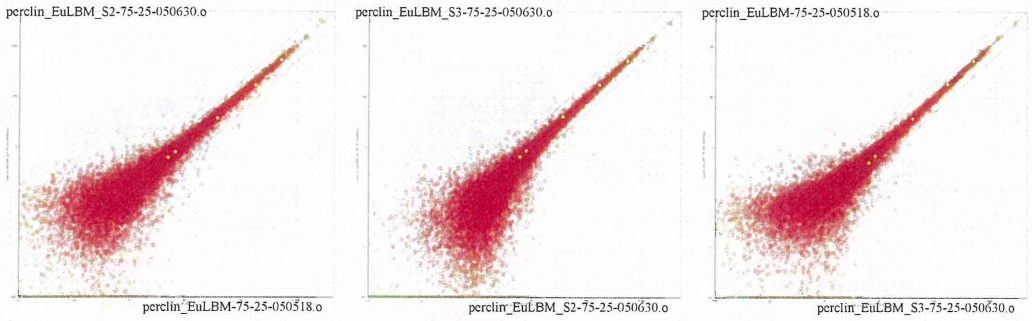
7.背景ターゲットによるアプローチ
散布図による確認

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

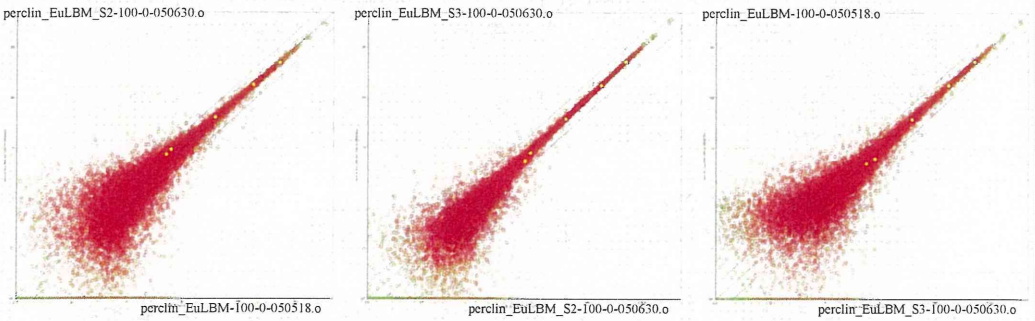


EuLBM_S3-75-25-050630が最も低く出ている。その差は1コピー以下である

7.背景ターゲットによるアプローチ
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



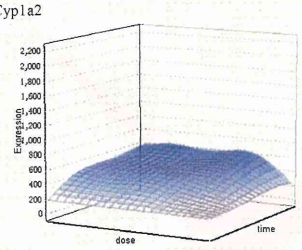
EuLBM_S3-25-75-050518が小さく、EuLBM-100-0-050518が最も大きいが、その差は、1コピー以下である

7.背景ターゲットによるアプローチ
TTG020(TCDD処置)補正結果

TTG020(TCDD処置の結果を用いて、実データに対する補正の精度を確認した。

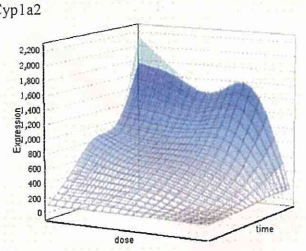
MAS5、MLANG、QPCRによる測定結果を比較した。

TTG020-L_SpNC_0 450715_at
Cyp1a2



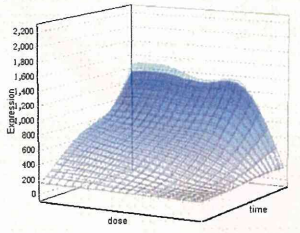
MAS5

perclin_TTG020-L_20110126
Cyp1a2



MLANG

TTG020-L_QPCR_SpNC_0
Cyp1a2



QPCR

MAS5では飽和により見えなかった値を含め再現されている。

8.まとめ

- べき乗分布で光学的背景レベルを導き出すのは、数値計算上の安定性に問題があった。
- 一定比で添加しているRNAを用いて、背景レベルを導き出すのは、最低濃度が高めに出てしまい、直線性が失われて、安定性が低下していると考えられた。
- 仮想的に全プローブと弱いハイブリダイゼーションが発生する背景となるターゲットRNAを導入することで、低発現の異常な変動を抑えることができた。
  - 肝臓、脳で、半数のプローブセットが0と推定された。
  - 三重化された実験において、偏差は1コピー以内に抑えられた。
- 試行した三手法で、背景ターゲットを導入する方法が、安定した値を算出することができた。

国立医薬品食品衛生研究所　毒性部　御中

委託研究報告書（STEP10）

次世代シークエンサによるRNA定量の数値化アルゴリズムに関する業務コンサルティング

平成24年2月24日
株式会社NTTデータ

テーマ

大量の核酸配列解読性能を持つ次世代シークエンサを応用し、エクソンレベルの遺伝子発現情報や非翻訳領域の情報など、マイクロアレイでは高精度に測定することが難しかった対象を含む、網羅的定量解析技術（RNA-Seq）が実用化されつつあるが、トキシコゲノミクスへの応用の可能性を評価するための予備的な検討では、数値化アルゴリズムの性能が不十分であることが判明している。本業務では、より高精度の網羅的遺伝子発現解析を行うための基盤技術としてRNA-Seqの可能性を探り、現行の数値化アルゴリズムの問題点を検討する。

