2018.05.29版

機能ゲノム学 第4回

[1]大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究プログラム
[2]微生物科学イノベーション連携研究機構
門田幸二（かどた　こうじ）
kadota@iu.a.u-tokyo.ac.jp
http://www.iu.a.u-tokyo.ac.jp/~kadota/

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

おさらい

マッピングは、マップする側の仮想RNA-seqデータが①sample_RNAseq1.fa、マップされる側のリファレンス配列が② ref_genome.faとして行われた

②

ref_genome.fa メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

①

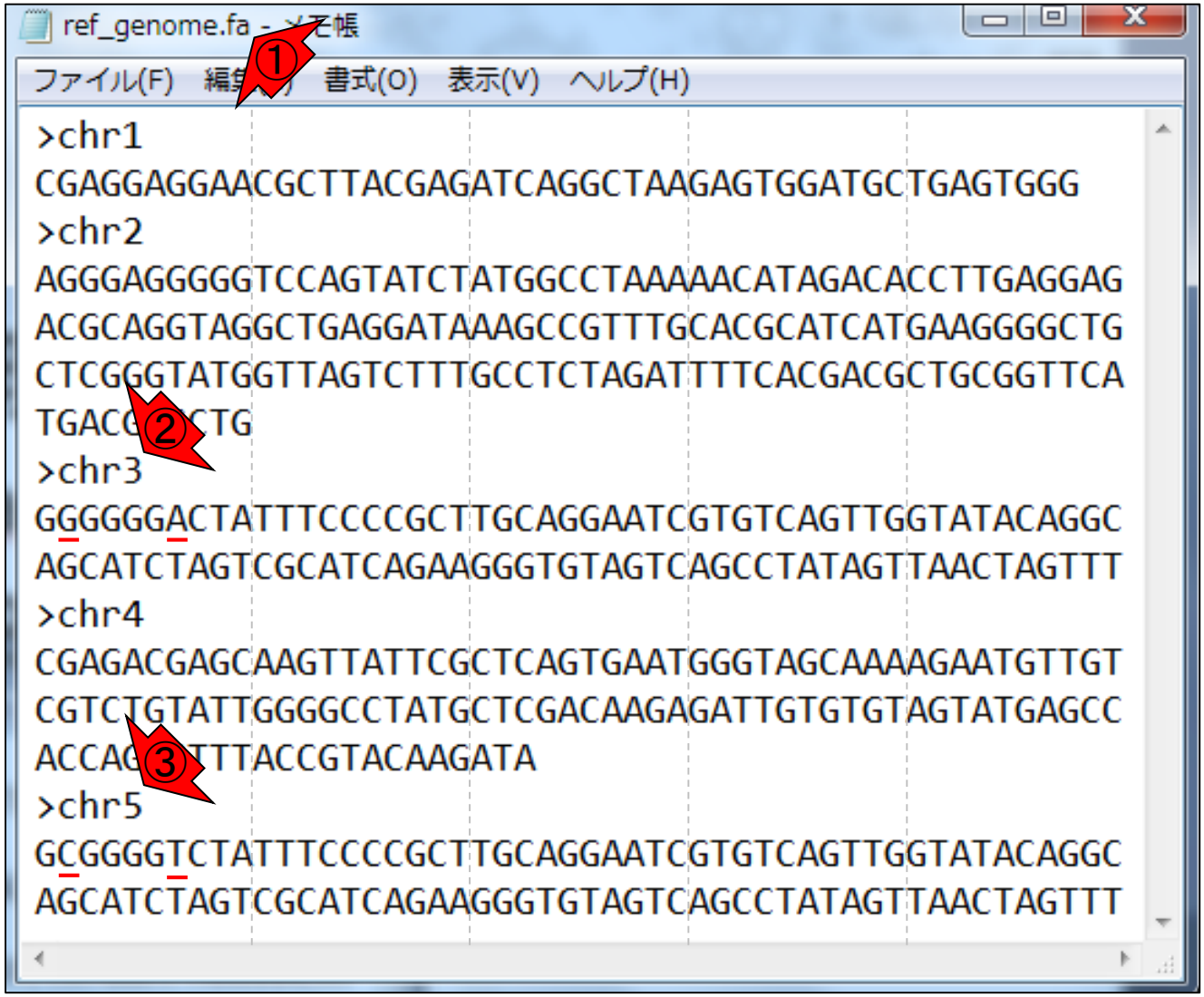
sample_RNAseq1.fa メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

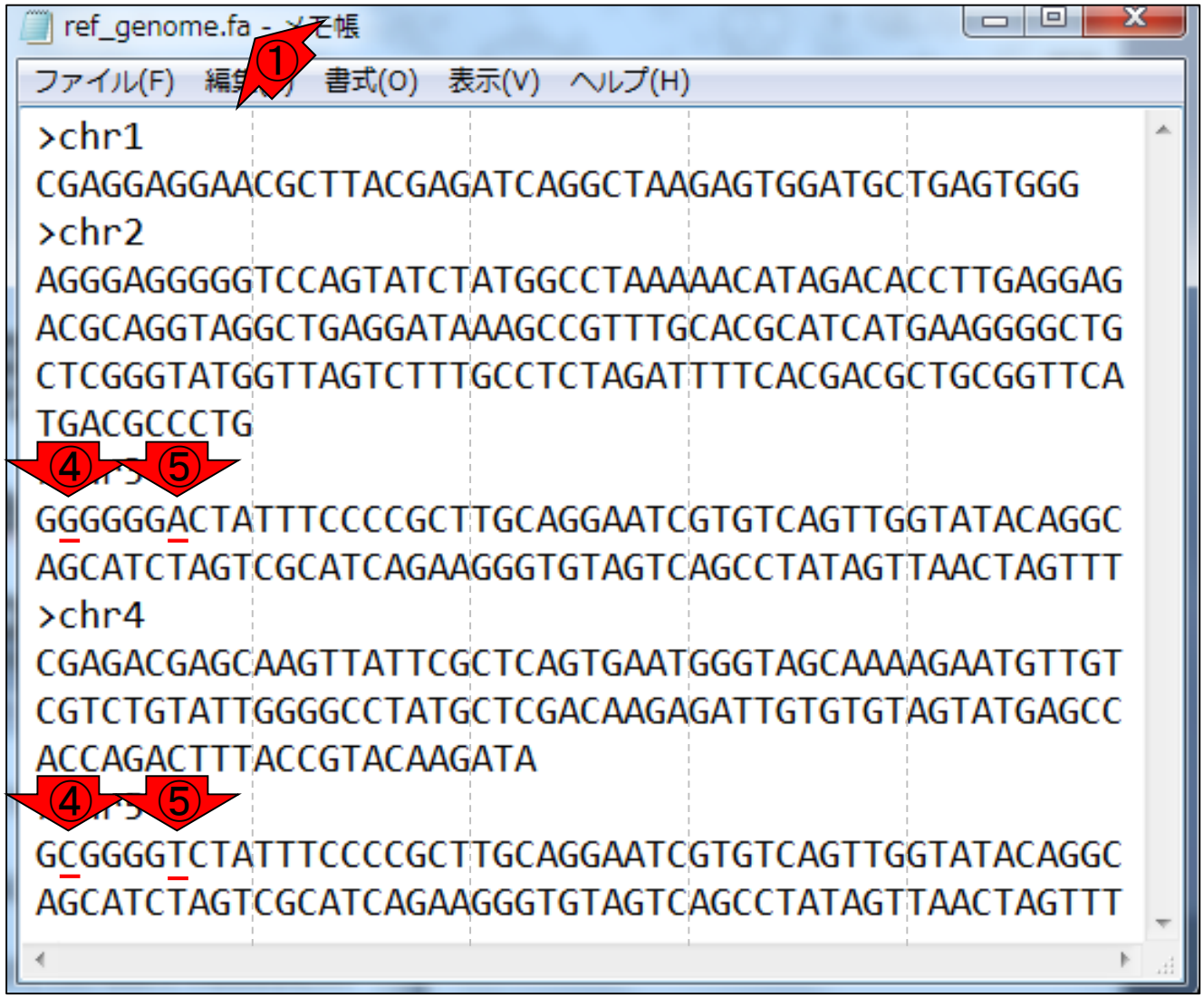
①ref_genome.faのおさらい。②chr3と③chr5の違いは…

マップされる側



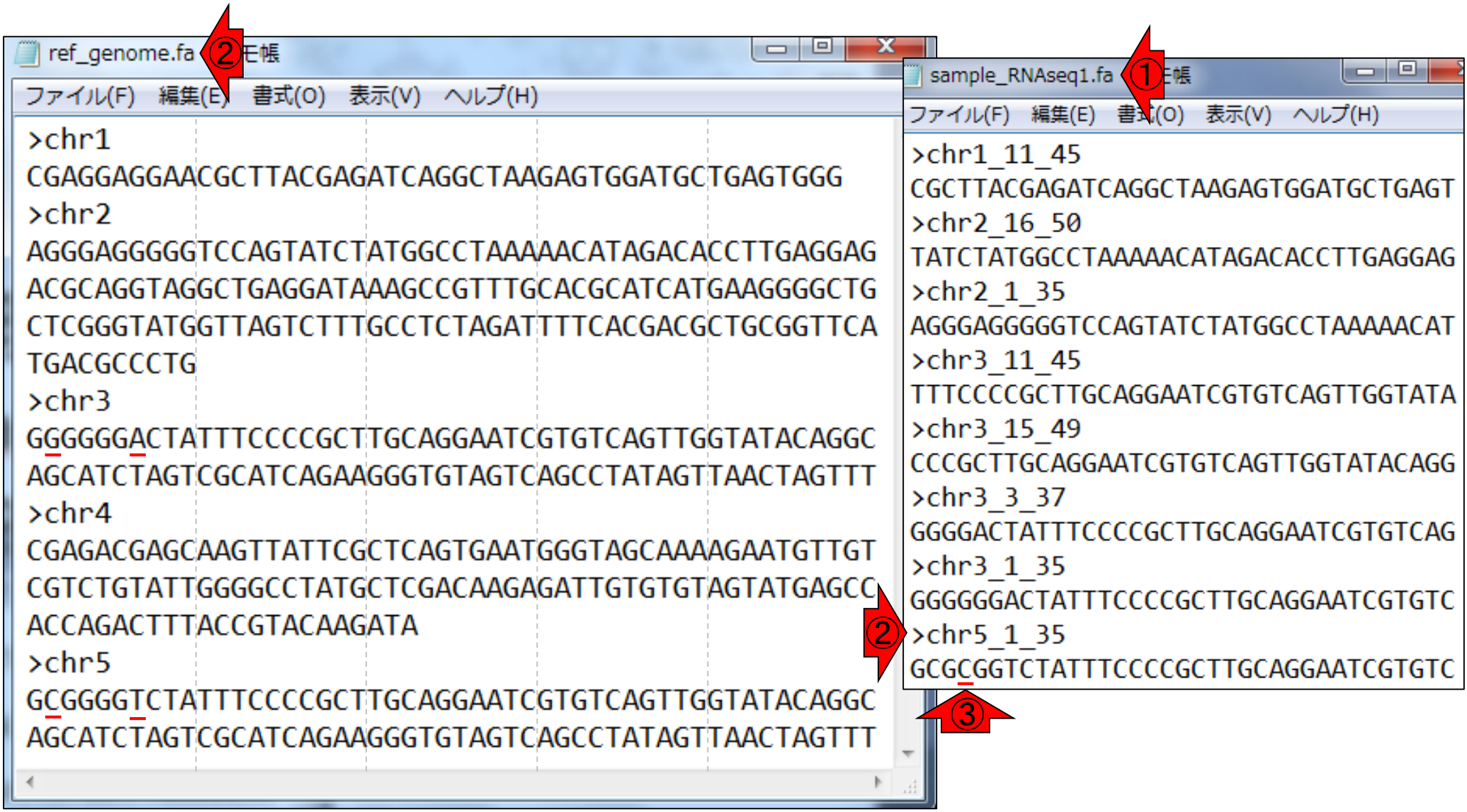
マップされる側

①ref_genome.faのおさらい。②chr3と③chr5の違いは、④2番目と⑤7番目の塩基のみ。従って、8番目の塩基以降は全く同じ



マップする側

①sample_RNAseq1.faのおさらい。全8リードのうち、②最後のリード（chr5_1_35）のみ、③4番目の塩基を変えている



Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

QuasRのマッピング結

①使用したオプション。計8リードのうち、マップされなかったのは赤枠の3リード。Bowtie ver.1に相当するRbowtieというパッケージを利用

- “-m 1 --best --strata -v 0”：0ミスマッチで1か所にのみマップされるリードを出力 ①

ref_genome.fa - メモ帳

ファイル(F)

```
>chr1
CGAGGAGGA                                   TGGATGCTGAGTGGG
>chr2
AGGGAGGGC                                   ATAGACACCTTGAGGAG
ACGCAGGTA                                   GCATCATGAAGGGGCTG
CTCGGGTAT                                   CACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

chr1	11	45
chr2	1	35
chr2	16	50
chr3	1	35
chr3	3	37

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

QuasRのマッピング結果

完全一致でも複数個所にマップされるために落とされたのは2リード

- “-m 1 --best --strata -v 0”：0ミスマッチで1か所にのみマップされるリードを出力

ref_genome.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

QuasRのマッピング結果

1か所にのみマップされるがミスマッチのため落とされたのは1リード

- “-m 1 --best --strata -v 0”：0ミスマッチで1か所にのみマップされるリードを出力

ref_genome.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

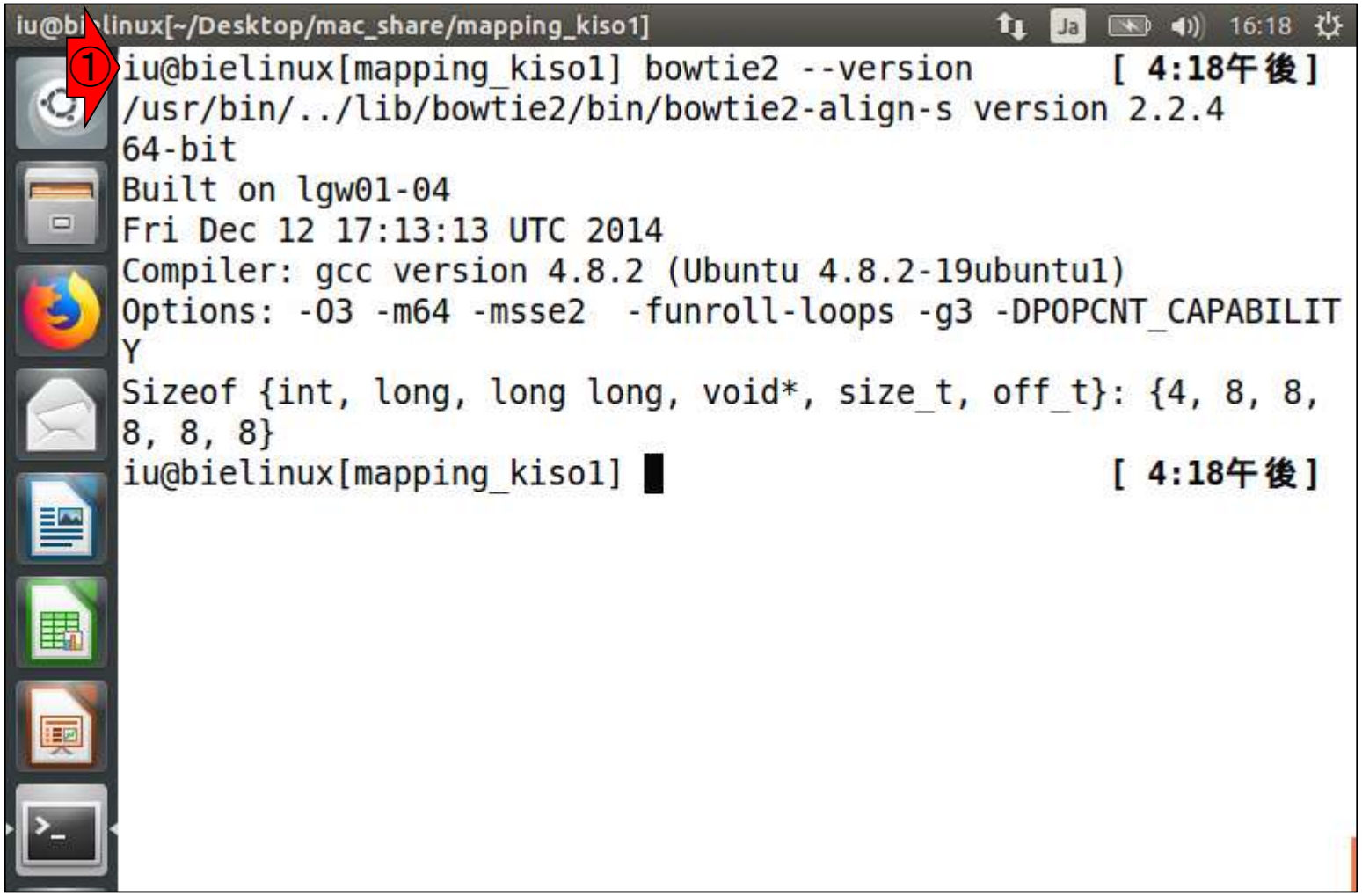
```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

