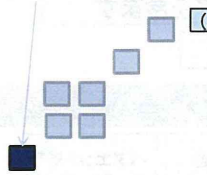
5. 4. ペアエンド処理:アルゴリズム②

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。

(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、この分岐を除外する。
    - (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、この分岐を除外する。
    - 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のケースは除外する。

5. 5. ペアエンド処理の効果

NTT DATA

ペアエンド処理を行うことにより次の効果が期待できる。

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**

測定中の振動などによる品質低下に対する抑止
測定中の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。
このケースでは読み始めから、何塩基目かに集中する。
逆方向からの読み取りでは、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。

5. 6. マッピングアルゴリズムのトライアル①

NTT DATA

ペアエンド処理によるどのように変化したかを確認した。

対象　Liver100%サンプル　ミトコンドリアを含まず

ロード時間(ペアエンド処理を含む):22時間34分

対象配列数	11,795,262	
	ペアエンド処理を行った場合	ペアエンド処理を行っていない場合
対象配列数	10,238,689	8,764,801
マイクロサテライト配列除外数	5	2,461
マッピングできた配列	5,802,090	4,591,681
マッピングできたペア(累計)	20,549,885	16,750,719

一方だけを使用した場合と比べ、マッピング対象とできる配列数が増加し、精度が高まったことを示している。

5. 7. マッピングアルゴリズムのトライアル②

NTT DATA

ペアエンド処理によるどのように変化したかを確認した。

対象　Liver100%サンプル　ペアエンド処理を実施

ロード時間(ペアエンド処理を含む):22時間34分

対象配列数(ロード前)	11,795,262	
	ミトコンドリア参照配列を含む	ミトコンドリア参照配列を含まない
対象配列数(ペアエンド考慮ロード後)	10,238,689	
マイクロサテライト配列除外数	5	
マッピングできた配列	6,183,665	5,802,090
マッピングできたペア(累計)	21,153,800	20,549,885

ミトコンドリアを参照配列として使用することにより、マッピングできた断片が増加した。

5. 8. エラー断片の状況

NTT DATA

ペアエンド処理で、不整合となったケース数(断片数)

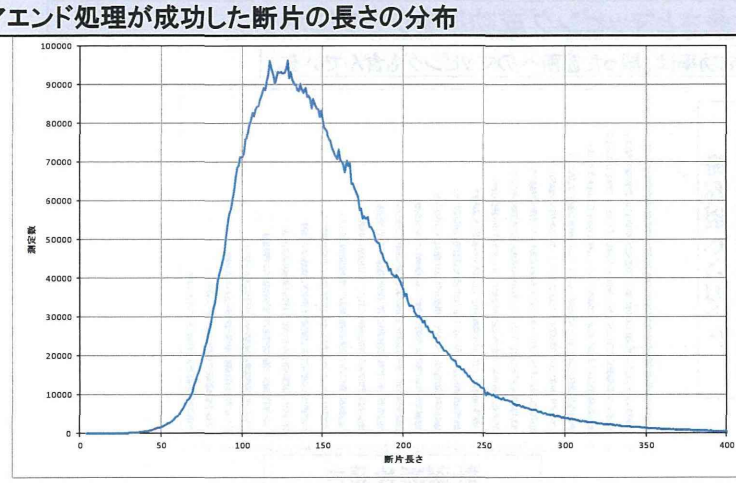
No	エラー理由	原因など	断片数
1	両方ともアダプターが見つからず、スキャンしたが、有効な経路が見つからない	もっと長いなど	375,517
2	ペア1だけでアダプターが見つかったが、有効な経路が見つからない	3連続不一致など	54,096
3	ペア2だけでアダプターが見つかったが、有効な経路が見つからない	3連続不一致など	11,980
4	両方のアダプターが見つかって、長さ同じだが、有効な経路が見つからない	3連続不一致など	8,373
5	両方のアダプターが見つかったけど、長さが3以上違う		267
6	有効な経路が途中で見つからなかった1	3連続不一致など	586,269
7	有効な経路が途中で見つからなかった2	3連続不一致など	510,352
8	マイクロサテライトの微妙な崩れ1		3,281
9	マイクロサテライトの微妙な崩れ2		5,749
	合計		1,555,884