1.シークエンスデータ

- 肝臓(Liver)と脳(Brain)を、細胞数をベースに、5パターン(肝臓100%、肝臓75%:脳25%、肝臓50%:脳50%、肝臓25%:脳75%、脳100%)の割合で混合したサンプル(Liver-Brain-Mixture)を、次世代シークエンサ(イルミナ社Genome Analyzer IIx)で読み込んだ。
- これらのシークエンスデータを用いて、RNA-Seqの可能性を探る

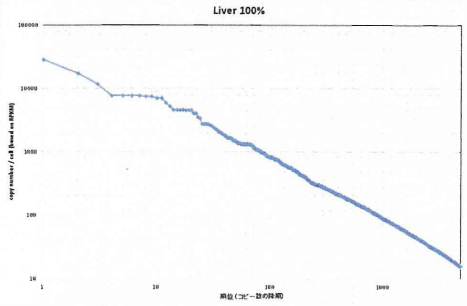
2.既存のRNA−Seq数値化アルゴリズムの性能検討

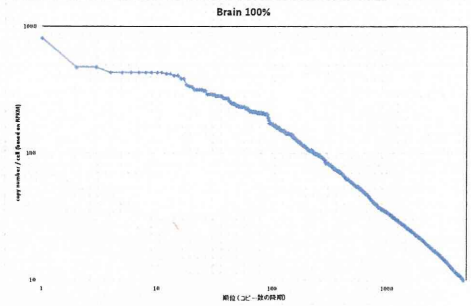
- 次世代シークエンスデータからRNAの量を求める方法が2008年に発表されている。この方法では、RPKMという単位を用いる
- RPKMとは、読み込んだタグ配列が、その遺伝子に割り振られた塩基数の合計を、遺伝子配列の塩基配列長で除した値の1000分の1である。
- このRPKMを用いて、既存の割り振り方法がよいのか検討する

RPKM = reads per kilobase of exon per million mapped sequence reads

2.1.次世代シークエンサによる読込結果の傾向把握

細胞あたりコピー数を多い順に遺伝子(Affymetrix Moe430v2 Probesetに変換)を並べた。

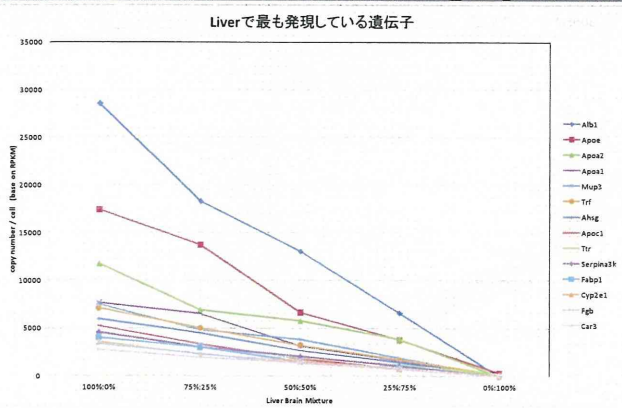




自然発生的に出来上がったものは、その数がべき乗法則(両対数グラフ上で直線となる)に従うことが多い。本結果も100位までは、同一遺伝子を複数プローブセットで表現しているなどの理由があり、従っていないが、それ以降はべき乗法則にしたがっていると考えられる

2.2.RPKM直線性比較

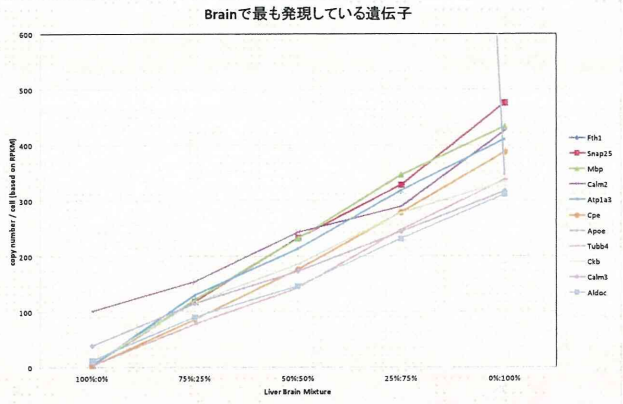
肝臓で多く発現している順に遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。

