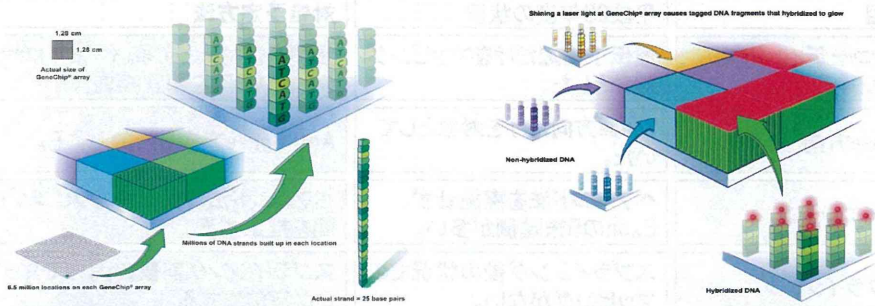
背景:マイクロアレイの原理

NTT DATA

マイクロアレイ上にRNAの各領域に相補鎖を作成し、対象サンプルを蛍光で標識する。その領域でどの程度の蛍光輝度があるかで、結合したRNA量(発現量)を測定する。



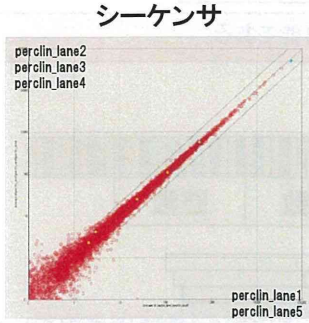
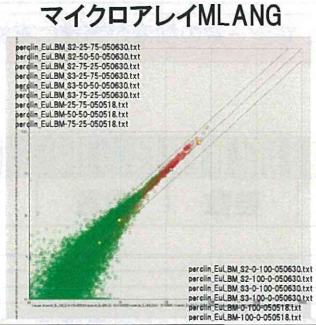
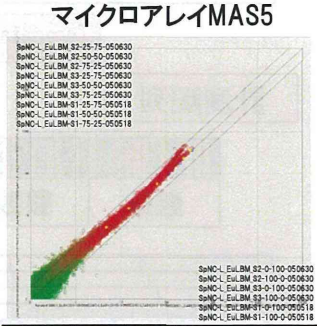
マイクロアレイの問題点

- 未知の配列を検出できない。
- 飽和現象を起こしやすい。
- 類似配列への誤結合(クロスハイブリダイゼーション)を起こすことがある。

試験管内混合と数値的混合の比較による測定性能の検証

NTT DATA

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縦軸に、Liver100%とBrain100%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。

シーケンサ	マイクロアレイMLANG	マイクロアレイMAS5
		
高発現から低発現まで対角線上にある	高発現は対角線上に乗っている	高発現で分岐する状況がある

マイクロアレイでは飽和現象があり、MAS5では、Liver側とBrain側で絶対量が異なるため、飽和パターンが異なり、分岐したと思われる。MLANG(特許第5177712号)では、飽和をキャンセルできたが限界がある。

2. 平成23年度における検討と本年度の実施項目

NTT DATA

平成23年度委託研究で明らかになった課題と対処予定方法

項目	平成23年度の状況	対処予定方法
ノンコーディングRNAへの対応	遺伝子周辺だけをマッピング対象とした。	全DNAを対象とする。(マイクロサテライトの排除を適宜実施)
Sense方向	Sense方向だけを対象としていた。	Anti-Sense方向も対象とする。
ペアエンド法	ペアエンド法を実施せず、Exonの5末端側が多い。	ペアエンド法の両端をマッピングし、間を推定する。
スプライシング	スプライシング後の状況でのマッピングがない。	スプライシング前後の両方でマッピング可能にする。
複数箇所へのマッチ	複数箇所にマッチした場合は平均按分。	周囲の状況を考慮した按分を実施する。