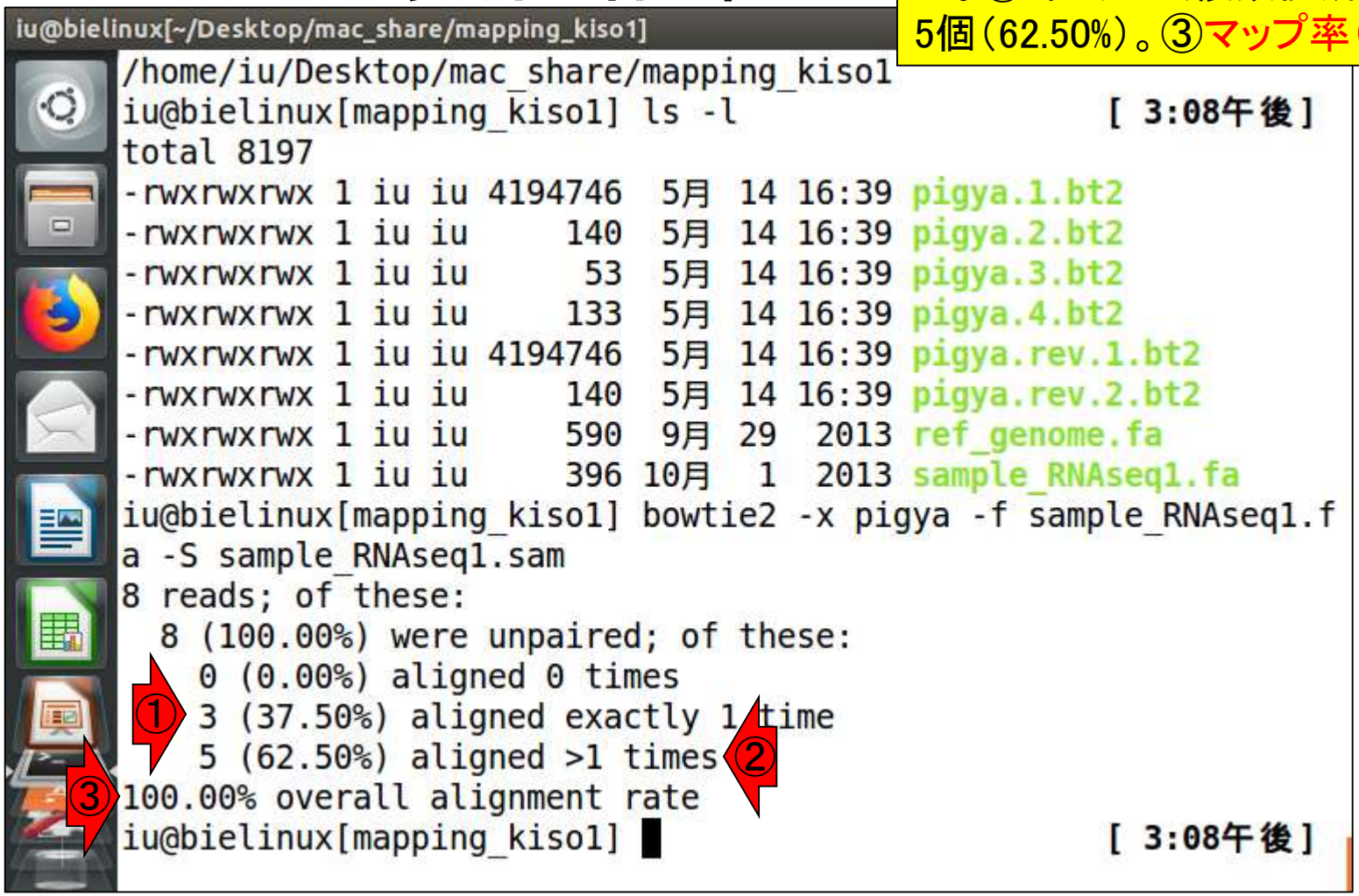
①Bowtie2のバージョンは2.2.4です

bowtie2実行結果



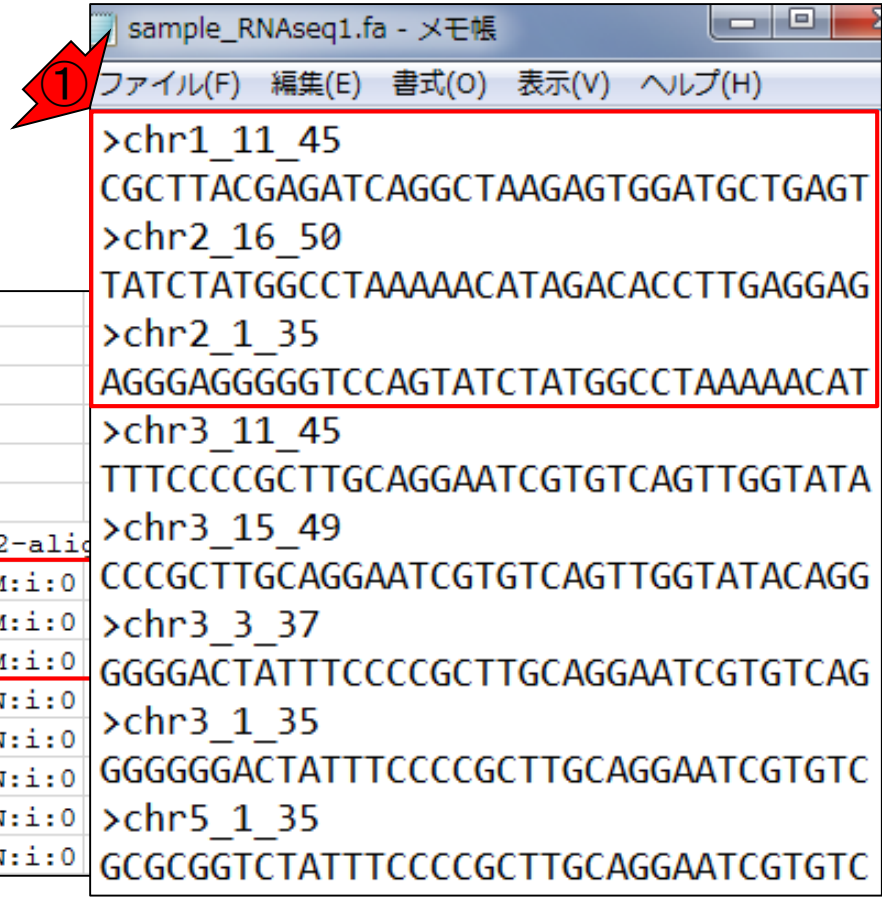
bowtie2実行結果

①1回だけマップされたリードは3個（37.50%）。これは1か所にのみマップされたリードと解釈すればよい。②2回以上（複数個所に）マップされたリードは5個（62.50%）。③マップ率（alignment rate）は100%



Bowtie2実行結果

①1か所にのみマップされた
3リードの、②マッピング結果



②

@HD	VN:1.0	SO:unsorted										
@SQ	SN:chr1	LN:48										
@SQ	SN:chr2	LN:160										
@SQ	SN:chr3	LN:100										
@SQ	SN:chr4	LN:123										
@SQ	SN:chr5	LN:100										
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-alig								
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0

Bowtie2実行結果

①複数個所にマップされた5リードの、②マップピング結果

@HD	VN:1.0	SO:unsorted										
@SQ	SN:chr1	LN:48										
@SQ	SN:chr2	LN:160										
@SQ	SN:chr3	LN:100										
@SQ	SN:chr4	LN:123										
@SQ	SN:chr5	LN:100										
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-alig								
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0

