国立医薬品食品衛生研究所　毒性部　御中

委託研究報告書（STEP9）

発現データ補正のための係数最適化研究

1.テーマ

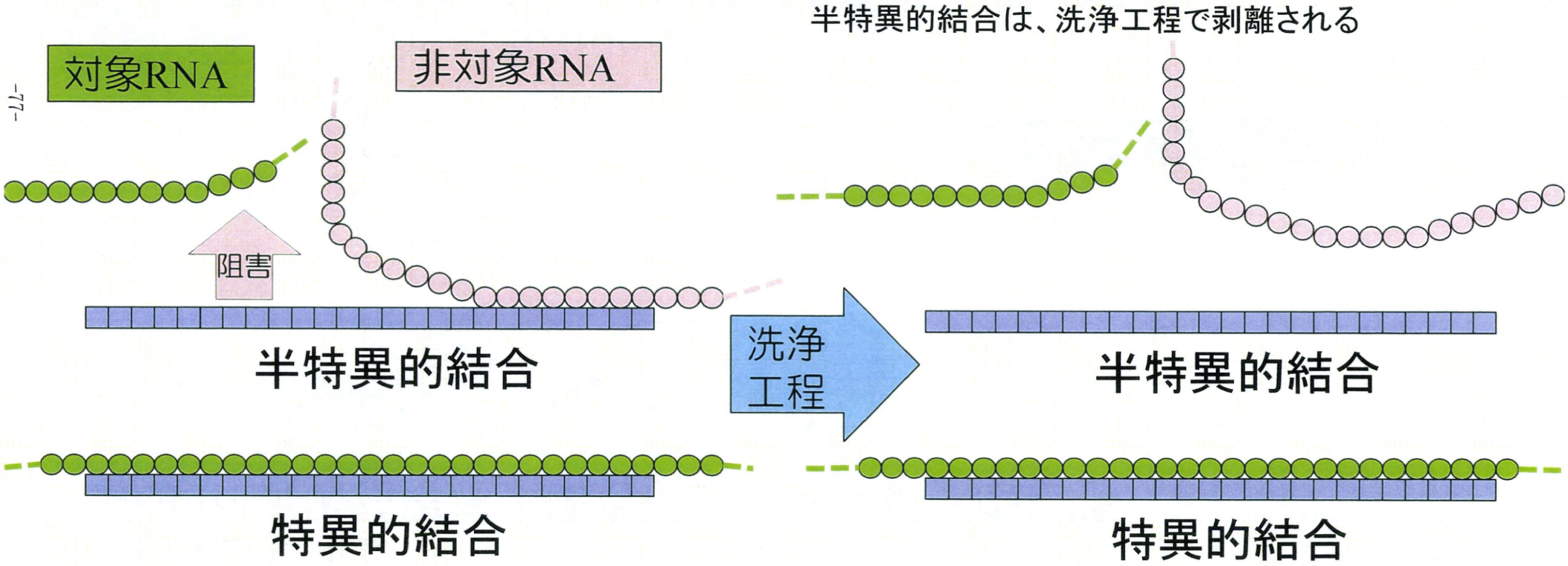
発現データ補正のために、実験データを使用した学習により係数を求めている。様々な種類の実験データを用いて学習することにより係数の精度が向上するため、ここでは出来るだけ多くのプローブについて適切な係数を得られるように学習用の実験データを選別、もしくは必要に応じて調整したサンプルを新たに測定し、学習処理と係数の最適化を進めた。

2.MLANG
Langmuir吸着等温平衡方程式式の多溶質への拡張

MLANGの基本概念およびアルゴリズムは、平成20年度の研究において確立した。

半特異的結合(Semi-specific Hybridization)という概念により多溶質吸着等温平衡方程式を構築

半特異的結合は、プローブと非対称RNAとの間に弱い結合力が働くことに基づく。
この結合力のため、非対称RNAもある程度の時間とどまり、平衡状態に於いても特異的結合の割合は一定以上にならない。



2.MLANG
Langmuir吸着等温平衡方程式の多溶質への拡張

各RNAの真の濃度が既知の場合の蛍光強度を導き出すモデル

HybridizationとWashingの2ステップ

非対象RNA

結合自由

対象RNA

Hybridization
多溶質による
熱平衡方程式

酸　温度　時間

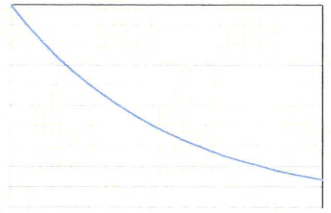
対象物質：観測したい対象

対象外物質

対象RNAのみとなる

Washing
減衰現象

酸　温度　時間



2.MLANG
Langmuir吸着等温平衡方程式の多溶質への拡張

多溶質による吸着のモデル化、Langmuir吸着等温式と同じ仮定

仮定

吸着速度は、吸着体の残り個数と溶質の濃度に比例すると仮定する

$$V' = K \cdot \left(N - \sum_{i=1}^{n} N_i \right) \left(\sum_{i=1}^{n} C_i \right)$$

個々の溶質の吸着速度は、溶質の濃度の比に比例する（多溶質用に新規導入）

$$V_i' = \frac{C_i}{\sum_{j=1}^{n} C_j} V'$$

個々の溶質の離脱速度は、吸着した個数に比例する

$$V_i = \alpha_i \cdot N_i$$

式変形

プローブiに吸着したターゲットjの個数

$$N_{ij} = \frac{C_j}{\alpha_{ij}} \cdot K \cdot \left(N - \sum_{j=1}^{m} N_{ij} \right) + \varepsilon_{ij}$$

プローブiに吸着したターゲットjの離脱速度係数
吸着速度係数は1とし、αで調整する

$$\alpha_{ij} \geq 0$$

2.MLANG
Langmuir吸着等温平衡方程式の多溶質への拡張

各RNAの真の濃度がわかっている場合の計測値を導き出すモデル(数式表現)

