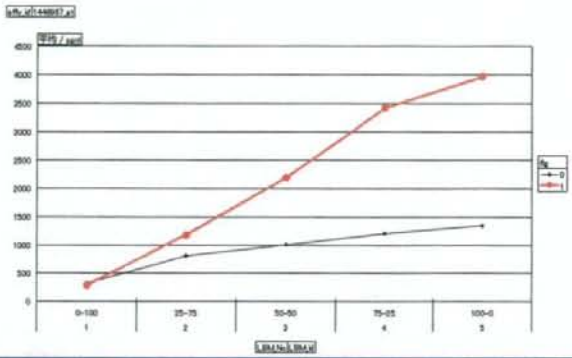
全プローブ計算③検証①

- 対象プローブセット:50%:50%において、1000程度の値を持つ
- 線形に引き延ばされ、高い値で不安定になっており、補正がある程度成功したと考えられる例(1448967_at:Nipsnap3a)

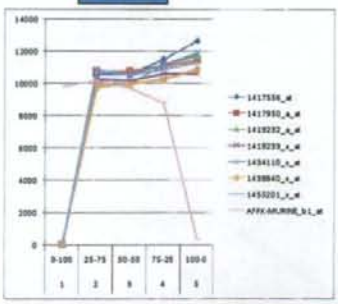


Liver側飽和のプローブは直線に引き延ばされ、高い値で、不安定になっている。

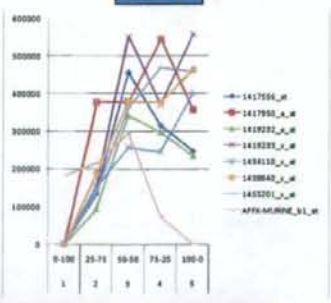
全プローブ計算③検証①

- 対象プローブセット:50%:50%において、10000以上の値を持つ

補正前



補正後

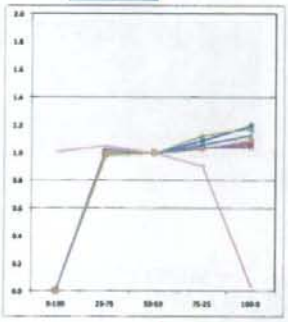


0%から50%まではすこし上向きだが直線に見えなくもない。しかし、50%を超えると、崩れている。Lngmuirの最大値推定に問題がありそうである

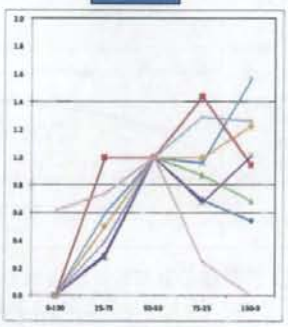
全プローブ計算③検証①

- 50%:50%において、1に正規化

補正前



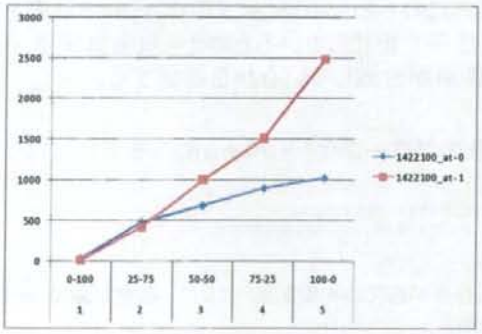
補正後



50%を超えるものでは、適切な補正ができているとは言い難い

全プローブ計算③検証①蝶の羽(7)