①

②

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

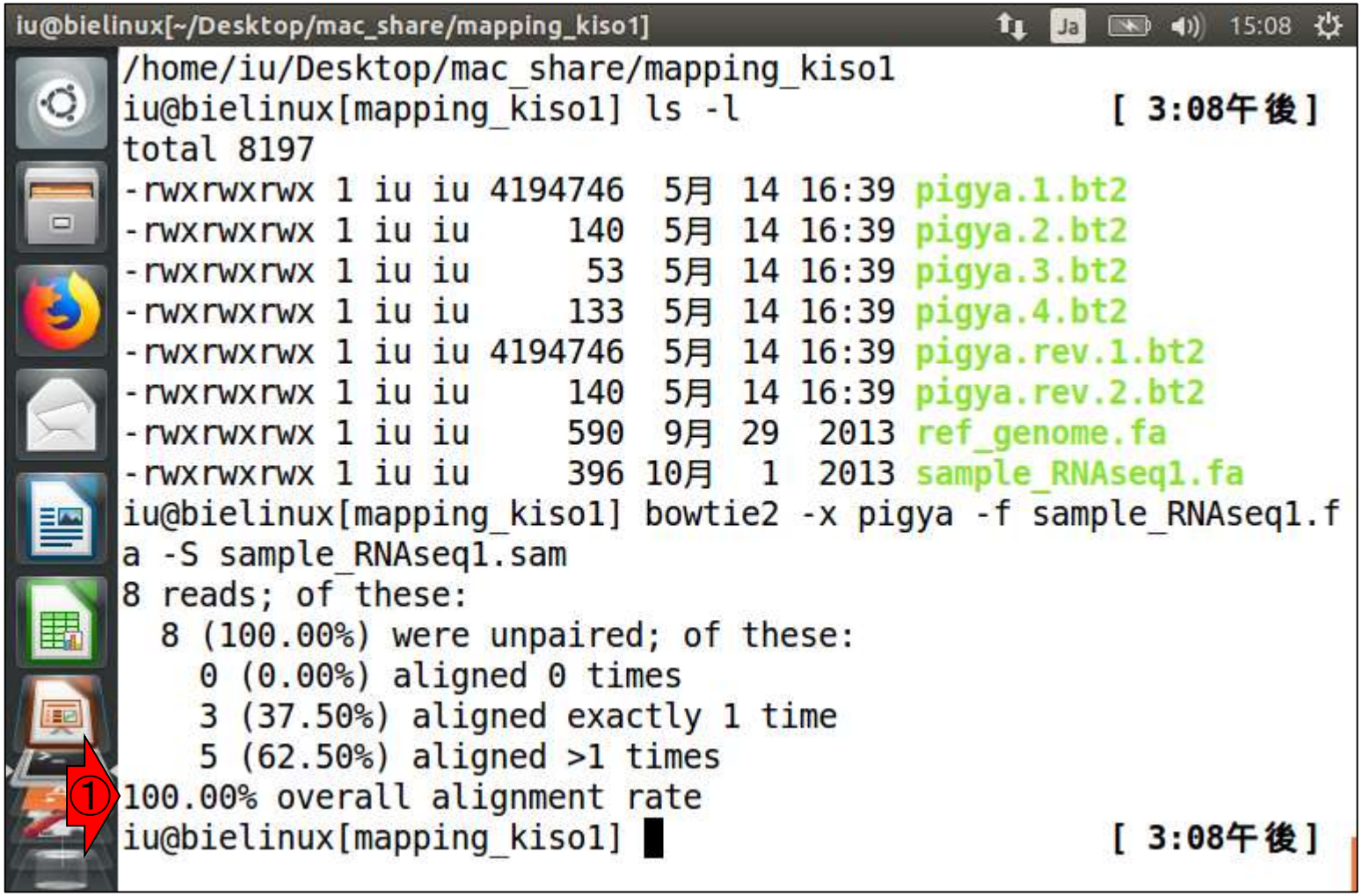
```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

bowtie2実行結果

背景：①Bowtie2では、予期せず全リードがマップされてしまった



マップされなかったリード

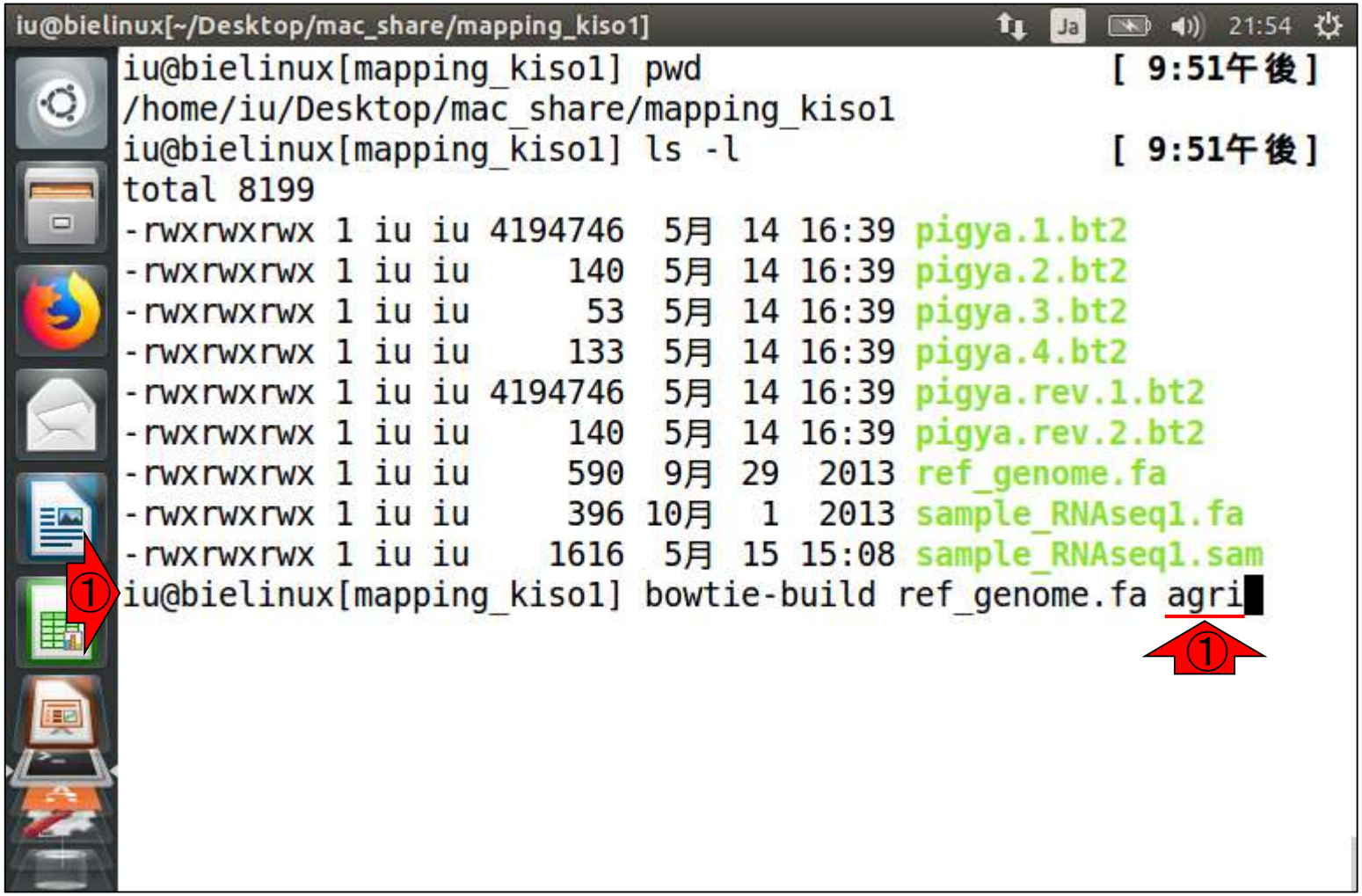
マップされなかったリードは、SAMファイル上でどのような扱いになるのだろう?! マッピングプログラムによって結果が異なることを示すべく、①Bowtie ver. 1をまずはデフォルトオプションで実行してみる

@HD	VN:1.0	SO:unsorted																
@SQ	SN:chr1	LN:48																
@SQ	SN:chr2	LN:160																
@SQ	SN:chr3	LN:100																
@SQ	SN:chr4	LN:123																
@SQ	SN:chr5	LN:100																
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAs														
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU

① Bowtie (Langmead et al., Genome Biol., **10**: R25, 2009)

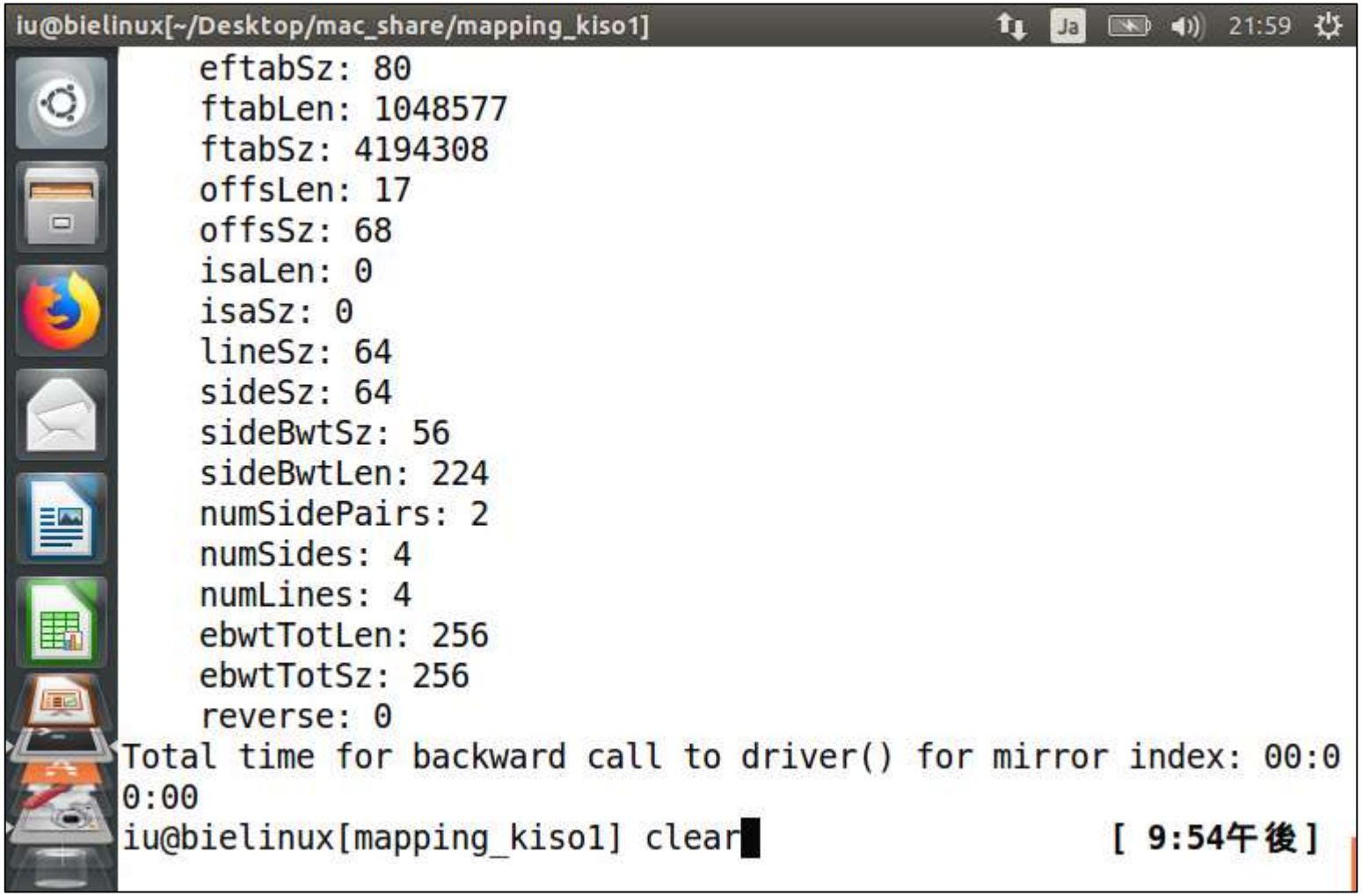
リファレンス配列の前処理

①basenameをagriにして、
②bowtie-buildを実行



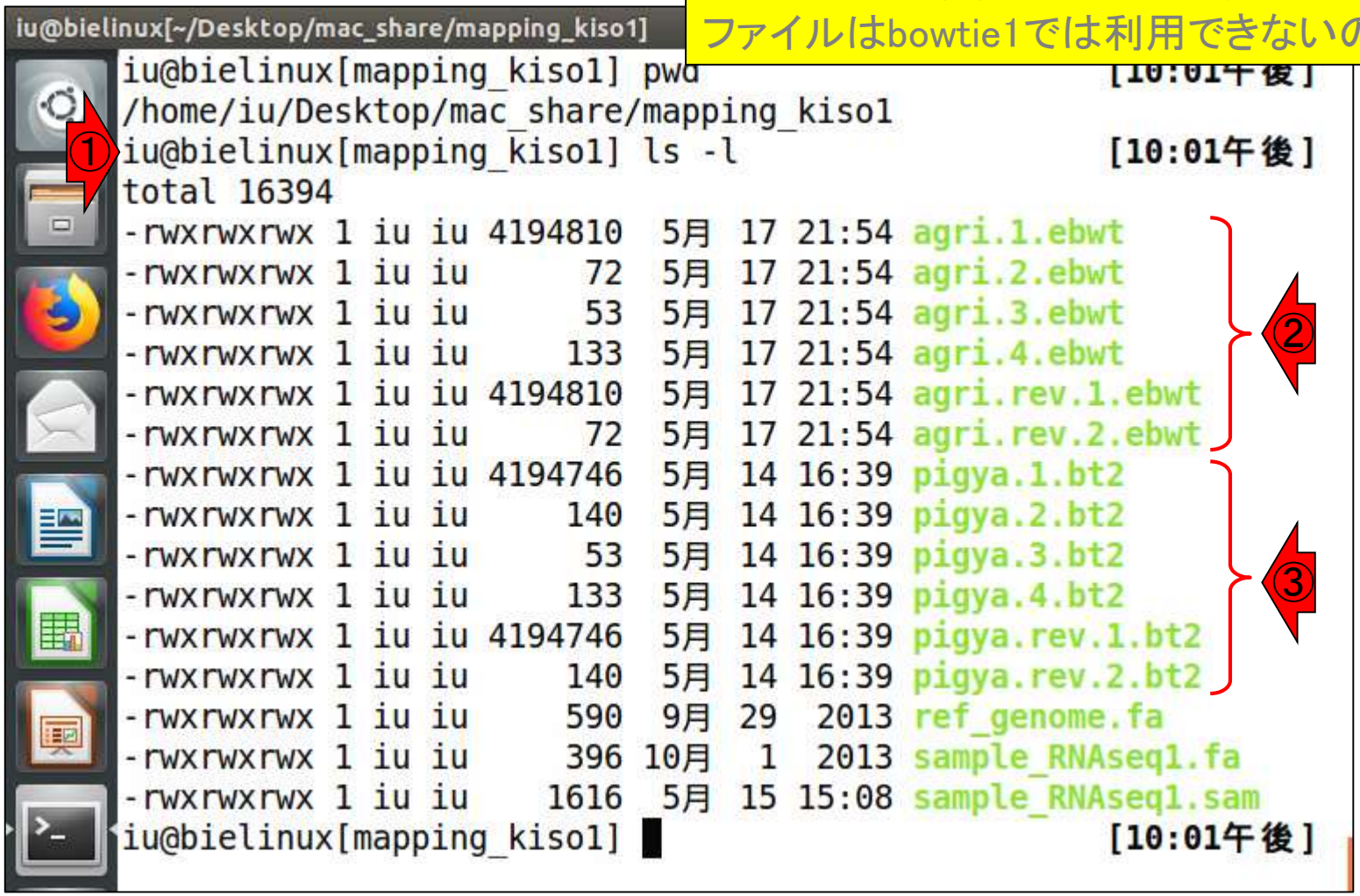
bowtie-build完了

Bowtie-build実行が無事完了したようだ。①clearして…



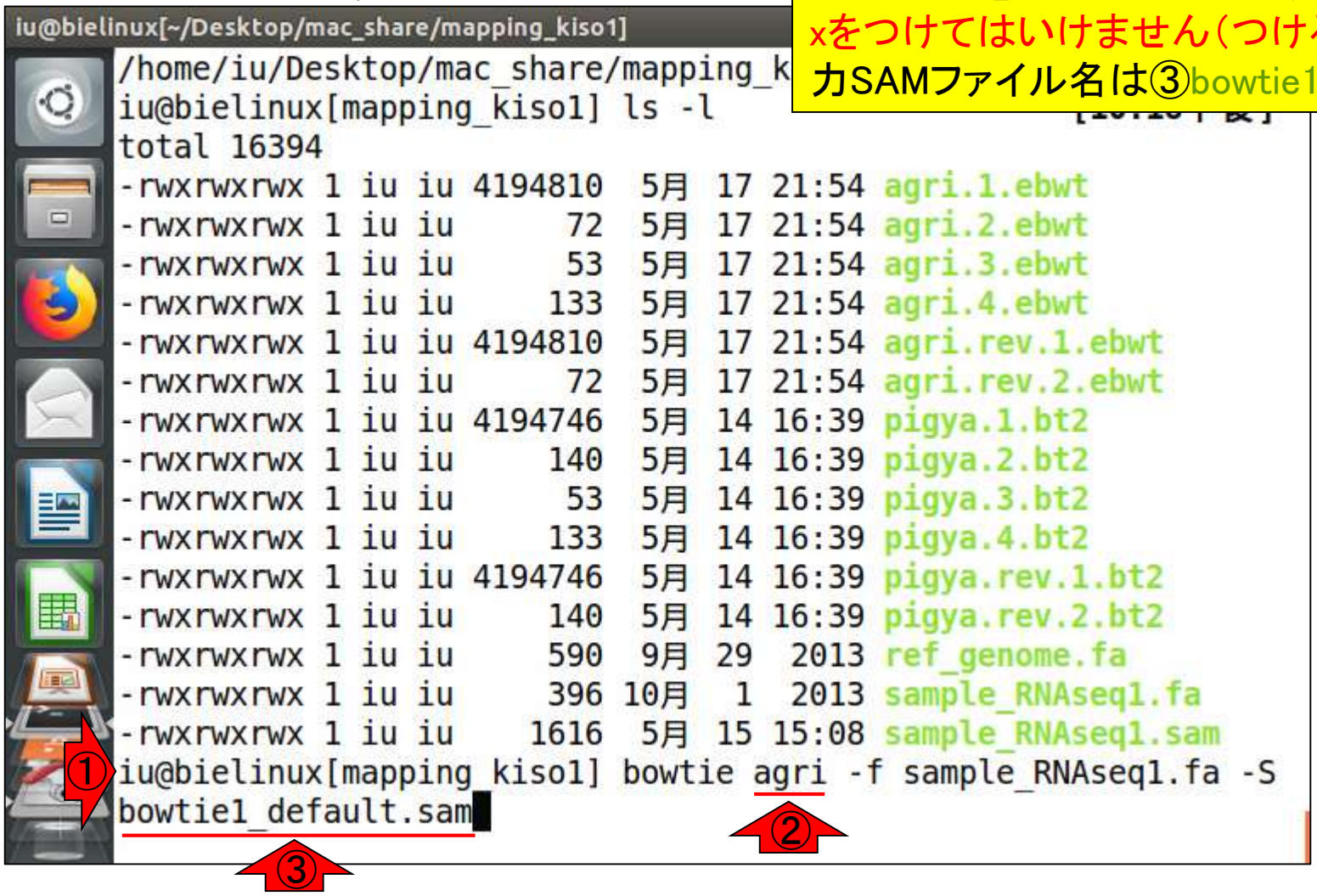
lsで確認

①lsで確認。作成されたインデックスファイルの拡張子部分が②bowtie1ではebwtとなっており、③bowtie2のbt2とは異なることが分かります。このことから、bowtie2で作成したインデックスファイルはbowtie1では利用できないのだろうと思ったりします



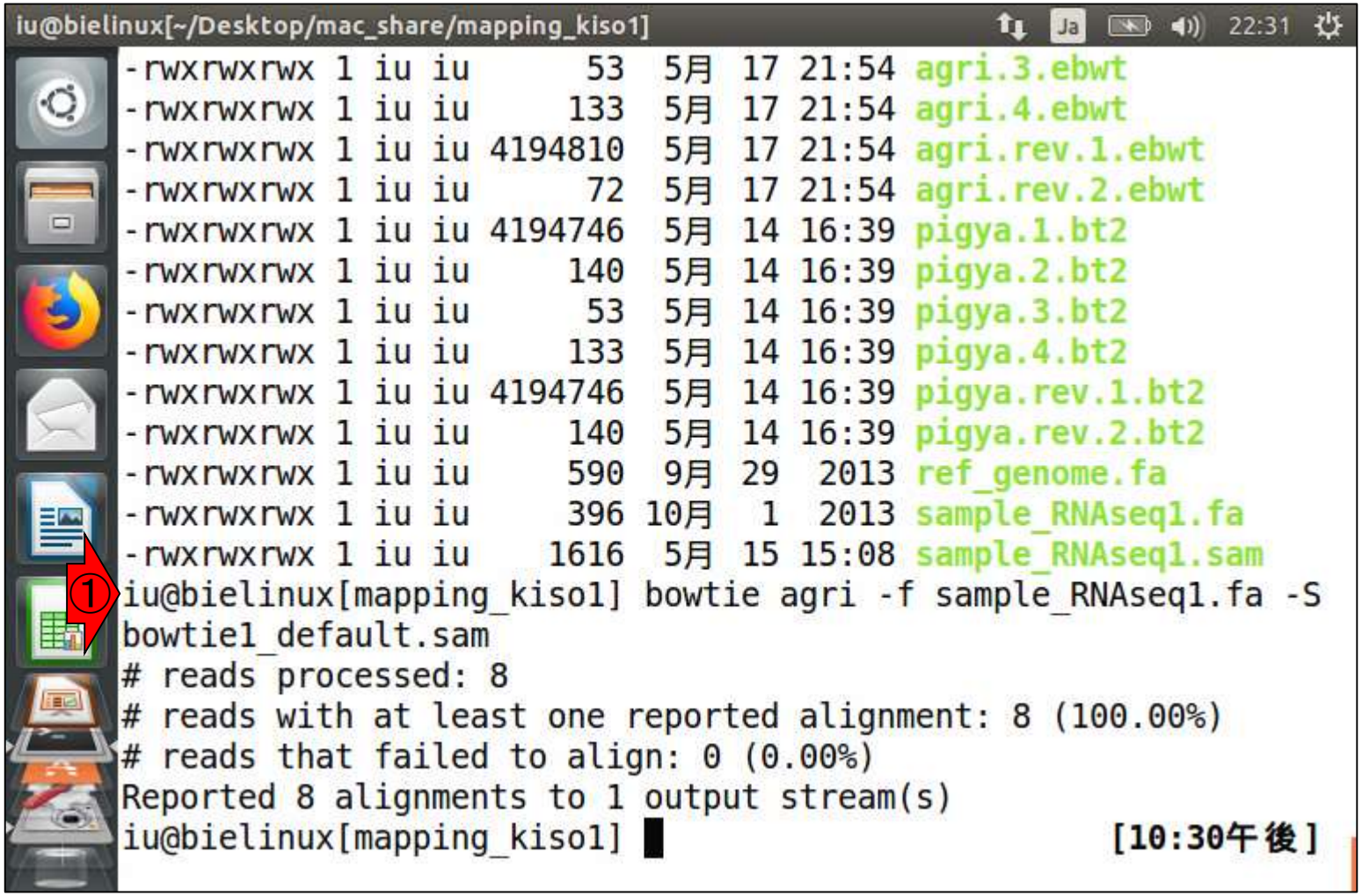
Bowtie実行

①bowtie実行コマンド。Bowtie2の時との違いは、②マップされる側のbasename情報(agri)を指定する際に-xオプションをつけていない点。Bowtie (ver. 1)のときは-xをつけてはいけません(つけると動きません)。尚、出力SAMファイル名は③bowtie1_default.samとしています



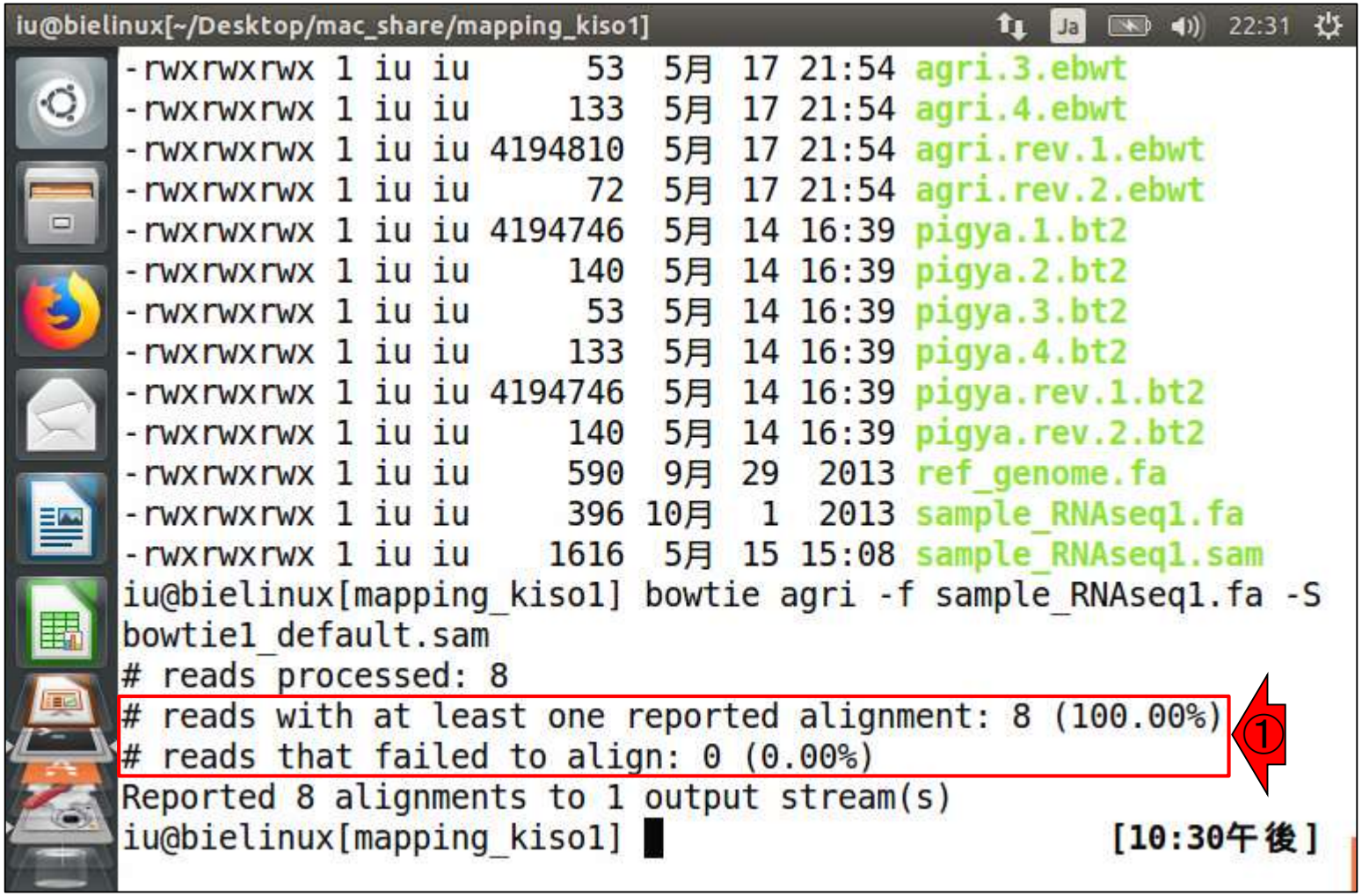
Bowtie実行

①bowtie実行コマンドが無事終了しました。Bowtie2のときとは、出力のされかたが若干異なりますね



Bowtie実行

①bowtieもデフォルトオプションで実行すると、全リードがマップされてしまいました…



bowtie1_default.sam

①デフォルトオプションで実行した結果ファイル（bowtie1_default.sam）

@HD	VN:1.0	SO:unsorted														
@SQ	SN:chr1	LN:48														
@SQ	SN:chr2	LN:160														
@SQ	SN:chr3	LN:100														
@SQ	SN:chr4	LN:123														
@SQ	SN:chr5	LN:100														
@PG	ID:Bowt:	VN:1.1	CL:"bowtie --wrapper basic-0 agri -f sample_RNAseq1.fa -S bowtie1_default.sam"													
chr1_11_45	0	chr1	11	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_16_50	0	chr2	16	255	35M	*	0	0	TA	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_1_35	0	chr2	1	255	35M	*	0	0	AC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_11_45	0	chr5	11	255	35M	*	0	0	TT	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_15_49	0	chr5	15	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_3_37	0	chr3	3	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_1_35	0	chr3	1	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr5_1_35	0	chr5	1	255	35M	*	0	0	GC	II	XA:i:1	MD:Z:3G31	NM:i:1			

bowtie1_default.sam

①5列目はマッピングクオリティ（MAPQ）だが、、Bowtie2のそれとはおそらく計算式が異なる。また、②完全一致で複数個所にマップされる2リードのマッピング結果中に、Bowtie2で複数個所にマップされることを示すXS:がないこともわかる

@HD	VN:1.0	SO:unsorted														
@SQ	SN:chr1	LN:48														
@SQ	SN:chr2	LN:160														
@SQ	SN:chr3	LN:100														
@SQ	SN:chr4	LN:123														
@SQ	SN:chr5	LN:100														
@PG	ID:Bowt:	VN:1.1	CL:"bowtie --wrapper basic-0 agri -f sample_RNAseq1.fa -S bowtie1_default.sam"													
chr1_11_45	0	chr1	11	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_16_50	0	chr2	16	255	35M	*	0	0	TA	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_1_35	0	chr2	1	255	35M	*	0	0	AC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_11_45	0	chr5	11	255	35M	*	0	0	TT	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_15_49	0	chr5	15	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_3_37	0	chr3	3	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_1_35	0	chr3	1	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr5_1_35	0	chr5	1	255	35M	*	0	0	GC	II	XA:i:1	MD:Z:3G31	NM:i:1			

sample_RNAseq1.sam

おさらい。①Bowtie2をデフォルトオプションで実行した結果ファイル（sample_RNAseq1.sam）。①と②が対応する箇所。それに加えて、③のマップされた配列名も異なっていることがわかる。しかしこのこと自体はどちらにマップされててもいいので大した問題ではない

@HD	VN:1.0	SO:unsorted																	
@SQ	SN:chr1	LN:48																	
@SQ	SN:chr2	LN:160																	
@SQ	SN:chr3	LN:100																	
@SQ	SN:chr4	LN:123																	
@SQ	SN:chr5	LN:100																	
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAs															
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU

①

②

③

おまけ：最後のリード

①最後のリードchr5_1_35は、②bowtie、③bowtie2の結果ともに、多くの部分で他のリードと異なっていることがわかる。このリードは、リファレンス配列上に1塩基ミスマッチを許容しないとマップされないようにデザインされていたことに起因すると思われる。そのような視点でXA:, MD:, NM:, AS:, XMなどの説明を読むと理解しやすいかも…

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr1 LN:48
@SQ     SN:chr2 LN:160
@SQ     SN:chr3 LN:100
@SQ     SN:chr4 LN:123
@SQ     SN:chr5 LN:100
@PG     ID:Bowt VN:1.1 CL:"bowtie --wrapper basic-0 agri -f sample_RNAseq1.fa -S bowtie1_default.sam"
chr1_11_45      0  chr1   11 255 35M * 0 0 CC II XA:i:0 MD:Z:35 NM:i:0
chr2_16_50      0  chr2   16 255 35M * 0 0 TA II XA:i:0 MD:Z:35 NM:i:0
chr2_1_35       0  chr2    1 255 35M * 0 0 AC II XA:i:0 MD:Z:35 NM:i:0
chr3_11_45      0  chr5   11 255 35M * 0 0 TT II XA:i:0 MD:Z:35 NM:i:0
chr3_15_49      0  chr5   15 255 35M * 0 0 CC II XA:i:0 MD:Z:35 NM:i:0
chr3_3_37       0  chr3    3 255 35M * 0 0 GC II XA:i:0 MD:Z:35 NM:i:0
chr3_1_35       0  chr3    1 255 35M * 0 0 GC II XA:i:0 MD:Z:35 NM:i:0
chr5_1_35       0  chr5    1 255 35M * 0 0 GC II XA:i:1 MD:Z:3G31 NM:i:1
@SQ     SN:chr5 LN:100
@PG     ID:bowt PN:bow VN:2 CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAs
chr1_11_45      0  chr1   11  42 35M * 0 0 CC II AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr2_16_50      0  chr2   16  42 35M * 0 0 TA II AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr2_1_35       0  chr2    1  42 35M * 0 0 AC II AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr3_11_45      0  chr5   11   1 35M * 0 0 TT II AS:i:0 XS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr3_15_49      0  chr3   15   1 35M * 0 0 CC II AS:i:0 XS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr3_3_37       0  chr3    3  31 35M * 0 0 GC II AS:i:0 XS:i:-6 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr3_1_35       0  chr3    1  35 35M * 0 0 GC II AS:i:0 XS:i:-12 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:Z:UU
chr5_1_35       0  chr5    1  16 35M * 0 0 GC II AS:i:-6 XS:i:-18 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:3G31 YT:Z:UU
```

