Garuda Alliance (snapshot May 2013)

まとめ

- 新型反復暴露解析:
  - 過渡反応と基線反応の基本的な関連性を見いだした。毒性学的に新規性が高くエピジェネティクス等の機序の関与が示唆される。上流にEif2、RICTORなどが想定された。
  - 反復毒性の分子毒性学的理解の促進、及び、単回暴露データからの反復毒性予測に重要。
- 胎児発生過程におけるマスター遺伝子を基軸とした遺伝子発現ネットワークの網羅的解析:
  - 微分解析手法により、網羅的・効率的に発生マスター遺伝子の解析に成功した。
  - 未報告の上流制御系が抽出される可能性が見いだされた。
- Percellome 3次元データ等の為の専用解析ソフトウエアの開発研究
  - 異種動物データの統合技術
  - 絶対量化されていない非Percellomeデータの絶対量推定技術
  - RSort改良
  - WebAPI改良
  - その他
- システムトキシコロジー解析基盤の研究開発:
  - 状態制御遺伝子群推定アルゴリズムAGCT
  - オルソローグ制御領域分析ソフトウエアSHOE
  - 遺伝子制御関係推定アルゴリズム
  - Web公開／Garuda Platform実装／AOP適合／Networkによるスクリーニングパイプライン
  - 次世代Computational Tox Screening System の創出

国立医薬品食品衛生研究所　毒性部　御中

委託研究報告書（STEP13）
次世代シークエンサを利用した遺伝子発現解析の高度化～スプライシングバリアントの識別アルゴリズムの改良

1.平成26年度研究テーマ

次世代シークエンサを利用した遺伝子発現解析の高度化
スプライシングバリアントの識別アルゴリズムの改良

先行研究より、次世代シークエンサを利用した遺伝子発現解析の高度化は、マッピングアルゴリズムの性能向上が欠かせないことが明らかになってきた。特に、未報告のエクソン構造・スプライシングパターンを持つメッセンジャーRNAの検出・定量には、高度なマッピングアルゴリズムの開発が必須である。

本研究では、マッピング精度を高め、スプライシングバリアントの推定・識別を可能とするための技術開発を行う。

2.1.昨年度の成果

マッピングアルゴリズムを作成し、
次世代シーケンサのリード結果をマッピングした。

マウスの全染色体を対象にマッピングを実施した。

ミトコンドリアのRNAが多く見つかった。

↓

ミトコンドリアの配列もマッピング対象とした。

遺伝子以外の領域に大量にマッピングされた。

↓

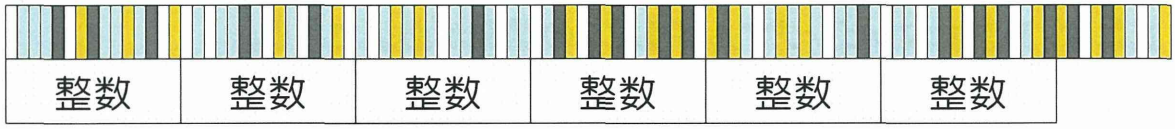
未報告の遺伝子の可能性

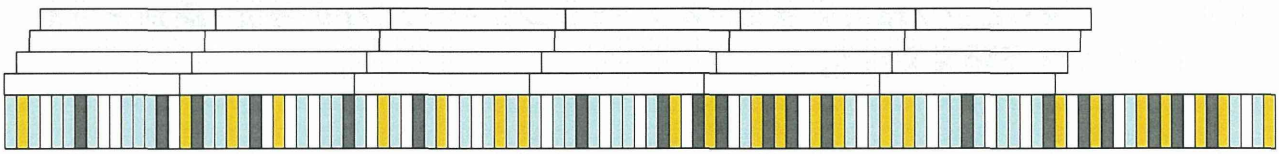
2.2.Teradataを用いたマッピングの基本アイデア

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 RDBMSは、ハッシュインデックスを用いて、完全一致の検索を高速に実行可能である。