2.2.RPKM直線性比較

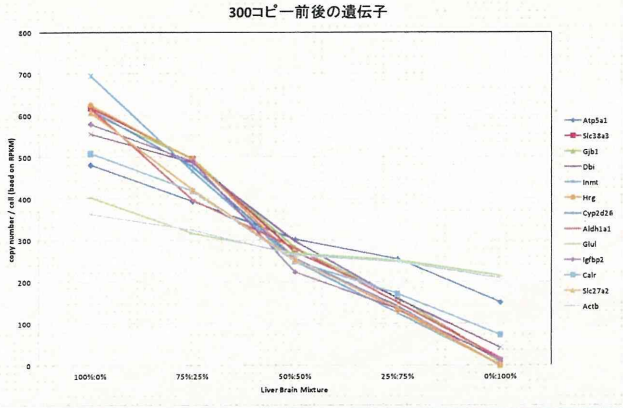
脳で多く発現している順に遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。
Apoeは、肝臓と脳の両方で大量に発現している。細胞あたりでは肝臓のほうが多いため、脳100%で、一部しか見えていない

2.2.RPKM直線性比較

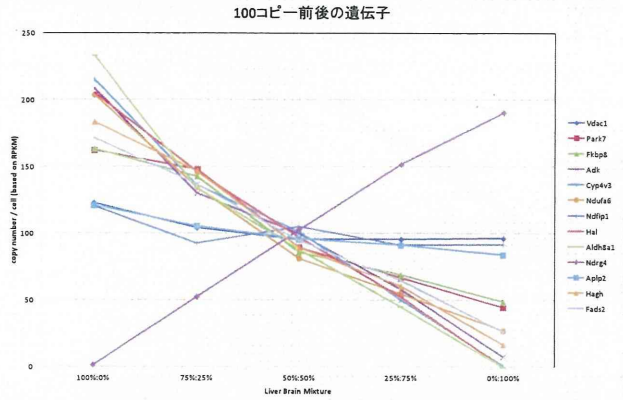
平均で300コピー前後発現している遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。

2.2.RPKM直線性比較

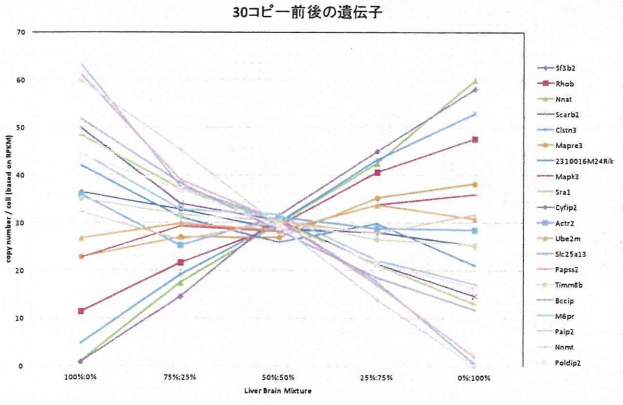
平均で100コピー前後発現している遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。

2.2.RPKM直線性比較

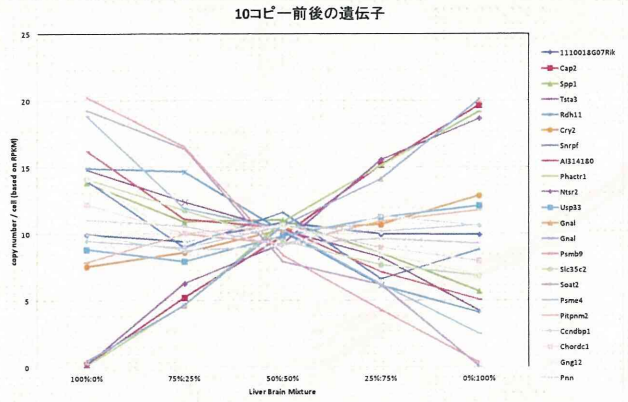
平均で30コピー前後発現している遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。

2.2.RPKM直線性比較

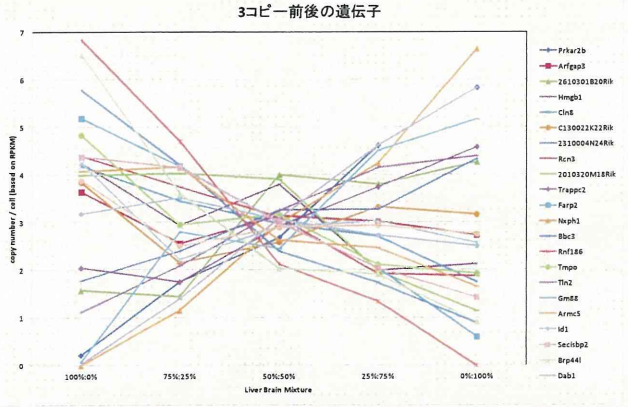
平均で10コピー前後発現している遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。