①

②

③

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

QuasR

QuasRは、内部的にBowtie (ver.1)を用いている。①で示すオプションを指定してマッピングを行った結果として、8リード中3リードがマップされなかった。その結果を再現すべく、①のオプションを付けて、Bio-Linux上でBowtie (ver.1)を実行する

- マッピング | (ESTレベルの長さの)contig (last modified 2014/06/
- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paire
- マッピング | paire
- マッピング | paire
- マップ後 | につい
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ
- マップ後 | 出力フ

マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)

QuasRパッケージを用いてsingle-end RNA-seqデータのリファレンスゲノム配列へのマッピングを行うやり方を示します。basic alignerの1つであるBowtie (Langmead et al., Genome Biol., 2009)を実装した Rbowtieパッケージを内部的に使っています。Bowtie自体は、複数個所にマップされるリードの取り扱い(uniquely mapped reads or multi-mapped reads)を"-m"オプションで指定したり、許容するミスマッチ数を指定する"-v"などの様々なオプションを利用可能ですが、「基礎」のところではやり方を示しませんでした。ここでは、マッピングのオプションをいくつか変更して挙動を確認したり、複数のRNA-seqファイルを一度にマッピングするやり方を示します。尚、出力ファイルは、"*.bam", "*_QC.pdf", "*.bed"の3つです。それ以外のファイルは基本無視で大丈夫です。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

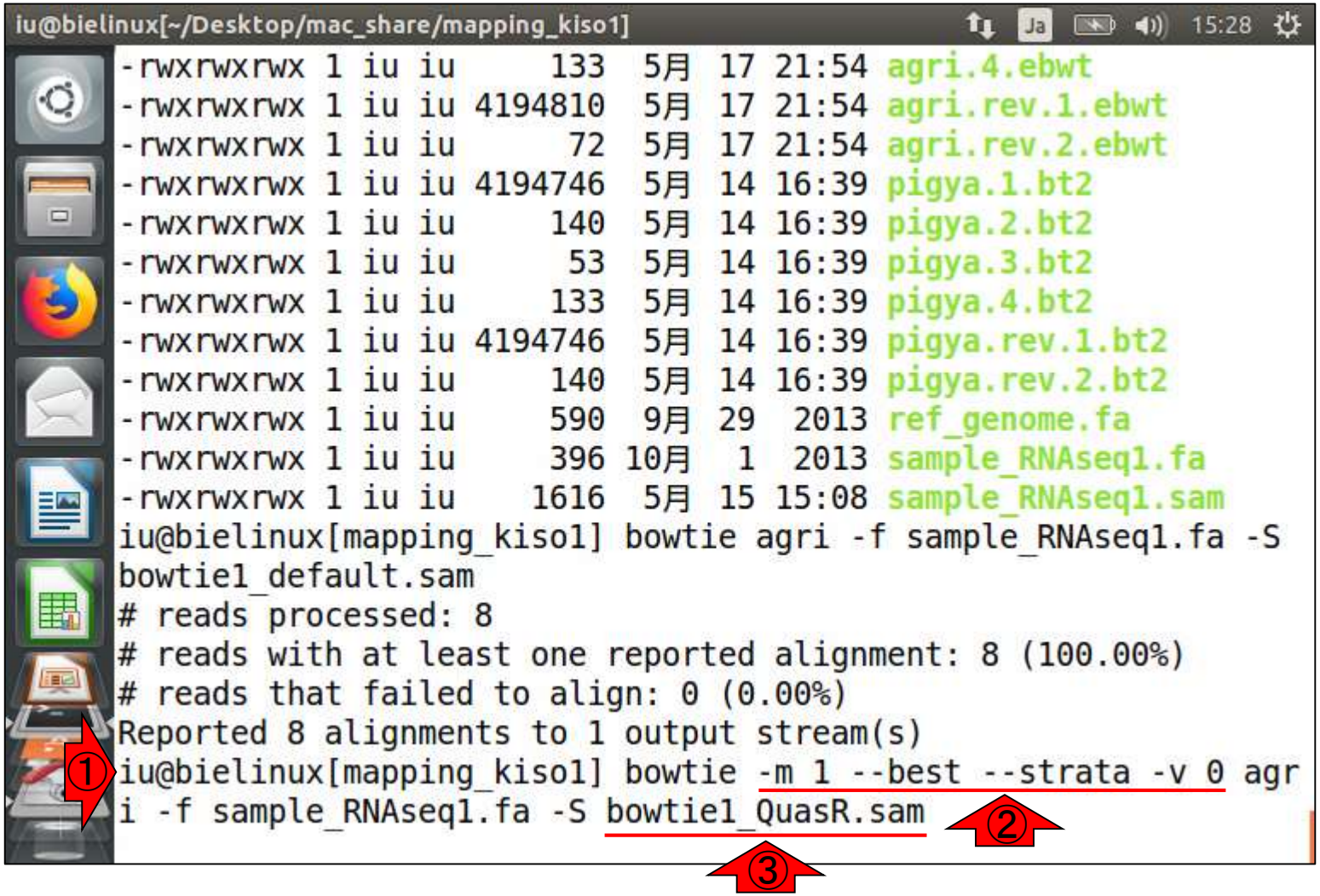
オプションを"-m 1 --best --strata -v 0"とした例です。sample_RNAseq1.faでマップされないのは計3リードです。2リード("chr3_11_45"と"chr3_15_49")はchr5にもマップされるので、"-m 1"オプションで落とされます。1リード("chr5_1_35")は該当箇所と完全一致ではない(4番目の塩基にミスマッチをいれている)ので落とされます。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"               #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0"#マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicAlignments)             #パッケージの読み込み