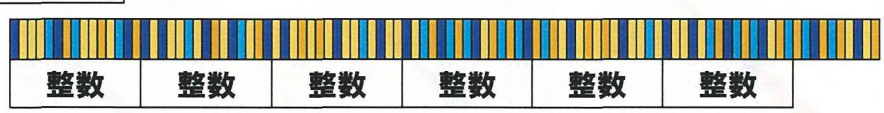
3. 1. Teradataによるマッピングの基本アイデア①

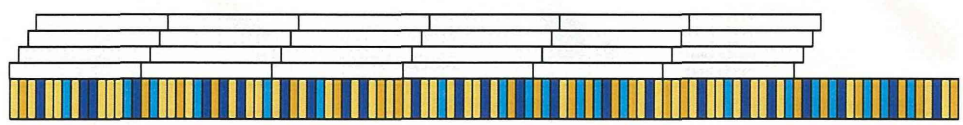
NTT DATA

読込配列と参照配列を、15塩基ずつに分割し、それぞれ、30ビット(4バイト)整数で表現する。

Teradataは、完全一致の検索が高速に実施可能である。

計測配列





参照配列

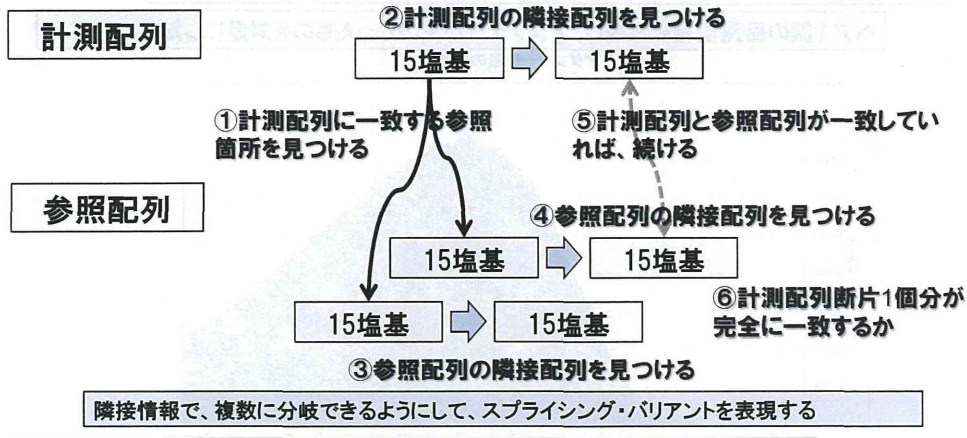
1塩基ずらしたパターンを生成する ⇨ **全塩基数の情報を生成する**

スプライシング前と後の2種類を生成する

3．2．マッピングの基本アイデア②

NTT DATA

①➔⑥の順番で処理を実施することにより、計測断片全長で、参照配列上と一致するかを判定可能である



隣接情報で、複数に分岐できるようにして、スプライシング・バリアントを表現する

課題点

- 参照配列上に同じパターンが多数存在すると、処理しきれなくなる
- 完全一致以外では、工夫が必要

3．3．参照配列上の未定配列

NTT DATA

参照配列上に、未定塩基(N)である箇所が何箇所か存在する。

連続個数	存在数
1	47
2	8
3	2
4〜7	0

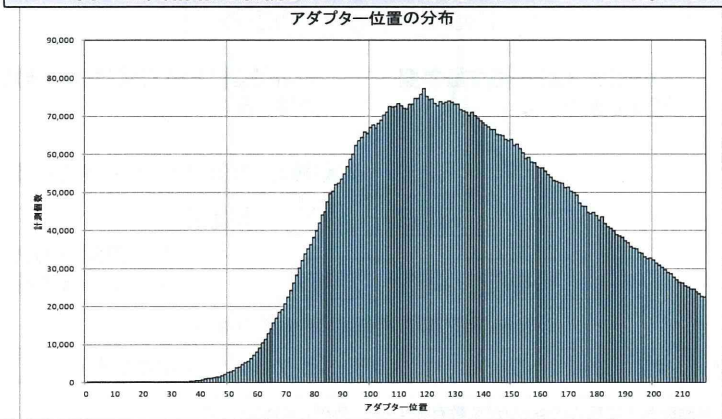
3連続以下の未定塩基は対処すべきである。8個以上連続の未定塩基列へのマッピングは困難であり、今回は対象外とする。