2.2.RPKM直線性比較

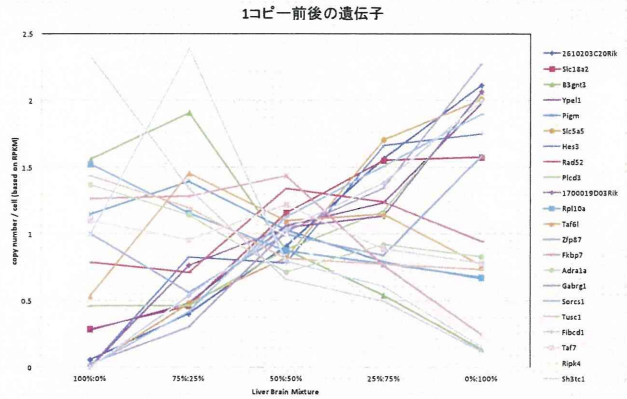
平均で3コピー前後発現している遺伝子を選び直線性を確認した



多少の誤差はあるが、線形と考えられる。

2.2.RPKM直線性比較

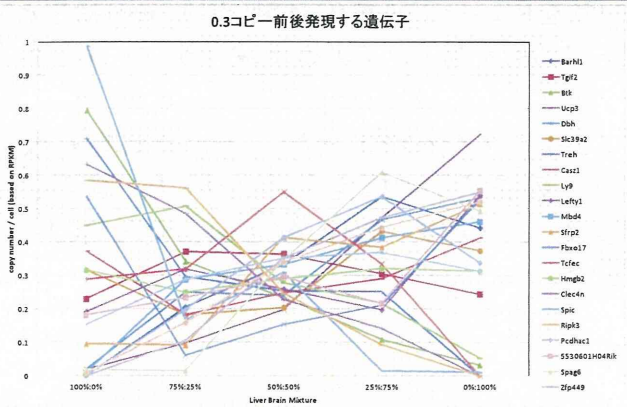
平均で1コピー前後発現している遺伝子を選び直線性を確認した



誤差は大きいが、線形性は認められる。

2.2.RPKM直線性比較

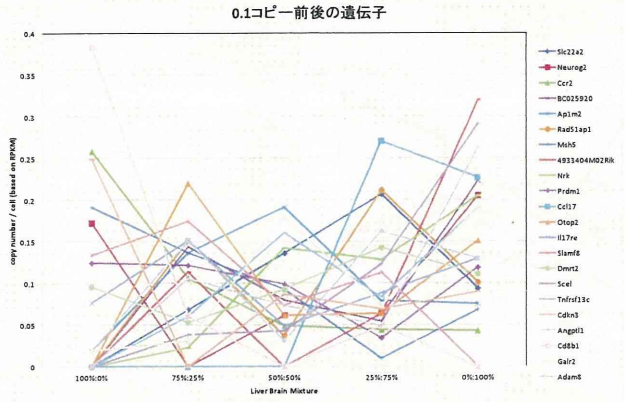
平均で0.3コピー前後発現している遺伝子を選び直線性を確認した



誤差は大きいが、線形性は認められる。

2.2.RPKM直線性比較

平均で0.1コピー前後発現している遺伝子を選び直線性を確認した



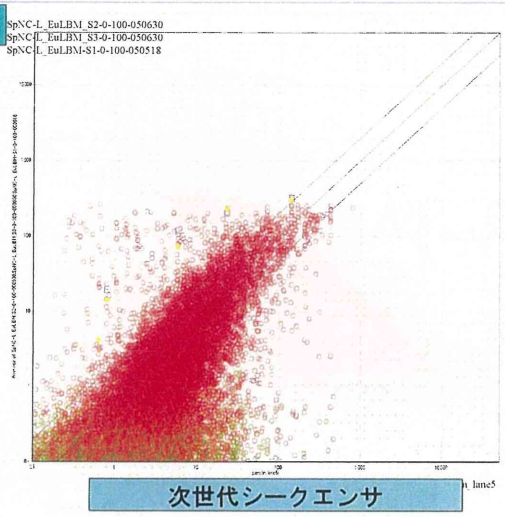
誤差は大きいが、線形性は認められる。

2.3.次世代シークエンサとマイクロアレイの結果比較

- **次世代シークエンサを用いたRNA量の計測(RNA-Seq)結果とマイクロアレイによる計測結果(MAS5による正規化を実施)の比較を実施した。**
- **横軸をRNA-Seq、縦軸をマイクロアレイとし、それぞれを対数軸で表わしたグラフ上に、各遺伝子の推定をプロットした。適切な結果を出力されていれば、その値は一致し、散布図上で対角線にプロットされるはずである。**