#本番(マッピング)
time_s <- proc.time()                  #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping)#マッピングを行うqAlign関数を実行した結果をout
time_e <- proc.time()                  #計算時間を計測するため
time_e - time_s                        #計算時間を表示(一番右側の数字。単位はsecond)
out                                    #マッピングに用いたパラメータや入力ファイルの情報などを表示
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス
```

①

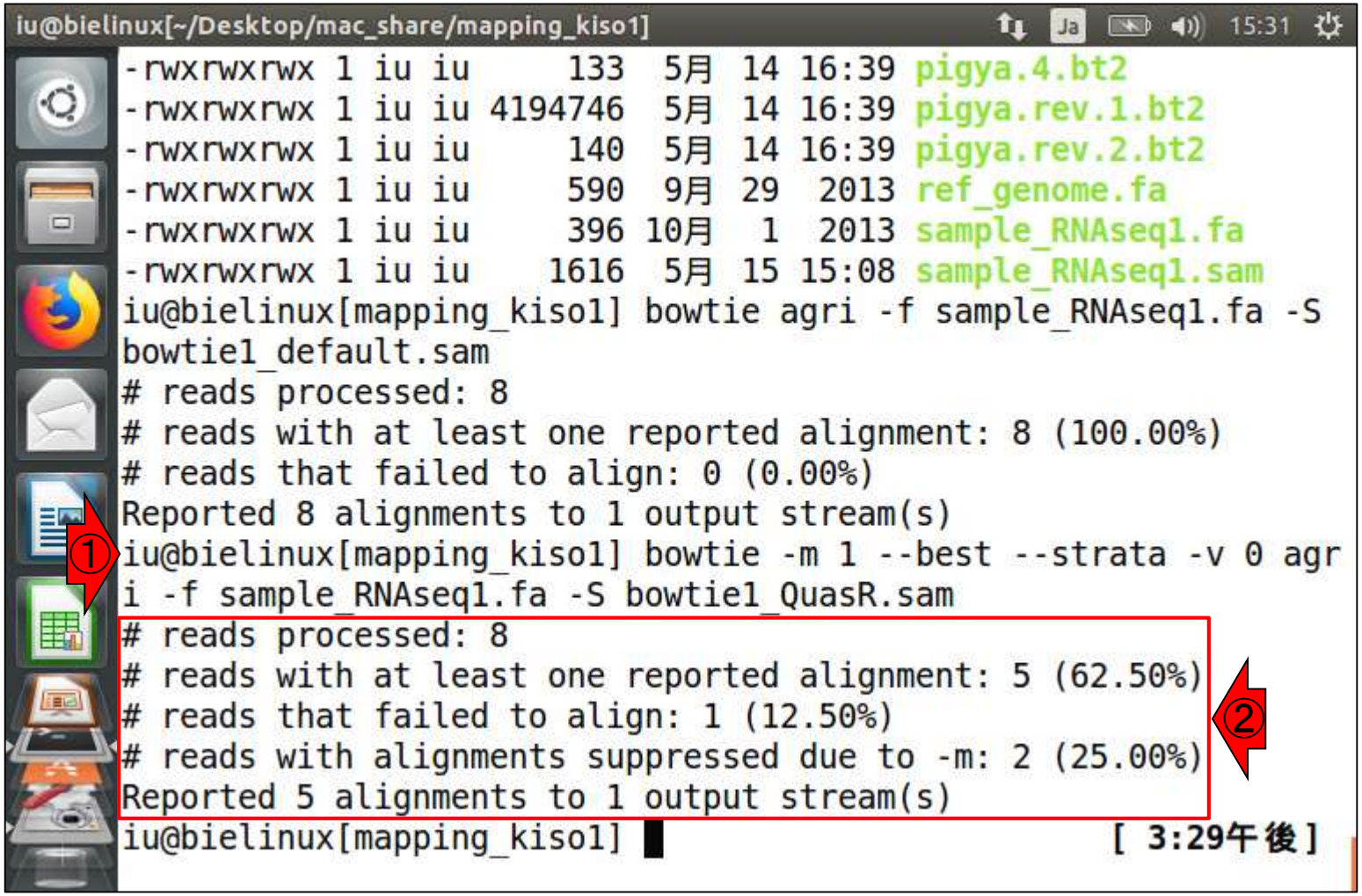
実行コマンド

①が、②オプション（-m 1 --best --strata -v 0）つきの実行コマンド。③出力ファイル名はbowtie1_QuasR.sam



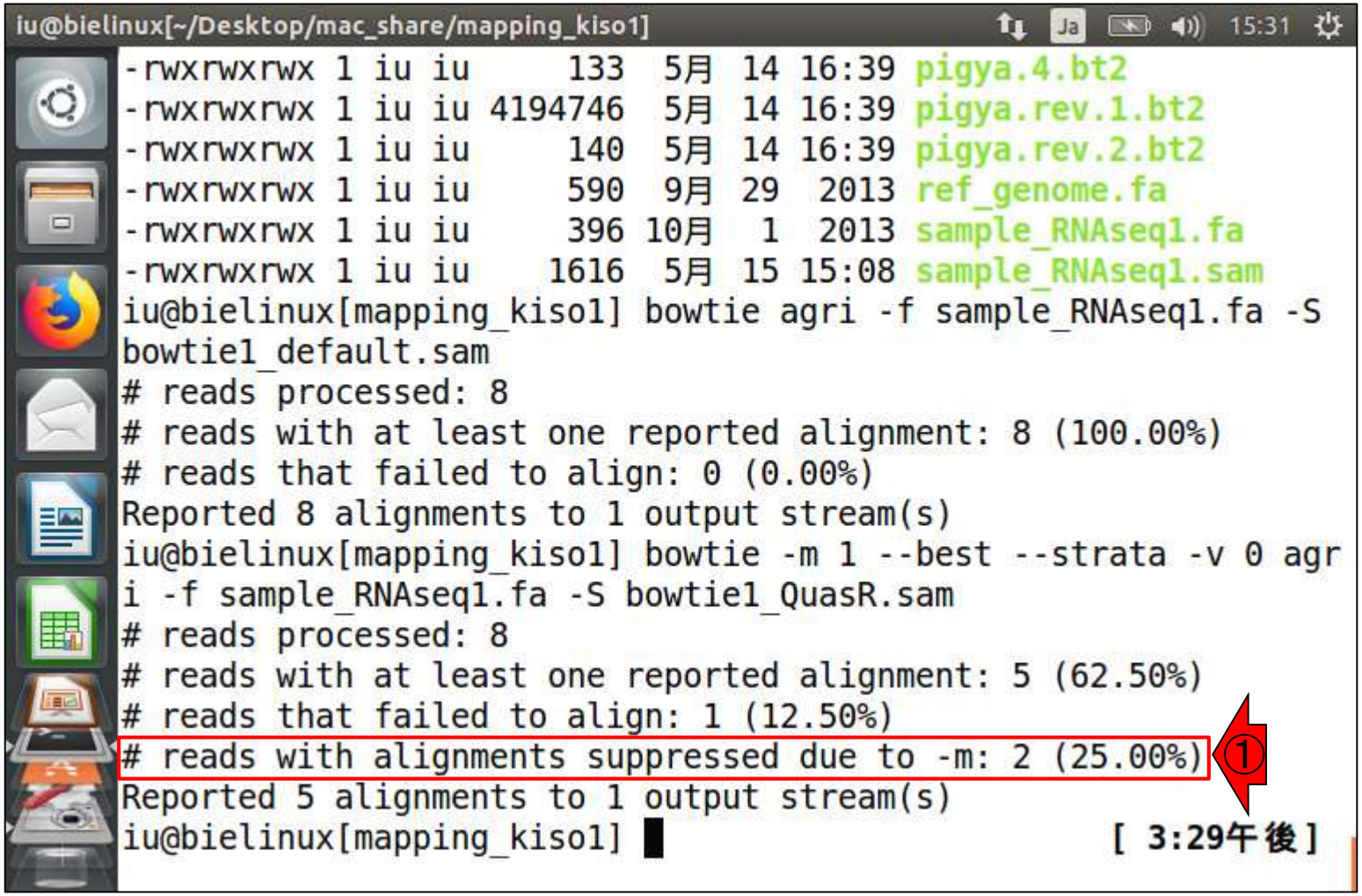
実行結果

①コマンド実行結果。②ぱっと見でQuasR上での
Bowtie (ver.1)実行結果と同じだろうと安心する



実行結果

①この2リードが、-m 1オプション（1か所にのみマップされたリードを出力）という条件を満たさなかった…



QuasRのマッピング結果

複数個所にマップされるこの2つのリードなのだろう

- “-m 1 --best --strata -v 0”：0ミスマッチで1か所にのみマップされるリードを出力

ref_genome.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

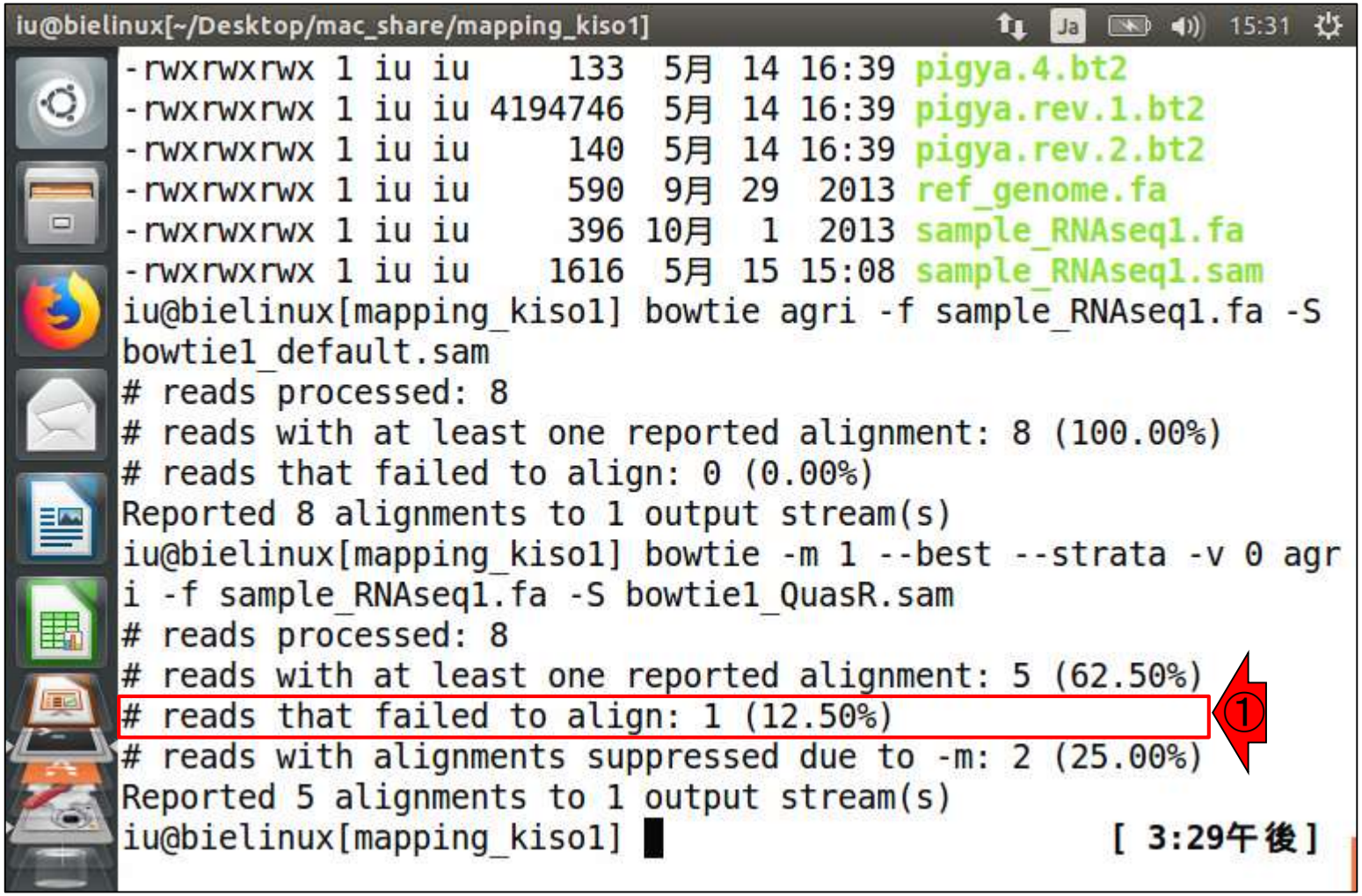
sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

実行結果

①この1リードが、-v 0オプション（許容するミスマッチ数は0）という条件を満たさなかった…



QuasRのマッピング結果

（1か所にのみマップされうるが）①の箇所でミスマッチがある、このリードなのだろう

- “-m 1 --best --strata -v 0”：0ミスマッチで1か所にのみマップされるリードを出力

ref_genome.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

sample_RNAseq1.fa - メモ帳

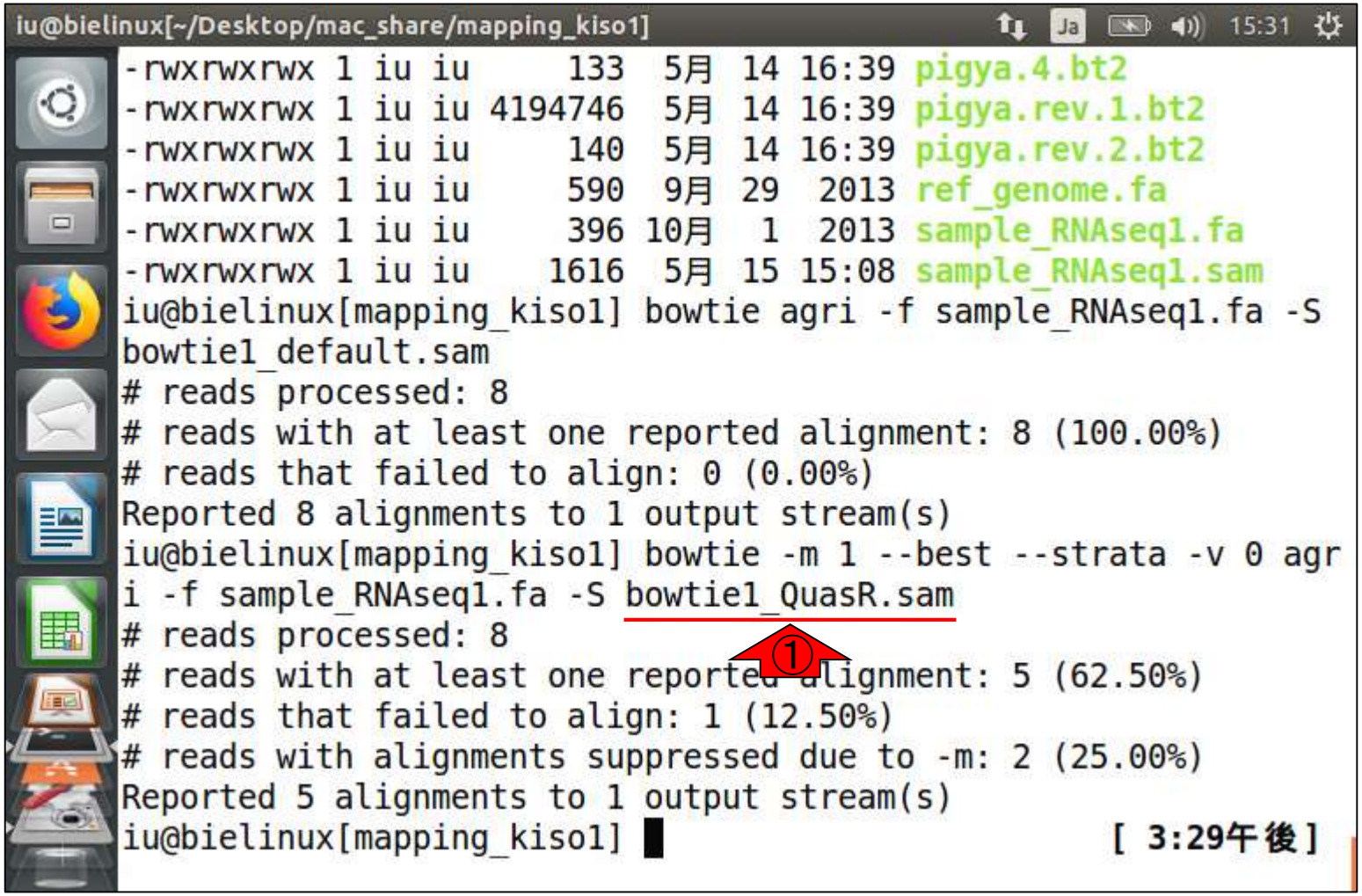
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

①

実行結果

オプション（-m 1 --best --strata -v 0）つきの、①実行結果ファイル（bowtie1_QuasR.sam）の中身が…



これです

bowtie1_QuasR.sam

@HD	VN:1.0	SO:unsorted														
@SQ	SN:chr1	LN:48														
@SQ	SN:chr2	LN:160														
@SQ	SN:chr3	LN:100														
@SQ	SN:chr4	LN:123														
@SQ	SN:chr5	LN:100														
@PG	ID:Bowt	VN:1.1	CL:"bowtie --wrapper basic-0 -m 1 --best --strata -v 0 agri -f sample_RNAseq1.													
chr1_11_45	0	chr1	11	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_16_50	0	chr2	16	255	35M	*	0	0	TA	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_1_35	0	chr2	1	255	35M	*	0	0	AC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_11_45	4	*	0	0	*	*	0	0	TT	II	XM:i:1					
chr3_15_49	4	*	0	0	*	*	0	0	CC	II	XM:i:1					
chr3_3_37	0	chr3	3	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_1_35	0	chr3	1	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr5_1_35	4	*	0	0	*	*	0	0	GC	II	XM:i:0					

bowtie1_QuasR.s

赤枠がマップされなかった3リード。①の列を眺めることで、マップされる側のリファレンス配列名（RNAME）が、マップされないものについては*になるのだと学習する。また、②が-m 1オプション（1か所にのみマップされたリードを出力）という条件を満たさなかったリードであり、③が-v 0オプション（許容するミスマッチ数は0）という条件を満たさなかったリードであることは、リード数の違いから明らか

@HD	VN:1.0	SO:unsorted														
@SQ	SN:chr1	LN:48														
@SQ	SN:chr2	LN:160														
@SQ	SN:chr3	LN:100														
@SQ	SN:chr4	LN:123														
@SQ	SN:chr5	LN:100														
@PG	ID:Bowt	VN:1.1	CL:"bowtie --wrapper basic-0 -m 1 --best --strata -v 0 agri -f sample_RNAseq1.													
chr1_11_45	0	chr1	11	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_16_50	0	chr2	16	255	35M	*	0	0	TA	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_1_35	0	chr2	1	255	35M	*	0	0	AC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_11_45	4	*	0	0	*	*	0	0	TT	II	XM:i:1					
chr3_15_49	4	*	0	0	*	*	0	0	CC	II	XM:i:1					
chr3_3_37	0	chr3	3	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_1_35	0	chr3	1	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr5_1_35	4	*	0	0	*	*	0	0	GC	II	XM:i:0					

①

②

③

bowtie1_QuasR.

①マップされなかったリードの違いを表しているのが赤枠部分。②-m 1オプション（1か所にのみマップされたリードを出力）という条件を満たさなかったリードと、③-v 0オプション（許容するミスマッチ数は0）という条件を満たさなかったリードであることが既知の状態で、④XM:のところの数値の説明を読むとわかりやすい。逆に言えば、そういう実例をいくつか知ったうえでないと、いきなり説明を読んでもチンプンカンプンである場合が多い（個人の感想です）

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr1 LN:48
@SQ     SN:chr2 LN:160
@SQ     SN:chr3 LN:100
@SQ     SN:chr4 LN:123
@SQ     SN:chr5 LN:100
@PG     ID:Bowt VN:1.1 CL:"bowtie --wrapper basic-0 -m 1 --best --strata -v 0 agri -f sample_RNAseq1.
chr1_11_45      0       chr1    11      255     35M     *       0       0       CC      II      XA:i:0  MD:Z:35 NM:i:0
chr2_16_50      0       chr2    16      255     35M     *       0       0       TA      II      XA:i:0  MD:Z:35 NM:i:0
chr2_1_35       0       chr2    1       255     35M     *       0       0       AC      II      XA:i:0  MD:Z:35 NM:i:0
chr3_11_45      4       *       0       0       *       *       0       0       TT      II      XM:i:1
chr3_15_49      4       *       0       0       *       *       0       0       CC      II      XM:i:1
chr3_3_37       0       chr3    3       255     35M     *       0       0       GC      II      XA:i:0  MD:Z:35 NM:i:0
chr3_1_35       0       chr3    1       255     35M     *       0       0       GC      II      XA:i:0  MD:Z:35 NM:i:0
chr5_1_35       4       *       0       0       *       *       0       0       GC      II      XM:i:0
```

①

④

②

③

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

SAM形式の解説

①SAM形式のBowtie出力ファイルの説明。②の部分が12となるところまでページ下部に移動

4. 平成30年05月29日
   講義資料PDF
   bowtie1_default.sam(Bowtieのデフォルトオプション実行結果)
   bowtie1_QuasR.sam(-m 1 --best --strata --0での実行結果)
   Bowtieマニュアル中の**SAM bowtie output** ①

SAM `bowtie` output ①

Following is a brief description of the SAM format as output by `bowtie` when the `-S/--sam` option is specified. For more details, see the SAM format specification.

When `-S/--sam` is specified, `bowtie` prints a SAM header with `@HD`, `@SQ` and `@PG` lines. When one or more `--sam-RG` arguments are specified, `bowtie` will also print an `@RG` line that includes all user-specified `--sam-RG` tokens separated by tabs.

Each subsequnt line corresponds to a read or an alignment. Each line is a collection of at least 12 fields separated by tabs; from left to right, the fields ②

1. Name of read that aligned
2. Sum of all applicable flags. Flags relevant to Bowtie are:

1	The read is one of a pair
2	The alignment is one end of a proper paired-end alignment
4	The read has no reported alignments
8	The read is one of a pair and has no reported alignments

erse reference strand

Bowtie (Langmead et al., Genome Biol., **10**: R25, 2009)

SAM形式の解説

①12の、②のあたりにXM:についての説明があります。③赤下線部分が今回指定した-m 1（1か所にのみマップされたリードを出力）と関連しています

12. Optional fields. Fields are tab-separated. For descriptions of all possible optional fields, see the SAM format specification. `bowtie` outputs some of these optional fields for each alignment, depending on the type of the alignment:

`NM:i:<N>`	Aligned read has an edit distance of `<N>`.
`CM:i:<N>`	Aligned read has an edit distance of `<N>` in colorspace. This field is present in addition to the `NM` field in `-C/--color` mode, but is omitted otherwise.
`MD:Z:<S>`	For aligned reads, `<S>` is a string representation of the mismatched reference bases in the alignment. See SAM format specification for details. For colorspace alignments, `<S>` describes the decoded *nucleotide* alignment, not the colorspace alignment.
`XA:i:<N>`	Aligned read belongs to stratum `<N>`. See Strata for definition.
`XM:i:<N>`	For a read with no reported alignments, `<N>` is 0 if the read had no alignments. If `-m` was specified and the read's alignments were supressed because the `-m` ceiling was exceeded, `<N>` equals the `-m` ceiling 1, to indicate that there were at least that many valid alignments (but all were suppressed). In `-M` mode, if the alignment was randomly selected because the `-M` ceiling was exceeded, `<N>` equals the `-M` ceiling 1, to indicate that there were at least that many valid alignments (of which one was reported at random).

`XM:i:<N>`	For a read with no reported alignments, `<N>` is 0 if the read had no alignments. If `-m` was specified and the read's alignments were supressed because the `-m` ceiling was exceeded, `<N>` equals the `-m` ceiling 1, to indicate that there were at least that many valid alignments (but all were suppressed). In `-M` mode, if the alignment was randomly selected because the `-M` ceiling was exceeded, `<N>` equals the `-M` ceiling 1, to indicate that there were at least that many valid alignments (of which one was reported at random).