3. 4. 計測配列の品質

NTT DATA

測定断片の長さ(アダプターの位置)の分布を確認した。

ペア1側の品質情報を無視し、アダプターが見つかったものを対象にした。



長さ120前後でピークとなる分布形状であった。

4. 1. マッピングアルゴリズムのトライアル①

NTT DATA

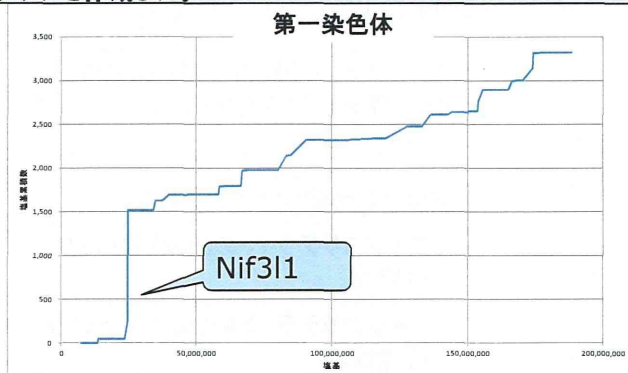
マッピングアルゴリズムの基本部分のみを作成し、課題の洗い出しを行った。

対象	ペアエンドの一方だけを使用	
	ランダムサンプル1/10000	
	対象配列数	1009
	計測配列中に未定塩基を含まない アダプターが見つかった	
	対象配列数	798
	マッピングできた配列	219
	マッピングできたペア(累計)	2057

4. 2. 第一染色体におけるマッピング結果

NTT DATA

マッピングした配列を、塩基アドレスと、マッピング累積塩基数の積み上げ数でグラフを作成した。



マッピング結果のペアで、測定配列と参照ゲノムの配列は一致しており、基本的なアルゴリズムが動作している事が確認できた。
Nif3l1近傍(遺伝子領域外)でマッピングされるリードが集中して表れた。

4. 3. マイクロサテライト

NTT DATA

短い配列の繰り返しは、ゲノム上に大量に存在し、マッピング処理の処理性能を著しく低下させる。マイクロサテライトの状況を確認した。

	ゲノム上での存在数	配列パターン	基本パターン番号	マイクロサテライトパターン	開始位置
1	3270514	ACACACACACACACA	6	CA	1
2	3269104	TGTGTGTGTGTGTGT	11	TG	0
3	3184366	CACACACACACACAC	6	CA	0
4	3182110	GTGTGTGTGTGTGTG	11	TG	1
5	1792080	TCTCTCTCTCTCTCT	15	TC	0
6	1782146	AGAGAGAGAGAGAGA	5	GA	1
7	1761334	CTCTCTCTCTCTCTC	15	TC	1
8	1751890	GAGAGAGAGAGAGAG	5	GA	0
9	1083204	TTTTTTTTTTTTTTT	3	T	0
10	1073202	AAAAAAAAAAAAAAA	0	A	0
11	958494	ATATATATATATATA	7	TA	1
12	957864	TATATATATATATAT	7	TA	0
13	391122	AAAGAAAGAAAGAAA	85	GAAA	1
14	387138	TTTCTTTCTTTCTTT	275	TTTC	0
15	382322	AGAAAGAAAGAAAGA	85	GAAA	3
16	379226	AAGAAAGAAAGAAAG	85	GAAA	2
17	379226	GAAAGAAAGAAAGAA	85	GAAA	0
18	377732	TCTTTCTTTCTTTCT	275	TTTC	2
19	374620	TTCTTTCTTTCTTTC	275	TTTC	1
20	374620	CTTTCTTTCTTTCTT	275	TTTC	3