2.3.次世代シークエンサとマイクロアレイの結果比較

Brain100%



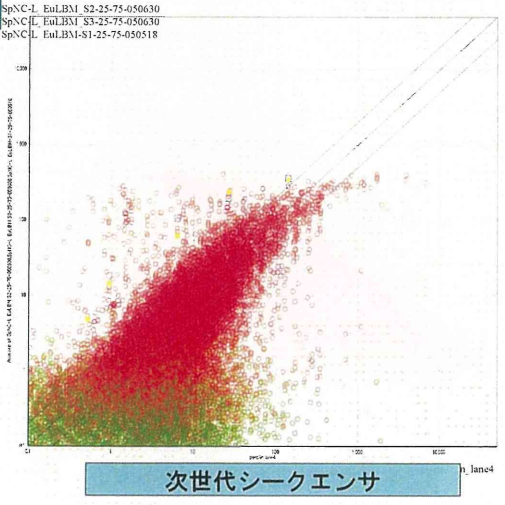
マイクロアレイ
MAS5

次世代シークエンサ

対角線より少しずれている

2.3.次世代シークエンサとマイクロアレイの結果比較

25%:75%



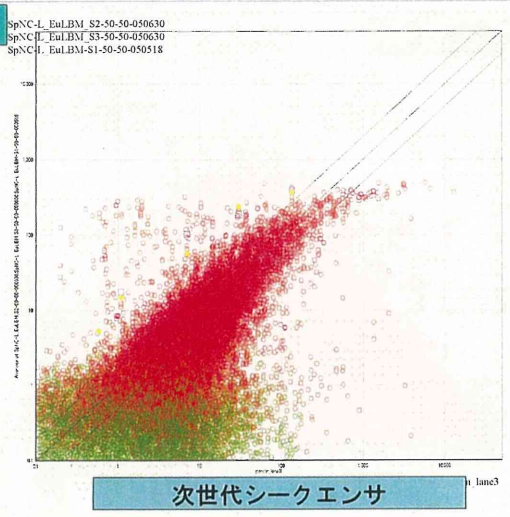
マイクロアレイ
MAS5

次世代シークエンサ

対角線より少しずれている

2.3.次世代シークエンサとマイクロアレイの結果比較

50%:50%



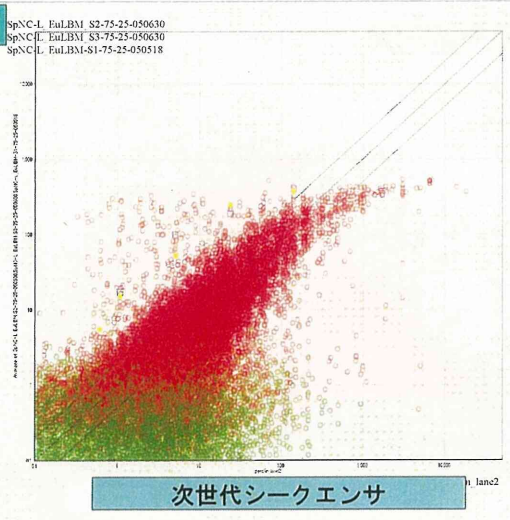
マイクロアレイ
MAS5

次世代シークエンサ

対角線より少しずれている

2.3.次世代シークエンサとマイクロアレイの結果比較

75%:25%



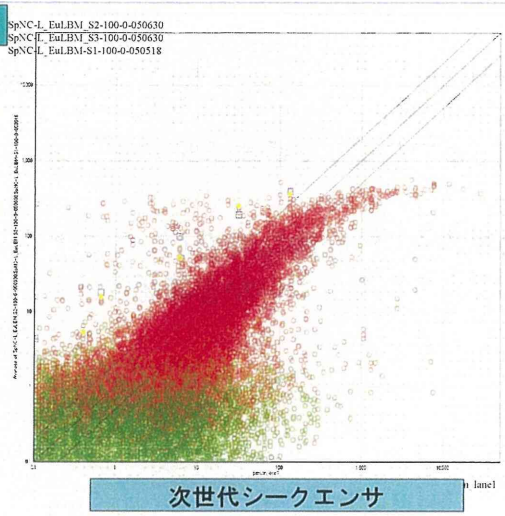
マイクロアレイ
MAS5

次世代シークエンサ

対角線より少しずれている

2.3.次世代シークエンサとマイクロアレイの結果比較

Liver100%



マイクロアレイ
MAS5

次世代シークエンサ

対角線より少しずれている

2.3.次世代シークエンサとマイクロアレイの結果比較　まとめ

- 全体を通して次のような現象がみられた

- マイクロアレイ(MAS5)では、測定値は数百コピー程度で頭打ちとなる。
  - プローブへの吸着を原理とするため、飽和現象が発生していると考えられる。
- マイクロアレイ(MAS5)は、低発現領域で、次世代シークエンサより大きな値を示す。
  - MAS5は計算の中で、PMよりもMMが明るいプローブペアを捨てており、偶然小さくなった場合も捨てている。このため低発現で大きめに計算されていると考えられる。
- 対角線から大きく外れた遺伝子が存在した。
  - マイクロアレイも次世代シークエンサも鋳型となる遺伝子配列を基準としており、これらの配列に誤りやSNPなどの問題があった場合には、誤った値を算出する。次世代シークエンサは全遺伝子情報を用いるため、エラー発生の可能性は大きい。

