
横軸と縦軸はそれぞれ1回目と2回目の実験における発現強度

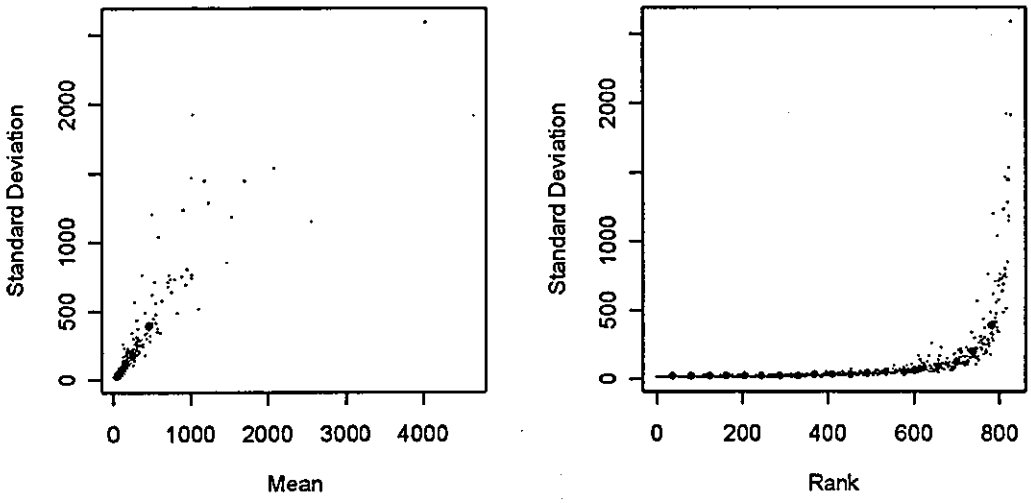


図2　各遺伝子の平均強度と標準偏差の散布図
左図の横軸は平均強度、右図の横軸はその順位、大きな点は系列の移動平均

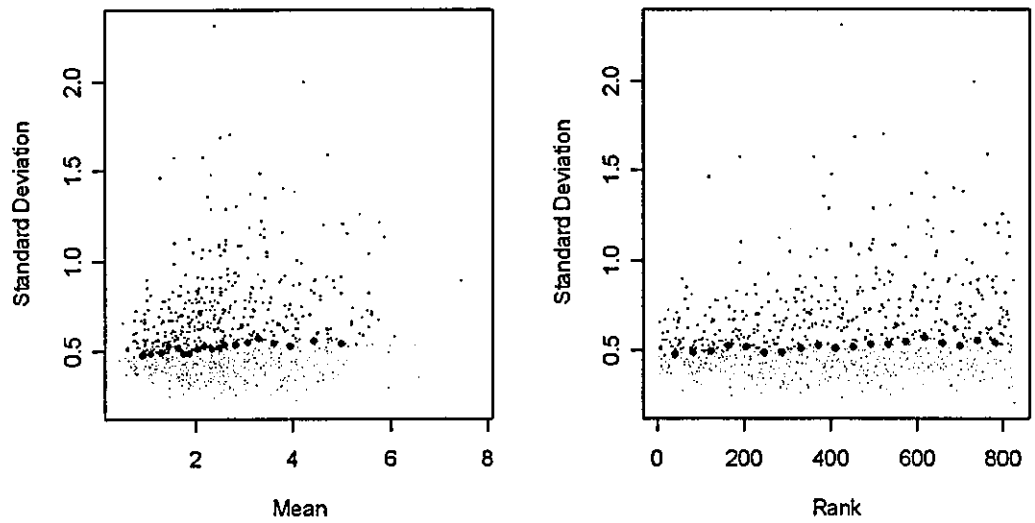