➡ **測定配列全体が、マイクロサテライトであるならば、その測定断片は、対象外とする。**

4．4．マイクロサテライト以外の非常に多い配列

NTT DATA

No	件数	配列パターン
1	122743	GTGTGTGTGTGTGTA
2	122743	ATGTGTGTGTGTGTG
3	122143	TACACACACACACAC
4	122143	CACACACACACACAT
5	93657	TGTGTGTGTGTGTAT
6	93657	TATGTGTGTGTGTGT
7	92357	ATACACACACACACA
8	92357	ACACACACACACATA
9	80888	CTTTTTTTTTTTTTT
10	80888	TTTTTTTTTTTTTTC
11	80257	GAAAAAAAAAAAAAA
12	80257	AAAAAAAAAAAAAAG
13	79363	GTTTTTTTTTTTTTT
14	79363	TTTTTTTTTTTTTTG
15	79329	AAAAAAAAAAAAAAC
16	79329	CAAAAAAAAAAAAAA
17	73675	AACAAACAAACAAAA
18	73675	AAAACAAACAAACAA
19	72783	TTTTGTTTGTTTGTT
20	72783	TTGTTTGTTTGTTTT
21	72518	TGTGTGTATGTGTGT
22	70994	ACACACATACACACA
23	68584	TGTGTATGTGTGTGT
24	68584	TGTGTGTGTATGTGT
25	68353	GTGTGTGTGTGTGTT
26	68353	TTGTGTGTGTGTGTG
27	68235	CACACACACACACAA
28	68235	AACACACACACACAC
29	67362	ACACATACACACACA
30	67362	ACACACACATACACA

マイクロサテライト以外で、非常に多いパターンがどのような配列か確認した。

マイクロサテライト以外で多い配列は、マイクロサテライトから繰り返しパターンがすこし変形したものであった。

次のような可能性が考えられる。

- **マイクロサテライトの読み間違いの可能性。**
- **マイクロサテライトのつなぎ目という可能性。**

このようなパターンとマイクロサテライトだけから構成される測定断片は、対象外とする。

4．5．マッピングアルゴリズムのトライアル②

NTT DATA

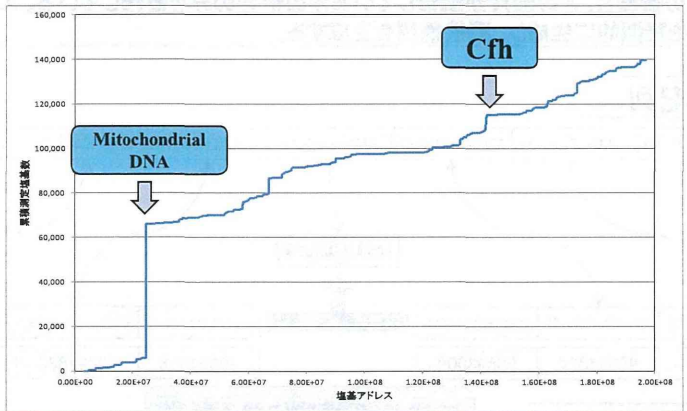
マッピングアルゴリズムの基本部分のみを作成し、全測定断片を用いて、課題の洗い出しを行った。

対象	ペアエンドの一方だけを使用	
	Liver100%サンプル	
	対象配列数	11,795,262
	アダプター配列が見つかった配列&15塩基以上	
	対象配列数	8,764,801
	マイクロサテライト配列除外数	2,461
	マッピングできた配列	4,591,681
	マッピングできたペア(累計)	16,750,719