計測配列





参照配列

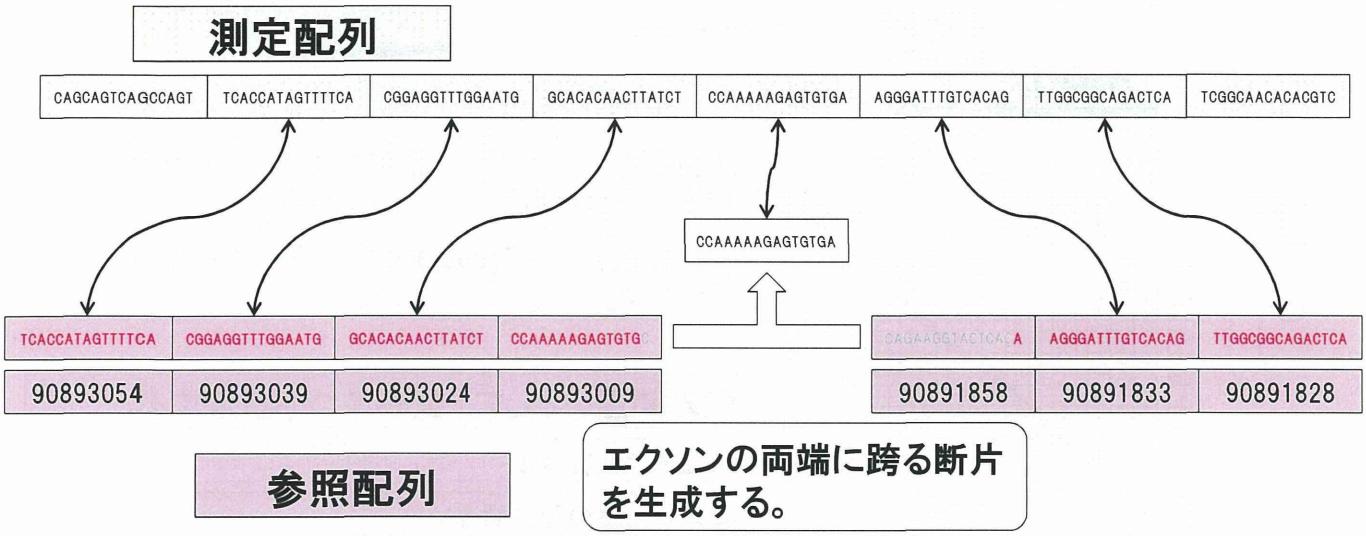
1塩基ずらしたパターンを生成する。

スプライシング前と後の2種類を生成する。

2.3.スプライシング後のRNAをマッピングさせる方法

エクソンに跨る配列のマッピングマッピング方法

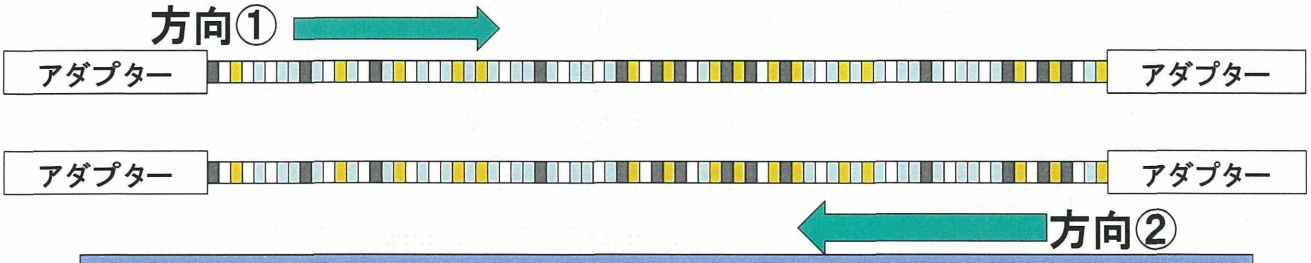
15塩基の断片の情報と、どの断片が隣接しているかの情報のみを蓄積している。エクソンに跨る断片を疑似的に生成し、隣接情報を生成する。



