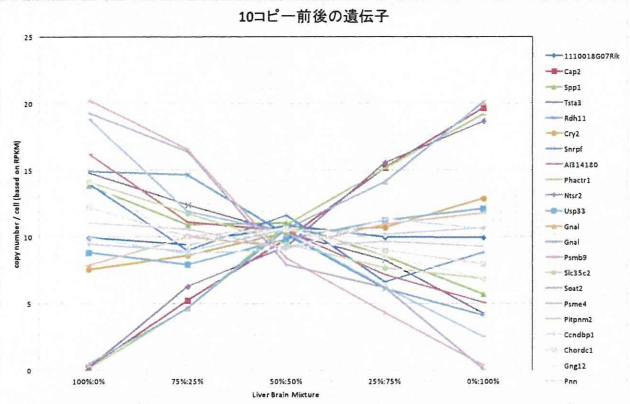
2.2.RPKM直線性比較

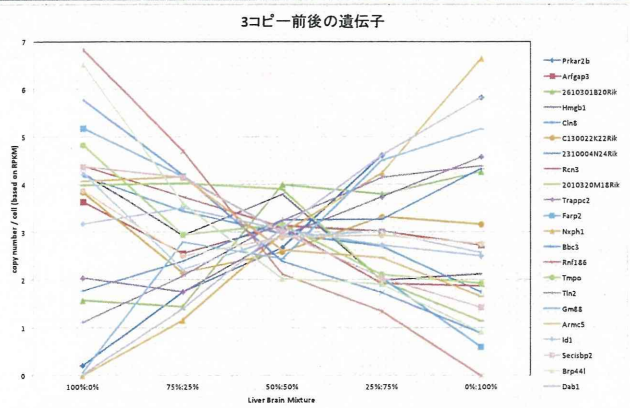
平均で10コピー前後発現している遺伝子を選び直線性を確認した



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

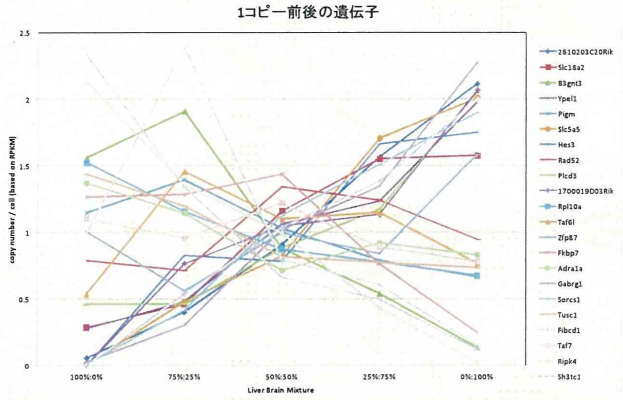
平均で3コピー前後発現している遺伝子を選び直線性を確認した



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

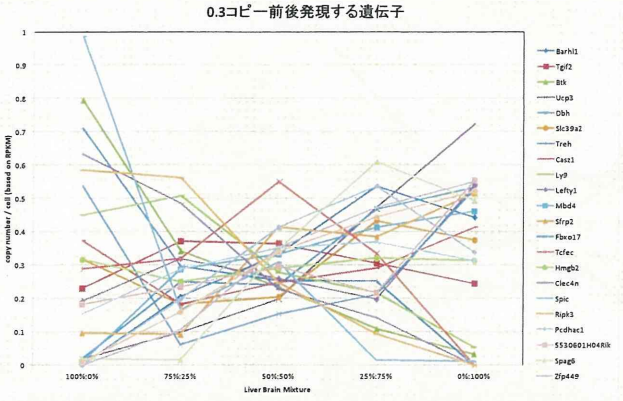
平均で1コピー前後発現している遺伝子を選び直線性を確認した



誤差は大きいが、線形性は認められる。

2.2.RPKM直線性比較

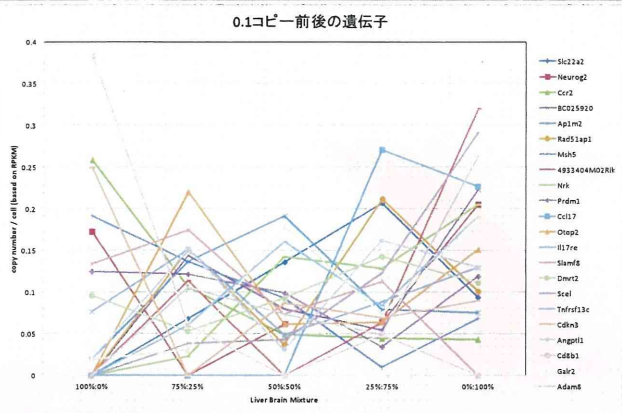
平均で0.3コピー前後発現している遺伝子を選び直線性を確認した



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。

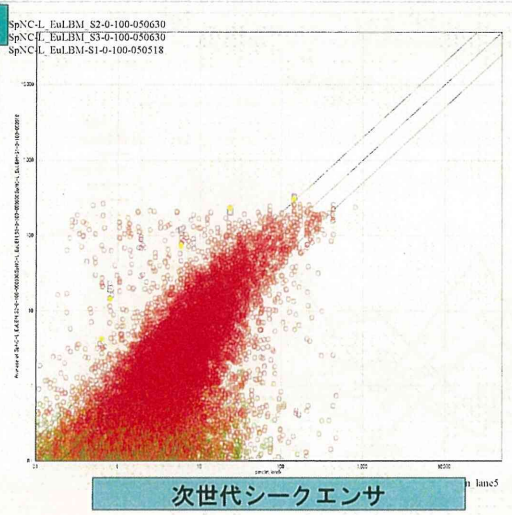