ペアエンド処理で失敗した分も、何らかの方法を考えるべきであるが、ペアエンド処理を通過した精度の高い断片を優先して対処法を構築すべきである。

5. 9. ペアエンド処理済み断片長さの分布

NTT DATA

ペアエンド処理が成功した断片の長さの分布

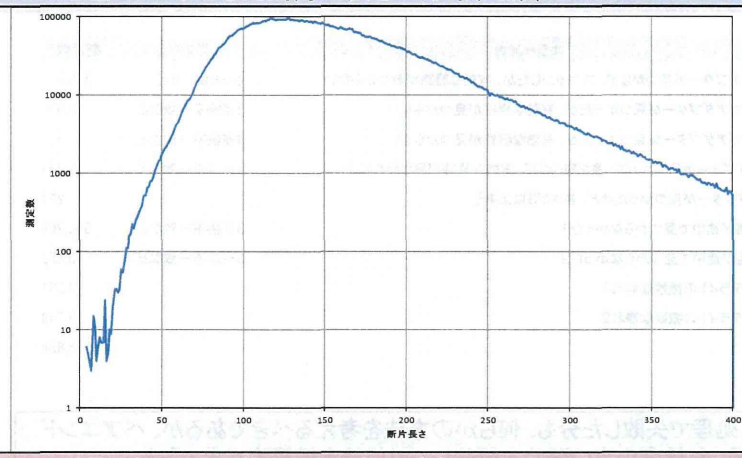


120塩基を中心に分布している。

5. 9. ペアエンド処理済み測定長さの分布 (対数)

NTT DATA

ペアエンド処理で、マッチした断片の長さの分布(対数)



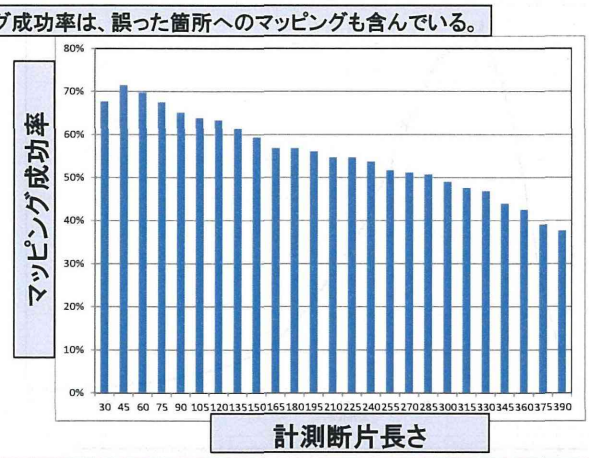
長さ200より長いものは、対数上で直線的に減少しており、べき乗分布にしたがっていると考えられる。このまま継続すると、約2万断片が存在する可能性が高い。

5. 10. 測定断片長さとマッピング成功率

NTT DATA

測定断片長さとマッピング成功率をグラフに示す。

マッピング成功率は、誤った箇所へのマッピングも含んでいる。

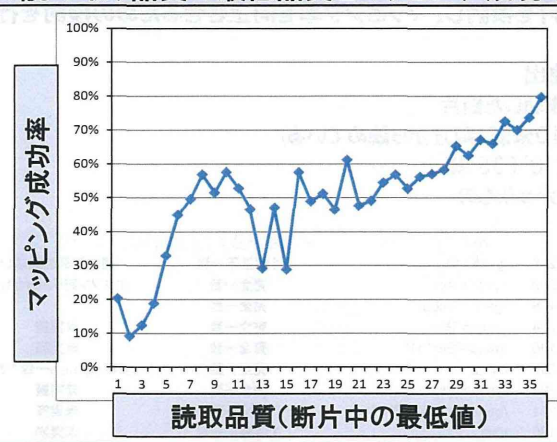


マッピング成功率は、計測断片が長くなると低下する。

5. 10. 読み取り品質とマッピング成功率

NTT DATA

計測断片の読み取り品質の最低品質とマッピング成功率を示す。

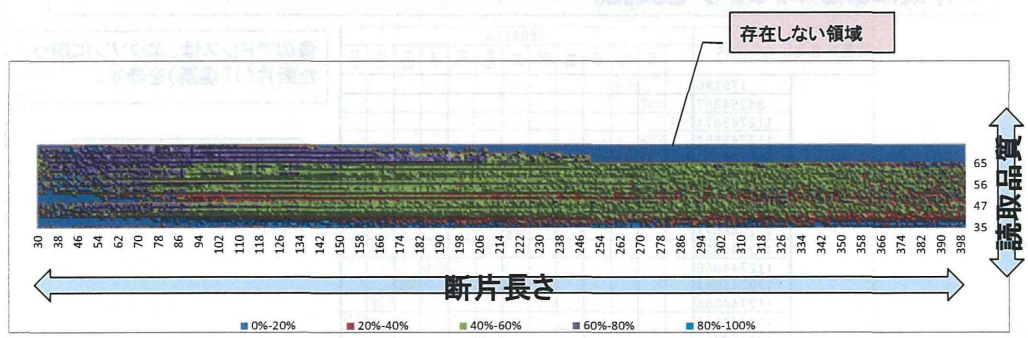


品質がきわめて低い領域では、マッピング成功率は低い。それ以降は徐々に上昇している。品質値13近辺でいったん減少しているのは、測定上の何らかの特異性の現れではないかと思われる。