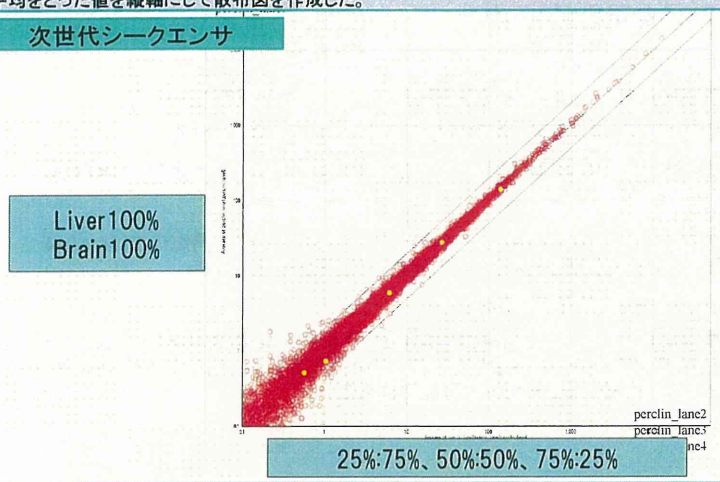
2.4.試験管内混合と数値的混合の比較

- Liver-Brain-Mixtureは、肝臓と脳の試料を試験管内(in-vitro)で、混合している。その混合が、他の要因による影響を受けていないのであれば、数値的な混合(平均処理/in-silico)と一致するはずである。
- 試験管内での混合と、数値的な混合(平均処理)の比較を実施し、混合以外の影響を受けていないかを確認する

2.4.試験管内混合と数値的混合の比較

試験管内で混合した25%:75%、50%:50%、75%:25%を数値的に平均をとった値を横軸に、Liver100%とBrain100%を数値的に平均をとった値を縦軸にして散布図を作成した。

次世代シークエンサ



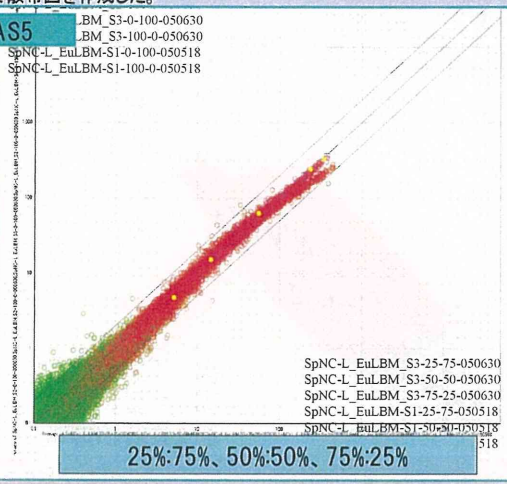
25%:75%、50%:50%、75%:25%

高発現から低発現までほぼ対角線上にある

2.4.試験管内混合と数値的混合の比較

試験管内で混合した25%:75%、50%:50%、75%:25%を数値的に平均をとった値を横軸に、Liver100%とBrain100%を数値的に平均をとった値を縦軸にして散布図を作成した。

マイクロアレイ MAS5



Liver100%
Brain100%

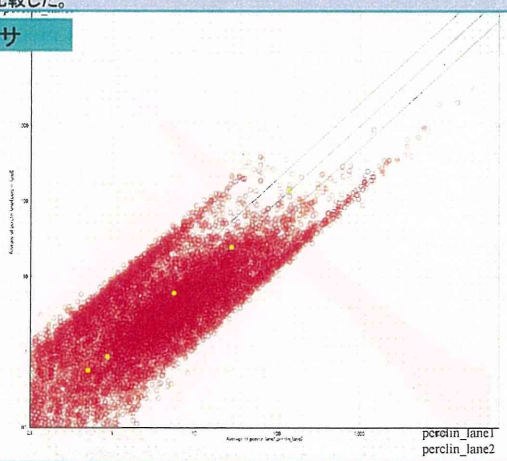
25%:75%、50%:50%、75%:25%

高発現において2分岐している

2.4.試験管内混合と数値的混合の比較

試験管内で混合した75%:25%とLiver100%を数値的に平均化した値を横軸に、25%:75%とBrain100%を数値的に平均化したものを縦軸にして散布図で比較した。

次世代シークエンサ

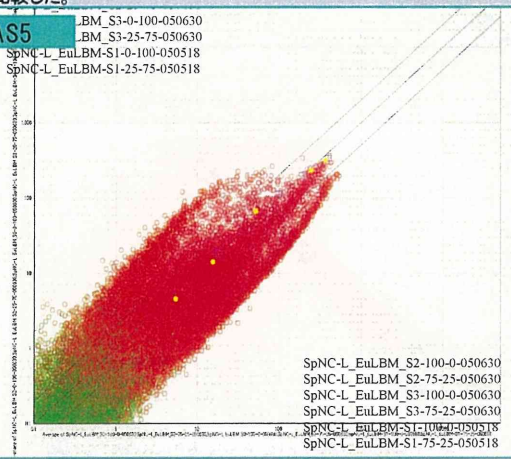


数値的に混合することにより、擬似的に、87.5%:12.5%と12.5%:87.5%のサンプルの比較を実施していることになる。対角線から8倍離れた位置にある遺伝子は、一方の臓器だけで発現していると考えられる。また、高発現の端の線が直線であり、線形性に優れている

2.4.試験管内混合と数値的混合の比較

試験管内で混合した75%:25%とLiver100%を数値的に平均化した値を横軸に、25%:75%とBrain100%を数値的に平均化したものを縦軸にして散布図で比較した。