2.3.次世代シークエンサとマイクロアレイの結果比較

- **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



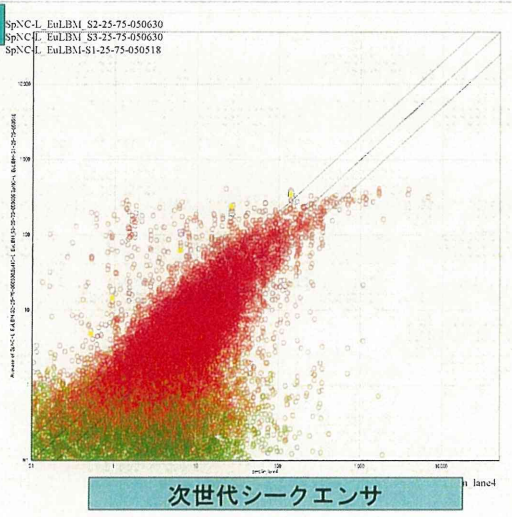
次世代シークエンサ

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

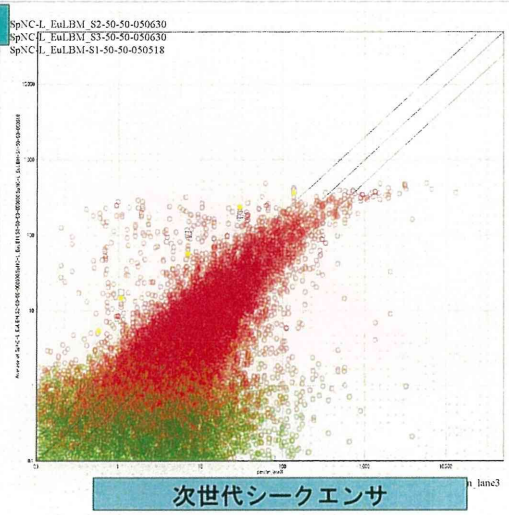


次世代シークエンサ

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%



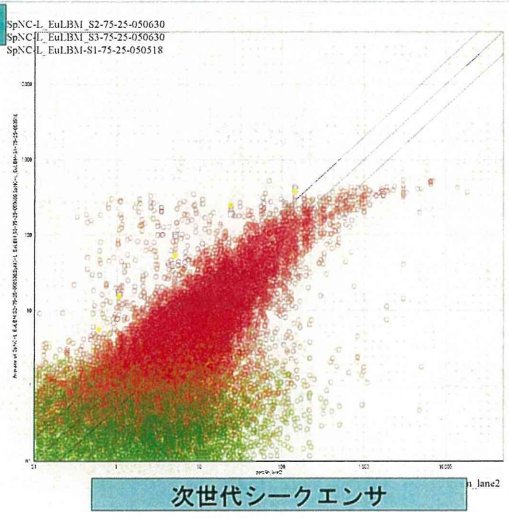
マイクロアレイ
MAS5

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

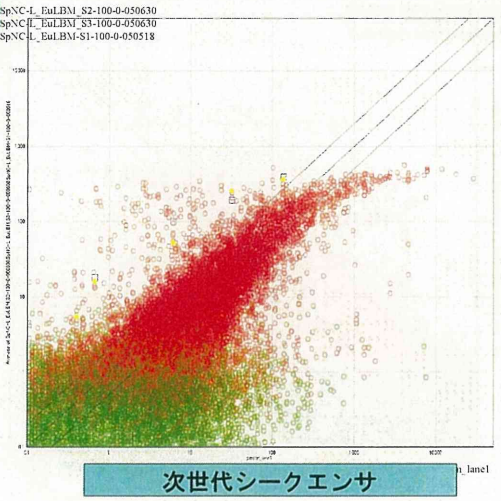
次世代シークエンサ

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

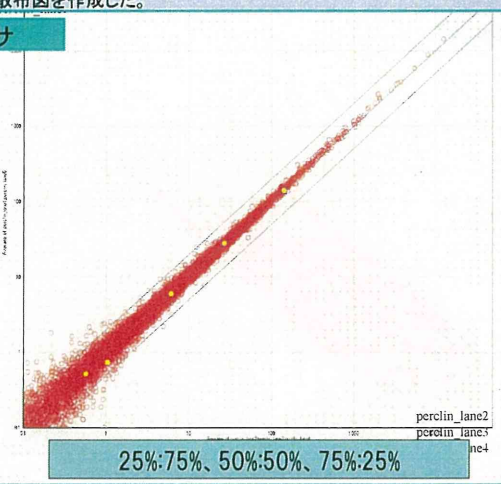