4. 6. 第一染色体に対するマッピング結果

NTT DATA

マッピングした配列を、塩基アドレスと、マッピング累積塩基数の積み上げ数でグラフを作成した。

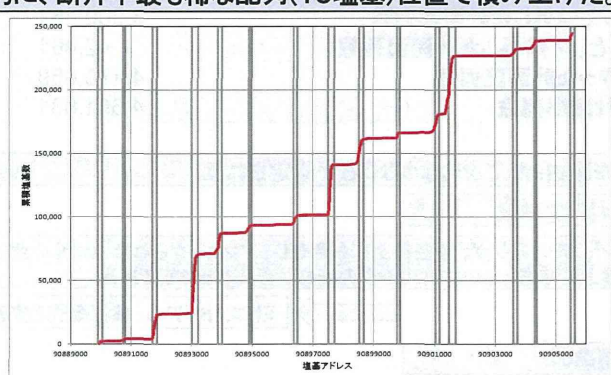


多く合致している箇所の配列を調べたところ、染色体上の遺伝子ではなく、ミトコンドリアのDNAと一致した。

4. 7. 遺伝子Albのマッピング結果

NTT DATA

**Alb近傍のマッピング断片約10万中、2千で累積分布を作成した。
簡易的に、断片中最も稀な配列(15塩基)位置で積み上げた。**



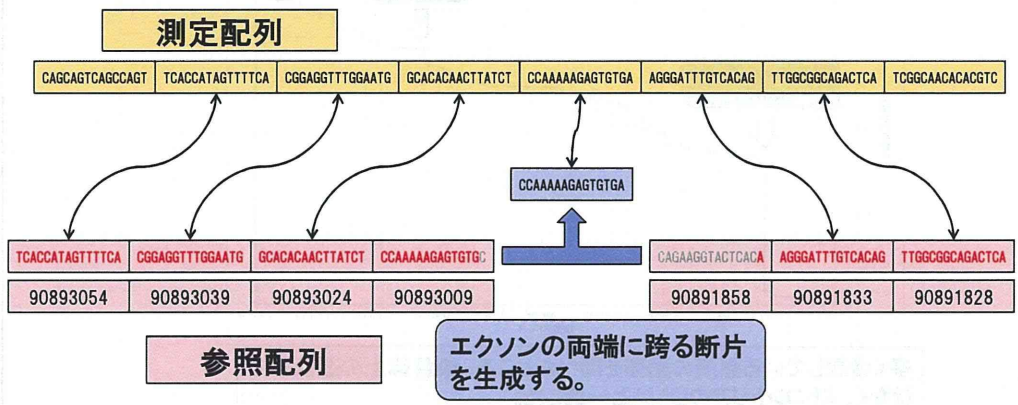
**エクソンで、積み上がっていく様子がうかがえる。
一部、イントロン領域にマッピングされている断片も存在した(スプライシング前のRNAと思われる)。**

4. 8. スプライシング後のRNAを　マッピングさせる方法

NTT DATA

エクソンに跨る配列のマッピングマッピング方法

15塩基の断片の情報と、どの断片が隣接しているかの情報のみを蓄積している。エクソンに跨る断片を疑似的に生成し、隣接情報を生成する。



4. 9. マッピングアルゴリズムのトライアルの考察

NTT DATA

対象計測配列数	11,795,262	
アダプターミスとして除外した計測配列数	3,030,461	25.69%
サイクロサテライトとして除外した計測配列数	2,461	0.02%
マッピングできなかった計測配列数	4,170,659	35.36%
マッピングできた計測配列数	4,591,681	38.93%

