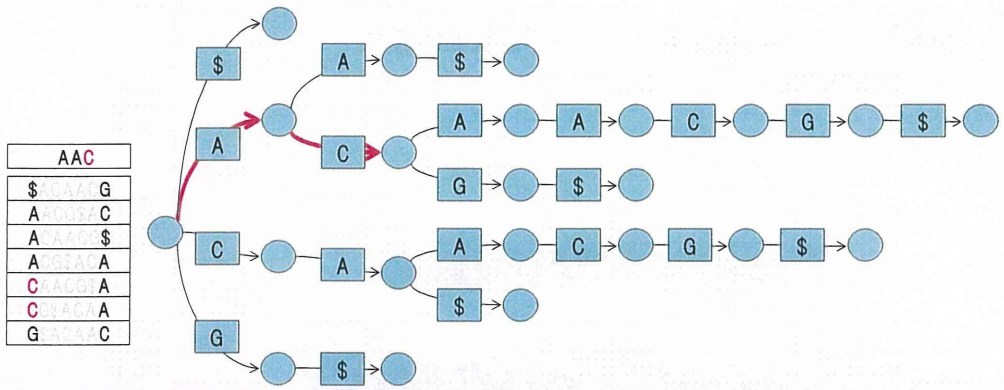
3.4.Burrows-Wheeler Transformの原理④

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

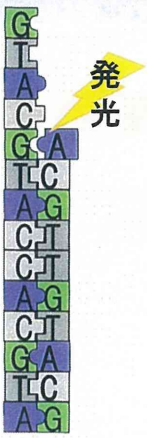
Teradataは、ハッシュキーによるデータ分散を基本構造として持っている Relational Database Management Systemである。

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

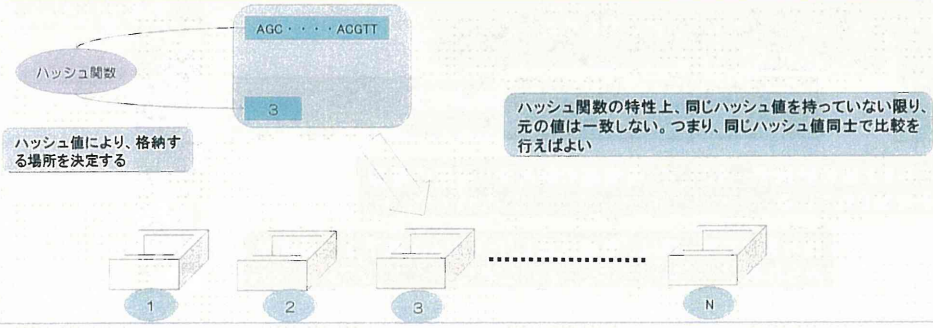
4.4.アライメント手順概略

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

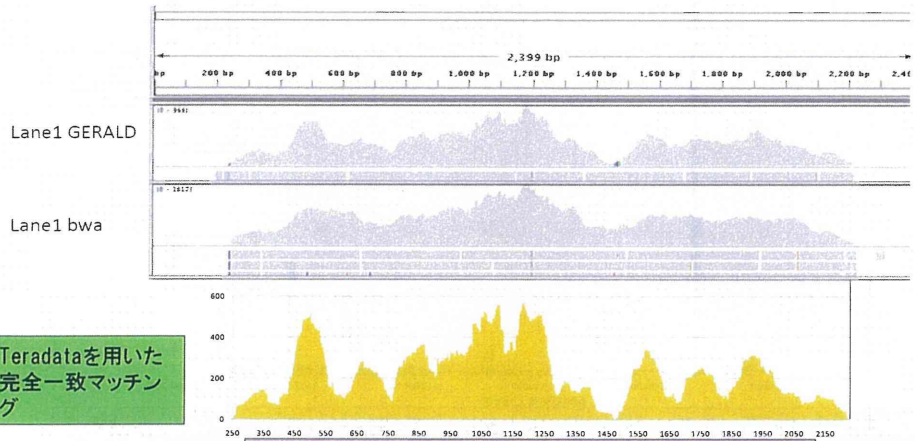
読取不可塩基なし数	37,873,323
マイクロサテライト無タグ数	37,429,628
101塩基完全一致タグ数	6,678,287