2.4.ペアエンド処理

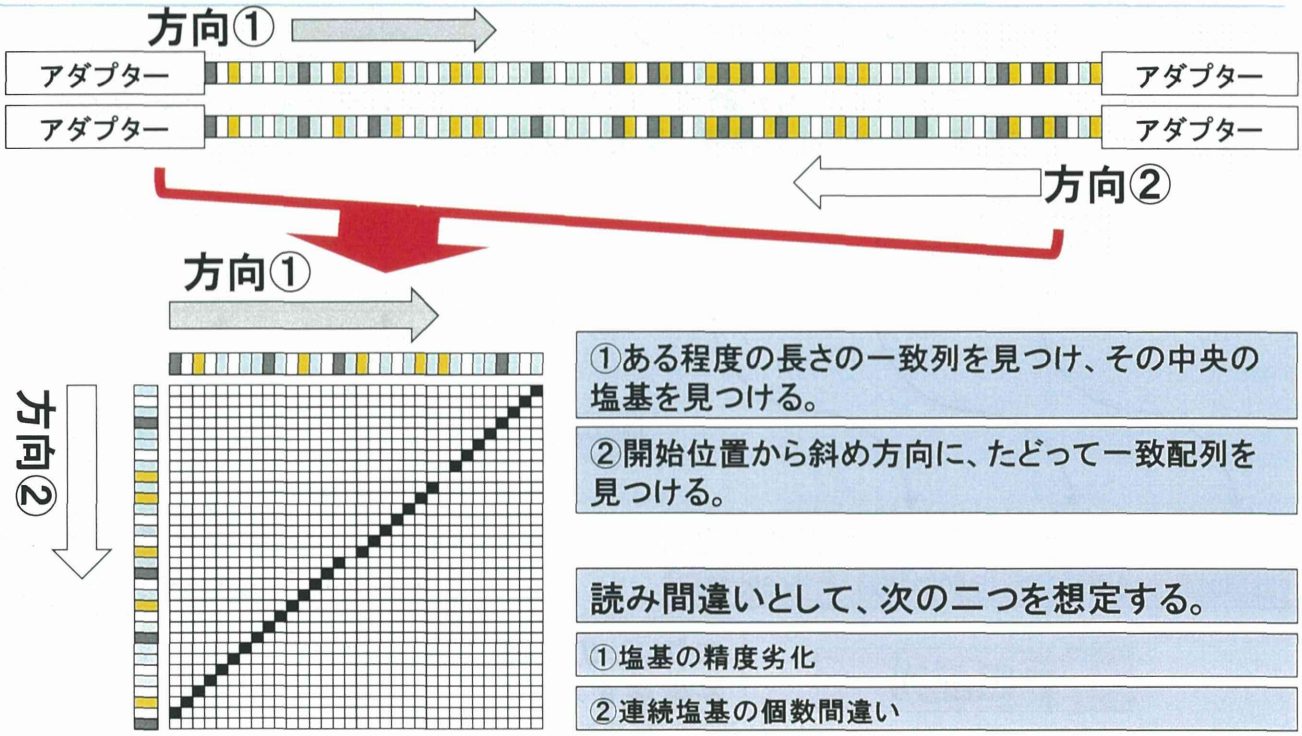
ペアエンド法の両側から読むことで、信頼性の高い配列を推定する。

ペア両方のファイルを読み込んで、処理しながらTeradataにロードしていく。



ペアエンド法は、同一の断片を逆側のアダプターを固定し、それぞれ読む。

2.5.ペアエンド処理：アルゴリズム概要



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。

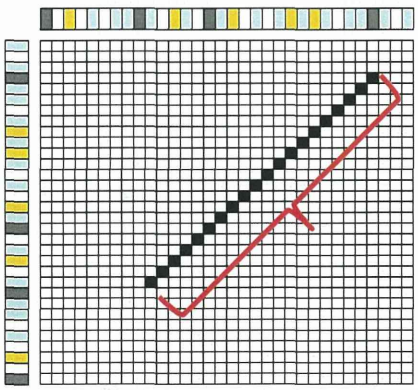
①塩基の精度劣化

②連続塩基の個数間違い

2.6.ペアエンド処理：アルゴリズム①

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。
ある程度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：諦める。

片方だけでアダプターが見つかった。

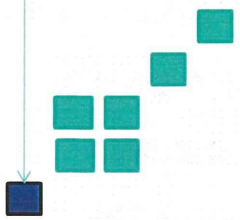
- 長さを変えながら、一致配列を探す。

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。

2.6.ペアエンド処理：アルゴリズム②

(a,b)までマッチしているとする



- (a+1,b+1)をチェックする。
  - 一致すれば、他はチェックせず、進む。
  - 一致しなければ、3か所チェックし、分岐する。
    - (a+2,b+2)をチェックする。
      - 一致すれば、コスト＋1で、進む。
      - 一致しなければ、対角線上でチェックしつづける。
    - (a+2,b+1)をチェックする。
      - 一致すれば、コスト＋3.4で、進む。
      - 一致しなければ、この分岐を諦める。
    - (a+1,b+2)をチェックする。
      - 一致すれば、コスト＋3.4で、進む。
      - 一致しなければ、この分岐を諦める。
    - 3分岐のコストが一番安いものを採用する。