マイクロアレイ　MAS5



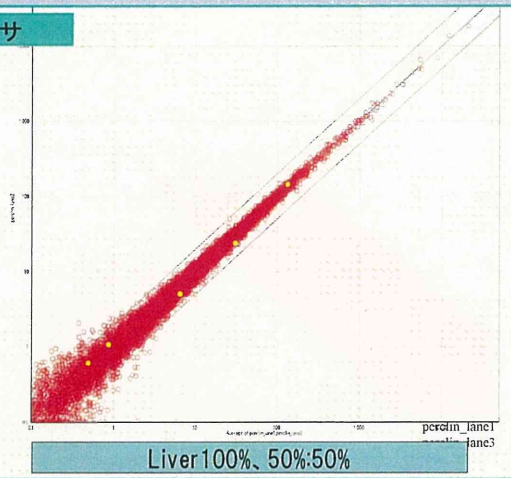
数値的に混合することにより、擬似的に、87.5%:12.5%と12.5%:87.5%のサンプルの比較を実施していることになる。対角線から8倍離れた位置にある遺伝子は、一方の臓器だけで発現していると考えられる。高発現の端の線が中央に向かっており、飽和していることが読み取れる

2.4.試験管内混合と数値的混合の比較

試験管内で混合したLiver100%と50%:50%を数値的に平均をとった値を縦軸に、75%:25%の値を横軸にして散布図を作成した。

次世代シークエンサ

75%:25%



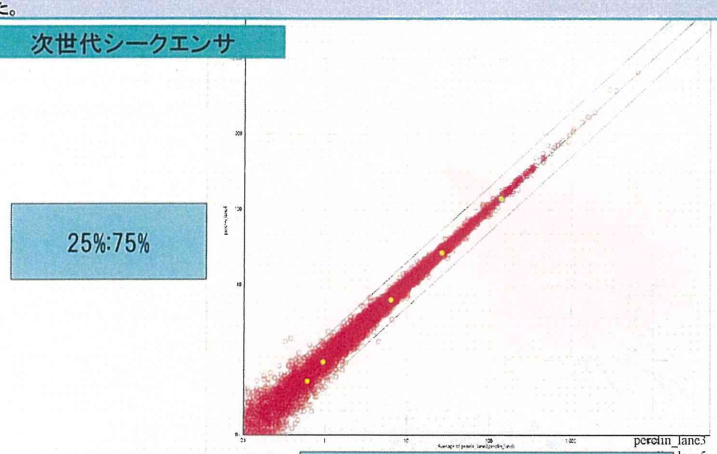
Liver100%、50%:50%

高発現から低発現まで対角線上にある

2.4.試験管内混合と数値的混合の比較

25%:75%の値を縦軸に、試験管内で混合したBrain100%と50%:50%を数値的に平均をとった値を横軸にして散布図を作成した。

次世代シークエンサ

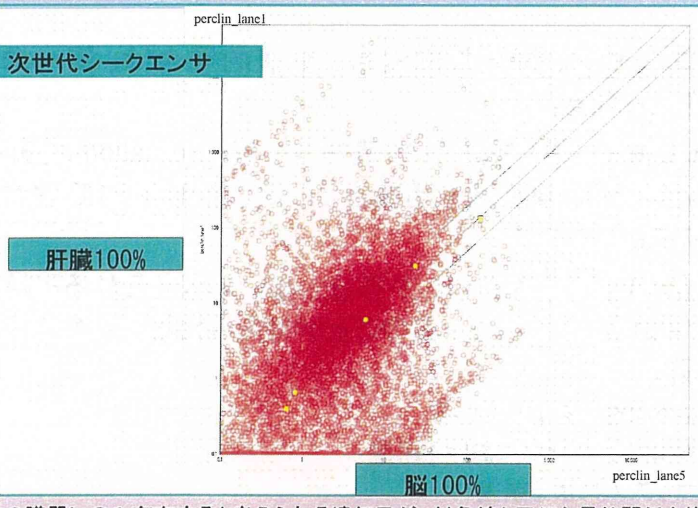


高発現から低発現まで対角線上にある

2.4.試験管内混合と数値的混合の比較

Liver100%とBrain100%を散布図で比較した。

次世代シークエンサ

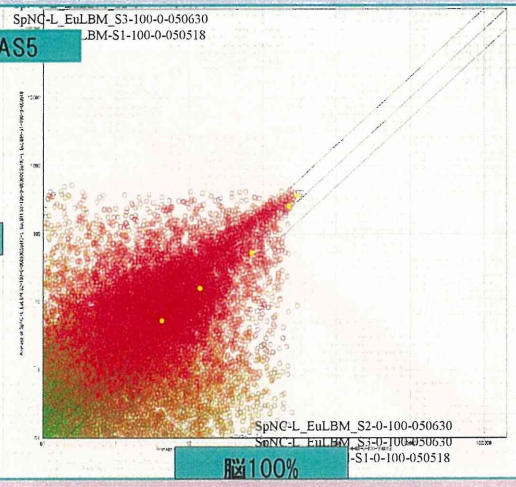


片方の臓器にのみ存在すると考えられる遺伝子が、対角線と平行な量的関係を伴っていると見える。これは、何らかの混入と考えられる。脳で多いもので二桁以上、肝臓で多い遺伝子では四桁近いので、大きな問題となってこなかったと思われる。発生原因の究明は次の課題となると考えられる。