ステップ1　Hybridization
プローブiに吸着したターゲットjの個数

$$N_{ij} = \frac{N - \sum_{k \neq i}^{m} N_{ik}}{\alpha_{ij} + C_j} C_j$$

ステップ2　Washing
プローブiに吸着したターゲットjの個数

$$n_{ij} = N_{ij} \cdot e^{-\gamma \cdot \alpha_{ij}}$$

ステップ3　Observation
プローブiで観測される値

$$n_i = \sum_{j=1}^{m} n_{ij} + \bar{\varepsilon}_i$$

N：吸着体個数（全プローブ同一）
C_j：ターゲットjの濃度
N_{ij}：プローブiに結合したターゲットjのRNA数
α_{ij}：プローブiからターゲットjが離脱する速度
γ：洗浄係数変換定数
n_{ij}：洗浄後にプローブiに結合しているターゲットjのRNA数
$\bar{\varepsilon}_i$：観測誤差(正規分布と仮定)

3.係数学習方針

- プローブとターゲットの全組み合わせに対して、物理化学的に塩基配列に基づいた結合力(離脱速度係数)を求めることは、いくつかの課題があり、困難である。
  - ターゲット範囲外での結合
  - 複数個所における結合
  - 完全平衡状態へ未到達
  - 断片化
  - 増幅効率
  - RNA同士の結合
  - RNA2次構造
- MLANGでは、観測値から離脱速度係数を求めている。

- 半特異的結合に対する離脱速度係数は、対象以外のRNAとの結合力である。種々のRNAが色々な割合で混合されたサンプルの計測データを用いて学習を行うことにより、離脱速度係数を求め、精度を高める。

3.係数学習方針

次の実験条件で実験を用いて学習を行う

検体なし

Affymetrix社の標準実験プロトコルにおいてコントロールとして、添加するRNAが存在する。添加RNAのみのデータを用いることで添加RNAの影響を係数に反映させる

臓器およびGSCの直交表による混合計測　＆　LBM

Percellome用の外部スパイクとして、6種のRNAを添加している(GSC)。これらのRNAによる影響および複数の臓器サンプルの混合により、相互作用に対する係数を学習する

臓器基本発現TTG関連Vehicle (GSC-dilution, TTG122-Y, TTG160-G, TTG160-H, TTG040-C)

各臓器で発現するRNAが関与する係数を学習する。特に、臓器特有な遺伝子および共通する遺伝子間の係数を学習する

TTG020-L(TCDD処置)

化合物を投与した際に特有に変動する遺伝子の係数を学習する

3.係数学習方針
直交表による混合計測

品質検査などで使用される、複数事象の組み合わせ試験のための手法として、直交表が用いられる。この直交表を応用して、各臓器およびスパイクを混合し、実験することにより、これらの間の関係を示す係数を求める

L32直交表を用いて、
・臓器濃度2水準(無/有)
・GSC濃度4水準(無/飽和無/飽和小/飽和大)
として、実験を実施する

L32直交表
31因子2水準に対して、32回の実験を行う。
GSCに関して2因子をまとめて、4水準とする方法を採用した

臓器(19)
腎臓
精巣
脾臓
小腸上皮
全脳
膵臓
心臓
骨格筋
肝臓
卵巣
骨髄
肺
子宮
胸腺
胚（E8.5）
胚（E9.5）
脂肪
腸間膜
ES細胞

GSC(6)
THR
Lys
Phe
Dap
Trp
Lambda

臓器は、臓器間で発現パターンが異なるように、NIHS毒性部様に選定いただいた。

4.異常係数削除

いくつかの要因により、係数学習中に係数の値が異常な値を示すことがある。全体に対して悪影響を及ぼす可能性があるため、学習の途中で、異常係数を見つけ出し削除した。

想定される異常係数発生の要因	
異常計測値	デブリなど、計測上の異常な値が含まれている場合には、異常な係数を生み出す可能性がある。一定の範囲内のエラーであれば、学習により他の実験結果により打ち消されるが、打ち消されず残る可能性がある。
係数不足	塩基配列からプローブとmRNAの間の離脱速度を示す係数を使用する。しかし、計算時間を実用時間に収めるために閾値を決めてあり、一定の長さ以下の関係しか見いだせなかった組み合わせは、係数を作成していない。塩基配列の間違いなど係数が作成されなかった組み合わせで、強い関係があった場合には、異常な係数となる可能性がある