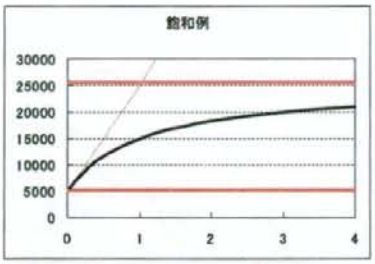
- Cyp7a1(1422100_at)で検証した



直線に引き延ばされたと考えられる。しかし、100%-0%では、行き過ぎた感がある

全プローブ計算③検証①課題

- 高い値を示す場合に、Langmuirの最大値を超える状況が発生した。推定されるLangmuirの最大値を引き上げれば、状況は少し改善する。



27000の場合にどうするか

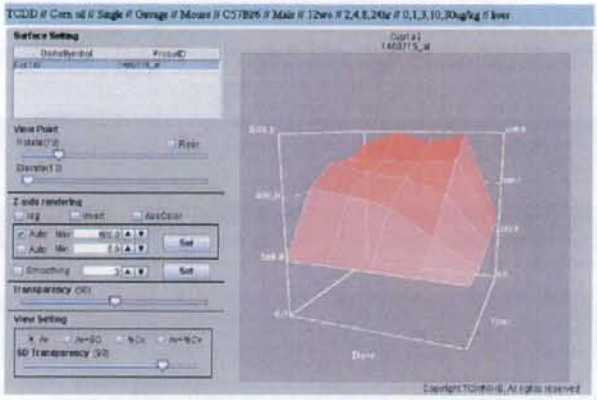
- 現在、誤差項の計算を対数領域で行っており、同一条件内で平均をとると、算術平均よりも小さな値となる。算術平均に置き換えるだけで大きな値側にぶれる。

全プローブ計算③ 検証②

- 基本的確認
  - Percellome論文(BMC Genomics 2006, 7:64 doi:10.1186/1471-2164-7-64)において、QPCR計測を行っている遺伝子で飽和していると考えられる遺伝子が、本方法と、結果が合致しているかを確認する。
- 課題
  - QPCR対象のプローブセットがLangmuirモデルになっているか？
    - Cyp7a1は、Langmuirモデル対象となっている。
    - サーフェースまで描いた遺伝子はLangmuir対象は他になし。
- 懸念事項
  - Percellome自身が飽和の影響を受けていて、高発現域でBiasを発生していないか？
- 解析対象データ
  - QPCRと同じ化合物を用いた実験のCELファイル

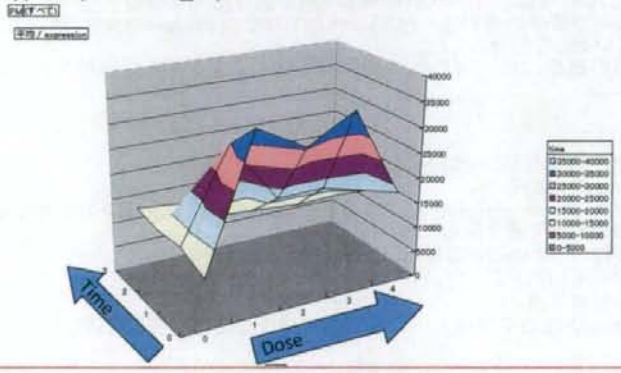
全プローブ計算③検証②Cyp1a2

- Cyp1a2(1450715_at)のTTG20の結果(ToxicOmics Database)



全プローブ計算③検証②Cyp1a2

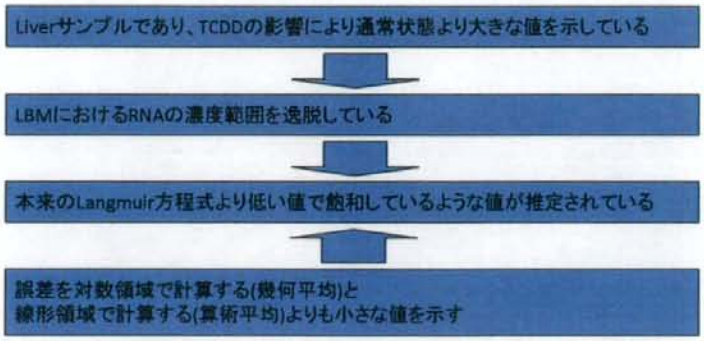
- Cyp1a2(1450715_at)のTTG20の変換結果



前ページの図とは、Dose軸の正負が異なるグラフである。大きな値に対して、小さな値を生み出している

全プローブ計算③見つかった課題：変換範囲

- Cyp1a2において、変換結果として良好な形状を示さなかった。
- 考察



全プローブ計算③見つかった課題：Summarize (Probe ➔ ProbeSet)