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。

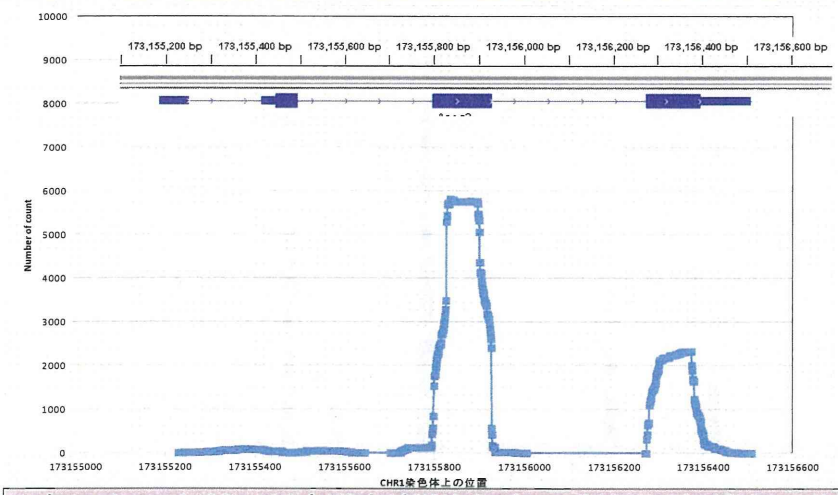
GSC1_thr　　2400 bp



Teradataを用いた完全一致マッチング

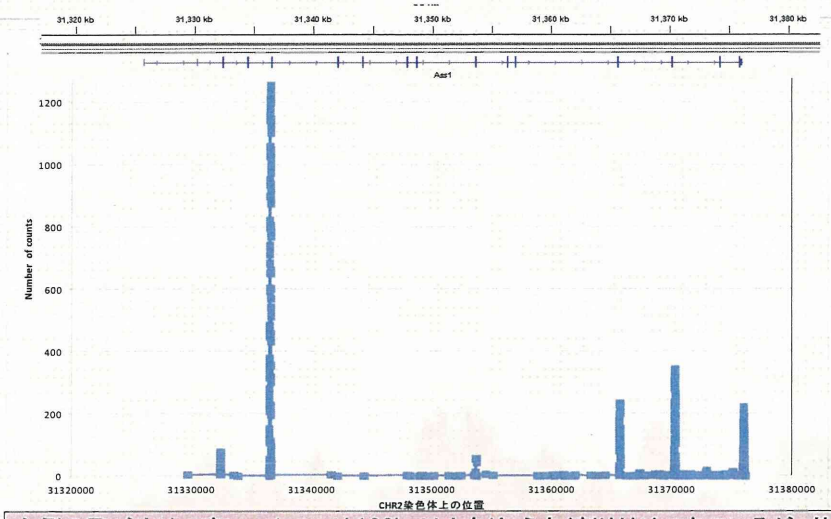
似たような形状を示している。

4.5.2.Apoa2の割り付け結果



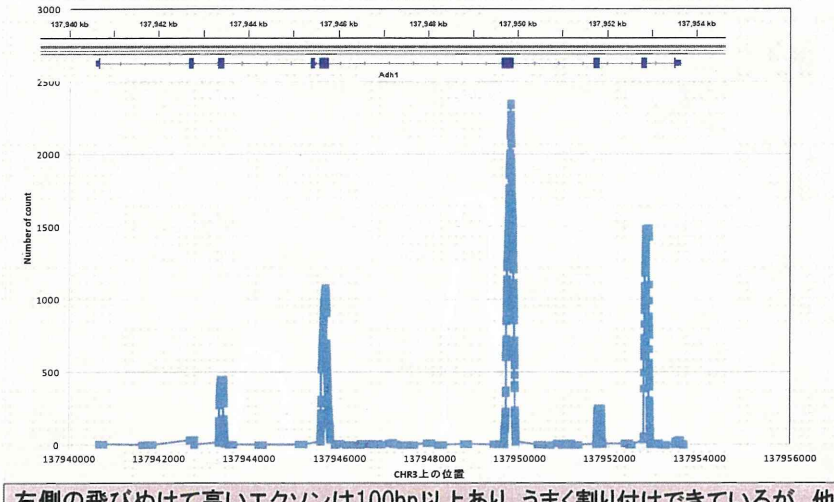
中央のエクソンは100bp以上ありうまく割り付けできているが、他は短く落ちているものが多いと思われる

4.5.3.Ass1の割り付け結果



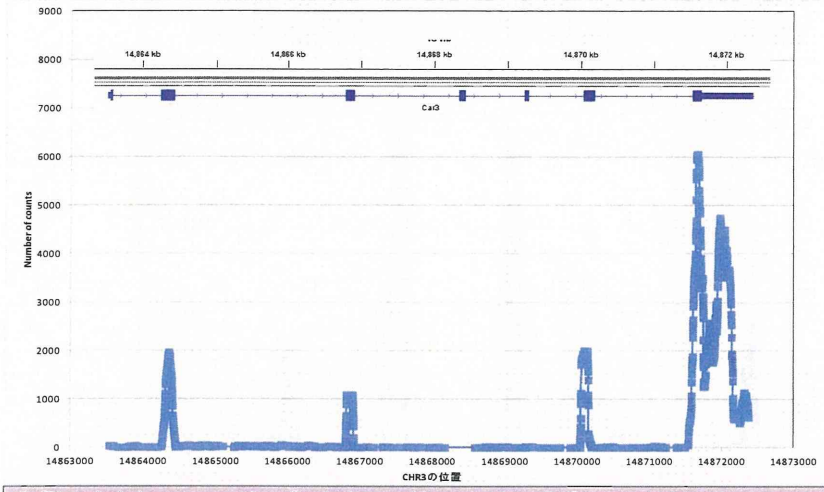
左側の飛びぬけて高いエクソンは100bp以上あり、うまく割り付けできているが、他は短く落ちているものが多いと思われる