- Liver-Brain-Mixtureは、肝臓と脳の試料を試験管内(in-vitro)で、混合している。その混合が、他の要因による影響を受けていないのであれば、数値的な混合(平均処理/in-silico)と一致するはずである。
- 試験管内での混合と、数値的な混合(平均処理)の比較を実施し、混合以外の影響を受けていないかを確認する

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

Liver100%
Brain100%



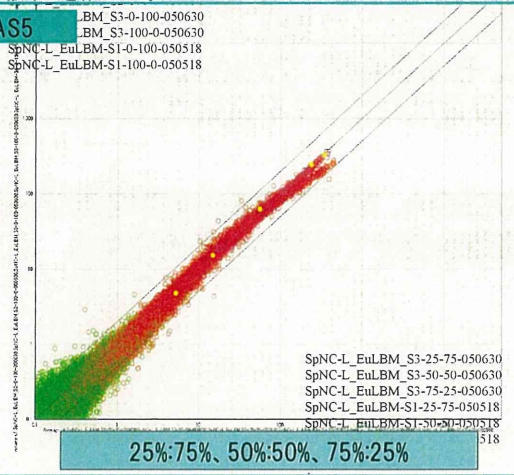
25%:75%、50%:50%、75%:25%

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%

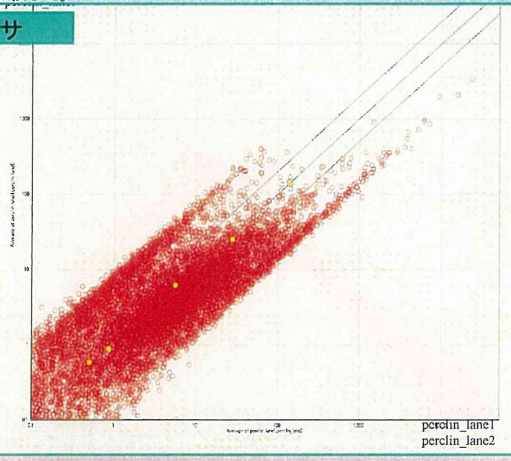
高発現において2分岐している

24

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



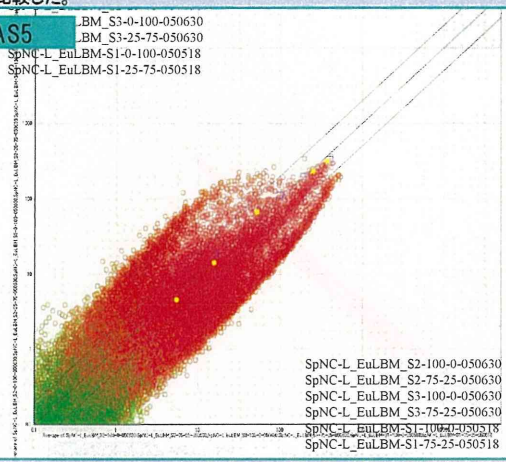
数値的に混合することにより、擬似的に、87.5%:12.5%と12.5%:87.5%のサンプルの比較を実施していることになる。対角線から8倍離れた位置にある遺伝子は、一方の臓器だけで発現していると考えられる。また、高発現の端の線が直線であり、線形性に優れている