- MAS5は、Tukey's biweight algorithmを使用している
- GCOSとSDK(=Expression Console)の間で違う結果となる。
  - Absentで多少の値が違うというレベルではなく、Presentで大きく値がずれているものが存在した。
  - 調査の結果、コンパイラのバージョンが違いによる計算誤差が原因と判明した。
- アルゴリズム
  1. ProbePair値計算:x=PM-MM
  2. Medianを求める
  3. Median absolute deviation:標準偏差計算の中で、2乗計算と平均を絶対値とMedianで置き換えたもの
  4. 標準化：MedianとMADを用いたZScore化:t=(x-M)/(c*MAD)
  5. 重みw=((1-t^2)^2)
  6. 重み付き平均
  - Median近辺の平均値を用いることで、極端な値の影響を排除している
- 今後対処方法を検討する

全プローブ計算④

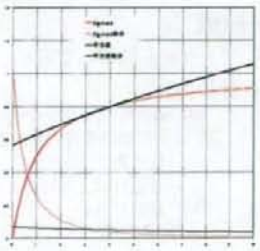
- 数値計算のテクニックを付加することで、安定した補正を行えるか検討する

全プローブ計算④数値計算上の技巧

- 数値計算を行う上で、Langmuirの方程式は上に有界な数式である。次の2点で数値計算上問題がある。
  - 濃度に対する係数に鈍感な関数で、完全飽和に近い状態で適切な収束が得られない。
    - 飽和していない場合には、AICにより、排除される
  - 今回は逆関数にするので、逆関数の際に右に有界で定義域が制限される。
- 高濃度において、平方根にフィットするものとみなす。
  - この領域に入った場合には詳細な検討が必要である。

$$f(x)=\begin{cases}\dfrac{x}{x+1} & for\ \ x\le 4\\ \dfrac{1}{5\sqrt{0.5}}\sqrt{x+4} & for\ \ x\ge 4\end{cases}$$

$$I_M=\begin{cases}I_p\dfrac{k_p c_j}{k_p c_j+1}+bg_p & for\ \ k_p c_j\le 4\\ \dfrac{I_p}{5\sqrt{0.5}}\sqrt{k_p c_j+4}+bg_p & for\ \ k_p c_j\ge 4\end{cases}$$



全プローブ計算④
Teradataによる線形領域モデル計算を用いた絞込

- Teradataを用いた線形計算を用いて、対象プローブの絞り込みを実施
- 次の条件を用いた
  - 上に凸(線形領域2次モデルの2次係数が負)
  - 線形領域線形モデルの傾きが平均値の5%以上
  - 線形領域線形モデルの傾きが標準偏差の5%以上

終了状態	Brain側飽和	Liver側飽和
勾配ベクトル判定収束	101,432	82,443
移動量判定収束	35,281	23,624
局所的最小値	13,051	13,989
繰り返し上限を超えた	5	0
総計	149,769	120,056

AICでLangmuirは選択されない

全プローブ計算④非線形最適化の課題(Langmuirモデル)

- Rによる収束計算における課題について記す
- 収束結果判定
  - 収束判定係数Steptol=1.0e-17として、完全不動の状態でしか収束させないようにする。
- 初期値
  - データから計算して求め、収束しやすいようにする。
  - 計算方法
    - Ipは、最大値と最小値の差を与える
      - 増加の場合、100:0の値に100:0と25:75の差を加えて、最大値のおおよその値を見つける
      - 数値計算上の技巧を付け加えたことにより、若干小さいほうが収束しやすい可能性がある
    - Bgpは、最小値の半分