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

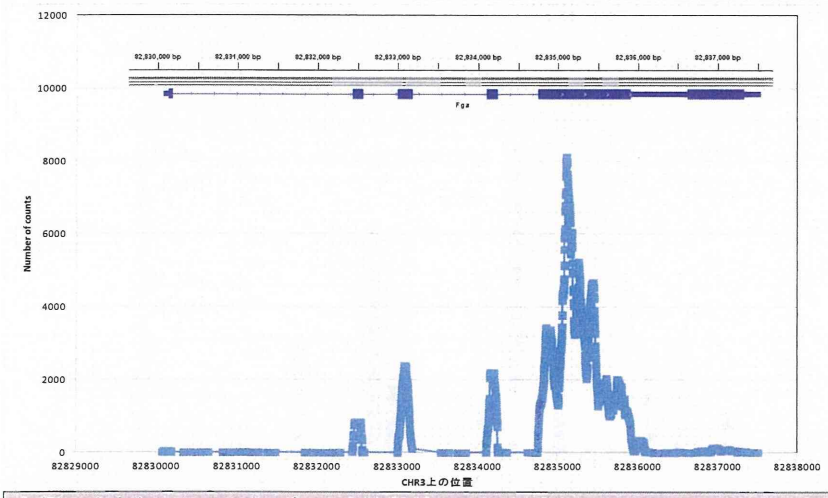
4.5.5.Car3の割り付け結果



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

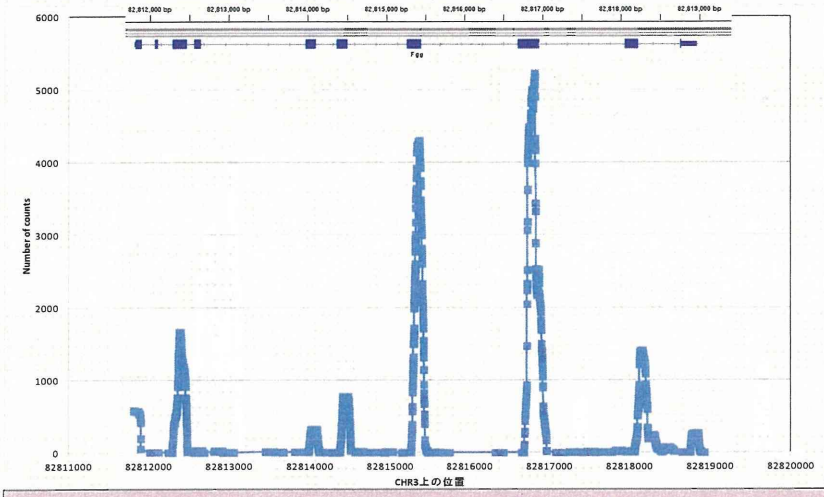
4.5.6.Fgaの割り付け結果



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？

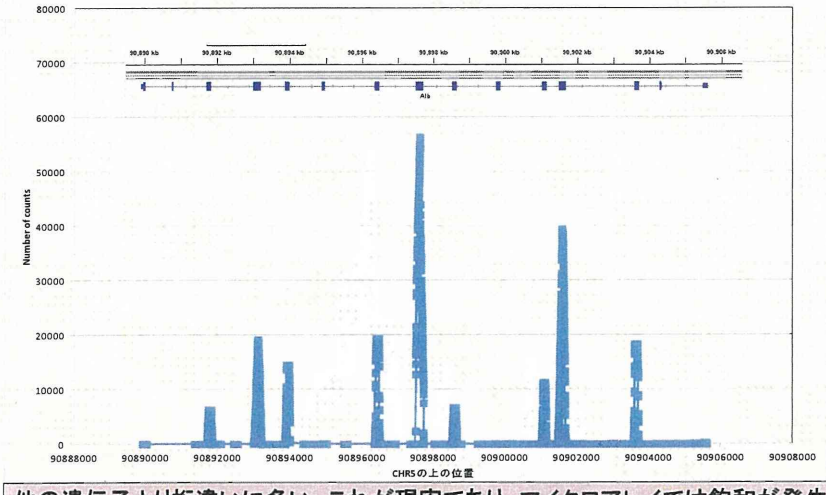