25

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

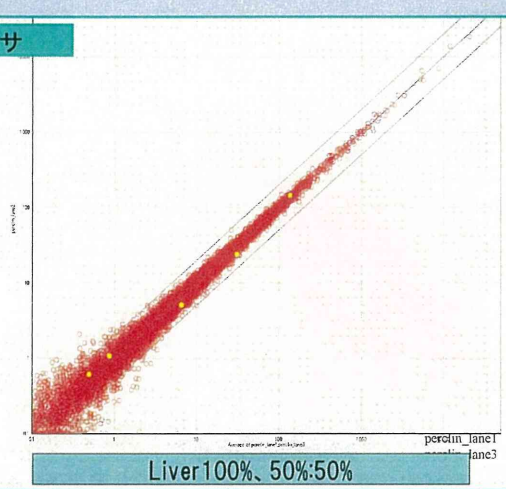
26

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%

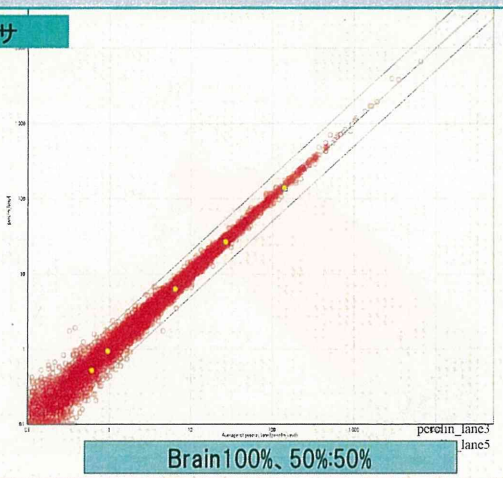
高発現から低発現まで対角線上にある

27

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



25%:75%

Brain100%、50%:50%

高発現から低発現まで対角線上にある

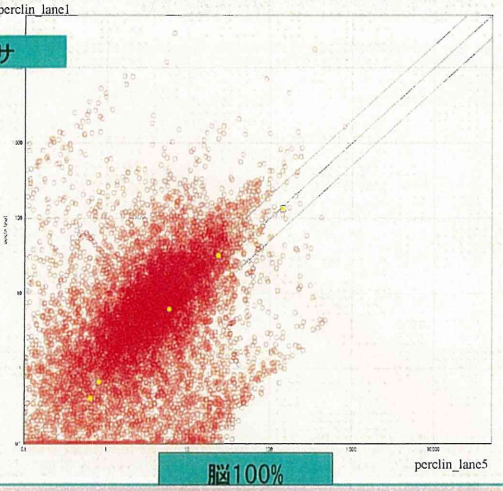
28

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。

perclin_lane1

次世代シークエンサ



肝臓100%

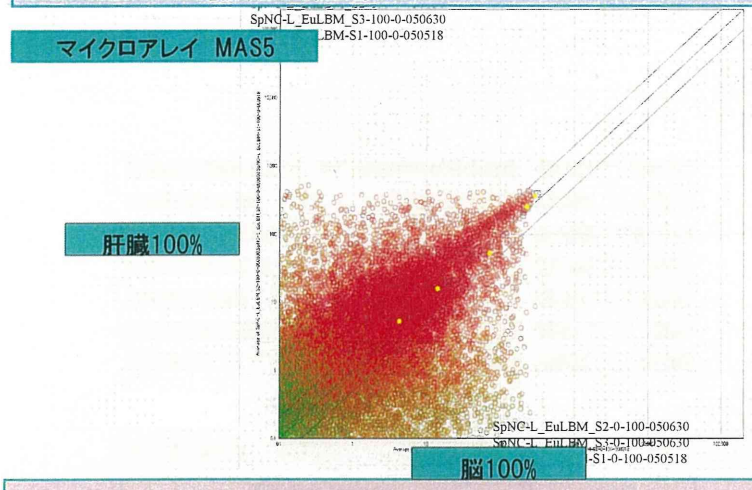
脳100%

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。

29

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

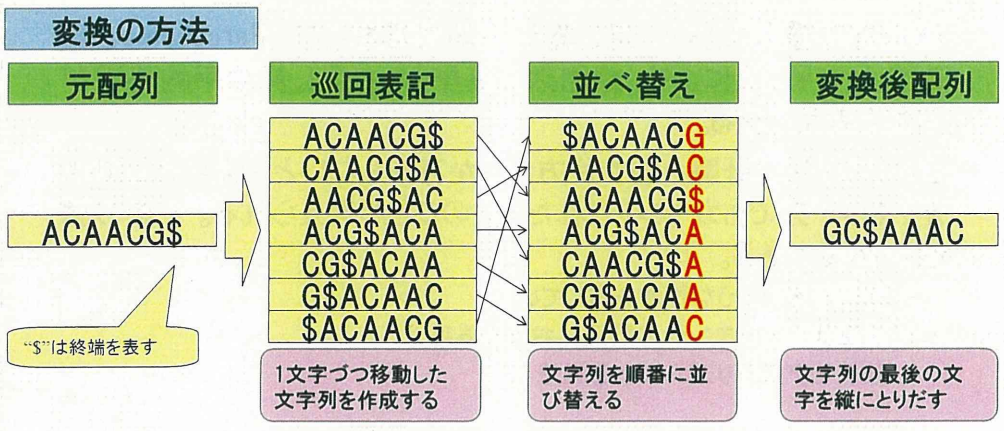
3.次世代シークエンサのアライメント用代表的アルゴリズム

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.totronto.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