全プローブ計算④非線形最適化の結果

- Liver側飽和では多くのプローブでLangmuirモデルが採用された。
- Brain側飽和では、Langmuirモデルが採用されたプローブは限られていた

終了状態	Brain側飽和	Brain側採用個数	Liver側飽和	Liver側採用個数
勾配ベクトル判定収束	101,432	364	82,443	29,109
移動量判定収束	35,281	155	23,624	8,390
局所的最小値	13,051	109	13,989	7,775
繰り返し上限を超えた	5	0	0	0
総計	149,769	628	120,056	45,274

全プローブ計算④ Langmuir Model採用個数

- ProbeSet中で何個のProbeがLangmuirモデルを採用しているかを検討した
- Brain側飽和

MM

PM

合計 / Count(*) 行ラベル	列ラベル 0	1	2	11	総計
0		152	3		155
1	255	6			261
2	30				30
3	5	2			7
4	1	1			2
5	3	1	1		5
6	2	3	1		6
11	1				1
20				1	1
総計	297	165	5	1	468

PM側のLangmuirモデル採用個数が多いProbeSetが補正の影響を受けるはずである

全プローブ計算④ Langmuir Model採用個数

- Liver側飽和

MM

PM

合計 / Count(*) 行ラベル	列ラベル 0	1	2	3	4	5	6	7	8	9	10	11	総計
0		3565	191	15									3771
1	3231	930	112	12									4285
2	546	285	55	7									893
3	263	133	52	9	2								459
4	160	101	47	20	7								335
5	120	94	53	26	15	2			1				311
6	57	88	58	32	18	8	1						363
7	34	75	65	56	17	15	2	1					265
8	32	57	45	51	32	19	15	5					256
9	8	31	35	59	49	41	33	24	5				285
10	1	17	16	54	61	45	58	56	40	24	14	3	389
11	2	6	14	18	22	40	75	72	75	85	110	96	615
16								1					1
17											1		1
総計	4454	5382	743	359	223	171	184	159	121	109	125	99	12129

PM側のLangmuirモデル採用個数が多いProbeSetが補正の影響を受けるはずである

全プローブ計算④検証①

- LBMでの線形性チェック
  - 50:50を1に正規化したグラフを作成し、線形性のチェック
  - Langmuir変換を含むプローブセットを全てプロットし、目視で線形性の確認する。値の低いもの以外で変な値を示すものがないことを確認する

全プローブ計算④検証①

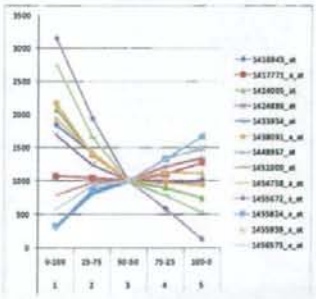
- 対象プローブセット：50%:50%において、1000程度の値を持つ

affy_id	Intensity
1424886_at	1008.53
1433934_at	1004.20
1424005_at	1002.17
1417771_a_at	1005.90
1451000_at	1000.37
1456573_x_at	1005.87
1455824_x_at	1006.07
1416943_at	1007.17
1454758_a_at	1007.60
1438091_a_at	1008.33
1448967_at	1003.97
1455939_x_at	1006.53
1455672_s_at	1004.00