49

4.5.7.Fggの割り付け結果



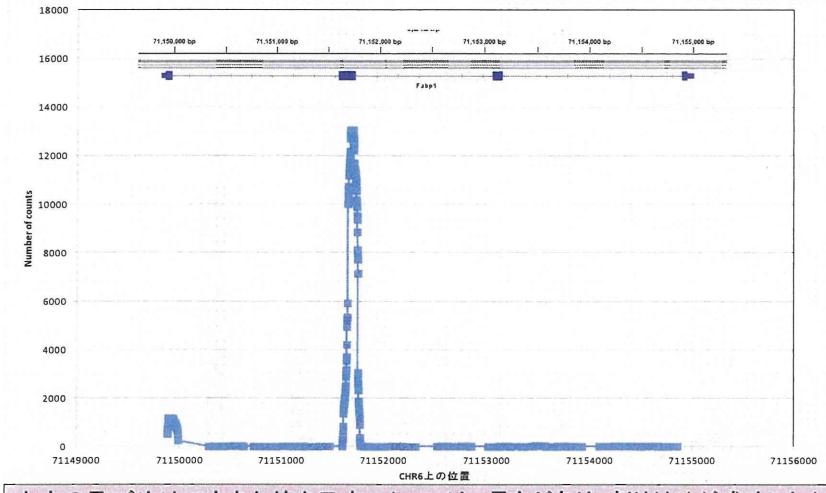
長いエクソンほどピークが高くなっている

4.5.8.Albの割り付け結果



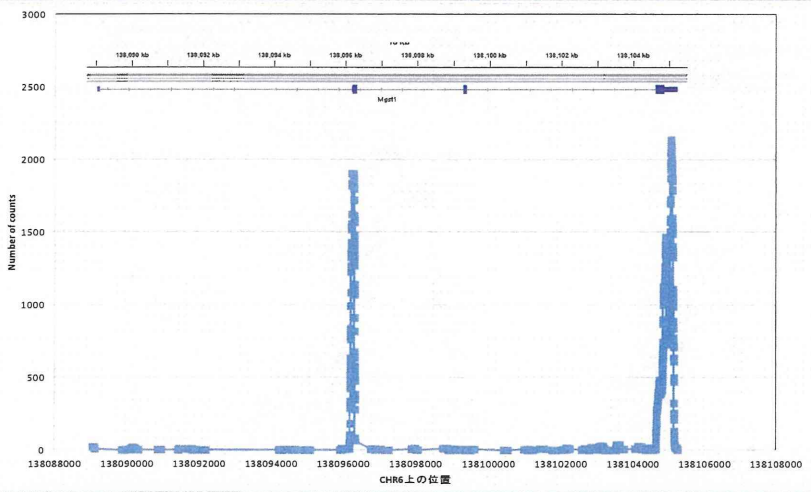
他の遺伝子より桁違いに多い。これが現実であり、マイクロアレイでは飽和が発生し、濃度を推定しにくい遺伝子と思われる

4.5.9.Fabp1の割り付け結果



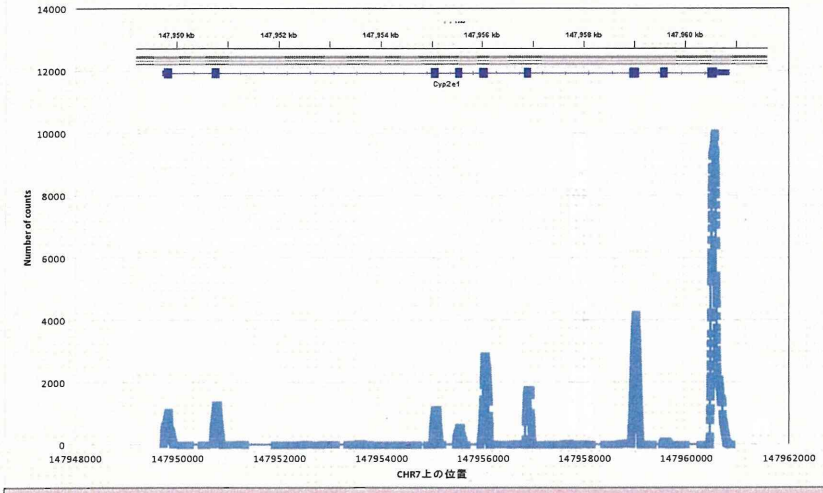
中央の飛びぬけて大きな値を示すエクソンは、長さがあり、割り付けが成功したと思われる