4.異常係数削除

学習の途中で次のものを削除した

係数値

- 解離係数300以上の通常ターゲットに関する係数
  - 非特異的結合の解離係数を計算の都合上350としている。半特異的結合の解離係数が、これに近い値は削除

異常プローブ

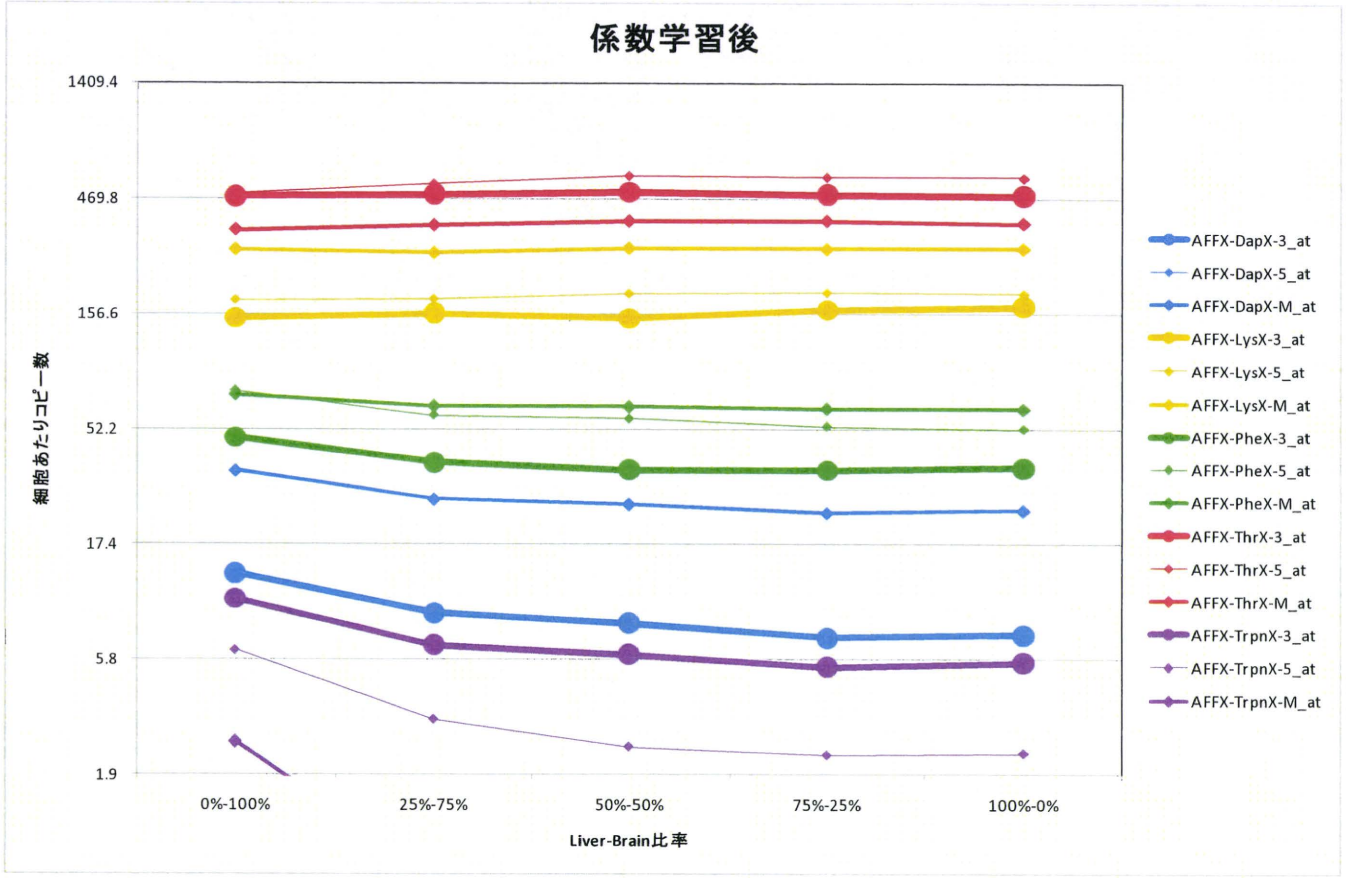
- PMプローブの設計上ターゲットとの解離係数が8以上
  - 目的とするRNAとの解離係数として１程度を目標としている。速く解離しすぎるので削除
- MMプローブの設計上ターゲットとの解離係数が１未満
  - 目的とするRNAとPMとの解離係数として１程度を目標としている。MMプローブと強く結合するのは削除
- 解離係数が0.3以下を含むプローブ
  - 強力すぎる結合なので、不整合と考えられる。（25mer一致が2か所存在することも考えられる）
- ターゲットとの結合が存在しないプローブ
  - 異常係数として削除されたものなど
- PMプローブ内の解離係数順位トップとターゲットとの解離係数の比率
  - 本来はターゲットがトップのはずだが、1.5倍を超えたら初期値として使用しない。4倍を超えたらプローブ削除
- PMプローブの存在しないMMプローブ
  - 設計上ターゲットを代表するものがPMプローブである。MMはいくつか存在しても構わないが、MMの方が主になると計算上不安定になる

ターゲット内

- ターゲットの初期化に使用しているプローブよりも、解離しにくいプローブが存在する
  - ターゲット特異性があるとみなされているプローブよりも、解離しにくいプローブが存在するとは考えにくい