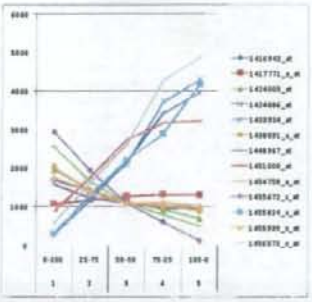
全プローブ計算④検証①

- 対象プローブセット：50%:50%において、1000程度の値を持つ

補正前



補正後

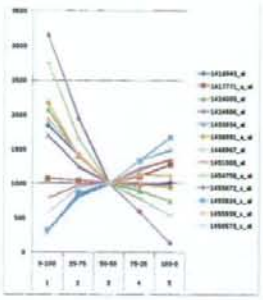


Liver側飽和のプローブは直線に引き延ばされ、高い値で、不安定になっている。

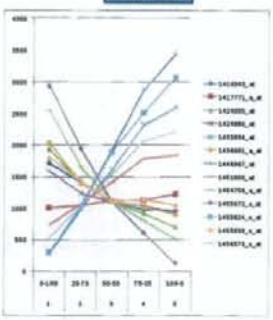
全プローブ計算④検証① 高値補正付き

- 対象プローブセット：50%:50%において、1000程度の値を持つ

補正前



補正後

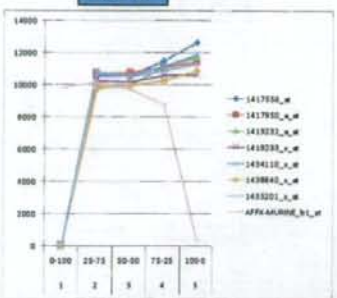


高値を補正する式にすると若干安定する

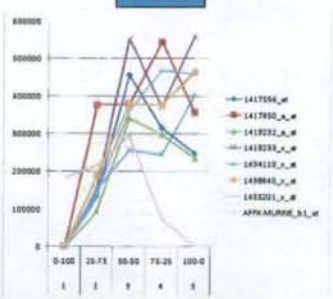
全プローブ計算④検証①

- 対象プローブセット：50%:50%において、10000以上の値を持つ

補正前



補正後

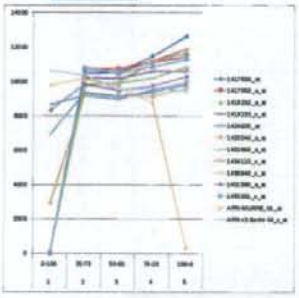


0%から50%まではすこし上向きだが直線に見えなくもない。しかし、50%を超えると、崩れている。Lngmuirの最大値推定に問題がありそうである

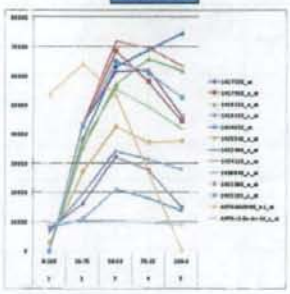
全プローブ計算④検証① 高値補正付き

- 対象プローブセット：50%:50%において、10000以上の値を持つ

補正前