対角できれいに並んでいる場合には、分岐せずに、コスト0で進める。

不一致と、塩基ずれの同時発生は諦める。

2.7.ペアエンド処理の効果

アダプターの読み間違いによる長さ不定による棄却の減少

塩基の読み間違いによる非マッチの減少

⇨ 読取品質の向上

振動などによる品質低下に対する抑止
- 振動などで特定位置での読取品質が低下する可能性がある。
- 読み始めから、何塩基目かで集中する。
- 逆方向からは、長さは集中しない。
- 品質の良い方を採用することで、全体の品質を向上させる。

3.1.今年度の方針

1. マッピング精度の向上
   - 非マッチリード配列の検証
     - エクソン情報の追加
     - 参照配列の補正　など

2. マッピングの高速化
   - 対象配列を絞り込むことで実施

→ 複数の参照配列を使用できるようにする。
- 新たな知見で、配列を修正する。
  - マッピング精度の向上
- 対象配列を絞り込むことで実施する。
  - マッピングの高速化

3.2.高精度化のためのマッピングエラー配列の特定

特定方法①

大量に存在するはずの遺伝子　Alb、Apoeなど

↓

エクソン中の塩基配列　30塩基程度

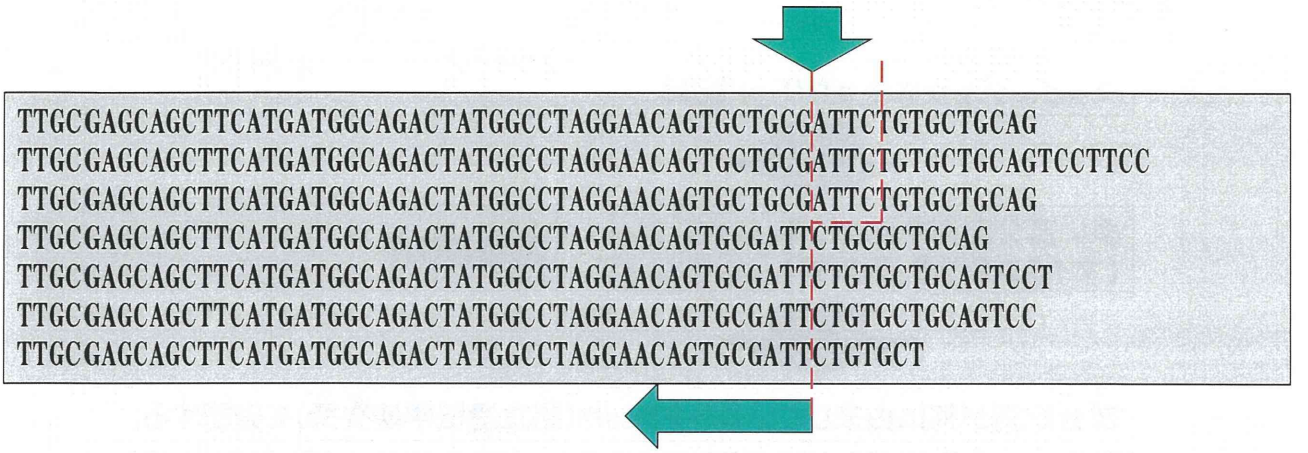
↓

一致したリード配列で、マッピングできなかった配列を見つけ出す。

どのような状況であれば、一致するのかを検討する。

3.3.大量に存在するはずの遺伝子例

Apoa2のエクソンに一致するリードの一部



ここより左は一致しているが、3塩基が挿入がエクソンの跨り部分で差異となっている。

3.3.大量に存在するはずの遺伝子例

Apoa2のエクソンに一致するリードの一部

マッチ失敗

TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCCTTCCCG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCCTTCCC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCCT
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCCT
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTCC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAGTC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTGTGCTGCAG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCGATTCTG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTCTG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGAT
TTCATGATGGCAGACTATGGCCTAG