5.係数学習後の補正性能
LBMデータGSC濃度

GSCの濃度が与えた濃度の比率通りの推定値になっているか確認した。

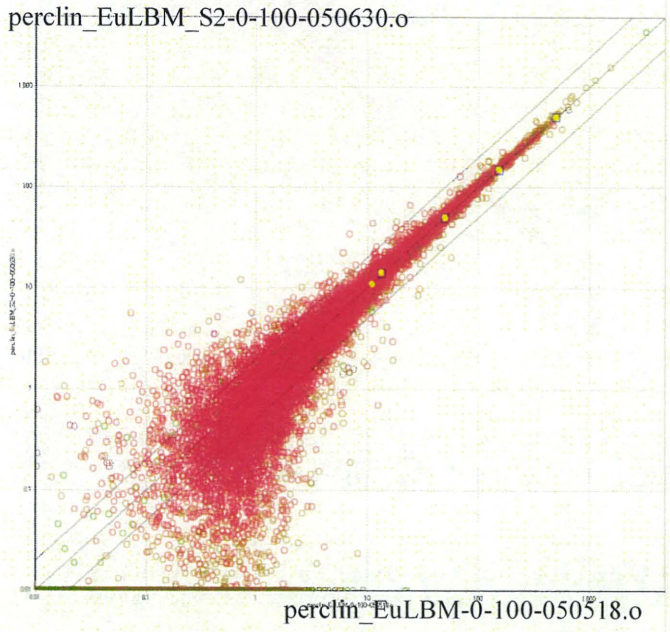


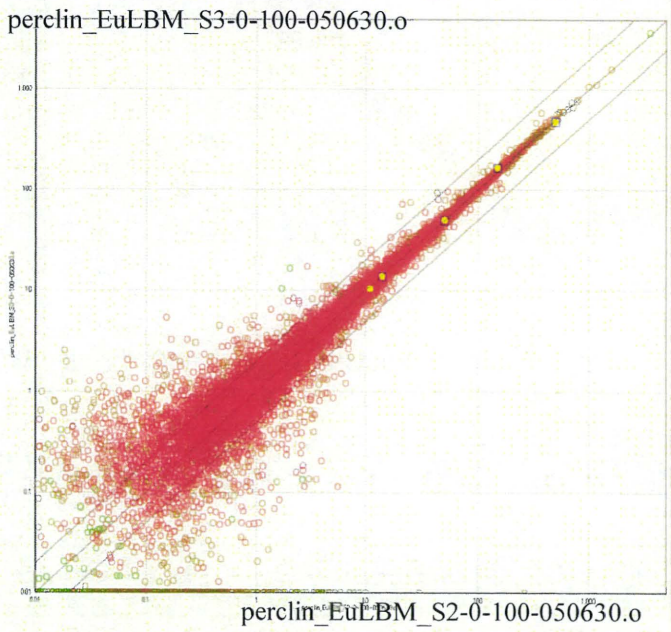
1番目THRと2番目Lysは、与えた濃度通りの補正結果となった。

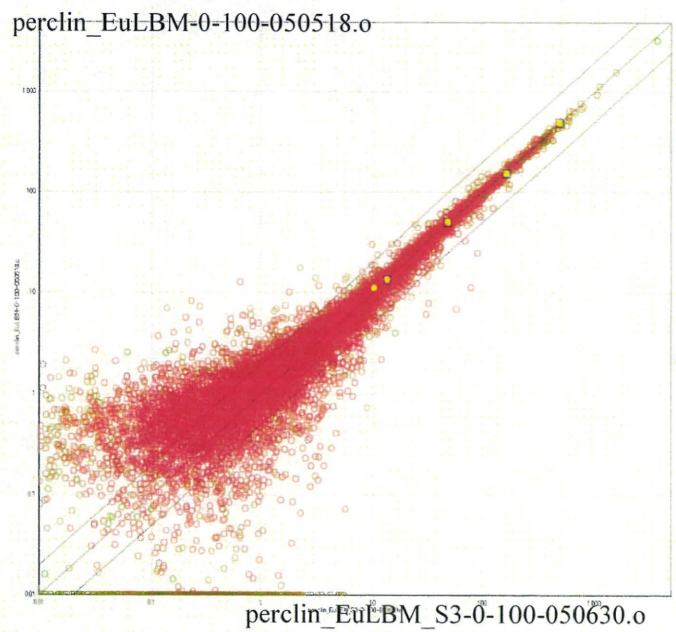
5.係数学習後の補正性能
LBMデータ散布図による確認

Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver0%-Brain100%　三重化データをサイクリックに散布図を作成した





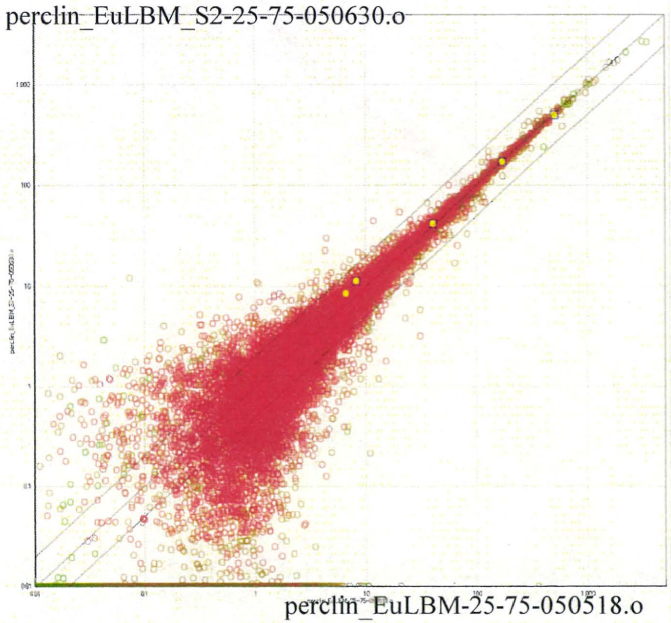


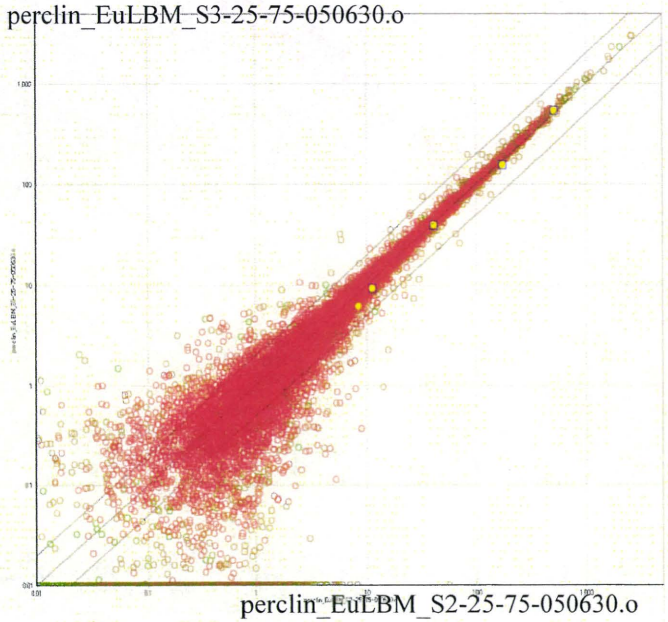
EuLBM-0-100-050518が他の2チップよりも、少し大きな値となっているが、1コピー以下である

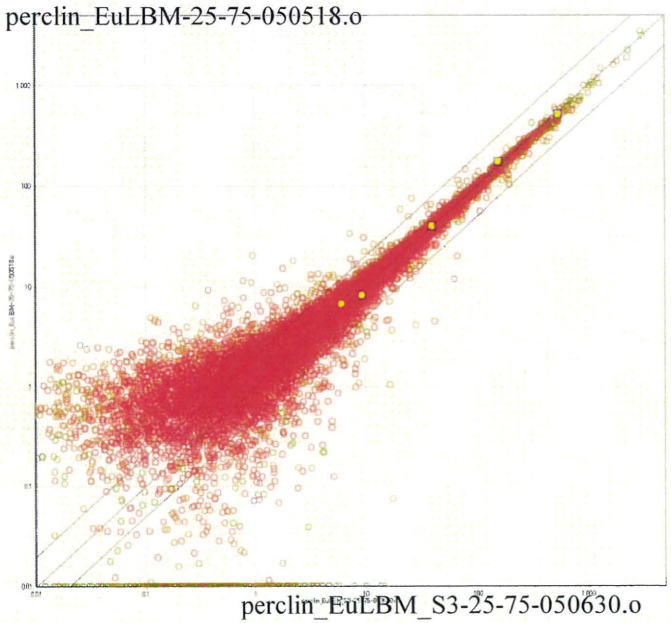
5.係数学習後の補正性能
LBMデータ散布図による確認

Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver25%-Brain75%　三重化データをサイクリックに散布図を作成した