4.5.10.Mgst1の割り付け結果



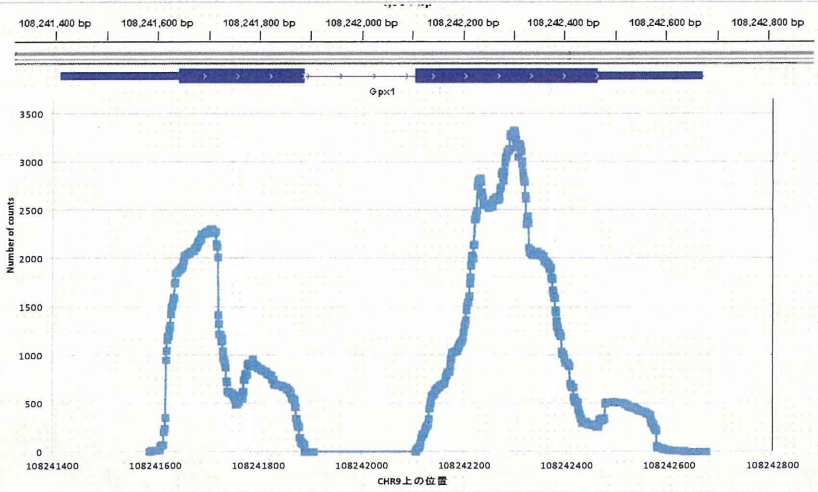
二つのピークとなっているエクソンは、長さがあり、割り付けが成功したと思われる

4.5.11.Cyp2e1の割り付け結果



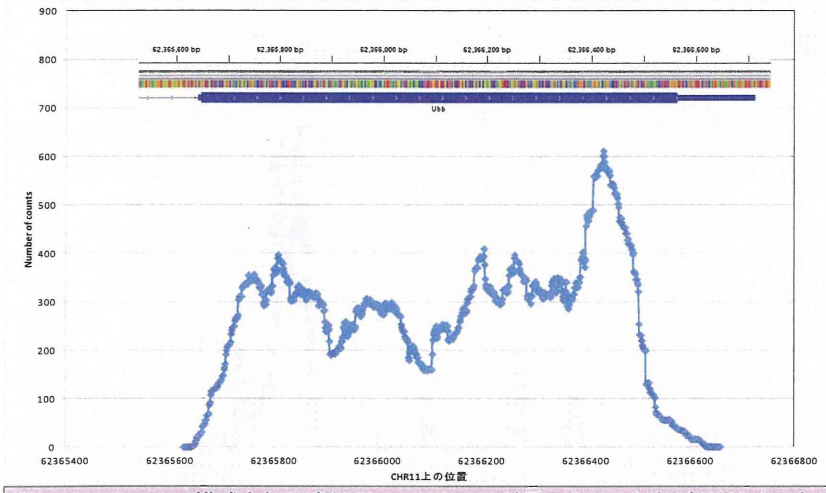
右側のエクソンは、長さがあり、割り付けが成功したと思われる

4.5.12.Gpx1の割り付け結果



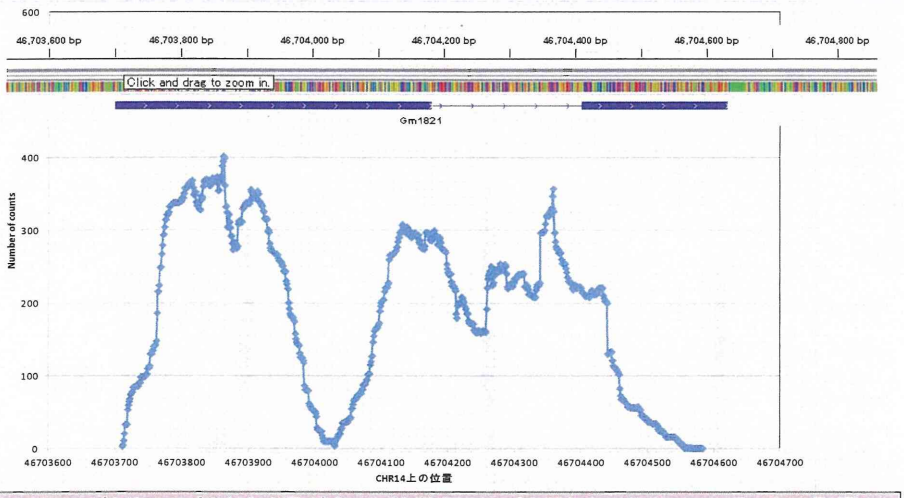
エクソン中に左側にピークがあるのは、ペアエンド法を実施しているためと考えられる。
割り付けアルゴリズムも対応させるべきである

4.5.13.Ubbの割り付け結果



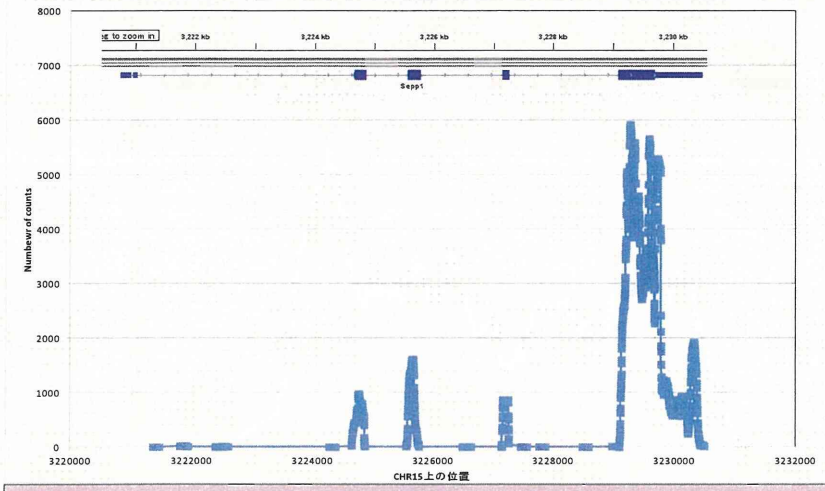
単一のエクソンで構成される遺伝子で、エクソンが長いため、うまく割り付けできていると思われる。

4.5.14.Gm1821の割り付け結果



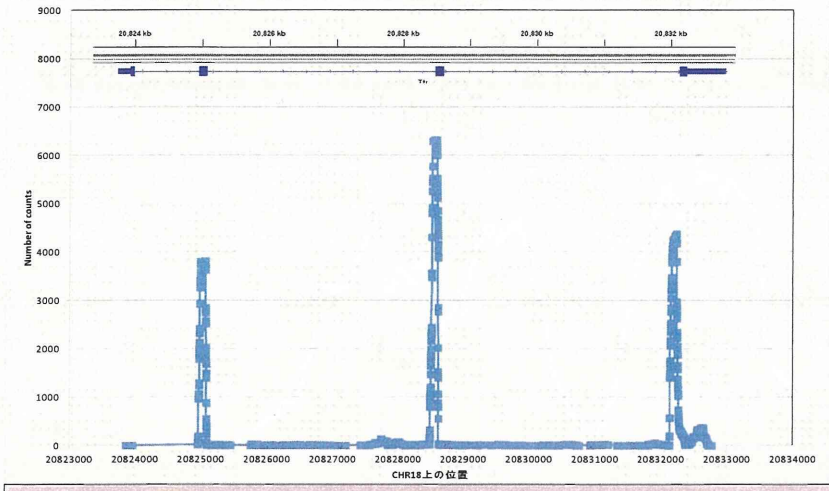
イントロン部分での割り付け量が多い。複数個所へ割り付け可能としていることの影響かもしれない。