2.4.試験管内混合と数値的混合の比較

Liver100%とBrain100%を散布図で比較した。

マイクロアレイ MAS5



マイクロアレイでは飽和が発生し、四角く区切られており、シーケンサで発生した対角と平行に存在した形状が発生するようなダイナミックレンジが存在しない。

2.4.試験管内混合と数値的混合のまとめ

- 次世代シークエンサは、試験管内混合と数値的混合は、同一の価値をもつと考えられる。
- 線形性に優れており、足し算や平均処理などを、Percellomeと組み合わせることにより、細胞数を基準とした足し算や平均処理が可能と考えられる。
- しかしながら、肝臓と脳で片方にしか存在しないと考えられるRNAが、もう一方でもカウントされた。次のことが考えられる。今後の調査が必要である。
  - 検体がそのような性質を有している
  - 次世代シークエンサの計測における誤差
  - 数値化アルゴリズムによる誤差

3.次世代シークエンサのアライメント用代表的アルゴリズム

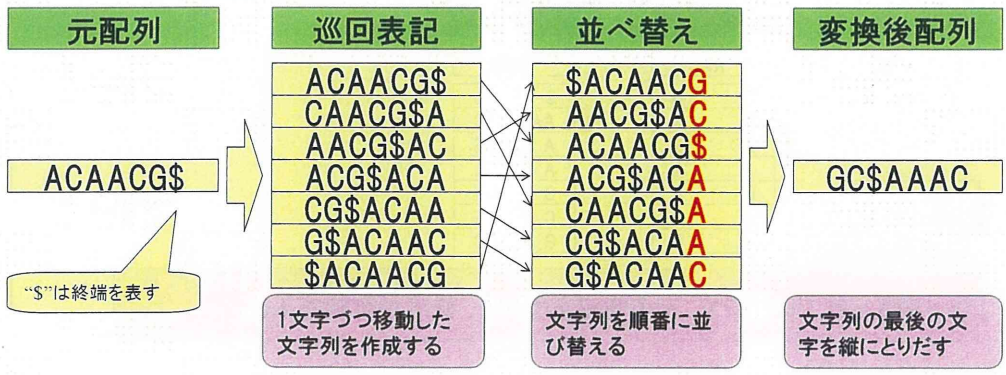
現在までに、次世代シークエンサ向けのアルゴリズムが発表されている。代表的なアルゴリズムとして次のようなものが発表されている。

第1世代	Hash-table-base	
	Eland	Cox 2007
	RMAP	Smith et al. 2008
	MAQ	Li et al. 2008a
	ZOOM	Lin et al. 2008
	SeqMap	Jiang and Wong 2008
	CloudBurst	Schatz 2009
	SHRiMP	http://compbio.cs.toronto.edu/shrimp
第2世代	BWT(Burrows-Wheeler Transform:1994)-base	
	SOAPv2	Ruiqiang Li et al. 2009
	Bowtie	Langmead et al. 2009
	BWA	Heng Li and Richard Durbin 2009

3.1.Burrows-Wheeler Transformの原理①

第2世代のアライメント用アルゴリズムとして用いられているBurrrows-Wheeler Transform(BWT:バーローウィーラー変換)がどのようなものかまとめる。

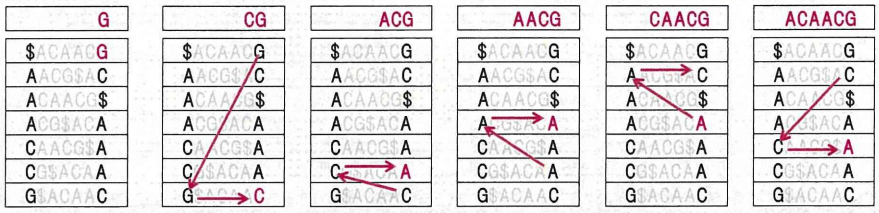
変換の方法



3.2.Burrows-Wheeler Transformの原理②

元文字列の復元

元の文字列を復元する。文字列の後ろから一文字づつ復元する



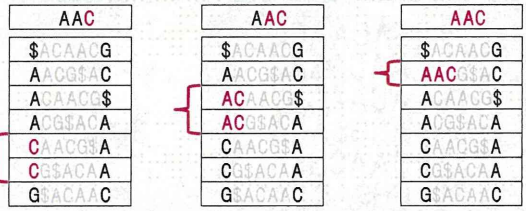
最初の縦列と最後の縦列を用いることにより、文字列の一致・検索処理が可能である

3.3.Burrows-Wheeler Transformの原理③

部分文字列の検索

部分文字列"AAC"を検索する。

後ろの1文字　後ろの2文字　後ろの3文字



並べ替えられているので、その情報を用いて、検索が可能である。