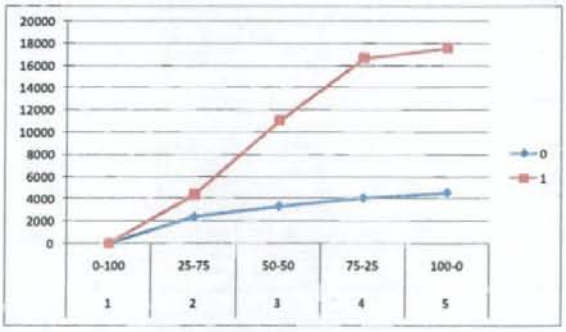
補正後



高値に対する補正を行う変形をすると、若干安定する

全プローブ計算④検証① 高値補正付

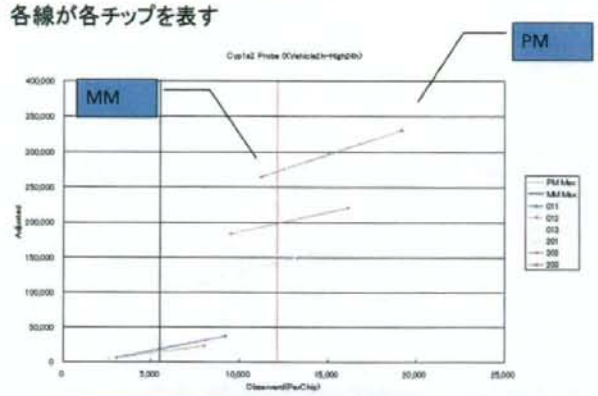
- Cyp1a2の結果



高値に対する補正を行う変形をすると、若干安定する

全プローブ計算④検証②プローブ単位

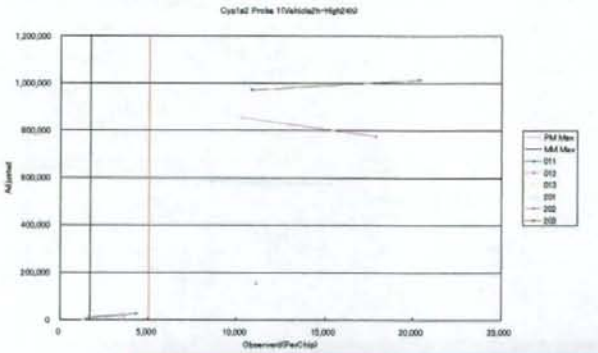
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

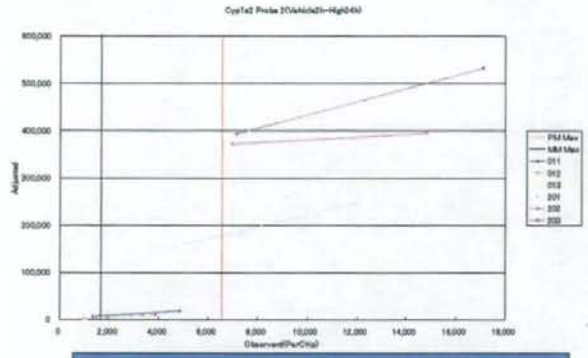
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。
補正後にPM/MMの大小関係が逆転しているペアが存在する

全プローブ計算④検証②プローブ単位

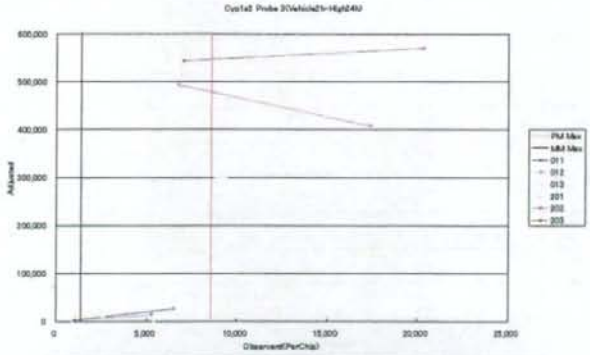
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

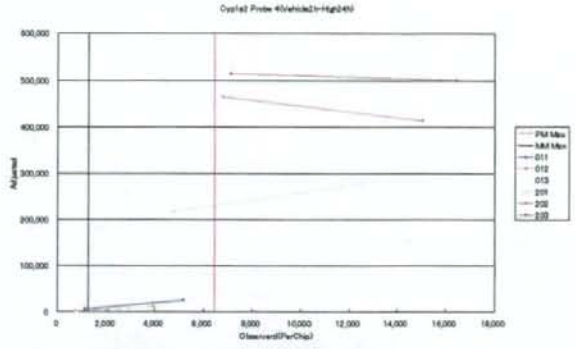
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

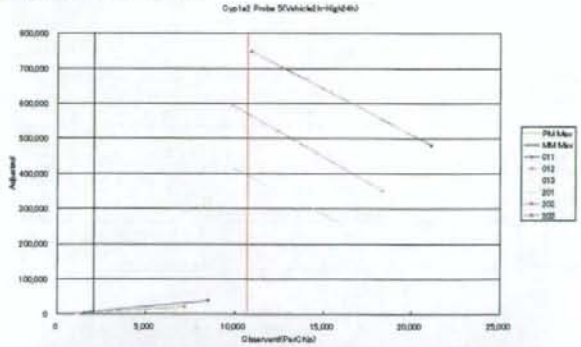
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