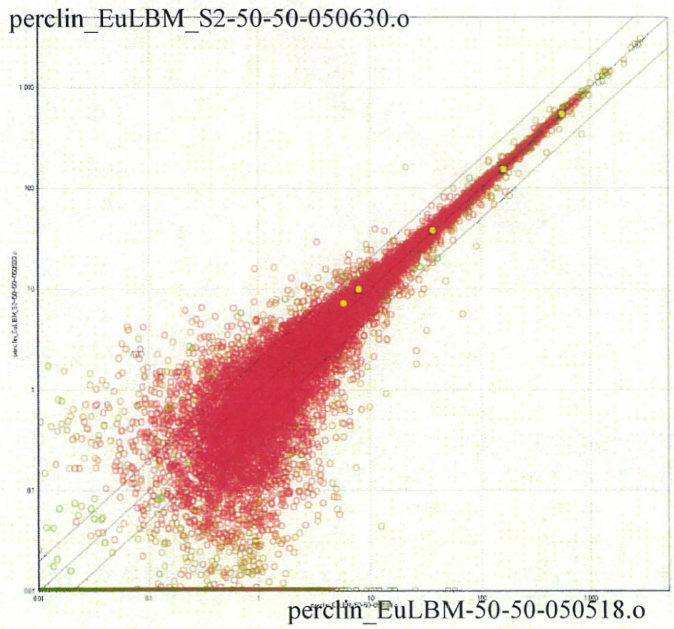


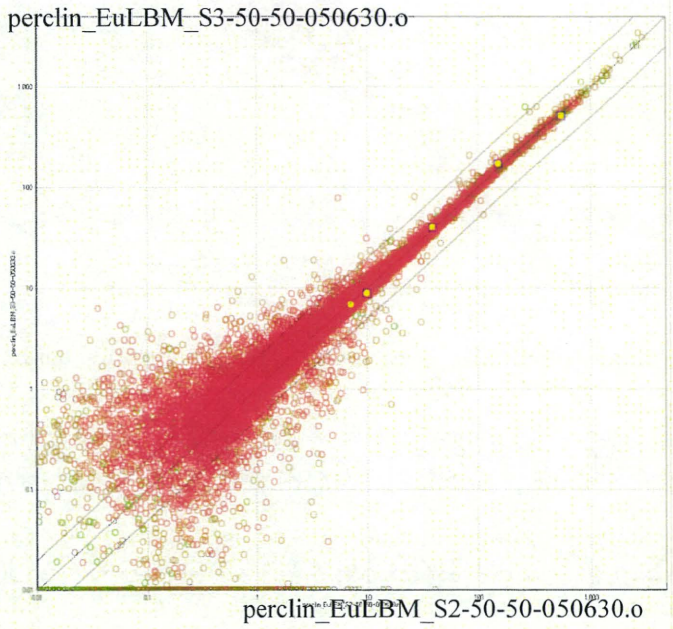
EuLBM-25-75-050518が他の2チップよりも、少し大きな値となっているが、1コピー以下である

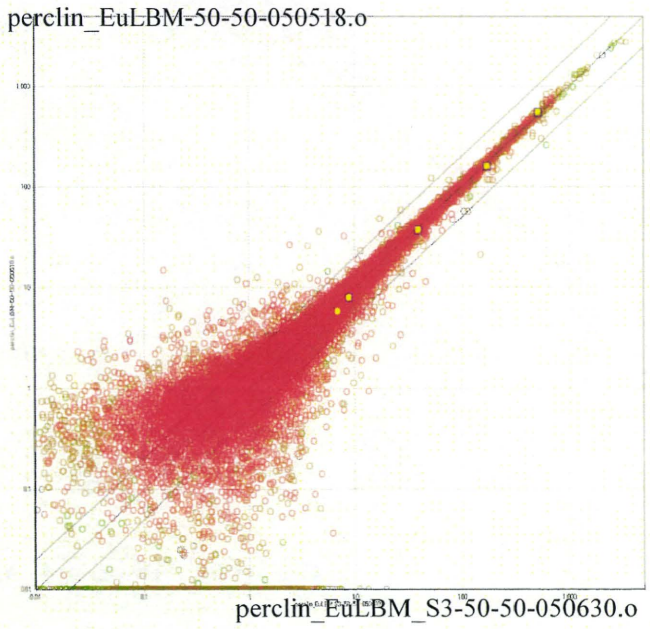
5.係数学習後の補正性能
LBMデータ散布図による確認

Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver50%-Brain50%　三重化データをサイクリックに散布図を作成した