マッピングできなかった理由として次のようなことが想定される。

理由①　**ミトコンドリア由来**

ミトコンドリアの配列が、染色体上にも存在し、マッピングされた。しかし、全ての配列が染色体上に存在しない。マッピングの漏れが存在するはずである。

⇒ ミトコンドリアも、参照配列とする。

理由②　**読み間違い**

次世代シーケンサは読み間違いを起こす可能性がある。ペアエンド法は、同じ断片を別方向から読んでいて、精度を向上させる可能性がある。

⇒ ペアエンド法を活用して品質を向上させる。

4. 10. 基本アルゴリズムトライアルにおける課題

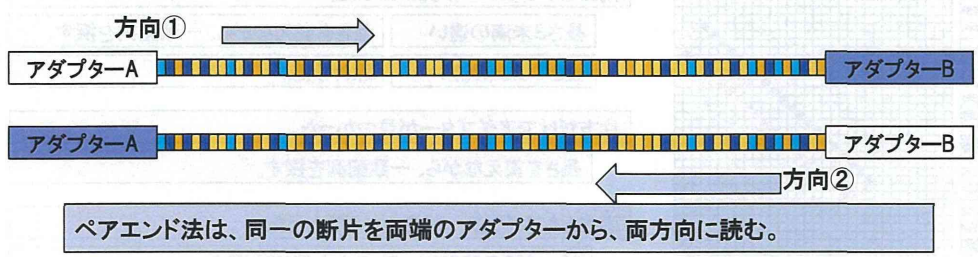
NTT DATA

- 基本アルゴリズムは、マッピングとして動作していることは確認できた。
- ミトコンドリア、リボゾーム配列
  - 染色体上の遺伝子と異なるので、別途参照配列を用意する。
- ペアエンド法
  - シークエンサーは、ペアエンド法を用いて、断片の両側から読んでいる。
  - トライアルにおいては、片方しか使用していない。
  - 両方を用いたものに改良する。

5. 1. ペアエンド処理

NTT DATA

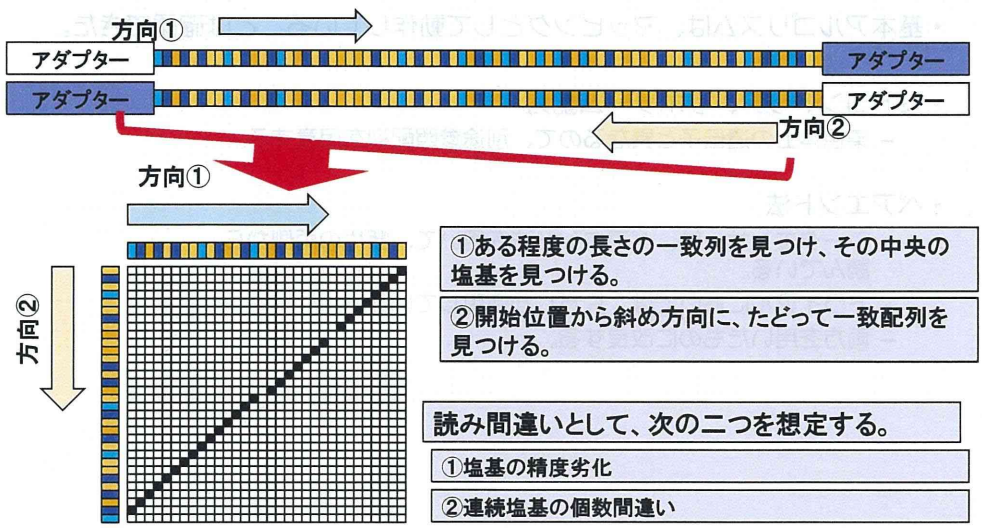
目的	ペアエンド法の両側から読むことで、信頼性の高い配列を推定する。
実装方法	ペア両方のファイルを読み込んで、処理しながらTeradataにロードする。



5. 2. ペアエンド処理：アルゴリズム概要

NTT DATA

二方向からの配列から一致する箇所を見つけ出し、その状況により、適切な断片を求める。



①ある程度の長さの一致列を見つけ、その中央の塩基を見つける。

②開始位置から斜め方向に、たどって一致配列を見つける。

読み間違いとして、次の二つを想定する。

①塩基の精度劣化

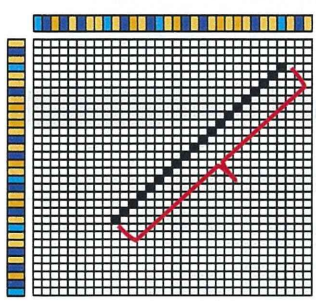
②連続塩基の個数間違い

5. 3. ペアエンド処理：アルゴリズム①

NTT DATA

アダプター検索で場合分けし、一致配列を探す開始位置を見つける。

制約条件なしに一致箇所を探そうとすると、非常に時間がかかる。
一定のレベルの精度で読み取れているのならば、ある程度の長さで一致しているはずである。
そこを起点として調べる。



両方でアダプターが見つかった。

長さ3未満の違い	長さを変えながら、一致配列を探す。
長さ3以上の違い	除外する。

片方だけでアダプターが見つかった。

長さを変えながら、一致配列を探す。

両方ともアダプターが見つからなかった。

30～400の範囲で一致の中心箇所を探す。