図3　分散安定化変換後の各遺伝子の平均強度と標準偏差の散布図
左図の横軸は平均強度、右図の横軸はその順位、大きな点は系列の移動平均

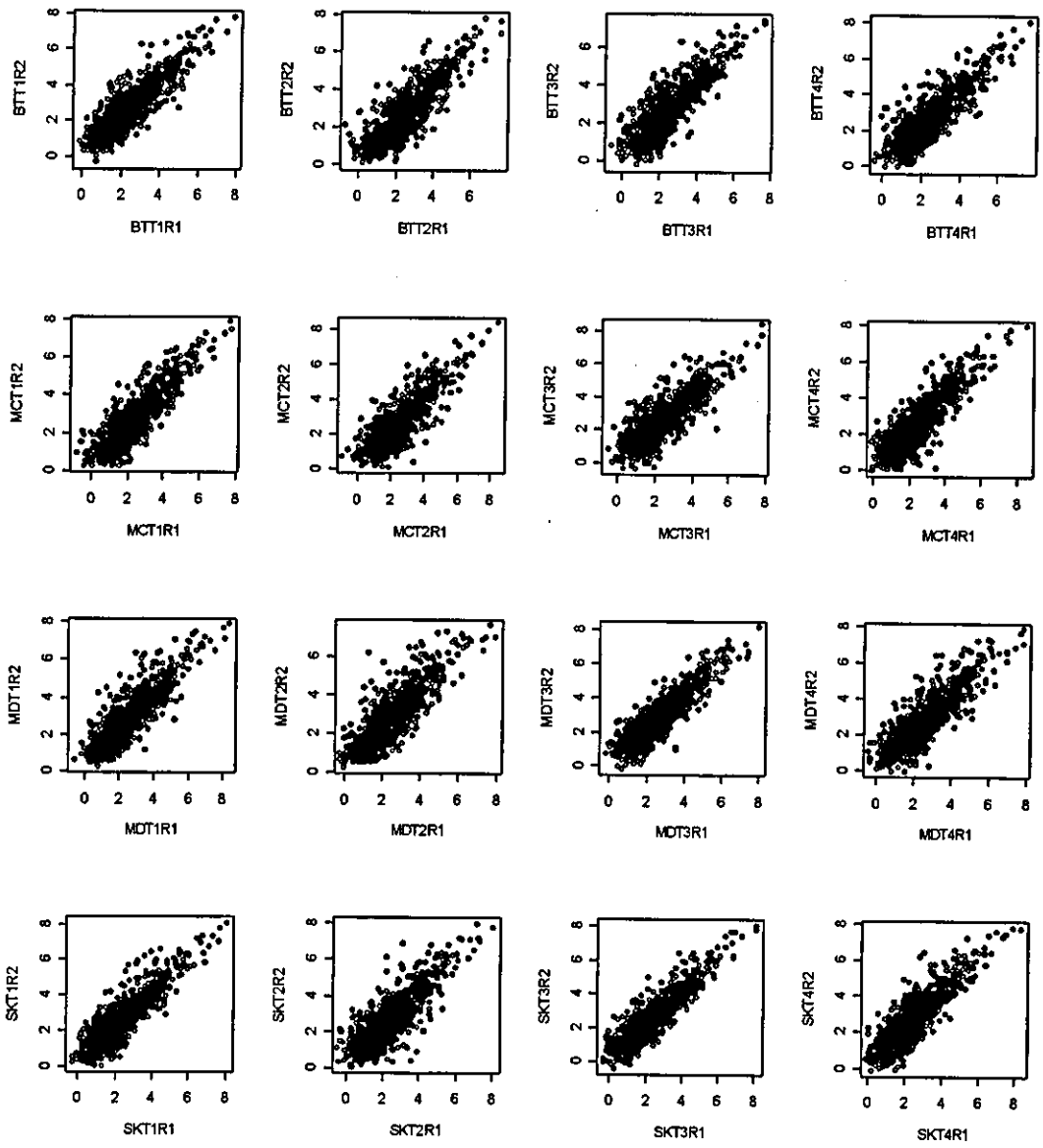


図4　分散安定化変換後の遺伝子発現強度の散布図

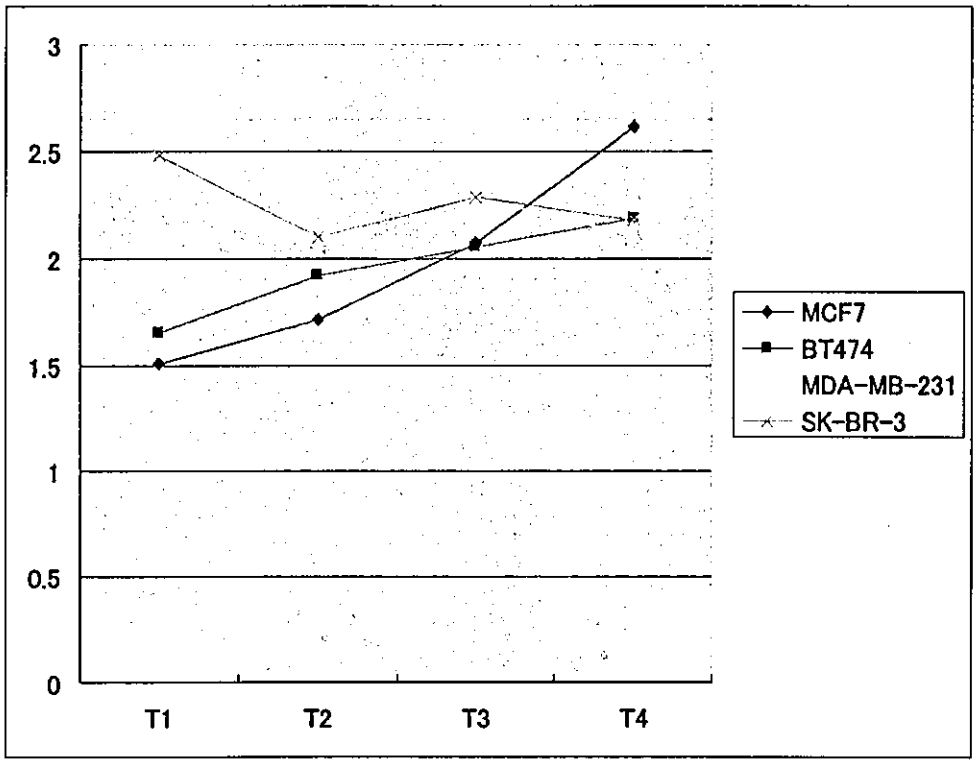