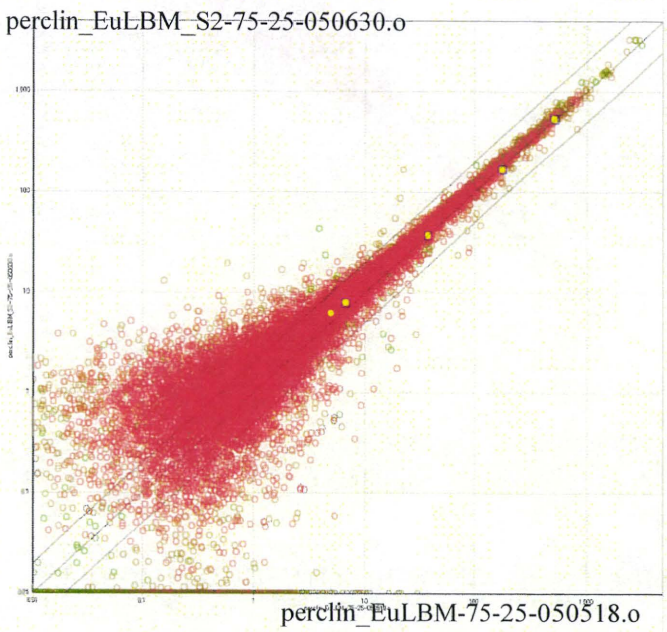


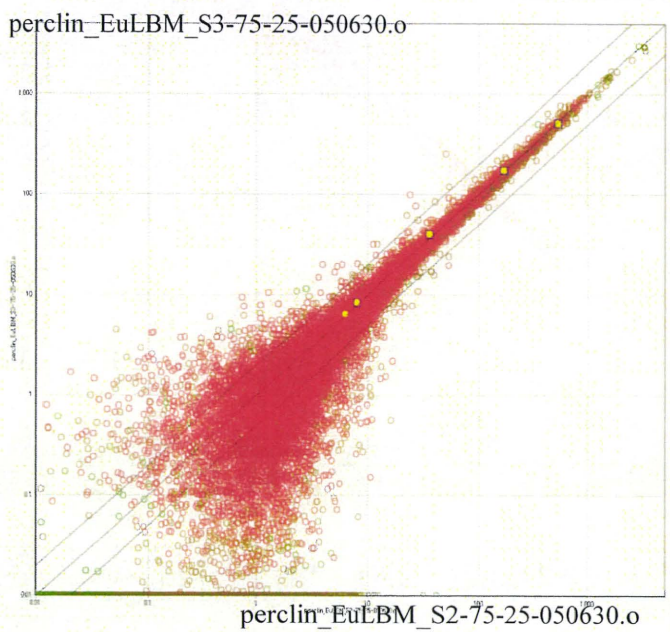
EuLBM-50-50-050518が最も高く、EuLBM_S3-50-50-050630が最も低く出ている。しかし、その差も1コピー以下である

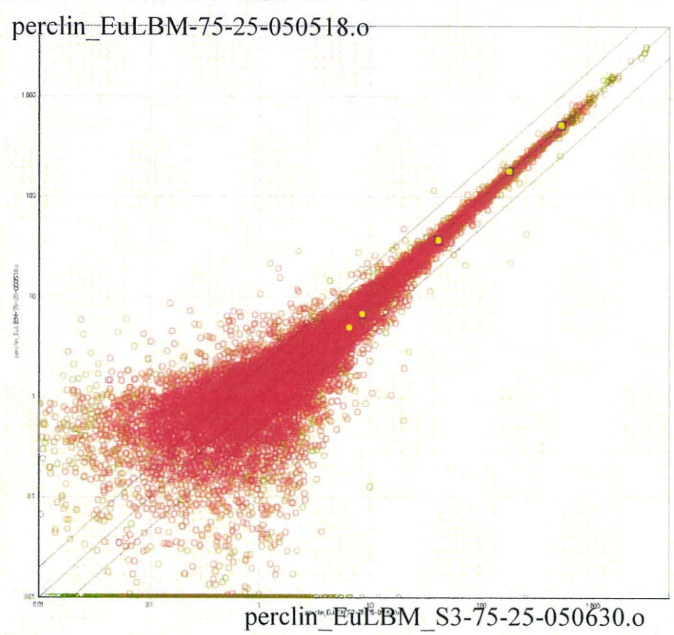
5.係数学習後の補正性能
LBMデータ散布図による確認

Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver25%-Brain75%　三重化データをサイクリックに散布図を作成した





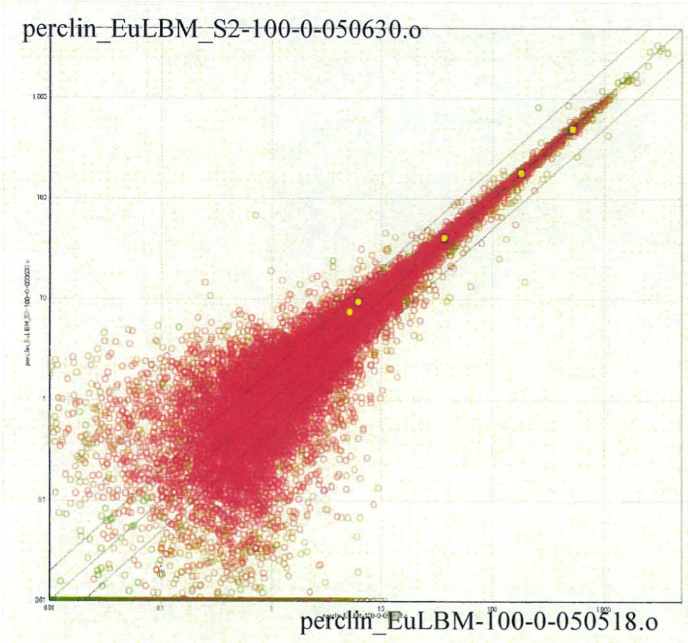


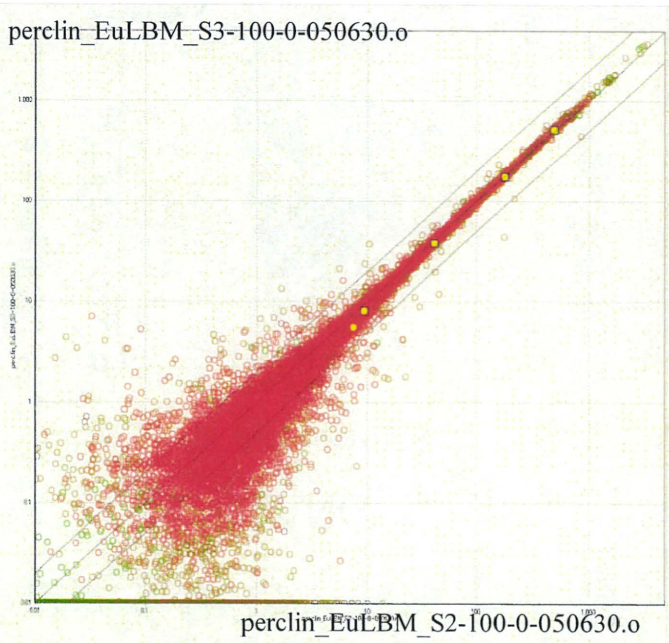
EuLBM_S3-75-25-050630が最も低く出ている。その差は1コピー以下である