これより左側は完全一致

マッチ成功

TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCGATTCTGTGCTGCAGTCCTTCC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCGATTCTGTGCTGCAGTCCTTC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCGATTCTGTGCTGCAG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCGATTCTG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCGATTC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTGCTGGTG
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATTC
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCGATT
TTCATGATGGCAGACTATGGCCTAGGAACAGTGCTG
TTCATGATGGCAGACTATGGCCTAGGAACAGTG
TTCATGATGGCAGACTATGGCCGATTCTGT
TTCATGATGGCAGACTATGGCCTAGGAAC
TTCATGATGGCAGACTATGGCCTAGGAAC
TTCATGATGGCAGACTATGGCCTAGGAAC
TTCATGATGGCAGACTATGGCC
TTCATGATGGCAGACTATGGCC
TTCATGATGGCAGACTATGG

失敗リードは、成功リードと比較して、3塩基少ないパターンが多い。

この手順により、一致しなかったリードのパターンを発見可能である。

3.4.高精度化のためのマッピングエラー配列の特定

特定方法②

マッピングできなかったリード配列

↓

多い部分配列を見つける。
（マイクロサテライト以外）

30塩基程度

↓

部分配列が何に由来しているかをblastn(国立遺伝学研究所)で確認する。

Teradata内に格納された染色体内の全配列との一致箇所を確認する。

どのような状況であれば、一致するのかを検討する。

4.5.高精度化のためのマッピングエラー配列の特定

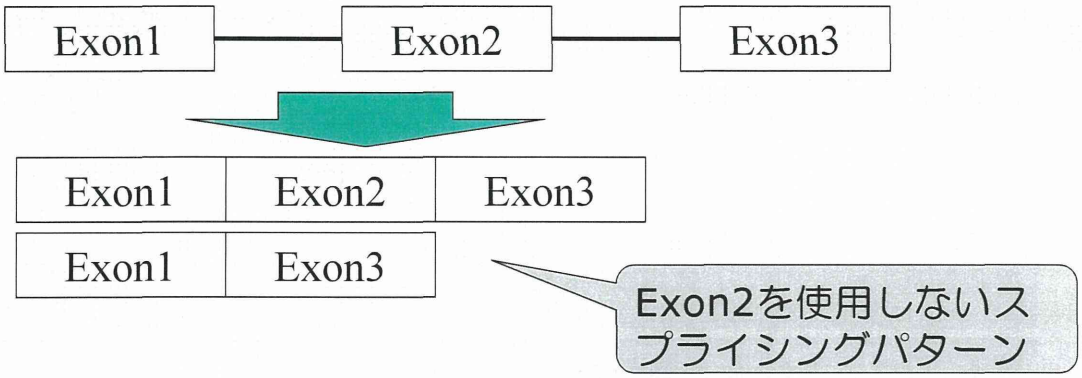
マッチしなかったリード配列の部分配列で多かった上位30

No	配列	個数	blastn description
1	CTTAGGTATAGTAAATGTATTGAATCCATCATACGTCACAGAATA	14340	Major Urinary Protein MUP8 MUP11 chr4
2	AGAATATTCACCAGCCTTTTCTGTTTTGTCAGCAACCATAGATAA	12217	Major Urinary Protein MUP8 MUP11 chr4
3	AGGAAGGGATGATGGTGGAGTCCTGGTGAGAAGTCTCCACTCAAC	11338	Major Urinary Protein MUP8 MUP11 chr4
4	GTCTCCACTCAACACTGGAGGCTCAGGCCATTCTTCATTCTCGGG	9811	Major Urinary Protein MUP8 MUP11 chr4
5	GGTGGAGTCCTGGTGAGAAGTCTCCACTCAACACTGGAGGCTCAG	9531	Major Urinary Protein MUP8 MUP11 chr4
6	TTGGCATTGGATAGGTCAATGATATTTTCTCTAAGGATTCCATGC	8315	Major Urinary Protein MUP8 MUP11 chr4
7	CTTGAACTCTCTCTTCAAAGTTCTTTTCAACTTTCCCTTACGGTA	7885	45S pre rRNA gene / 28S ribosomal RNA
8	GCCCTCTTGAACTCTCTCTTCAAAGTTCTTTTCAACTTTCCCTTA	7877	45S pre rRNA gene / 28S ribosomal RNA
9	CAGAATATTCACCAGCCTTTTCTGTTTTGTCAGCAACCATAGATA	7654	Major Urinary Protein MUP8 MUP11 chr4
10	CTGGGGATGCTGTATGGATAGGAAGGGATGATGGTGGAGTCCTGG	6890	Major Urinary Protein MUP8 MUP11 chr4
11	CTCTTGAACTCTCTCTTCAAAGTTCTTTTCAACTTTCCCTTACGG	6511	45S pre rRNA gene / 28S ribosomal RNA