図5　p21-activated kinase alphaの発現プロファイル
データ点は2回の実験繰り返しにおける平均値

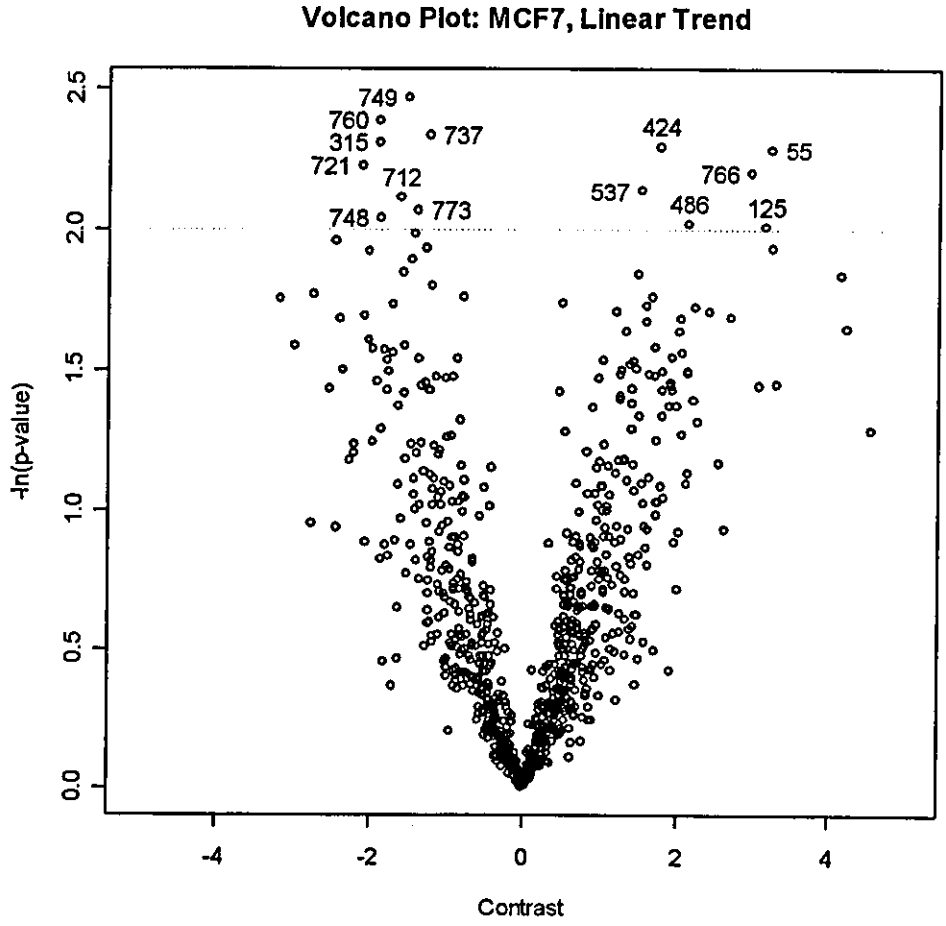


図6　MCF7細胞株におけるボルケーノプロット

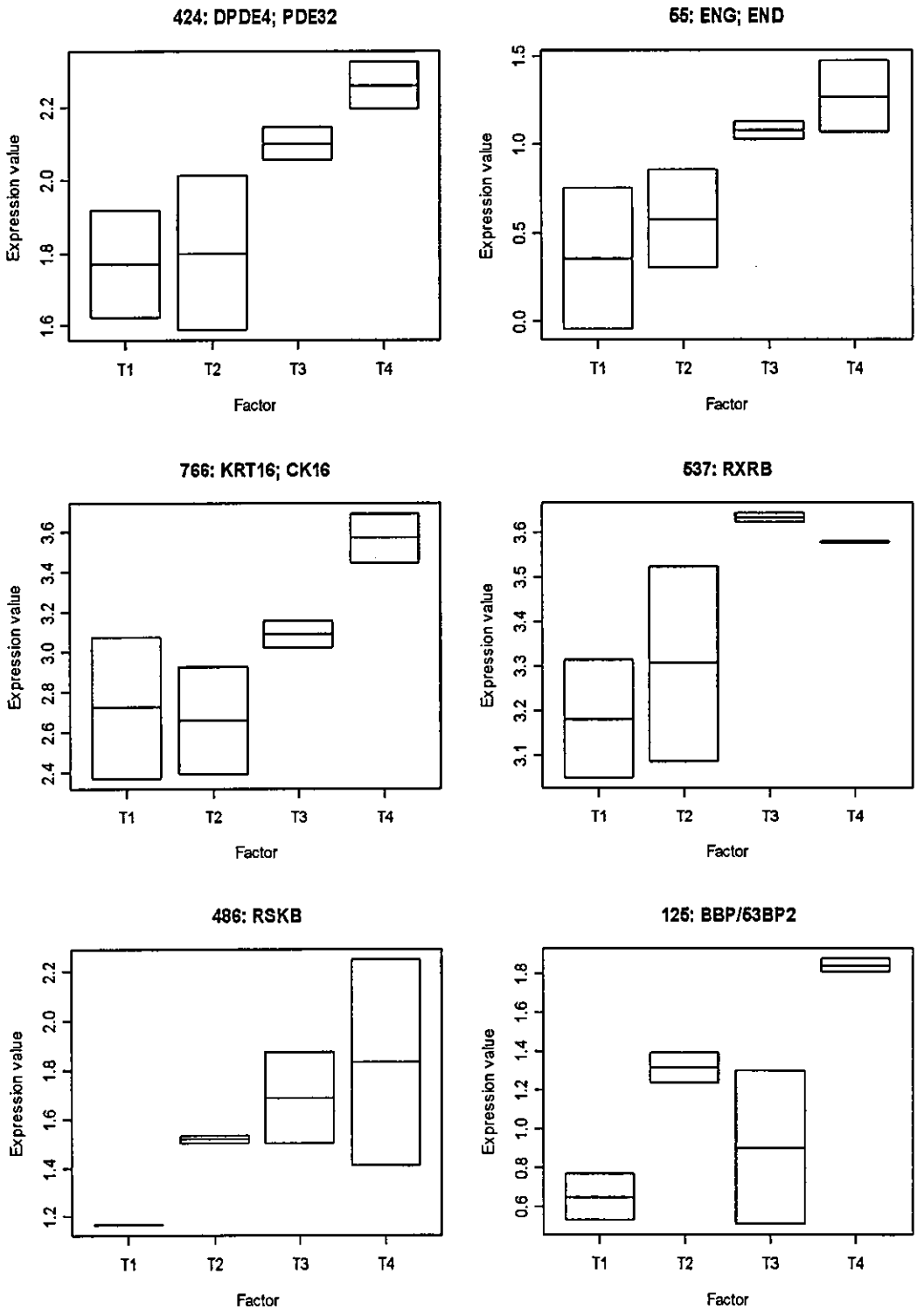