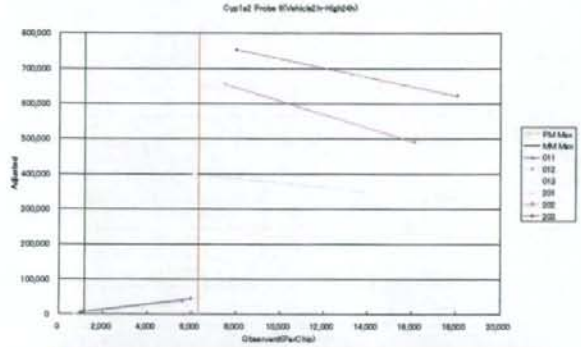
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



PM－MMで値が逆転しており、結果を不安定にしている

全プローブ計算④検証②プローブ単位

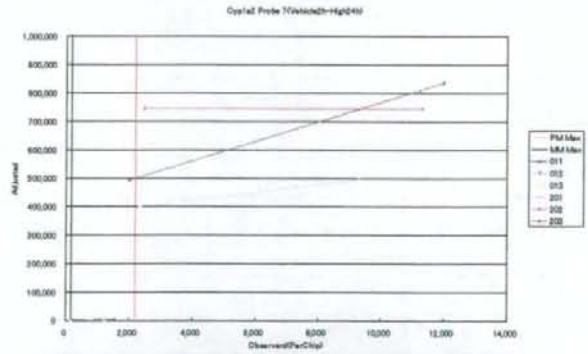
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

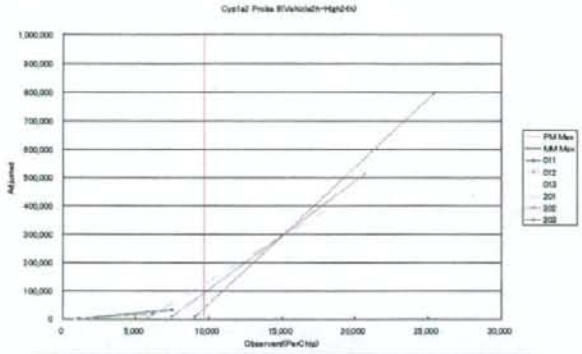
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

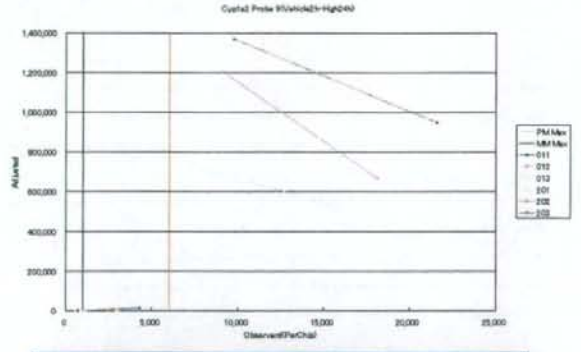
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



MMはLangmuirが選択されていない

全プローブ計算④検証②プローブ単位

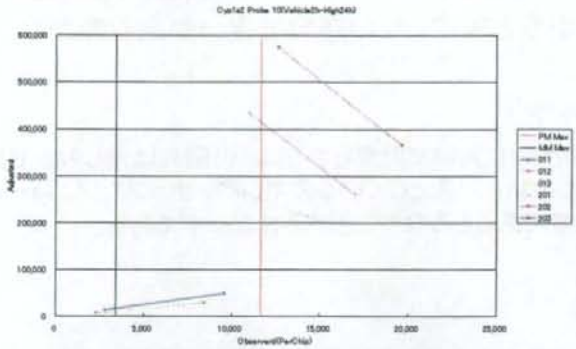
- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④検証②プローブ単位