Bowtie (Langmead et al., Genome Biol., **10**: R25, 2009)

SAM形式の解

①-m 1オプション（1か所にのみマップされたリードを出力）という条件を満たさなかったリードなので、確かに②の部分が1になっている。そして、③-v 0オプション（許容するミスマッチ数は0）という条件を満たさなかったリードは、-m 1オプションに該当するわけではないので、確かに④の部分が0になっている

@HD	VN:1.0	SO:unsorted														
@SQ	SN:chr1	LN:48														
@SQ	SN:chr2	LN:160														
@SQ	SN:chr3	LN:100														
@SQ	SN:chr4	LN:123														
@SQ	SN:chr5	LN:100														
@PG	ID:Bowt	VN:1.1	CL:"bowtie --wrapper basic-0 -m 1 --best --strata -v 0 agri -f sample_RNAseq1.													
chr1_11_45	0	chr1	11	255	35M	*	0	0	CC	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_16_50	0	chr2	16	255	35M	*	0	0	TA	II	XA:i:0	MD:Z:35	NM:i:0			
chr2_1_35	0	chr2	1	255	35M	*	0	0	AC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_11_45	4	*	0	0	*	*	0	0	TT	II	XM:i:1					
chr3_15_49	4	*	0	0	*	*	0	0	CC	II	XM:i:1					
chr3_3_37	0	chr3	3	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr3_1_35	0	chr3	1	255	35M	*	0	0	GC	II	XA:i:0	MD:Z:35	NM:i:0			
chr5_1_35	4	*	0	0	*	*	0	0	GC	II	XM:i:0					

②
①
③
④

read had no alignments. If -m was specified and the read's alignments were supressed because the -m ceiling was exceeded, <N> equals the -m ceiling 1, to

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

マッピング結果の違い

Bio-Linux環境で行ったマッピングは、①bowtie2のデフォルト、②bowtieのデフォルト、③bowtieのオプション（-m 1 --best --strata -v 0）つきの計3通りであった

```
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
```

①

```
iu@bielinux[mapping_kiso1] bowtie agri -f sample_RNAseq1.fa -S
bowtie1_default.sam
# reads processed: 8
# reads with at least one reported alignment: 8 (100.00%)
# reads that failed to align: 0 (0.00%)
Reported 8 alignments to 1 output stream(s)
```

②

```
iu@bielinux[mapping_kiso1] bowtie -m 1 --best --strata -v 0 agr
i -f sample_RNAseq1.fa -S bowtie1_QuasR.sam
# reads processed: 8
# reads with at least one reported alignment: 5 (62.50%)
# reads that failed to align: 1 (12.50%)
# reads with alignments suppressed due to -m: 2 (25.00%)
Reported 5 alignments to 1 output stream(s)
```

③

マッピング結果の違い

①bowtie2のデフォルトの結果は、②3リードが1個所に、③5リードが複数個所にマップされるというものであった

①

```
iu@bielinux[mapping_kisoi] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
```

②

③

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

マッピング結果

①bowtie2のデフォルトの結果は、②3リードが1個所に、③5リードが複数個所にマップされるというものであった。④bowtieのオプション（-m 1 --best --strata -v 0）つきの結果は、⑤-v 0オプション（許容するミスマッチ数は0）を満たさなかった1リードと、

①

```
iu@bielinux[mapping_kisoi] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
```

②

③

sample_RNAseq1.fa - メモ帳

ファイル(F)　編集(E)　書式(O)　表示(V)　ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

④

⑤

```
iu@bielinux[mapping_kisoi] bowtie -m 1 --best --strata -v
i -f sample_RNAseq1.fa -S bowtie1_QuasR.sam
# reads processed: 8
# reads with at least one reported alignment: 5 (62.50%)
# reads that failed to align: 1 (12.50%)
# reads with alignments suppressed due to -m: 2 (25.00%)
Reported 5 alignments to 1 output stream(s)
```

マッピング結果

①bowtie2のデフォルトの結果は、②3リードが1個所に、③5リードが複数個所にマップされるというものであった。④bowtieのオプション（-m 1 --best --strata -v 0）つきの結果は、⑤-v 0オプション（許容するミスマッチ数は0）を満たさなかった1リードと、⑥ -m 1オプション（1か所にのみマップ）を満たさなかったリード

```
iu@bielinux[mapping_kisoi] bowtie2 [...] -x pigyu -1 sample_RNAseq1.fa -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
```

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

```
iu@bielinux[mapping_kisoi] bowtie -m 1 --best --strata -v 1 -f sample_RNAseq1.fa -S bowtie1_QuasR.sam
# reads processed: 8
# reads with at least one reported alignment: 5 (62.50%)
# reads that failed to align: 1 (12.50%)
# reads with alignments suppressed due to -m: 2 (25.00%)
Reported 5 alignments to 1 output stream(s)
```

課題1の基礎情報

①6番目と7番目のリードは、②bowtie2では2回以上マップされ（aligned >1 times）、③bowtieでは少なくとも1回はマップ（at least one）されている

```
iu@bielinux[mapping_kiso1] bowtie2 -x pigya -f sample_RNAseq1.f
a -S sample_RNAseq1.sam
8 reads; of these:
  8 (100.00%) were unpaired; of these:
    0 (0.00%) aligned 0 times
    3 (37.50%) aligned exactly 1 time
    5 (62.50%) aligned >1 times
100.00% overall alignment rate
```

②

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

①

```
iu@bielinux[mapping_kiso1] bowtie -m 1 --best --strata -v
i -f sample_RNAseq1.fa -S bowtie1_QuasR.sam
# reads processed: 8
# reads with at least one reported alignment: 5 (62.50%)
# reads that failed to align: 1 (12.50%)
# reads with alignments suppressed due to -m: 2 (25.00%)
Reported 5 alignments to 1 output stream(s)
```

③

課題1

①6番目のリード（chr3_3_37）と②7番目のリード（chr3_1_35）について、③リファレンス配列上のどこにマップされたのか示せ。リファレンス配列名（例：chr1）とマップされた領域の左端の位置（例：8番目の塩基）のみでよい

ref_genome.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACGCCCTG
>chr3
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTG
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGC
ACCAGACTTTACCGTACAAGATA
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
```

sample_RNAseq1.fa - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

```
>chr1_11_45
CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2_16_50
TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
>chr2_1_35
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
>chr3_11_45
TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
>chr3_15_49
CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr3_3_37
GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
>chr3_1_35
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
>chr5_1_35
GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
```

課題2の基礎情報

Bowtie2実行結果のSAMファイル（sample_RNAseq1.sam）。赤枠内は、6番目のリード（chr3_3_37）と7番目のリード（chr3_1_35）のマッピング結果部分

sample_RNAseq1.sam

@HD	VN:1.0	SO:unsorted																
@SQ	SN:chr1	LN:48																
@SQ	SN:chr2	LN:160																
@SQ	SN:chr3	LN:100																
@SQ	SN:chr4	LN:123																
@SQ	SN:chr5	LN:100																
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAs														
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_11_45	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_49	0	chr3	15	1	35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1	16	35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU

課題2

①6番目のリード（chr3_3_37）と②7番目のリード（chr3_1_35）のマッピング結果「リファレンス配列名とマップされた領域の左端の位置」として、③「chr3上の3番目の塩基」と④「chr3上の1番目の塩基」が採用された理由について自由に考えを述べよ。

sample_RNAseq1.sam

@HD	VN:1.0	SO:unsorted																	
@SQ	SN:chr1	LN:48																	
@SQ	SN:chr2	LN:160																	
@SQ	SN:chr3	LN:100																	
@SQ	SN:chr4	LN:123																	
@SQ	SN:chr5	LN:100																	
@PG	ID:bowt	PN:bow	VN:2	CL:"/usr/bin/../lib/bowtie2/bin/bowtie2-align-s --wrapper basic-0 -x pigya -f sample_RNAs															
chr1_11_45	0	chr1	11	42	35M	*	0	0	CC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_16_50	0	chr2	16	42	35M	*	0	0	TA	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr2_1_35	0	chr2	1	42	35M	*	0	0	AC	II	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3_11_	0	chr5	11	1	35M	*	0	0	TT	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_15_4	0	chr3	15		35M	*	0	0	CC	II	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_3_37	0	chr3	3	31	35M	*	0	0	GC	II	AS:i:0	XS:i:-6	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3_1_35	0	chr3	1	35	35M	*	0	0	GC	II	AS:i:0	XS:i:-12	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr5_1_35	0	chr5	1		35M	*	0	0	GC	II	AS:i:-6	XS:i:-18	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:3G31	YT:Z:UU

①

②

③

④

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

他のBowtie2実行手段

①と②については、基本的な使い方の解説があります

- DDBJ Pipeline (Nagasaki et al., DNA Res., **20**: 383–90, 2013) ①
  - DDBJが提供するクラウド解析環境
- Galaxy (Goecks et al., Genome Biol., **11**: R86, 2010) ②
  - Galaxy projectが提供するクラウド解析環境。Galaxy mainというサイトが有名。
- Illumina BaseSpace
  - Illumina社が提供するクラウド解析環境。
- …

他のBowtie2実行手段

乳酸菌学会誌NGS連載の、①第6回がDDBJ Pipeline、②第11-12回がGalaxyについての解説。但し、マッピングについては書かれていない

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページのR関連部分は、インストール | について の推奨手順 (Windows2018.03.12版とMacinto
トール済みであるという前提で記述しています。初心者の方は基本的な利用法(Windows2015.04.03版と
的にまとめた書籍もあります。(2015/04/03)

What's new?

- Silhouetteスコアの新たな使い道提唱論文(Zhao et a
- Silhouetteスコアの新たな使い道提唱論文(Zhao et a
- 「平成29年度NGSハンズオン講習会」の動画が公開

- 門田からメール返信をもらえない場合は (last modif
- はじめに (last modified 2015/03/31)
- 参考資料 | 書籍、学会誌 (last modified 2017/11/13)
- 参考資料 | 講習会、講義、講演資料 (last modified

- 書籍 | トランスクリプトーム解析 | 4.3.3 2群間比較 (last modified 2014/04/28)
- 書籍 | トランスクリプトーム解析 | 4.3.4 他の実験デザイン(3群間) (last modified 2014/04/28)
- 書籍 | 日本乳酸菌学会誌 | について (last modified 2018/05/10) NEW
- 書籍 | 日本乳酸菌学会誌 | 第1回イントロダクション (last modified 2016/12/22)
- 書籍 | 日本乳酸菌学会誌 | 第2回GUI環境からコマンドライン環境へ (last modified 2015/11/26)
- 書籍 | 日本乳酸菌学会誌 | 第3回Linux環境構築からNGSデータ取得まで (last modified 2017/07/02)
- 書籍 | 日本乳酸菌学会誌 | 第4回クオリティコントロールとプログラムのインストール (last modified 201
- 書籍 | 日本乳酸菌学会誌 | 第5回アセンブル、マッピング、そしてQC (last modified 2017/06/25)
- 書籍 | 日本乳酸菌学会誌 | 第6回ゲノムアセンブリ (last modified 2017/06/21)
- 書籍 | 日本乳酸菌学会誌 | 第7回ロングリードアセンブリ (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | 第8回アセンブリ後の解析 (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | 第9回ゲノムアノテーションとその可視化、DDBJへの登録 (last modified 2
- 書籍 | 日本乳酸菌学会誌 | 第10回DDBJへの塩基配列の登録(後編) (last modified 2017/06/28)
- 書籍 | 日本乳酸菌学会誌 | 第11回統合データ解析環境Galaxy (last modified 2017/11/13)
- 書籍 | 日本乳酸菌学会誌 | 第12回Galaxy:ヒストリーとワークフロー (last modified 2018/03/23)
- イントロ | 一般 | ランダムに行を抽出 (last modified 2014/07/17)
- イントロ | 一般 | 任意の文字列を行の最初に挿入 (last modified 2014/07/17)

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

カウント情報取得は、③の部分に相当

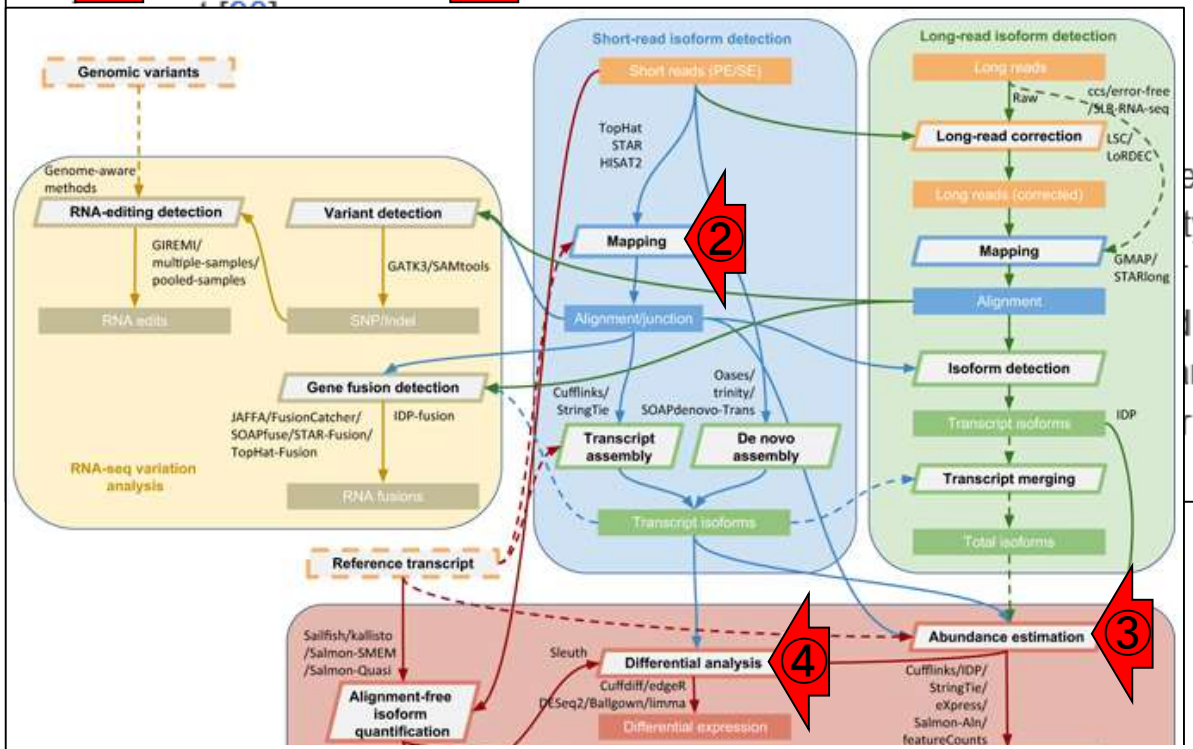
全体像のおさらい

RNA-Seq data analysis

最近の総説 (Lowe et al., PLoS Comput. Biol., **13**: e1005457, 2017)

processed to yield useful information. Data analysis usually requires a combination of bioinformatics software tools that vary according to the experim① design ②als. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a com③d-line interface, eith④ a Unix environment or within the R/Bioconductor statistical

ence needs to be
ty scores for base
representation of
l duplication rate [85].
nd FaQCs software
tagged for special

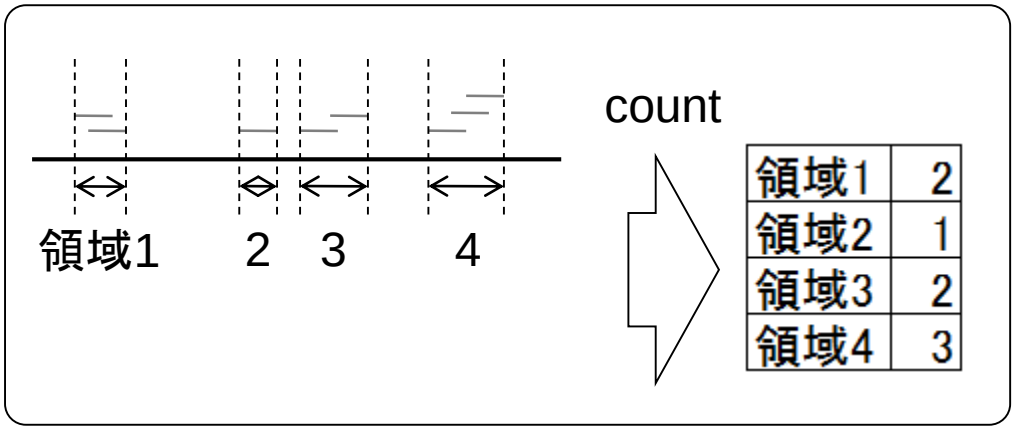


RNACocktail (Sahraeian et al., Nat Commun., **8**: 59, 2017)

単一サンプル

Rでアノテーション情報を利用する場合は、TxDbが基本。アノテーション情報がない場合は、マップされたリードの領域をたよりに転写領域を決める

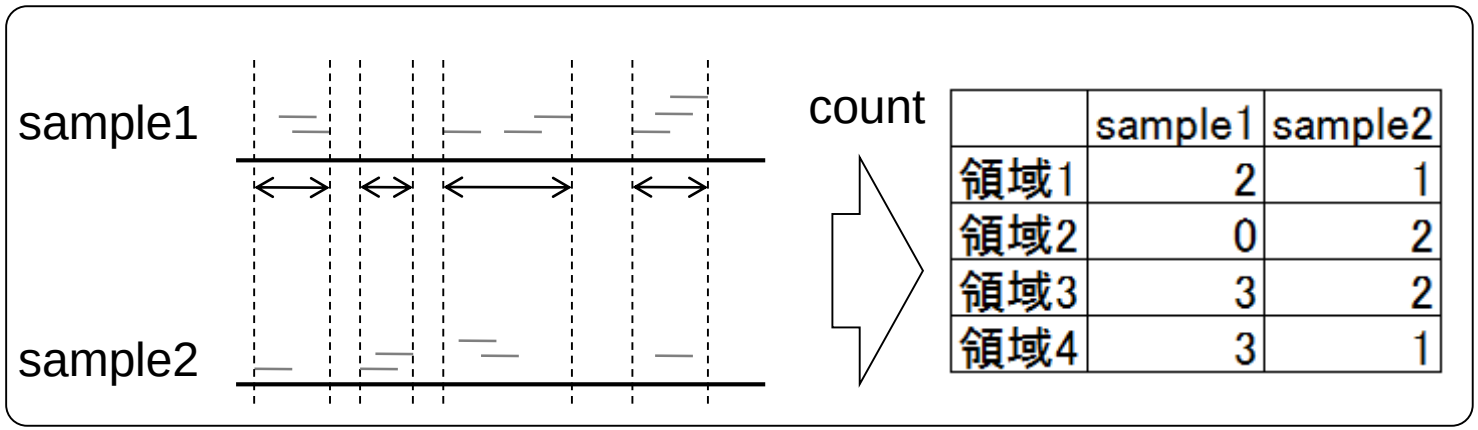
- アノテーション情報を利用する場合
  - UCSC known Genes, Ensembl Genesなど様々なテーブル名を指定可能
  - gene, exon, promoter, junctionなど様々なレベルを指定可能
- アノテーション情報がない場合
  - マップされたリードの和集合領域を同定したのち、領域ごとのリード数をカウント
  - BEDtools (Quinlan et al., 2010)中のmergeBedプログラムを実行して和集合領域同定後、intersectBedプログラムを実行してリード数をカウントする作業に相当



複数サンプル

アノテーション情報がない場合の戦略は、複数サンプルの場合には領域が変わりうる。Cufflinks（最近ではStringTie）を知っているヒトはcuffmergeと同じイメージだと思えばよい

- アノテーション情報を利用する場合
  - UCSC known Genes, Ensembl Genesなど様々なテーブル名を指定可能
  - gene, exon, promoter, junctionなど様々なレベルを指定可能
- アノテーション情報がない場合
  - マップされたリードの和集合領域を同定したのち、領域ごとのリード数をカウント
  - BEDtools (Quinlan et al., 2010)中のmergeBedプログラムを実行して和集合領域同定後、intersectBedプログラムを実行してリード数をカウントする作業に相当



	sample1	sample2
領域1	2	1
領域2	0	2
領域3	3	2
領域4	3	1

単一サンプル

利用可能なアノテーション情報がなく、単一サンプルで転写領域を定め、その領域にマップされるリード数をカウントするやり方を示します。①single-endのアノテーション無のところです

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページのR関連部分は、インストール||についての推奨手順(Windows2018.03.12版とMacinto
トール済みであるという前提で記述しています。初心者の方は基本的な利用法(Windows2015.04.03版と
的にまとめた書籍もあります。(2015/04/03)

What's new?

- Silhouetteスコアの新たな使い道提唱論文(Z
- Silhouetteスコアの新たな使い道提唱論文(Z
- 「平成29年度NGSハンズオン講習会」の動画

- 門田からメール返信をもらえない場合は (las
- はじめに (last modified 2015/03/31)
- 参考資料 | 書籍、学会誌 (last modified 2017
- 参考資料 | 講習会、講義、講演資料 (last m

- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | トランスクリプトーム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/10)
- マッピング | paired-end | トランスクリプトーム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2016/02/10)
- マップ後 | について (last modified 2013/06/19)
- マップ後 | 出力ファイル形式について (last modified 2013/11/05)
- マップ後 | 出力ファイルの読み込み | BAM形式 | について (last modified 2016/09/14)
- マップ後 | 出力ファイルの読み込み | BAM形式 | rbamtools(Kaisers_2015) (last modified 2016/09/14)
- マップ後 | 出力ファイルの読み込み | BAM形式 | GenomicAlignments(Lawrence_2013) (last modified 2016/09/14)
- マップ後 | 出力ファイルの読み込み | Bowtie形式 (last modified 2013/06/18)
- マップ後 | 出力ファイルの読み込み | SOAP形式 (last modified 2013/06/19)
- マップ後 | 出力ファイルの読み込み | htSeqTools(Planet_2012) (last modified 2013/06/19)
- マップ後 | カウント情報取得 | について (last modified 2017/01/11)
- マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) (last modified 2015/02/26)
- マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015) (last modified 2014/06/22)
- マップ後 | カウント情報取得 | paired-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) (last modified 2016/02/13)
- マップ後 | カウント情報取得 | paired-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015) (last modified 2015/07/02)
- マップ後 | カウント情報取得 | paired-end | トランスクリプトーム | QuasR(Gaidatzis_2015) (last modified 2016/02/12)
- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- マップ後 | 配列長とカウント数の関係 (last modified 2015/07/03)
- 正規化 | について (last modified 2014/06/22)
- 正規化 | 基礎 | RPK or CPK (配列長補正) (last modified 2015/07/04)

①

単一サンプル

①例題1をやります。②作業ディレクトリは「Desktop – hoge – mapping_kiso1」。③最低限この3つのファイルがあれば動きます

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015)

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt): ①

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"               #入力ファイル
param_mapping <- "-m 1 --best --strata -v 0"#マッピ

#必要なパッケージをロード
library(QuasR)                         #パッケージの
library(GenomicAlignments)             #パッケージの

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計
out <- qAlign(in_f1, in_f2, alignmentParameter=para
time_e <- proc.time()                  #計算時間を計
time_e - time_s                        #計算時間を表
out                                    #マッピングに
alignmentStats(out)                    #マッピング結

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(
tmpsname <- out@alignments[,2]         #サンプル名(
for(i in 1:length(tmpfname)){          #サンプル数(
  if(i == 1){
    k <- readGAlignments(tmpfname[i])  #BAM形式ファ
```