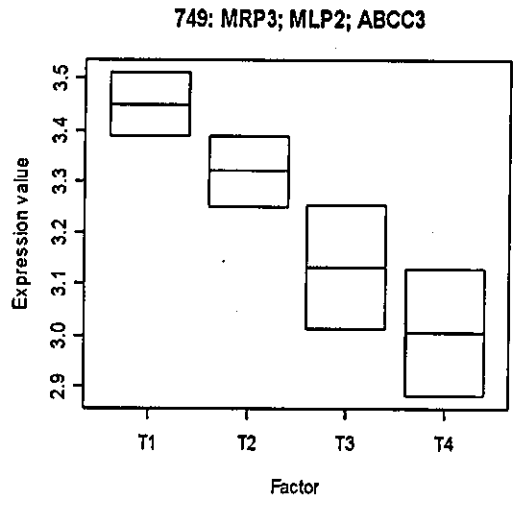
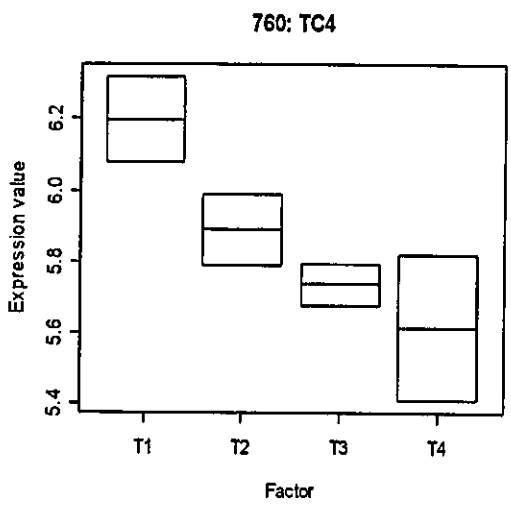
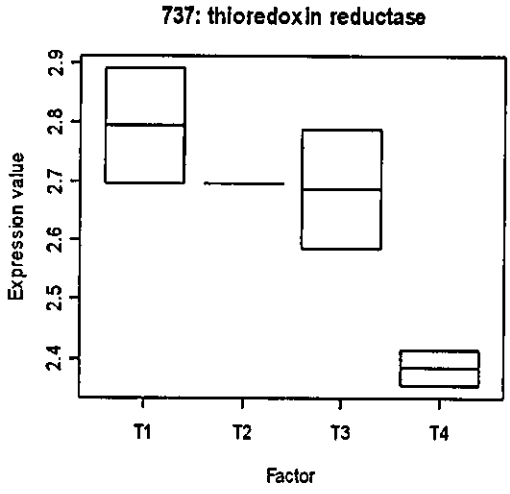
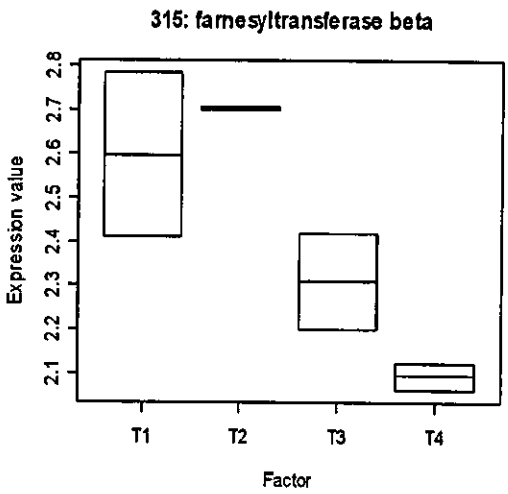