4.5.14.Sepp1の割り付け結果



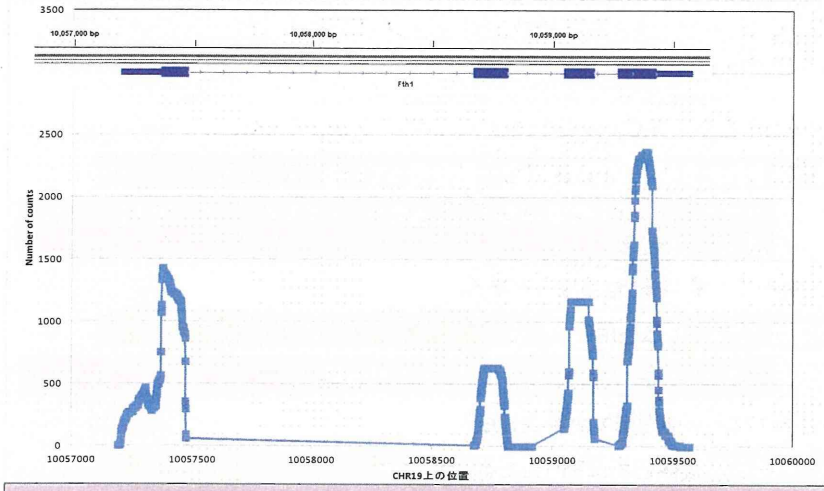
右側の長いエクソン部分はうまく割り付けできていると思われる。

4.5.15.Ttrの割り付け結果



各エクソンが適度な長さがあり、割り付けがうまくいっていると思われる。

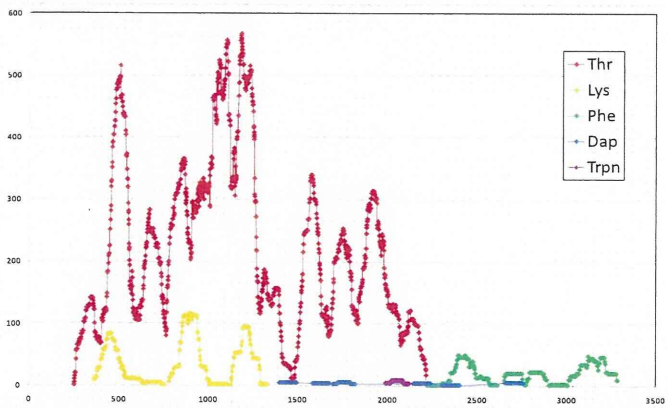
4.5.16.Fth1の割り付け結果



各エクソンが適度な長さがあり、割り付けがうまくいっていると思われる。

4.5.17.GSC-RNAの発現量

GSCの5種類のRNAの割り当てを確認した



同一RNAの中で、配列位置により、計測されたタグの数が異なっている。分配係数で補正可能なものか？何らかのバイアスが発生していないか？ ➔ 今後検討する

4.6.アライメント試行のまとめ

以下にあげる現象がみつかり、本格的なアライメント用アルゴリズムのためには対処が必要である

1. 長いエクソンは割当量が多くなり、短いエクソンは割当量が少なくなる
   - 読み込まれたタグの100塩基完全一致を行っている
     - スプライシングの影響が考えられるので、スプライシングに対応したアルゴリズムが必要
2. 3' 末端側の割り当て量が多くなる場合がある。
   - 計測はペアエンド法を用いている
     - ペアエンド法に対応したアルゴリズムが必要
3. イントロン部分で大量の割り当てが存在した
   - 単一のタグが複数個所に割り当てるようにした
     - 複数個所の割り当てを適切に分配させるアルゴリズムが必要