R Console

```
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> list.files()
[1] "mapping_single_genome1.txt"
[2] "ref_genome.fa"
[3] "sample_RNAseq1.fa"
>
>
>
>
>
>
>
>
>
>
>
```

②

③

単一サンプル

①以前のマッピング結果が残っていても問題ありません。②このパラメータ(マッピングオプション)は前回指定したものと同じです。③マッピングを行う関数もコード内に記述されていますが、QuasRは③マッピングを実行する前に、④作業ディレクトリ内の⑤実行ログファイル(QuasR_log_...)をまず見ます

マップ後 | カウント情報取得 | single-end | ゲ

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンス
の一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignmentsパッケージを利用して、マップされた
リードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"   #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"                #入力ファイル
param_mapping <- "-m 1 --best --strata -v 0"#コピ

#必要なパッケージをロード
library(QuasR)                          #パッケージの
library(GenomicAlignments)              #パッケージの

#前処理(マッピング)
time_s <- proc.time()                   #計算時間を計
out <- qAlign(in_f1, in_f2, alignmentParameter=para
time_e <- proc.time()                   #計算時間を計
time_e - time_s                         #計算時間を表
out                                     #マッピングに
alignmentStats(out)                     #マッピング結

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]          #ファイル名(
tmpsname <- out@alignments[,2]          #サンプル名(
for(i in 1:length(tmpfname)){           #サンプル数(
  if(i == 1){
    k <- readGAlignments(tmpfname[i])   #BAM形式ファ
```

②

③

```
R Console
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> list.files()
 [1] "mapping_single_genome1.txt"
 [2] "QuasR_log_146c2f3d4e64.txt"
 [3] "ref_genome.fa"
 [4] "ref_genome.fa.fai"
 [5] "ref_genome.fa.md5"
 [6] "ref_genome.fa.Rbowtie"
 [7] "sample_RNAseq1.fa"
 [8] "sample_RNAseq1_146c6c6d54aa.bam"
 [9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
>
```

④

⑤

①

単一サンプル

そして、以前マッピングを行った結果が今回のものと同じ入力ファイルで②同じパラメータあれば、今回改めてマッピングを行うことはせずに以前の結果を流用してくれるというありがたい機能があります。マッピングに数時間以上かかるような場合は、便利さを実感します

マップ後 | カウント情報取得 | single-end | ゲノム |

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配
の一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignmentsパッケージを利用して、マップされた
リードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"               #入力ファイル
param_mapping <- "-m 1 --best --strata -v 0"#コピ

#必要なパッケージをロード
library(QuasR)                         #パッケージの
library(GenomicAlignments)             #パッケージの

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計
```

②

R Console

```
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> list.files()
 [1] "mapping_single_genome1.txt"
 [2] "QuasR_log_146c2f3d4e64.txt"
```

⑤

C:¥Users¥kojik¥Desktop¥hoge¥mapping_kiso1¥QuasR_log_146c2f3d4e64.txt - EmEditor

ファイル(F) 編集(E) 検索(S) 表示(V) ツール(T) ウィンドウ(W) ヘルプ(H)

QuasR_log_146c2f3d4e64.txt

```
[1] "Executing bowtie on DESKTOP-3J8LKP8 using 1 cores. Parameters:"
[1] "¥"C:/Users/kojik/Desktop/hoge/mapping_kiso1/ref_genome.fa.Rbowtie/bowtieIndex¥" ¥"C:/Users/kojik/Desktop/hoge/mapping_kiso1/sample_RNAseq1
[1] "Converting sam file to sorted bam file on DESKTOP-3J8LKP8 : C:¥¥Users¥¥kojik¥¥AppData¥¥Local¥¥Temp¥¥Rtmpw5nBDI¥¥sample_RNAseq1.fa2dd079d64
[1] "Genomic alignments for sample 1 (namae) have been successfully created on DESKTOP-3J8LKP8"
```

598 バイト (598 バイト), 5 行。 Text 1行, 1桁 日本語 (シフト JIS)

```
    k <- readGAlignments(tmpfname[i])  #BAM形式ファ
```

```
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
>
```

①今回は、この状態でコピペ実行します

コピペ実行

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015)

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"               #入力ファイル
param_mapping <- "-m 1 --best --strata -v 0"#マッピ

#必要なパッケージをロード
library(QuasR)                         #パッケージの
library(GenomicAlignments)             #パッケージの

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計
out <- qAlign(in_f1, in_f2, alignmentParameter=para
time_e <- proc.time()                  #計算時間を計
time_e - time_s                        #計算時間を表
out                                    #マッピングに
alignmentStats(out)                    #マッピング結

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(
tmpsname <- out@alignments[,2]         #サンプル名(
for(i in 1:length(tmpfname)){          #サンプル数(
  if(i == 1){
    k <- readGAlignments(tmpfname[i])  #BAM形式ファ
```

R Console

```
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> list.files()
 [1] "mapping_single_genome1.txt"
 [2] "QuasR_log_146c2f3d4e64.txt"
 [3] "ref_genome.fa"
 [4] "ref_genome.fa.fai"
 [5] "ref_genome.fa.md5"
 [6] "ref_genome.fa.Rbowtie"
 [7] "sample_RNAseq1.fa"
 [8] "sample_RNAseq1_146c6c6d54aa.bam"
 [9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
>
```

①

こんな感じで、エラーメッセージが出てなければOK

コピペ実行後

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015)

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtie1によるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"               #入力ファイル
param_mapping <- "-m 1 --best --strata -v 0"#マッピ

#必要なパッケージをロード
library(QuasR)                         #パッケージの
library(GenomicAlignments)             #パッケージの

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計
out <- qAlign(in_f1, in_f2, alignmentParameter=para
time_e <- proc.time()                  #計算時間を計
time_e - time_s                        #計算時間を表
out                                    #マッピングに
alignmentStats(out)                    #マッピング結

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(
tmpsname <- out@alignments[,2]         #サンプル名(
for(i in 1:length(tmpfname)){          #サンプル数(
  if(i == 1){
    k <- readGAlignments(tmpfname[i])  #BAM形式ファ
```

R Console

```
+ }
> m <- reduce(granges(k))                    #GRanges$
>
> #本番(カウント情報取得)
> tmp <- as.data.frame(m)                    #出力ファ$
> for(i in 1:length(tmpfname)){              #サンプル$
+   tmpcount <- summarizeOverlaps(m, tmpfname[i])#$
+   count <- assays(tmpcount)$counts         #Summariz$
+   colnames(count) <- tmpsname[i]           #行列coun$
+   tmp <- cbind(tmp, count)                 #保存した$
+ }
>
> #ファイルに保存
> out_f <- sub(".bam", "_range.txt", tmpfname[i])#$
> write.table(tmp, out_f, sep="\t", append=F, quot$
>
```

*_range.txt

①*_range.txtが、主目的のカウント情報データのファイルです

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015)

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル
in_f2 <- "ref_genome.fa"               #入力ファイル
param_mapping <- "-m 1 --best --strata -v 0"#マッピ

#必要なパッケージをロード
library(QuasR)                         #パッケージの
library(GenomicAlignments)             #パッケージの

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計
out <- qAlign(in_f1, in_f2, alignmentParameter=para
time_e <- proc.time()                  #計算時間を計
time_e - time_s                        #計算時間を表
out                                    #マッピングに
alignmentStats(out)                    #マッピング結

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(
tmpsname <- out@alignments[,2]         #サンプル名(
for(i in 1:length(tmpfname)){          #サンプル数(
  if(i == 1){
    k <- readGAlignments(tmpfname[i])  #BAM形式ファ
```

R Console

```
> out_f <- sub(".bam", "_range.txt", tmpfname[i])#$
> write.table(tmp, out_f, sep="\t", append=F, quot$
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> list.files()
 [1] "mapping_single_genome1.txt"
 [2] "QuasR_log_146c2f3d4e64.txt"
 [3] "ref_genome.fa"
 [4] "ref_genome.fa.fai"
 [5] "ref_genome.fa.md5"
 [6] "ref_genome.fa.Rbowtie"
 [7] "sample_RNAseq1.fa"
 [8] "sample_RNAseq1_146c6c6d54aa.bam"
 [9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
[13] "sample_RNAseq1_146c6c6d54aa_range.txt"
> 
```

①

*_range.txt

コピペしたコードの下部に移動。出力ファイルは何も指定していませんが、①*_range.txtという名前のファイルが作成されます。これは、②.bamという名前のファイルを内部的に入力として読み込み、その文字列中の.bamを_range.txtに置換したものを出力ファイル名として自動作成しているからそうなります

マップ後 | カウント情報取得 | single-end |

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノムへのマッピングからカウントまでの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref

オプションを"-m 1 --best --strata -v 0"とした例です。

```
for(i in 1:length(tmpfname)){            #サンプル数(
  if(i == 1){
    k <- readGAlignments(tmpfname[i])    #BAM形式ファ
  } else{
    k <- c(k, readGAlignments(tmpfname[i]))#BAM形式
  }
}
m <- reduce(granges(k))                  #GRangesオブ

#本番(カウント情報取得)
tmp <- as.data.frame(m)                  #出力ファイル
for(i in 1:length(tmpfname)){            #サンプル数(
  tmpcount <- summarizeOverlaps(m, tmpfname[i])#GRa
  count <- assays(tmpcount)$counts       #SummarizedE
  colnames(count) <- tmpsname[i]         #行列countの
  tmp <- cbind(tmp, count)               #保存したい情
}

#ファイルに保存
out_f <- sub(".bam", "_range.txt", tmpfname[i])#変更
write.table(tmp, out_f, sep="\t", append=F, quote=F
```

R Console

```
> out_f <- sub(".bam", "_range.txt", tmpfname[i])#$
> write.table(tmp, out_f, sep="\t", append=F, quot$
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso1"
> list.files()
 [1] "mapping_single_genome1.txt"
 [2] "QuasR_log_146c2f3d4e64.txt"
 [3] "ref_genome.fa"
 [4] "ref_genome.fa.fai"
 [5] "ref_genome.fa.md5"
 [6] "ref_genome.fa.Rbowtie"
 [7] "sample_RNAseq1.fa"
 [8] "sample_RNAseq1_146c6c6d54aa.bam"
 [9] "sample_RNAseq1_146c6c6d54aa.bam.bai"
[10] "sample_RNAseq1_146c6c6d54aa.bam.txt"
[11] "sample_RNAseq1_146c6c6d54aa.bed"
[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf"
[13] "sample_RNAseq1_146c6c6d54aa_range.txt"
>
```

②

①

*_range.txt

.bedファイルと*_range.txtファイルを見比べると理解が深まるでしょう。*_range.txtファイルの一番右側の列がカウント情報です。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015)

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_ge

オプションを"-m 1 --best --strata -v 0"とした例です。

```
in_f1 <- "mapping_single_genome1.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "ref_genome.fa"               #入力ファイル名を指定してin_f2に格納(リファレンス配列)
param_mapping <- "-m 1 --best --strata -v 0"#マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicAlignments)             #パッケージの読み込み

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping)#
                                       を計測するため
                                       を表示(一番右側の
                                       グに用いたパラメ
                                       グ結果(alignment
#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(in_f1の1列目に相当)をtmpfnameとして取り扱いたいだけです
tmpsname <- out@alignments[,2]         #サンプル名(in_f1の2列目に相当)をtmpsnameとして取り扱いたいだけです
for(i in 1:length(tmpfname)){          #サンプル数(ファイル数)分だけループを回す
  if(i == 1){
    k <- readGAlignments(tmpfname[i])  #BAM形式ファイルを読み込んだ結果をkに格納(これはGAlignmentsオブジェ
```

	A	B
1	FileName	SampleName
2	sample_RNAseq1.fa	namae

*.bed

chr1	11	45
chr2	1	35
chr2	16	50
chr3	1	35
chr3	3	37

*_range.txt

seqnames	start	end	width	strand	namae
chr1	11	45	35	+	1
chr2	1	50	50	+	2
chr3	1	37	37	+	2

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

複数サンプル

①例題5をやってみましょう。「デスクトップ－hoge－mapping_kiso2」フォルダを作成し、必要な入力ファイルを揃えてコピペ実行

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015)

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報がない場合を想定しているので、GenomicAlignments パッケージを利用して、マップされたリードの和集合領域(union range)を得たのち、領域ごとにマップされたリード数をカウントしています。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ18,19のRNA-seqデータ(sample_RNAseq1.fa)のref_genome.faへのマッピングの場合(mapping_single_genome1.txt):

オプションを"-m 1 --best --strata -v 0"とした例で

```
in_f1 <- "mapping_single_genome1.txt"
in_f2 <- "ref_genome.fa"
param_mapping <- "-m 1 --best --strata

#必要なパッケージをロード
library(QuasR)
library(GenomicAlignments)

#前処理(マッピング)
time_s <- proc.time()
out <- qAlign(in_f1, in_f2, alignmentP
time_e <- proc.time()
time_e - time_s
out
alignmentStats(out)

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]
tmpsname <- out@alignments[,2]
for(i in 1:length(tmpfname)){
  if(i == 1){
    k <- readGAlignments(tmpfname[i])
```

①

5.サンプルデータ18-20の複数のRNA-seqデータ(sample_RNAseq1.faとsample_RNAseq2.fa)をref_genome.faにマッピングする場合(mapping_single_genome4.txt):

全部のマッピング結果をまとめて和集合領域を定め、カウント情報を得るやり方です。一般的なカウントデータ行列の形式(2列目以降がカウント情報)にし、配列長情報と別々のファイルにして保存するやり方です。