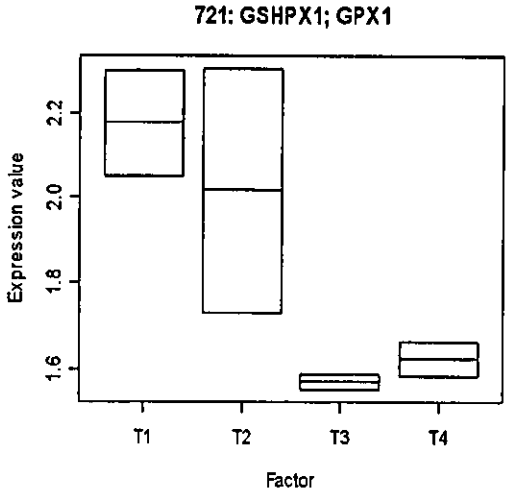
図7　MCF7細胞株における増加傾向遺伝子のプロット
（箱の上下はデータ点、中線は平均値）

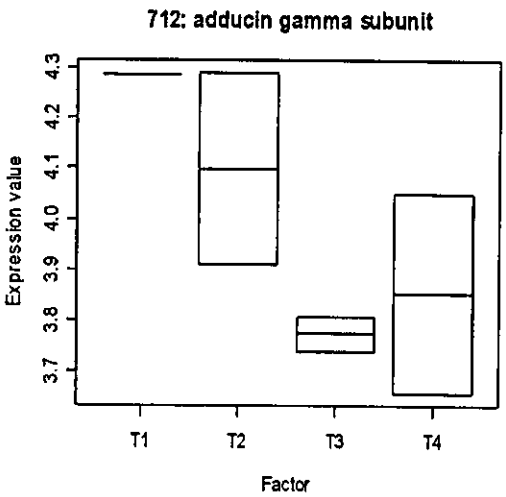

749: MRP3; MLP2; ABCC3
Expression value
2.9
3.0
3.1
3.2
3.3
3.4
3.5
T1
T2
T3
T4
Factor


760: TC4
Expression value
5.4
5.6
5.8
6.0
6.2
T1
T2
T3
T4
Factor


737: thioredoxin reductase
Expression value
2.4
2.5
2.6
2.7
2.8
2.9
T1
T2
T3
T4
Factor