5. 10. 断片長さと読み取り品質による、 マッピング成功率

NTT DATA

計測断片の読取品質の最低値と断片長さによるマッピング成功率を示す。



品質が高い領域と短い場合に、マッピング成功率が高い。

5. 11. マッピング失敗した断片の検討

NTT DATA

- マッピング失敗した断片を検討し、マッピング率を向上させるための検討を行った。
- 次の条件でサンプル抽出
  - ペアエンド処理に成功した断片
  - 長さ196以上～210未満（両方から読めている）
  - 最低読み取り品質 'C'（35）以上
  - マッピングができなかったもの

断片番号	最低読取品質	断片長さ	BLASTN		手作業によるマッピング
3605215	D	206	/gene="Hpxn"	1塩基不一致	測定配列の間違い
5730742	D	202	/gene="Alb"	完全一致	エクソン跨りがおかしい
1013566	C	198	/gene="Acbd5"	完全一致	
1223316	C	204	/gene="Rdx"	完全一致	未実施
1547480	C	196	/gene="Eef1a1"	完全一致	未実施
3545347	C	208	chr12	完全一致	Retrovirul sequence一致する配列なし
4168977	C	201	/gene="Plxnb1"	完全一致	未実施
5042606	C	201	/gene="alpha-1 PI-1"	完全一致	未実施
9845024	C	196	chr12	1塩基不一致	未実施
9993103	C	205	/gene="3110049J23Rik" /gene_synonym="Mawbp"	完全一致	未実施

5. 11. 1. 断片(3605215)

NTT DATA

手作業によるマッピングを実施。

染色体番号	アドレス	3605215												
		1	16	31	46	61	76	91	106	121	136	151	166	181
7	-175290			1										
	84354887		1											
	112743671	1												
	112743686		1											
	112743701			1										
	112743976				1									
	112743991					1								
	112744006						1							
	112744021							1						
	112744036								1					
	112744066										1			
	112744081											1		
	112744096												1	
	112744111													1
	131128386								1					

負のアドレスは、エクソンに跨った断片（15塩基）を表す。

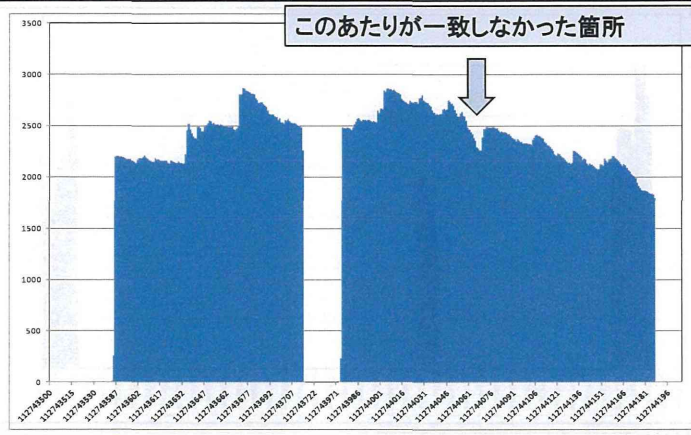
エクソンの結合位置

測定断片121で一致するものがなく、マッチング処理が失敗したと考えられる。

5. 11. 1. 断片(3605215)

NTT DATA

マッピング位置周辺の按分結果を確認した。



不一致箇所周辺で特異的な形状が見られなかったので、断片側の読み間違いと考えられる。

5. 11. 2. 断片(5730742)

NTT DATA

手作業によるマッピングを実施。

染色体番号	アドレス	5730742												
		1	16	31	46	61	76	91	106	121	136	151	166	181
5	18060209				1									
	22643248		1											
	32604959				1									
	37667266													1
	90903663													1
	90903678												1	
	90903693											1		
	90903708										1			
	90904307								1					
	90904322							1						
	90904337						1							
	90905490				1									
	90905505			1										
	90905520		1											
	90905535	1												