62

（資料　3）

マイクロアレイ補正アルゴリズム

TERADATA THE BEST DECISION POSSIBLE

マイクロアレイ補正アルゴリズム

Multi-adsorption Langmuir

日本テラデータ株式会社　松本伸哉

2012年2月9日

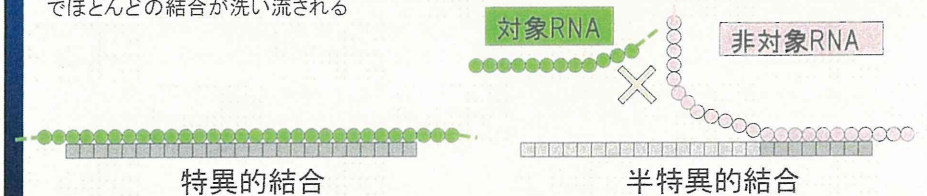
半特異的結合

プローブの最大塩基長で「特異的」に結合し、それ以外の結合は非特異的結合として取り扱われてきた。

1塩基短く一致している場合に、特異性を完全に喪失するとは考えにくい。短くなるにつれて徐々に特異性が低下するのではないだろうか？

非特異的結合は特異的結合に徐々に置き換わるといわれている。置き換わって平衡状態になったときに、全てが特異的結合に置き換わっているのか？

部分的に配列が一致する結合は、ハイブリダイゼーション工程で特異的結合を阻害し、洗浄工程でほとんどの結合が洗い流される



完全一致よりも短い長さの一致による結合を
「**半特異的結合** (semi-specific binding)」と名付けた

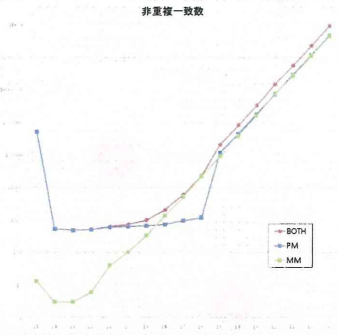
TERADATA THE BEST DECISION POSSIBLE

MOE430v2のターゲットとプローブ間の一致数

Affymetrix社のMOE430v2のターゲットRNAとプローブの間で、連続して一致するパターンを網羅的に調査した。

MMに25塩基一致が13個存在した。

連続長	PM	BOTH	MM
25	526,104	526,117	13
24	527	530	3
23	481	484	3
22	513	519	6
21	592	632	40
20	631	733	102
19	668	999	331
18	730	2,068	1,338
17	948	6,091	5,143
16	1,199	23,195	21,996
15	117,190	209,391	92,201
14	440,156	824,631	384,475
13	1,774,086	3,363,228	1,589,142
12	7,217,804	14,955,606	7,737,802
11	29,355,970	56,799,526	27,443,556
10	118,411,755	230,527,744	112,115,989
9	477,288,011	933,024,417	455,736,406



12mer連続一致では、MM側が若干多い。MMの設計で中央位置を除くというのが影響していると考えられる

16mer連続一致では、PM側が圧倒的に少ない。設計段階でチェックされていると思われる。

TERADATA THE BEST DECISION POSSIBLE

ラングミューアの吸着等温平衡方程式

- Irving Langmuirによって1918年に導出された理論的な吸着等温平衡方程式である。以下のような仮定を持っている。
  - 吸着媒には有限な数N個の吸着サイトがあり、そこだけで吸着質分子と結合する。
  - すべての吸着サイトは等価である。
  - 1個の吸着サイトは1個の吸着質分子としか結合しない。
  - 空の吸着サイトM、吸着質S、吸着サイトに結合した吸着質M-Sの間にM + S ⟷ M-Sの化学平衡が成立する。

$$K = \frac{N\theta}{N \cdot (1-\theta) \cdot C}$$

K:平衡定数
N:吸着サイト数
θ:吸着されているサイトの割合
C:濃度

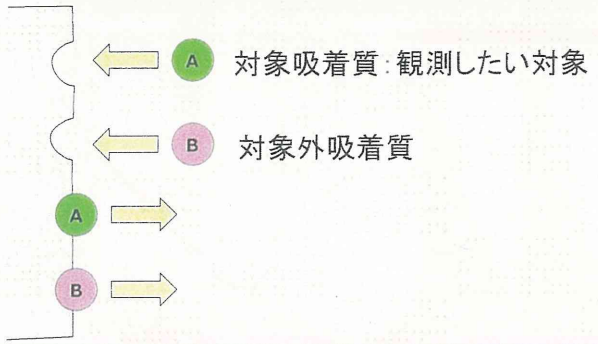
TERADATA THE BEST DECISION POSSIBLE

ラングミューアの吸着等温平衡方程式の複数吸着質への拡張

ラングミューアの吸着等温平衡方程式の基本に戻り、拡張を行う。

仮定の拡張

- 一つの吸着サイトは、複数種類の吸着質と結合し得る
- 一つの吸着サイトは、同時に一つの吸着質としか結合しない



吸着速度と離脱速度が平衡状態になると仮定して、方程式を解く

ラングミューアの吸着等温平衡方程式 2吸着質の場合の結合量

2吸着質の量が特殊な関係を持っている場合に、特殊な性質を有する

一方の濃度が、他方の濃度に比例する部分と、一定の濃度に分解できる場合

1吸着質の場合

$$N\cdot\theta = N\cdot\frac{K\cdot C}{1+K\cdot C}$$

2吸着質の場合

$$N\cdot\theta_A = N\cdot\frac{C_A}{\frac{\alpha}{K}+\frac{\alpha}{\beta}\cdot C_B + C_A}$$

仮定 $C_B = p+q\cdot C_A$

$$N\cdot\theta_A = N\cdot\frac{C_A}{\frac{\alpha}{K}+\frac{\alpha}{\beta}\cdot(p+q\cdot C_A)+C_A}$$

$$N\cdot\theta_A = N\cdot\frac{C_A}{\left(\frac{\alpha}{K}+\frac{\alpha}{\beta}\cdot p\right)+\left(\frac{\alpha}{\beta}\cdot q+1\right)\cdot C_A}$$

$$N\cdot\theta_A = N_A\cdot\frac{k_A\cdot C_A}{1+k_A\cdot C_A}$$

濃度が定数倍、結果の蛍光強度最大値が定数倍となったラングミューアの方程式と一致する

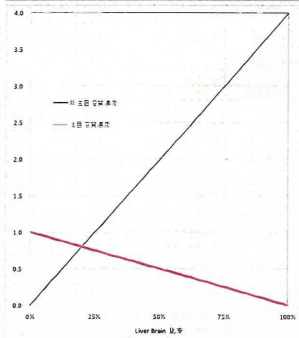
パラメータを求めるために、一連の実験を行う場合には、この仮定を満たすように設定する場合が多い。他のRNAの影響を受けているのかが分からなくなる。

複数吸着質におけるラングミューアの方程式から導かれる特殊な状況

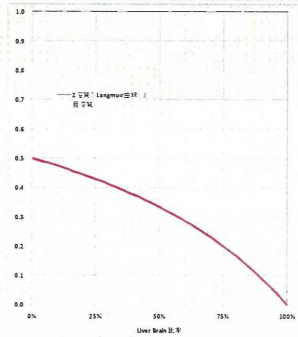
半特異的結合により、他のRNAの特異的結合を阻害していると仮定すると、脳で発現が大きくても、肝臓において、半特異的結合をしやすい配列のRNAが発現している場合には、下に凸な形状を示す。