315: farnesyltransferase beta
Expression value
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
T1
T2
T3
T4
Factor


721: GSHPX1; GPX1
Expression value
1.6
1.8
2.0
2.2
T1
T2
T3
T4
Factor


712: adducin gamma subunit
Expression value
3.7
3.8
3.9
4.0
4.1
4.2
4.3
T1
T2
T3
T4
Factor

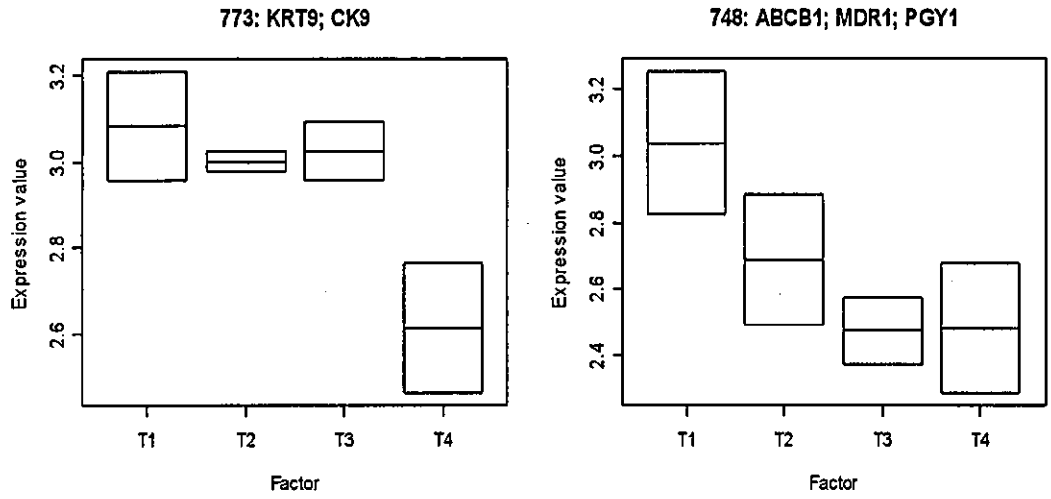


図8　MCF7細胞株における減少傾向遺伝子のプロット
（箱の上下はデータ点、中線は平均値）

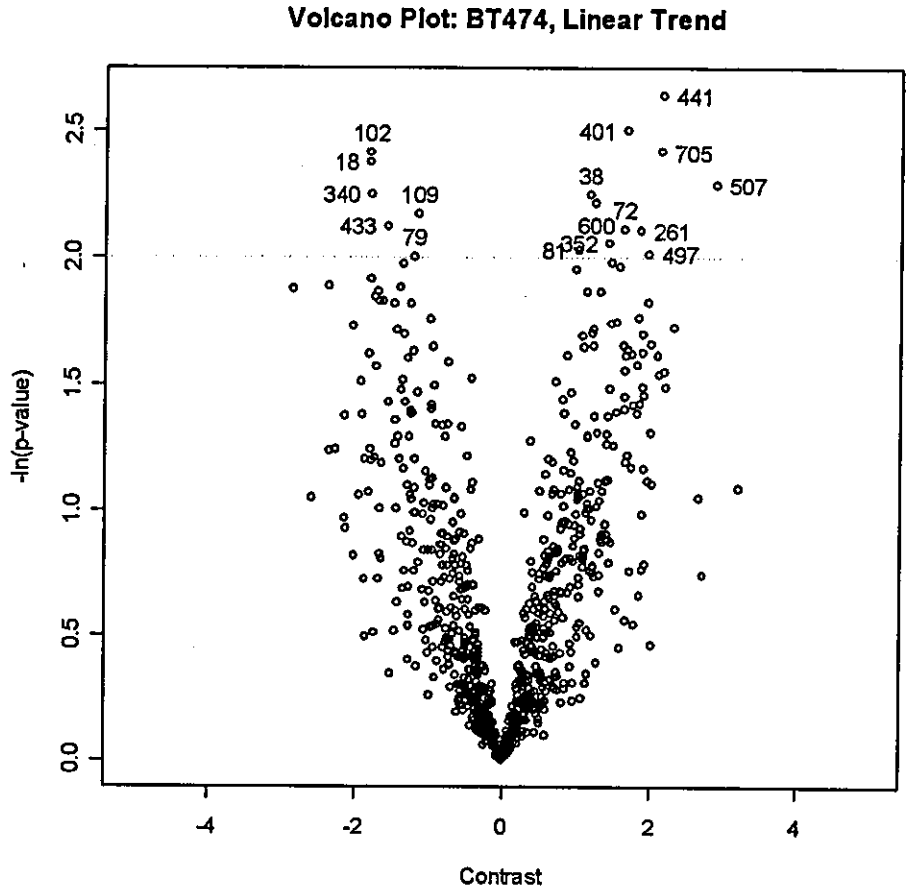
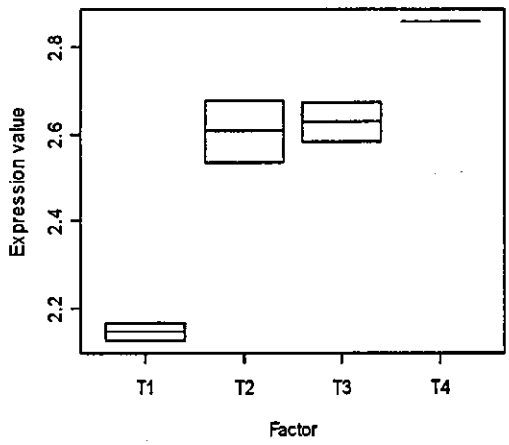
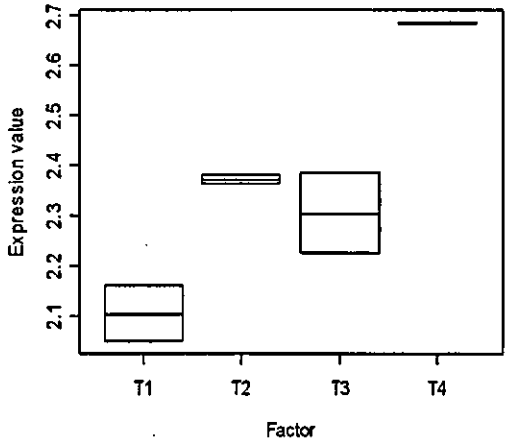