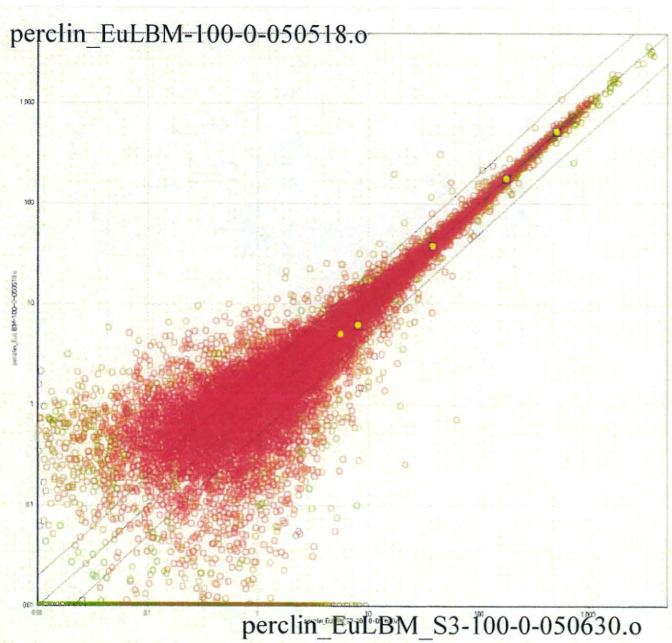
5.係数学習後の補正性能
LBMデータ散布図による確認

Liver-Brain-Mixtureの三重化実験データを用いて、低発現域における偏差を確認した。

Liver25%-Brain75%　三重化データをサイクリックに散布図を作成した







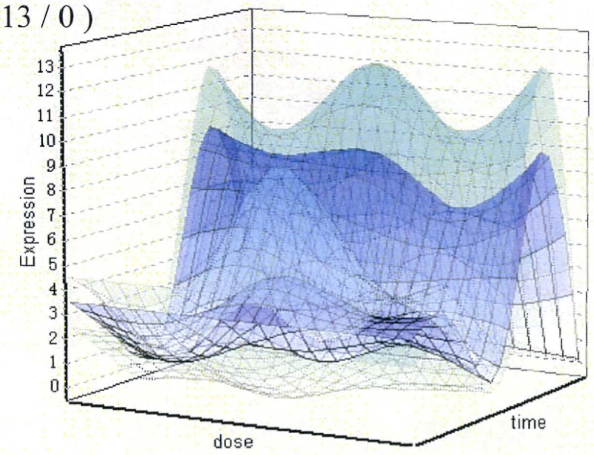
EuLBM_S3-25-75-050518が小さく、EuLBM-100-0-050518が最も大きいが、その差は、1コピー以下である

5.係数学習後の補正性能
MAS5/MLANG/QPCR結果比較

遺伝子Per1の比較

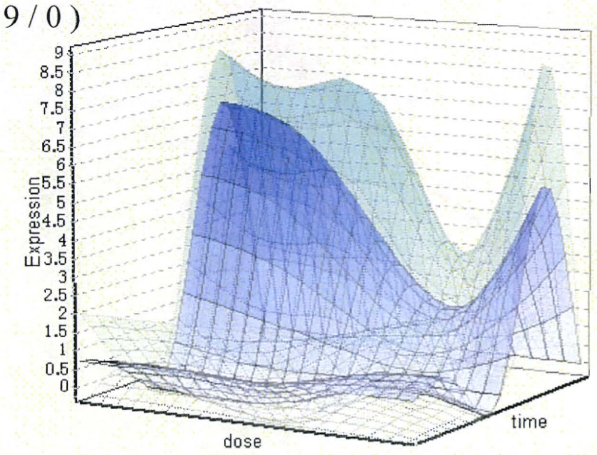
MAS5

TTG020-L_SpNC_0 1449851_at
Per1
(13 / 0)



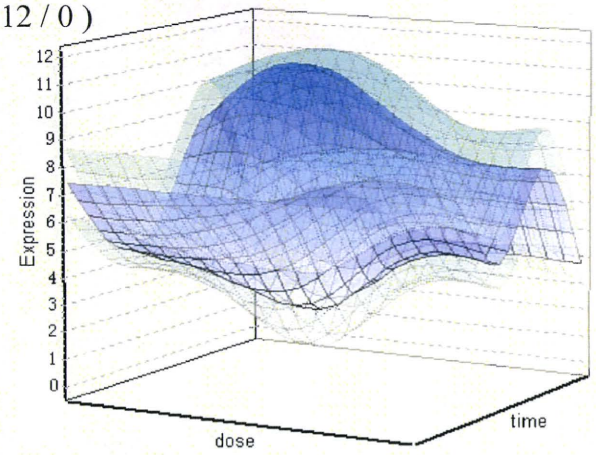
MLANG

perclin_TTG020-L_20110126 1449851_at
Per1
(9 / 0)



QPCR

TTG020-L_QPCR_SpNC_0 1449851_at
Per1
(12 / 0)



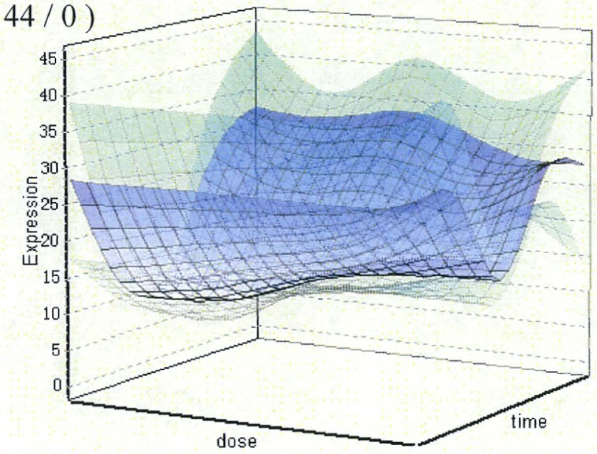
MAS5の2h,Mid-Lowにおける大きな変動は消えたが、QPCRとは異なる動きを示しており、別の遺伝子を観測している可能性がある。