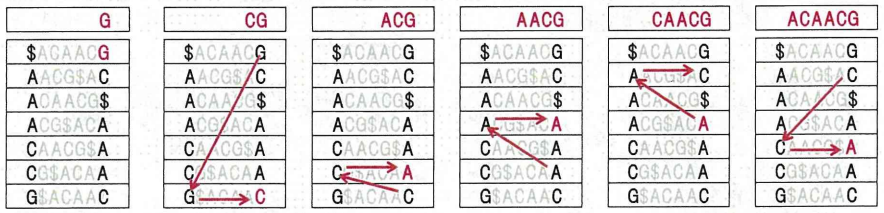
第2世代のアライメント用アルゴリズムとして用いられているBurrrows-Wheeler Transform(BWT:バーローウィーラー変換)がどのようなものかまとめる。



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



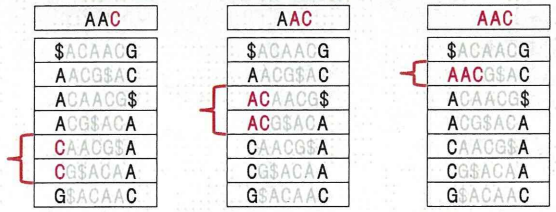
最初の縦列と最後の縦列を用いることにより、文字列の一致・検索処理が可能である

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。

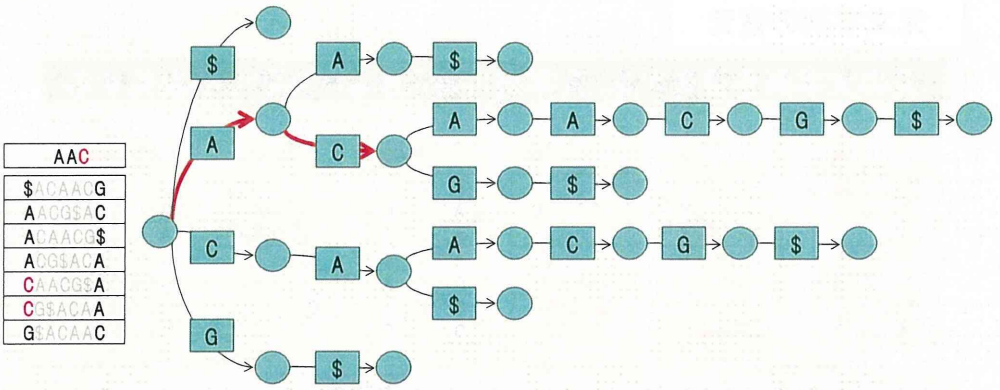
後ろの1文字　後ろの2文字　後ろの3文字



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。

3.4.Burrows-Wheeler Transformの原理④

ツリー構造による表現



並べ替え後の情報をツリー表示する。ツリーをたどることで、部分文字列を高速に検索可能である。例えば、文字列"AC"は、赤太線でたどり、最後までたどることで、2か所で一致することが分かる。

4.Teradataを用いたアライメント試行

- NIHS毒性部に導入済みのTeradata RDBMSを用いて、試験的にアライメントを実施する。
  - Teradata RDBMSは、ビジネス分野におけるデータウェアハウスなど大量データ(数十ペタバイト)処理に向いたデータベースエンジンである。
- 対象とする実験
  - LBM実験
  - 1条件あたり1回の計測を実施している
- 対象となる参照配列
  - ゲノム配列情報
    - 19個＋XY染色体
  - 遺伝子情報
  - GSC配列
    - 6種類

4.1.参照配列塩基数

染色体別

染色体	塩基数
1	197,195,437
2	181,748,092
3	159,599,788
4	155,630,125
5	152,537,264
6	149,517,042
7	152,524,558
8	131,738,876
9	124,076,177
10	129,993,261
11	121,843,862
12	121,257,536
13	120,284,318
14	125,194,870
15	103,494,980
16	98,319,156
17	95,272,657
18	90,772,037
19	61,342,436
X	166,650,301
Y	15,902,560

- 対象とするマウスの参照配列の塩基数を右表に示す。
- 合計26億塩基以上存在する

合計2,654,895,333 塩基

遺伝子は、ペアになっており、二つの方向が存在するが、アライメントの試行として、一方向だけを実施する。

4.2.Teradataを用いたアライメントアルゴリズム作成

既存アルゴリズムでの課題

複数個所に割り当てられた場合の取り扱いが単純すぎる。

対策

複数個所に割り当てられた配列の取り扱いを含めた割り当てを行うアルゴリズムの作成を試みる。

Teradata RDBMSの特徴

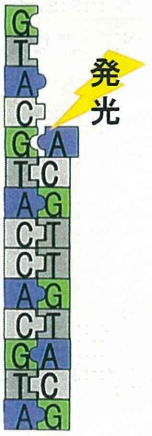
Teradataは、ハッシュキーによるデータ分散を基本構造として持っている Relational Database Management Systemである。