図9　BT474細胞株におけるボルケーノプロット

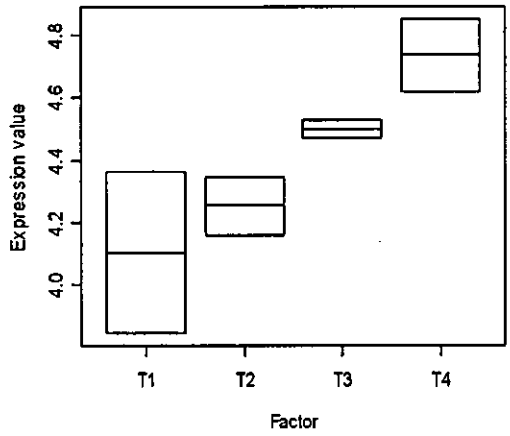
441: FRA1; FOSL1



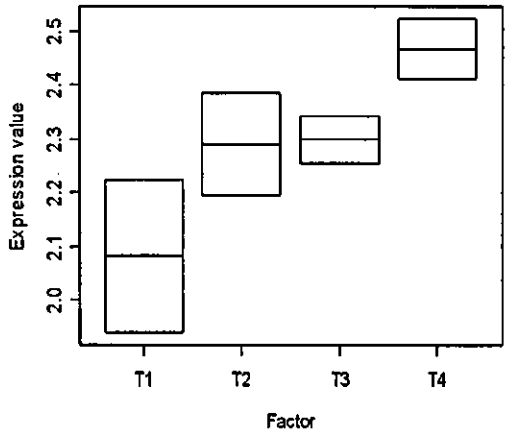
401: NEK3; NIMA-related protein kinase 3; HSPK 36