12	GAATATTCACCAGCCTTTTCTGTTTTGTCAGCAACCATAGATAAT	6406	Major Urinary Protein MUP8 MUP11 chr4
13	GCATTGGATAGGTCAATGATATTTTCTCTAAGGATTCCATGCTCC	6259	Major Urinary Protein MUP8 MUP11 chr4
14	CCATCAGCTGGAAGGTTTCCCCATCCTTTTCGTTAATGAGATGAG	5808	Major Urinary Protein MUP8 MUP11 chr4
15	GTCACAGAATATTCACCAGCCTTTTCTGTTTTGTCAGCAACCATA	5803	Major Urinary Protein MUP8 MUP11 chr4
16	GGAGGGGCGGCCCGGCCCCCACCCCCACGCCCGCCCGGGAGGCGG	5784	45S pre rRNA gene / 28S ribosomal RNA
17	AATATTCACCAGCCTTTTCTGTTTTGTCAGCAACCATAGATAATT	5585	Major Urinary Protein MUP8 MUP11 chr4
18	CTCTCTTCAAAGTTCTTTTCAACTTTCCCTTACGGTACTTGTTGA	5449	45S pre rRNA gene / 28S ribosomal RNA
19	CCGGGAGGCGGACGGGGGGAGAGGGAGAGCGCGGCGACGGGTATC	5447	45S pre rRNA gene / 28S ribosomal RNA
20	CAAACCTTTCCTTGATGTCTGAACTCAAATCTGGTTCTCGGCCAT	5275	Major Urinary Protein MUP8 MUP11 chr4
21	CGGGAGGCGGACGGGGGGAGAGGGAGAGCGCGGCGACGGGTATCT	5202	45S pre rRNA gene / 28S ribosomal RNA
22	GTGAGAAGTCTCCACTCAACACTGGAGGCTCAGGCCATTCTTCAT	5084	Major Urinary Protein MUP8 MUP11 chr4
23	GCCATTATCTTCTATCTTTTCTCTTTTGTCAGAGGCCAGGATAAT	5070	Major Urinary Protein MUP8 MUP11 chr4
24	TGCCATTATCTTCTATCTTTTCTCTTTTGTCAGAGGCCAGGATAA	5042	Major Urinary Protein MUP8 MUP11 chr4
25	TTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGTGGTC	4993	Mus musculus clone 28-385 MMuLV retroviral integration site genomic sequence.
26	CACTGGAGGCTCAGGCCATTCTTCATTCTCGGGCCTGGAGGCAGC	4784	Major Urinary Protein MUP8 MUP11 chr4
27	CCCTCTTGAACTCTCTCTTCAAAGTTCTTTTCAACTTTCCCTTAC	4728	45S pre rRNA gene / 28S ribosomal RNA
28	GTCCTAAGGCGAGCTCAGGGAGGACAGAAACCTCCCGTGGAGCAG	4726	45S pre rRNA gene / 28S ribosomal RNA
29	GGAACGGGGAGCGGGAAAGATCCGCCGGGACCACCGGCACGGCCG	4720	45S pre rRNA gene / 28S ribosomal RNA
30	GGATGATGGTGGAGTCCTGGTGAGAAGTCTCCACTCAACACTGGA	4654	Major Urinary Protein MUP8 MUP11 chr4

MUPとRibosomalRNAが多い、配列、エクソン情報の誤りの可能性がある。

4.6.昨年度のアルゴリズムの未達成事項

スプライシングバリアントの想定



いくつかの遺伝子では、エクソンパターンが多く存在しており、このようなエクソンを使用していないパターンも網羅しているケースも存在した。

エクソンを飛ばした、跨りのパターンを追加する。

5.参照用配列

参照配列として、以下の配列を用いる。

- **ダウンロードサイト**
  - Cufflinksサイト
  - http://cufflinks.cbcb.umd.edu/igenomes.html
  - Mus musculus/UCSC/mm10　　Jun 14 11:29
- **対象**
  - **Mm10内の情報**
  - **総配列**
    - Chromosome
  - **遺伝子アドレス情報**
    - Ref_text