導出された式より得られる値のグラフを示す

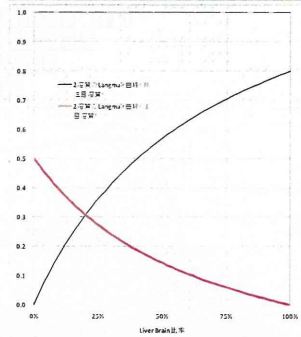
注目RNAと比注目RNAの濃度



半特異的結合なしの場合



半特異的結合ありの場合



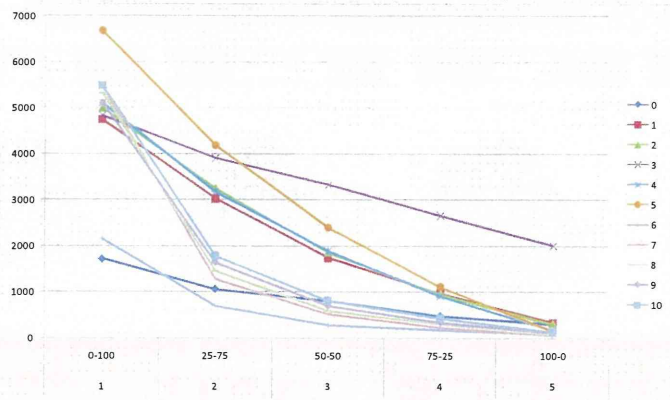
他のRNAの影響がなければ、上に凸になるはずだが、脳だけで発現する遺伝子で、肝臓で極端に発現が大きなRNAと配列が似ていた場合には、半特異的結合の阻害により、下に凸になる。下に凸な形状を見つければ、相手の臓器で発現量の多いRNAと似ているものが見つかると考えられる。

TERADATA THE BEST DECISION POSSIBLE

実験結果(LBM)の下に凸な状況

理論式から導かれる下に凸な状況を示すプローブが存在した

1436268_at : Ddn (Dendrin)



特に、Probe 7～10は、落ち込み方が激しい。

TERADATA THE BEST DECISION POSSIBLE

実験結果(LBM)の下に凸な状況

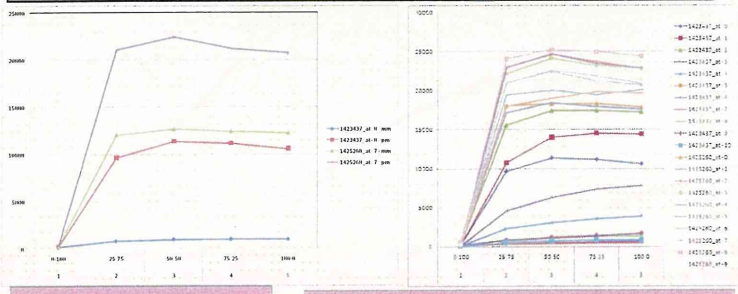
・対象プローブと配列が似ており、L側で飽和していると考えられるプローブ(プローブ同士の配列を比較した)

1423437_at	0	TTTCTAAGGCGTCTGTGTCAATTCA
1425260_at	7	GGATGACTTTGCACAGTTCCTGGAT
1436268_at	7	GTATATAAGGCGTCTTGCACAGTGC
1436268_at	8	ATATAAGGCGTCTTGCACAGTGCCT
1436268_at	9	ATAAGGCGTCTTGCACAGTGCCTGG
1436268_at	10	AAGGCGTCTTGCACAGTGCCTGGCA

Ssta3

Alb1

Ddn



1423437_at : Gsta3
1425260_at : Alb1

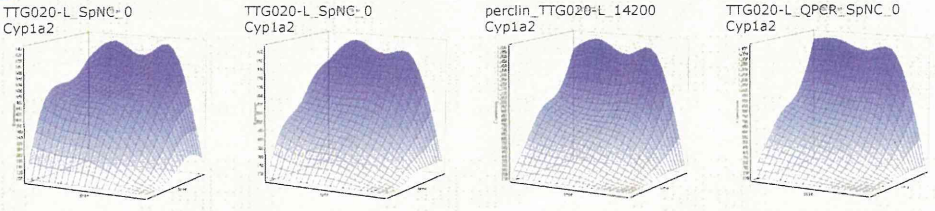
脳で大量に存在するRNAと似た配列を有していた

TERADATA THE BEST DECISION POSSIBLE

補正結果 TCDD Cyp1a2

Cyp(Cytochrome P450)は、ダイオキシン類によって誘導されることが知られている。

MAS5	RMA	MLANG	QPCR
TTG020-L_SpNC_0 Cyp1a2	TTG020-L_SpNC_0 Cyp1a2	perclin_TTG020-L_14200 Cyp1a2	TTG020-L_QPCR_SpNC_0 Cyp1a2



TCDDによって強く誘導され、マイクロアレイでは飽和現象が見え始めており、MAS5では飽和を処理しきれない。RMAでは、一部解消されている。MLANGではかなり飽和がキャンセルされている。

TERADATA THE BEST DECISION POSSIBLE