ハッシュキー構造を活かした、BWTの強みを生かしたアルゴリズムが望ましい

4.3.イルミナ社の次世代シークエンサの基本原理と読みとり精度

原理

1塩基ずつ、伸長反応を行っていく。塩基が結合する際に、4種類の塩基ごとに色の異なる光を出す。光の色(波長)を読み取ることにより、どの塩基が結合したかを検知する



塩基読み取り精度

- 同じ塩基が続くと長さを間違えやすい
- 塩基合成の速度が安定しないせいかもしれない
- 読み取り始めのほうが読み取り精度が高く、だんだんと精度が落ちる
- 溶媒の塩濃度などが変化し、塩基合成の適切な状態を保てないらしい

イルミナ社の計測データとして、各塩基の読み取り精度の情報が付加されている。

読み取りエラー率は、最初は高く、だんだんと低下する傾向がみられる。

4.3.読取エラー率

イルミナでは、各塩基の読み取りに関して、エラー率が記録されている。記録量削減のため、ASCII文字で記録している。

QualityScore=2であるとは、エラー率63%である。つまり、正解率37%で、偶然の可能性25%を僅かに上回っているに過ぎない

$$QualitySco\ re = -10\log_{10}(p)$$

$$p = 10^{-\frac{QualitySco\ re}{10}}$$

ASCIIコード	ASCII文字	Quality Score	error rate (p)
64	@	0	1.000000000
65	A	1	0.794328235
66	B	2	0.630957344
67	C	3	0.501187234
68	D	4	0.398107171
69	E	5	0.316227766
70	F	6	0.251188643
71	G	7	0.199526231
72	H	8	0.158489319
73	I	9	0.125892541
74	J	10	0.100000000
75	K	11	0.079432823
76	L	12	0.063095734
77	M	13	0.050118723
78	N	14	0.039810717
79	O	15	0.031622777
80	P	16	0.025118864
81	Q	17	0.019952623
82	R	18	0.015848932
83	S	19	0.012589254
84	T	20	0.010000000
85	U	21	0.007943282
86	V	22	0.006309573
87	W	23	0.005011872
88	X	24	0.003981072
89	Y	25	0.003162278
90	Z	26	0.002511886
91	[	27	0.001995262
92	¥	28	0.001584893
93	]	29	0.001258925
94	^	30	0.001000000
95	_	31	0.000794328
96	`	32	0.000630957
97	a	33	0.000501187
98	b	34	0.000398107
99	c	35	0.000316228
100	d	36	0.000251189
101	e	37	0.000199526
102	f	38	0.000158489
103	g	39	0.000125893
104	h	40	0.000100000

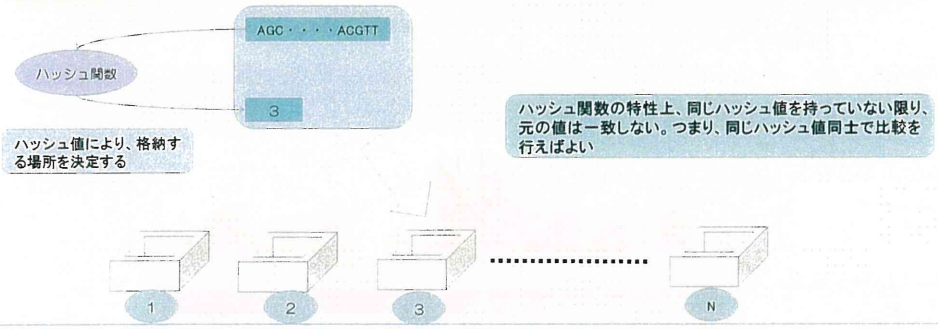
4.4.アライメント手順概略

詳細については特許出願に差し障りがあるため非公開とさせて頂きます

Teradata RDBMSの特徴

Teradataは、ハッシュキーによるデータ分散を基本構造として持っているRelational Database Management Systemである。

TERADATAのハッシュ分散構造



4.5.アライメント試行結果

アライメントの試行を、Liver100%を対象として実施した

Lane1(Liver100%)