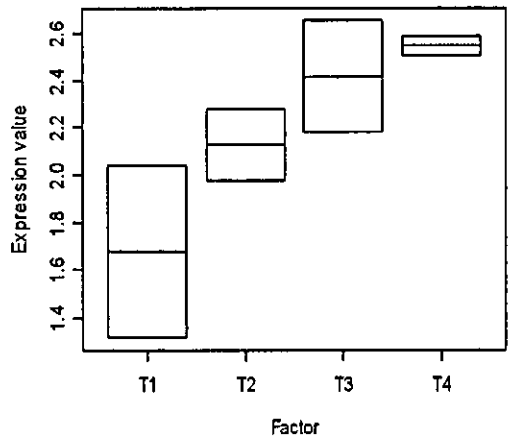
705: RAD23A; hHR23A



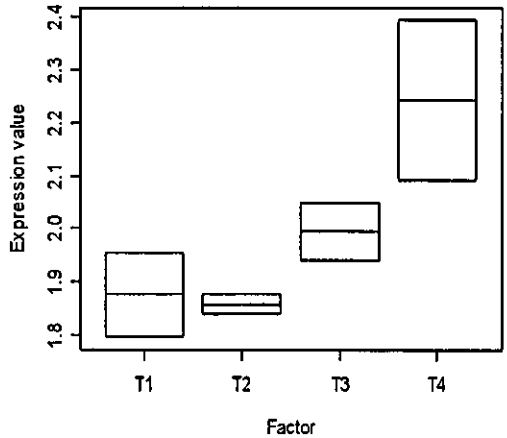
38: ECM2



507: RAP-1A; C21KG; SMG-p21A; G-22K



600: SH3P12



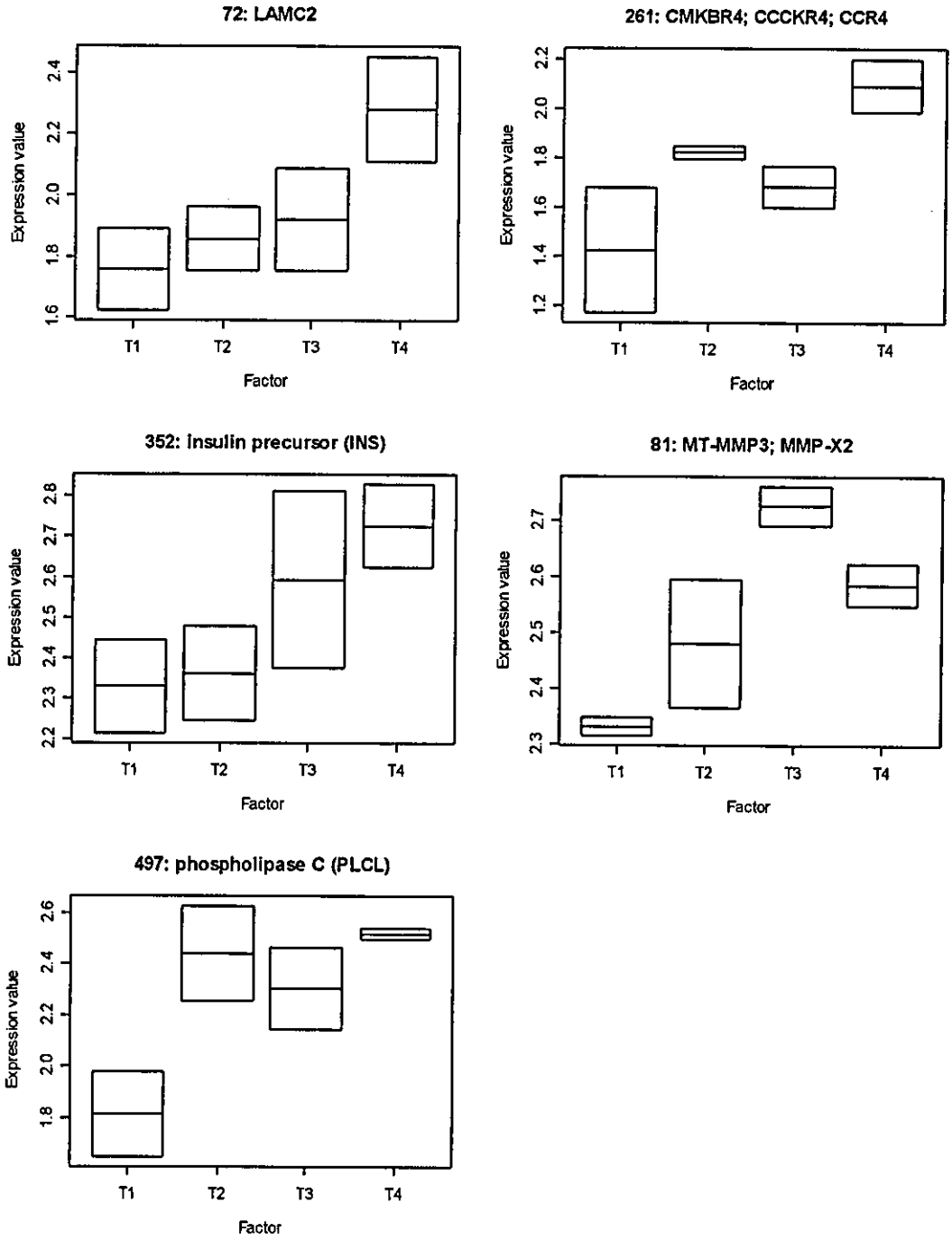


図10　BT474細胞株における増加傾向遺伝子のプロット
（箱の上下はデータ点、中線は平均値）