5.係数学習後の補正性能
MAS5/MLANG/QPCR結果比較

遺伝子Ahrの比較

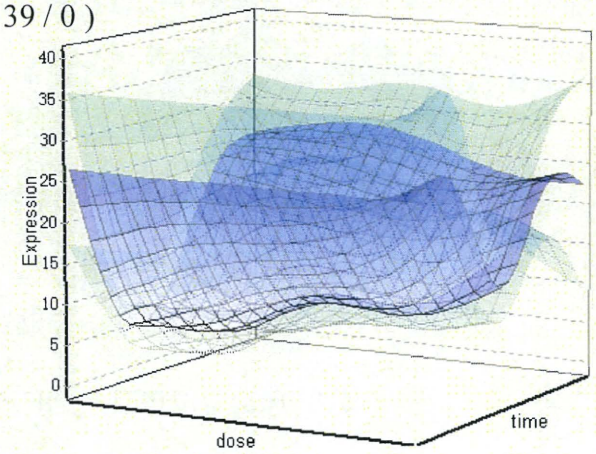
MAS5

TTG020-L_SpNC_0 1422631_at
Ahr
(44 / 0)



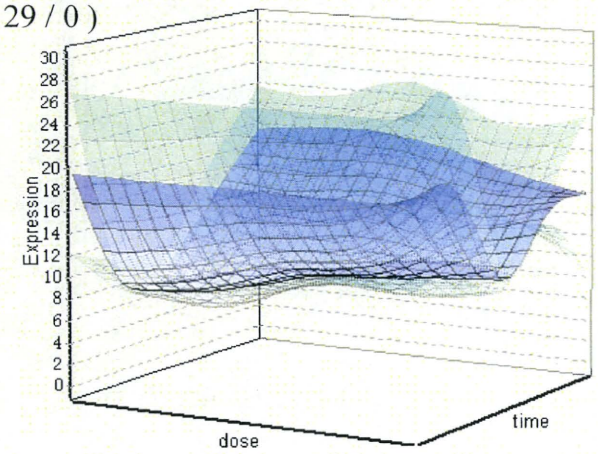
MLANG

perclin_TTG020-L_20110126 1422631_at
Ahr
(39 / 0)



QPCR

TTG020-L_QPCR_SpNC_0 1450655_at
Ahr
(29 / 0)



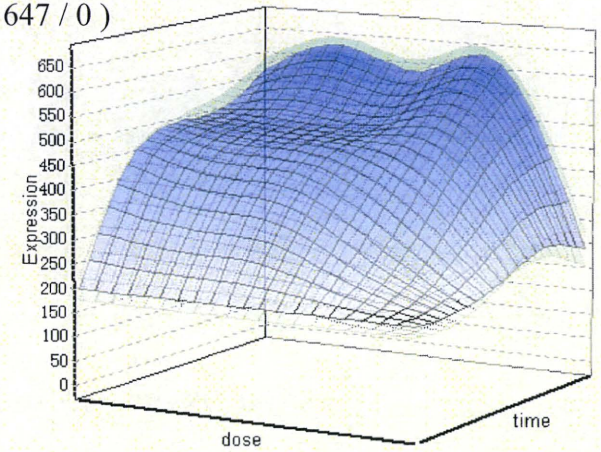
3手法の結果でほぼ似た形状が得られた。

5.係数学習後の補正性能
MAS5/MLANG/QPCR結果比較

遺伝子Cyp1a2の比較

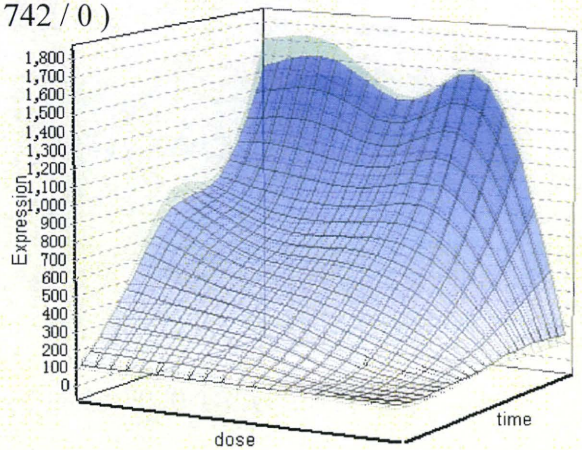
MAS5

TTG020-L_SpNC_0_1450715_at
Cyp1a2
(647 / 0)



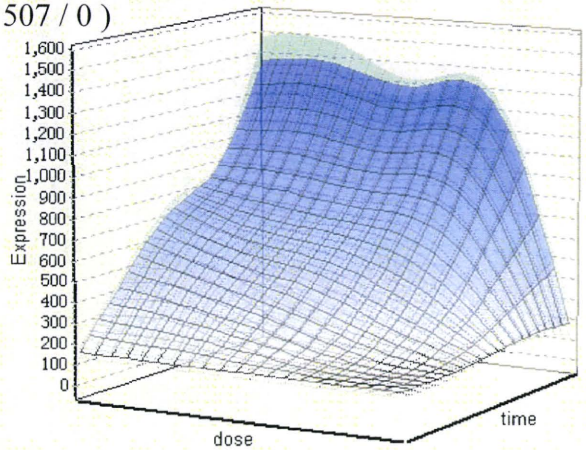
MLANG

perclin_TTG020-L_20110126_1450715_at
Cyp1a2
(1742 / 0)



QPCR

TTG020-L_QPCR_SpNC_0_1450715_at
Cyp1a2
(1507 / 0)



MAS5では、飽和により、8hや24hの形状が分からないが、MLANGは24hまで、mRNAが増加していることまで補正できている。