読取数	39,631,834

BWAによる解析処理

フィルターパス数	35,273,281
除外数	4,358,553

完全一致マッチング処理 参照配列の1方向のみで試行した

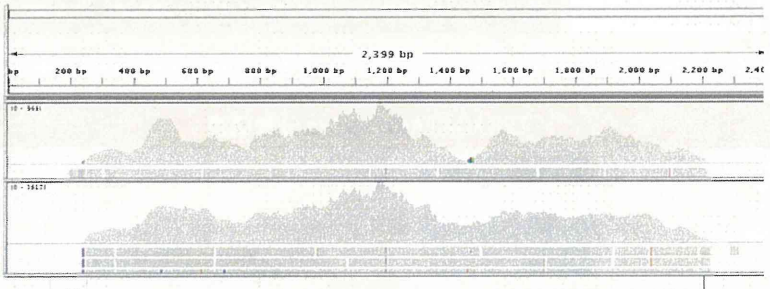
読取不可塩基なし数	37,873,323
マイクロサテライト無タグ数	37,429,628
101塩基完全一致タグ数	6,678,287

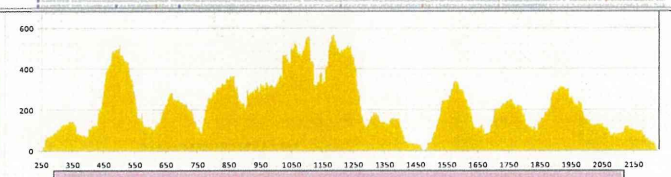
参照配列が1方向だけなので、半分が合致しないとしても、その半分にも満たない。何らかの現象が発生していると考えられる