- Cyp1a2のプローブ単位でチェック
  - 各線が各チップを表す



計測値はLangmuir曲線の最大値を大きく上回っている。

全プローブ計算④課題：変換範囲

- Cyp1a2において、変換結果として良好な形状を示さなかった。
- 考察

Liverサンプルであり、TCDDの影響により通常状態より大きな値を示している

↓

LBMにおけるRNAの濃度範囲を逸脱している

↓

本来のLangmuir方程式より低い値で飽和しているような値が推定されている

↑

誤差を対数領域で計算する(幾何平均)と
線形領域で計算する(算術平均)よりも小さな値を示す

全プローブ計算⑤

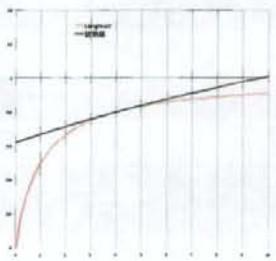
- 濃度平衡係数がPM/MMにおいて同じであると仮定する場合と仮定しない場合で違いがあるのか？
- アイデア
  - PM/MMにおいて対象とするRNAの配列は同じものである。同じRNAを対象としているので、同一チップでは、同一の濃度であるとみなすことが可能なはずである。

全プローブ計算⑤数値計算上の技巧

- 数値計算を行う上で、Langmuirの方程式は上に有界な数式である。次の2点で数値計算上問題がある。
  - 濃度に対する係数に鈍感な関数で、完全飽和に近い状態で適切な収束が得られない。
    - 飽和していない場合には、AICにより、排除される
  - 今回は逆関数にするので、逆関数の際に右に有界で定義域が制限される。
- 高濃度において、平方根にフィットするものとみなす。
  - この領域に入った場合には詳細な検討が必要である。

$$f(x)=\begin{cases}\dfrac{x}{x+1} & for\quad x\le 4\\ \sqrt{\dfrac{8}{125}}\sqrt{x+6} & for\quad x\ge 4\end{cases}$$

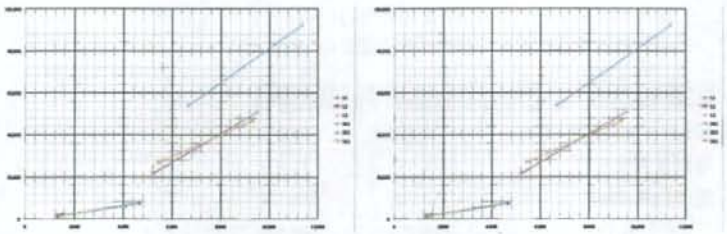
$$I_{pi}=\begin{cases}I_p\dfrac{k_p c_j}{k_p c_j+1}+bg_p & for\quad k_p c_j\le 4\\ I_p\sqrt{\dfrac{8}{125}}\sqrt{k_p c_j+6}+bg_p & for\quad k_p c_j\ge 4\end{cases}$$



全プローブ計算⑤結果

濃度平衡係数がPM/MMにおいて同じであると仮定する場合と仮定しない場合(Cyp1a2 Probe0)

PM/MM濃度連動モデル　　**PM/MM濃度独立モデル**



係数は若干異なるが変換結果に大きな影響はない

全プローブ計算⑥

- 最大値を固定する方法
  - 最大値をLBMだけではなく、別の実験から求めて使用する
- アイデア
  - 最大値は、次の欠点が考えられる
    - 観測誤差の影響を受けやすい。
    - 似た別の配列が存在する場合に、Cross-Hybridizationにより、Langmuir方程式の仮定を崩す状況が考えられる
  - 現在まで非常に多くの実験を行ってきたこの中で最大値を飽和値とみなす。
    - 今回は、ITG20の最大値を飽和値の95%とみなす
    - 将来的には、全チップの上位10%の+5σを最大値とみなすようなことを考える