エクソンの結合位置

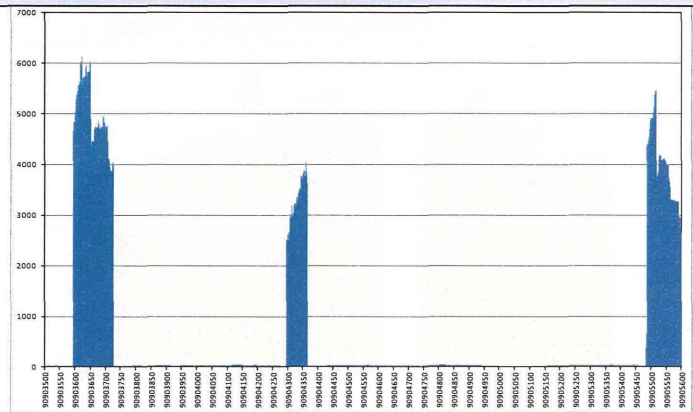
エクソンの結合位置

エクソン跨りの箇所で一致せず、一致しないと判断されたと思われる。

5. 11. 2. 断片(5730742)

NTT DATA

マッピング位置周辺の按分結果を確認した。



エクソン跨りの箇所において一致しなかった。
調査の必要がある。

5. 11. 3. 断片(1013566)

NTT DATA

手作業によるマッピングを実施した。

chr_name	stt1	1013566											
		1	16	31	46	61	76	106	121	136	151	166	181
2	4600535	1											
	22955188												1
	22955203											1	
	22955218										1		
	22955233									1			
	22955248								1				
	22955263							1					
	22965773						1						
	22965788					1							
	22966185			1									
	22966200		1										
	22966215	1											
	125939245			1									

エクソンの結合位置

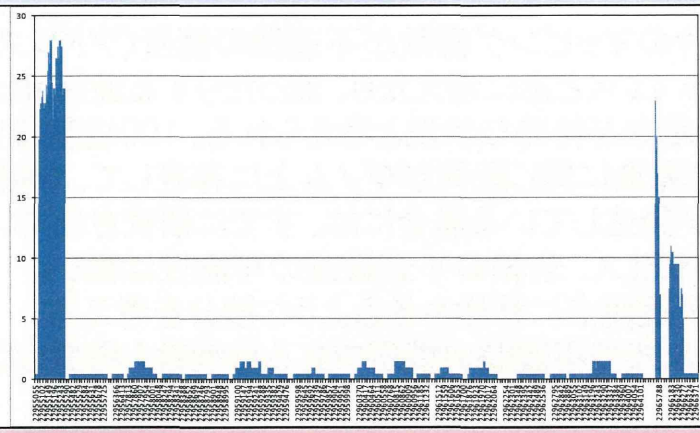
エクソンの結合位置

エクソン跨りの箇所で一致せず、一致しないと判断されたと思われる。

5. 11. 3. 断片(1013566)

NTT DATA

マッピング位置周辺の按分結果を確認した。



エクソン跨りの箇所において一致しなかった。
調査の必要がある。

5. 11. 4. マッチング処理の課題

NTT DATA

逆順におけるエクソン跨りが、適切にマッピングできていない可能性が考えられる。

エクソン跨りに関して、マッピング可能とするような追加が必要と思われる。

6. 1. 按分方法の方針

NTT DATA

測定断片のマッピング個数が不連続の箇所(アドレス方向にみていくと急に増えたり、減ったりする箇所)は、エクソン端など特殊な状況と考えられる。100塩基程度の長さで完全に同じ配列がゲノム上に存在して、特殊な状況が発生している場合には、すでに研究されている箇所が正しく、未知の不連続点の可能性は低い。
既知の不連続点に移動させるような按分を考える。
ただし、未知の不連続の箇所である可能性も考慮しなければならない。

6. 2. 按分方法のアルゴリズム

NTT DATA

次の手順で実施する。

1. 測定配列が参照配列のどの位置に一致するか、全ての一致箇所を挙げる。
2. 各測定配列が複数一致する場合には、按分割合を求める。
   - 2. 1. 初期の按分割合として、次のように決める。
     - 各箇所を平等に按分する。
     - エクソン＞イントロン＞非遺伝子領域で、重みを付ける。
   - 2. 2. 按分割合にしたがって、塩基単位で割り付ける。
   - 2. 3. 塩基割当量の染色体上の変化(アドレス方向の微分？)を、塩基単位で計算する。
   - 2. 4. 変化が大きくても許される箇所(エクソン端)など以外での、異常程度を評価する。
   - 2. 5. 計測配列の末端での、異常程度が、按分における異常程度とみなし、異常程度の大きい合致箇所の按分割合を下げる。