4.5.1.THRによる一致の確認

GSCの最大量であるTHRを用いて、BWAと今回のアライメント結果を比較した。

GSC1_thr　　2400 bp

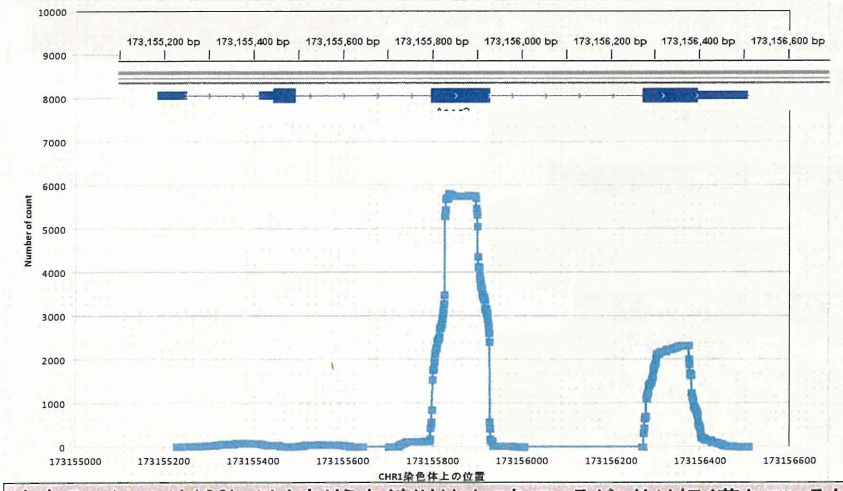




Teradataを用いた完全一致マッチング

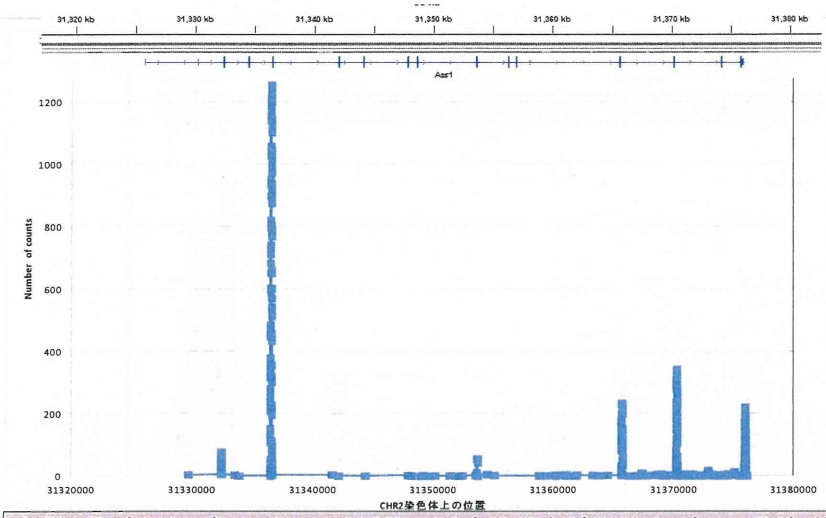
似たような形状を示している。

4.5.2.Apoa2の割り付け結果



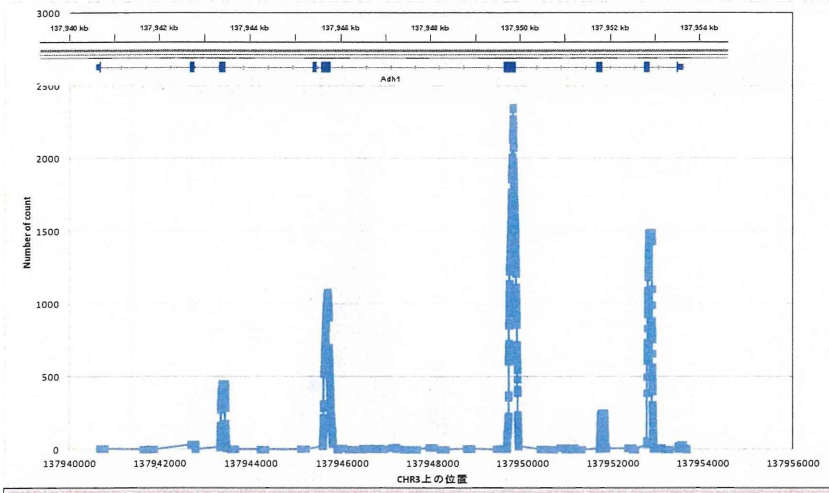
中央のエクソンは100bp以上ありうまく割り付けできているが、他は短く落ちているものが多いと思われる

4.5.3.Ass1の割り付け結果



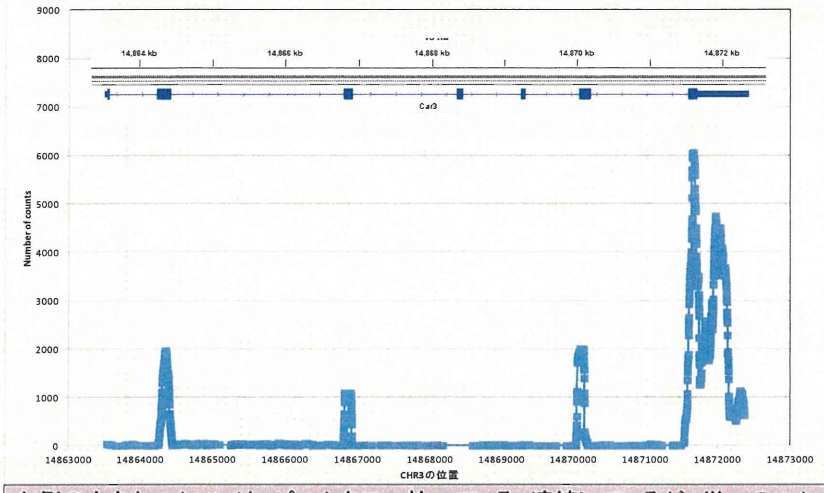
左側の飛びぬけて高いエクソンは100bp以上あり、うまく割り付けできているが、他は短く落ちているものが多いと思われる

4.5.4.Adh1の割り付け結果



右側の飛びぬけて高いエクソンは100bp以上あり、うまく割り付けできているが、他は短く落ちているものが多いと思われる

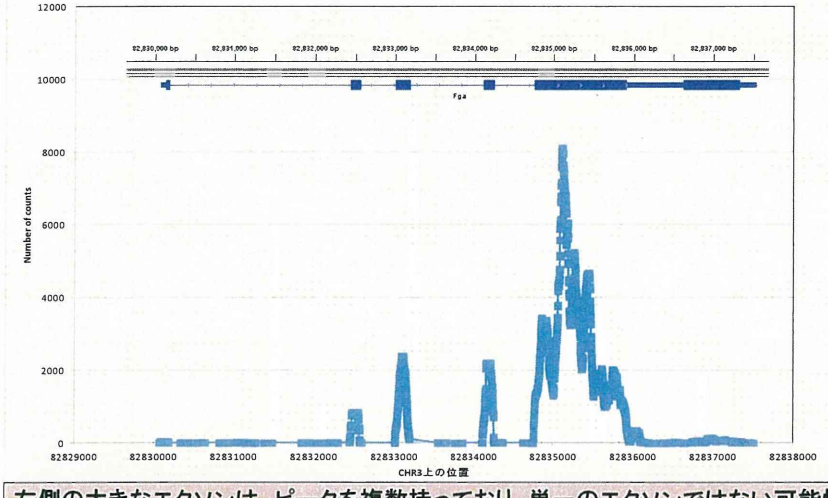
4.5.5.Car3の割り付け結果



右側の大きなエクソンは、ピークを二つ持っている。連続しているが、単一のエクソンではない可能性があるのではないか？

48

4.5.6.Fgaの割り付け結果



右側の大きなエクソンは、ピークを複数持っており、単一のエクソンではない可能性があるのではないか？

49