```
in_f1 <- "mapping_single_genome4.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqフ
in_f2 <- "ref_genome.fa"               #入力ファイル名を指定してin_f2に格納(リファレン
out_f1 <- "hoge5_count.txt"            #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge5_genelength.txt"       #出力ファイル名を指定してout_f2に格納
param_mapping <- "-m 1 --best --strata -v 1"#マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicAlignments)             #パッケージの読み込み

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping)#マッピングを行うqAlign
time_e <- proc.time()                  #計算時間を計測するため
time_e - time_s                        #計算時間を表示(一番右側の数字。単位はsecond)
out                                    #マッピングに用いたパラメータや入力ファイルの情報
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。s

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(in_f1の1列目に相当)をtmpfnameとして
tmpsname <- out@alignments[,2]         #サンプル名(in_f1の2列目に相当)をtmpsnameとして
for(i in 1:length(tmpfname)){          #サンプル数(ファイル数)分だけループを回す
```

複数サンプル

①例題5をやってみましょう。「デスクトップ－hoge－mapping_kiso2」フォルダを作成し、必要な入力ファイルを揃えてコピペ実行

5.サンプルデータ18-20の複数のRNA-seqデータ(sample_RNAseq1.faとsample_RNAseq2.fa)をref_genome.faにマッピングする場合(mapping_single_genome4.txt):

全部のマッピング結果をまとめて和集合領域を定め、カウント情報を得るやり方です。一般的なカウントデータ行列の形式(2列目以降がカウント情報)にし、配列長情報と別々のファイルにして保存するやり方です。

```
in_f1 <- "mapping_single_genome4.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqフ
in_f2 <- "ref_genome.fa"               #入力ファイル名を指定してin_f2に格納(リファレン
out_f1 <- "hoge5_count.txt"            #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge5_genelength.txt"       #出力ファイル名を指定してout_f2に格納
param_mapping <- "-m 1 --best --strata -v 1"#マッピン

#必要なパッケージをロード
library(QuasR)                         #パッケージの読
library(GenomicAlignments)             #パッケージの読

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計測
out <- qAlign(in_f1, in_f2, alignmentParameter=param_
time_e <- proc.time()                  #計算時間を計測
time_e - time_s                        #計算時間を表示
out                                    #マッピングに用
alignmentStats(out)                    #マッピング結果

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(in_
tmpsname <- out@alignments[,2]         #サンプル名(in_
for(i in 1:length(tmpfname)){          #サンプル数(フ
```

R Console

```
詳しくは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の$
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみ$
'q()' と入力すれば R を終了します。

> getwd()
[1] "C:/Users/kadota/Desktop/hoge/mapping_kiso2"
> list.files()
[1] "mapping_single_genome4.txt"
[2] "ref_genome.fa"
[3] "sample_RNAseq1.fa"
[4] "sample_RNAseq2.fa"
>
```

複数サンプル

①2つの出力ファイルのうち、主に取り扱うのはカウントデータを含むファイル(hoge5_count.txt)のほうです

5.サンプルデータ18-20の複数のRNA-seqデータ(sample_RNAseq1.faとsample_RNAseq2.fa)をref_genome.faにマッピングする場合(mapping_single_genome4.txt):

全部のマッピング結果をまとめて和集合領域を定め、カウント情報を得るやり方です。一般的なカウントデータ行列の形式(2列目以降がカウント情報)にし、配列長情報と別々のファイルにして保存するやり方です

```
in_f1 <- "mapping_single_genome4.txt"  #入力ファイル名を
in_f2 <- "ref_genome.fa"               #入力ファイル名を
out_f1 <- "hoge5_count.txt"            #出力ファイル名を
out_f2 <- "hoge5_genelength.txt"       #出力ファイル名を
param_mapping <- "-m 1 --best --strata -v 1"#マッピング

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み
library(GenomicAlignments)             #パッケージの読み

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計測す
out <- qAlign(in_f1, in_f2, alignmentParameter=param_m
time_e <- proc.time()                  #計算時間を計測す
time_e - time_s                        #計算時間を表示(
out                                    #マッピングに用い
alignmentStats(out)                    #マッピング結果(

#本番(マップされたリードの和集合領域同定)
tmpfname <- out@alignments[,1]         #ファイル名(in_f
tmpsname <- out@alignments[,2]         #サンプル名(in_f
for(i in 1:length(tmpfname)){          #サンプル数(ファ
```

① hoge5_count.txt / hoge5_genelength.txt

R Console

```
> tmp <- cbind(tmp, h$width)                    #和集$
> write.table(tmp, out_f2, sep="\t", append=F, $
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso2"
> list.files()
 [1] "hoge5_count.txt"
 [2] "hoge5_genelength.txt"
 [3] "mapping_single_genome4.txt"
 [4] "QuasR_log_17ecaca51c5.txt"
 [5] "ref_genome.fa"
 [6] "ref_genome.fa.fai"
 [7] "ref_genome.fa.md5"
 [8] "ref_genome.fa.Rbowtie"
 [9] "sample_RNAseq1.fa"
[10] "sample_RNAseq1_17ec48601f6d.bam"
[11] "sample_RNAseq1_17ec48601f6d.bam.bai"
[12] "sample_RNAseq1_17ec48601f6d.bam.txt"
[13] "sample_RNAseq2.fa"
[14] "sample_RNAseq2_17ec1390671e.bam"
[15] "sample_RNAseq2_17ec1390671e.bam.bai"
[16] "sample_RNAseq2_17ec1390671e.bam.txt"
>
```

複数サンプル

①リストファイル中で指定した②サンプル名が、③のカウントデータ行列の④列名となります

5.サンプルデータ18-20の複数のRNA-seqデータ(sample_RNAseq1.faとsample_RNAseq2.fa)をref_genome.faにマッピングする場合(mapping_single_genome4.txt):

全部のマッピング結果をまとめて和集合領域を定め、カウント情報を得るやり方です。一般的なカウントデータ行列の形式(2列目以降がカウント情報)にし、配列長情報と別々のファイルにして保存するやり方です。

```
in_f1 <- "mapping_single_genome4.txt"  ①入力ファイル名を指定してin_f1に格納(RNA-seqフ
in_f2 <- "ref_genome.fa"               入力ファイル名を指定してin_f2に格納(リファレン
out_f1 <- "hoge5_count.txt"  ③         #出力ファイル名を指定してout_f1に格納
out_f2 <- "hoge5_genelength.t          #出力ファイル名を指定してout_f2に格納
param_mapping <- "-m 1 --best --strata -v 1"#マッピング時のオプションを指定

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicAlignments)             #パッケージの読み込み

#前処理(マッピング)
time_s <- proc.time()                  #計算時間を計測するため
out <- qAlign(in_f1, in_f2, alignmentParameter=param_mapping)#マッピングを行うqAlign
time_e <- proc.time()                  #計算時間を計測するため
time_e - time_s                        #計算時間を表示(一番右側の数
out                                    #マッピングに用いたパラメー
                                       ignment
```

FileName	SampleName
sample_RNAseq1.fa	sample1
sample_RNAseq2.fa	sample2

②

④

tmp	sample1	sample2
chr1_11_45_35_+	1	0
chr2_1_60_60_+	2	1
chr3_1_37_37_+	2	0
chr4_6_65_60_+	0	1
chr5_1_35_35_+	1	0

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

アノテーション情報あり

乳酸菌のアノテーション情報ファイル（.gff3）で定義されたgene領域にマップされるリード数をカウントするやり方を示します。①
single-endのアノテーション有のところです

(Rで)塩基配列解析

(last modified 2018/05/01, since 2010)

このウェブページのR関連部分は、インストール||についての推奨手順(Windows2018.03.12版とMacinto
トール済みであるという前提で記述しています。初心者の方は基本的な利用法(Windows2015.04.03版と
的にまとめた書籍もあります。(2015/04/03)

What's new?

- Silhouetteスコアの新たな使い道提唱論文(Z
- Silhouetteスコアの新たな使い道提唱論文(Z
- 「平成29年度NGSハンズオン講習会」の動画

- 門田からメール返信をもらえない場合は (las
- はじめに (last modified 2015/03/31)
- 参考資料 | 書籍、学会誌 (last modified 2017
- 参考資料 | 講習会、講義、講演資料 (last mo

- マッピング | 基礎 (last modified 2013/06/19)
- マッピング | single-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2014/06/21)
- マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2015/06/28)
- マッピング | single-end | ゲノム | splice-aware aligner | QuasR(Gaidatzis_2015) (last modified 2014/06/21)
- マッピング | paired-end | ゲノム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2016/02/11)
- マッピング | paired-end | トランスクリプトーム | basic aligner(基礎) | QuasR(Gaidatzis_2015) (last modified 2016/02/10)
- マッピング | paired-end | トランスクリプトーム | basic aligner(応用) | QuasR(Gaidatzis_2015) (last modified 2016/02/10)
- マップ後 | について (last modified 2013/06/19)
- マップ後 | 出力ファイル形式について (last modified 2013/11/05)
- マップ後 | 出力ファイルの読み込み | BAM形式 | について (last modified 2016/09/14)
- マップ後 | 出力ファイルの読み込み | BAM形式 | rbamtools(Kaisers_2015) (last modified 2016/09/14)
- マップ後 | 出力ファイルの読み込み | BAM形式 | GenomicAlignments(Lawrence_2013) (last modified 2016/09/14)
- マップ後 | 出力ファイルの読み込み | Bowtie形式 (last modified 2013/06/18)
- マップ後 | 出力ファイルの読み込み | SOAP形式 (last modified 2013/06/19)
- マップ後 | 出力ファイルの読み込み | htSeqTools(Planet_2012) (last modified 2013/06/19)
- マップ後 | カウント情報取得 | について (last modified 2017/01/11)
- マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) (① modified 2015/02/26)
- マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015) (last modified 2014/06/22)
- マップ後 | カウント情報取得 | paired-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) (last modified 2016/02/13)
- マップ後 | カウント情報取得 | paired-end | ゲノム | アノテーション無 | QuasR(Gaidatzis_2015) (last modified 2015/07/02)
- マップ後 | カウント情報取得 | paired-end | トランスクリプトーム | QuasR(Gaidatzis_2015) (last modified 2016/02/12)
- マップ後 | カウント情報取得 | トランスクリプトーム | BEDファイルから (last modified 2014/06/21)
- マップ後 | 配列長とカウント数の関係 (last modified 2015/07/03)
- 正規化 | について (last modified 2014/06/22)
- 正規化 | 基礎 | RPK or CPK (配列長補正) (last modified 2015/07/04)

アノテーション情報あり

①例題10。マップされる側のリファレンス配列は、②乳酸菌ゲノム配列。③アノテーションファイル。④機能ゲノム学の第1回（5/8）で、転写物配列取得時に使ったものと同じです。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis 2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ7

mapping_single_gen
のサンプル名」(例:hum
す。hg19にマップした結
で、Entrez Gene IDに対

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

①

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa"  #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した
alignmentStats(out)                    #マッピング結果(alignment statistics

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

②

③

講義日程（平成30年度）

1. 平成30年05月08日
   - 講義資料PDF
   - .gff3ファイル (約1.3MB)
   - .faファイル (約2.2MB)
   - (Rで)塩基配列解析
   - (Rで)マイクロアレイデータ解析
   - plasmid1.gff3(課題用)
   - plasmid2.gff3(課題用)

④

おさらい

①GFF3ファイルの中身は、こんな感じでした。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=ge
Chromosome ena transcript 360 1676 . + . ID=tr
Chromosome ena exon 360 1676 . + . Pare
Chromosome ena CDS 360 1676 . + 0 ID=C
###
Chromosome ena gene 1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon 1852 2991 . + . Pare
Chromosome ena CDS 1852 2991 . + 0 ID=C
###
Chromosome ena gene 3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon 3233 3457 . + . Pare
Chromosome ena CDS 3233 3457 . + 0 ID=C
###
Chromosome ena gene 3467 4588 . + . ID=ge
```

ン有 | QuasR(Gaidatzis_2015) NEW

るマッピングから、カウントデータ取得までの一連の
bオブジェクトをネットワーク経由で取得するのを基本
ンの詳細については マッピング | single-end | ゲノム |

をコピペ。

ァイルを乳酸菌ゲノムにマッピングする場合:

STA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino
Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル
0829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式の
s_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウン

```
#入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
#出力ファイル名を指定してout_fに格納
#カウントデータ取得時のレベルを指定："gene", "exon", "prom", "juncti

#パッケージの読み込み
#パッケージの読み込み

"auto")#txdbオブジェクトの作成
#確認してるだけです

#マッピングを行うqAlign関数を実行した結果をoutに格納
#マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                                  #行数と列数を表示
head(count)                                 #確認してるだけです
```

①

geneレベルのカウントデータ

①例題10は、②GFF3ファイルから、③gene領域内にマップされたリード数をカウントするやり方

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeaturesパッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ7

mapping_single_gen
のサンプル名」(例:hum
す。hg19にマップした結
で、Entrez Gene IDに対

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

①

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "prom      er", "junct

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

②

③

geneレベルのカウン

geneレベルとはいっても、バクテリアなので例えばexonを指定しても同じ結果になります。geneを指定したときは、①のようなgeneと書かれた行の赤枠領域にマップされたリード数を調べることになります。

```
##gff-version  3
##sequence-region   Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome e  gene        360  1676  .  +  .  ID=ge
Chromosome en transcript  360  1676  .  +  .  ID=tr
Chromosome ena exon       360  1676  .  +  .  Pare
Chromosome ena CDS        360  1676  .  +  0  ID=C
###
Chromosome e  gene        1852 2991  .  +  .  ID=ge
Chromosome en transcript  1852 2991  .  +  .  ID=tr
Chromosome ena exon       1852 2991  .  +  .  Pare
Chromosome ena CDS        1852 2991  .  +  0  ID=C
###
Chromosome e  gene        3233 3457  .  +  .  ID=ge
Chromosome en transcript  3233 3457  .  +  .  ID=tr
Chromosome ena exon       3233 3457  .  +  .  Pare
Chromosome ena CDS        3233 3457  .  +  0  ID=C
###
Chromosome e  gene        3467 4588  .  +  .  ID=ge
```

①

ファイルを乳酸菌ゲノムにマッピングする場合:

STA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル 0829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式の s_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウン

```
#入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
#出力ファイル名を指定してout_fに格納
#カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

#パッケージの読み込み
#パッケージの読み込み

"auto")#txdbオブジェクトの作成
#確認してるだけです

#マッピングを行うqAlign関数を実行した結果をoutに格納
#マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                                              #行数と列数を表示
head(count)                                             #確認してるだけです
```

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

マップする側のファイル

マップする側のファイルは、①のリストファイル内に記載されている、②sample_RNAseq4.fa。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"                  #①入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"                            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

#必要なパッケージをロード
library(QuasR)                                         #パッケージの読み込み
                                                       #パッケージの読み込み

                                                       o")#txdbオブジェクトの作成
                                                       認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)                            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                                             #行数と列数を表示
head(count)                                            #確認してるだけです
```

FileName	SampleName
sample_RNAseq4.fa	Lacto

マップする側のファイル

マップする側のファイルは、①のリストファイル内に記載されている、②sample_RNAseq4.fa。③が中身

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ7のFAS

mapping_single_genome のサンプル名」(例:hum
す。hg19にマップした結
で、Entrez Gene IDに対

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング
out <- qAlig
alignmentSta

#前処理(TxDb
txdb <- make
txdb

#本番(カウントデー
count <- qCount(o
```

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノム

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_R et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosc アノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829 ト情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"   #①入力ファイル名を指定
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829
out_f <- "hoge10.txt"                   #出力ファイル名を指定
param_reportlevel <- "gene"             #カウントデータ取得時

#必要なパッケージをロード
library(QuasR)                          #パッケージの読み込み
                                        #パッケージの読み込み

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)             #マッピングを行うqAli
alignmentStats(out)                     #マッピング結果(align

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)                              #行数と列数を表示
head(count)                             #確認してるだけです
```

FileName	SampleName
sample_RNAseq4.fa	Lacto

```
o")#txdbオブジェク
認してるだけです
```

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

マップする側のファイル

①sample_RNAseq4.faは、②サンプルデータの例題47のコピペで作成しています。わざわざ見に行かなくてもよい

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファイ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

マップする側のファイル

①マップする側（sample_RNAseq4.fa）のリードは、②アノテーションファイル（.gff3）中のgene領域を参考にしながら作成しています。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

①

②

②

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa"#入力ファイル
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junction

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

マップする側のファイル

①マップする側（sample_RNAseq4.fa）のリードは、②アノテーションファイル（.gff3）中の③gene領域を参考にしながら作成しています。
全て完全一致でマップされるように設計。

②

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome  ena  gene        360  1676  .  +  .  ID=ge
Chromosome  ena  transcript  360  1676  .  +  .  ID=tr
Chromosome  ena  exon        360  1676  .  +  .  Pare
Chromosome  ena  CDS         360  1676  .  +  0  ID=C
###
Chromosome  ena  gene        1852  2991  .  +  .  ID=ge
Chromosome  ena  transcript  1852  2991  .  +  .  ID=tr
Chromosome  ena  exon        1852  2991  .  +  .  Pare
Chromosome  ena  CDS         1852  2991  .  +  0  ID=C
###
Chromosome  ena  gene        3233  3457  .  +  .  ID=ge
Chromosome  ena  transcript  3233  3457  .  +  .  ID=tr
Chromosome  ena  exon        3233  3457  .  +  .  Pare
Chromosome  ena  CDS         3233  3457  .  +  0  ID=C
###
Chromosome  ena  gene        3467  4588  .  +  .  ID=ge
```

③ ③ ③ ③

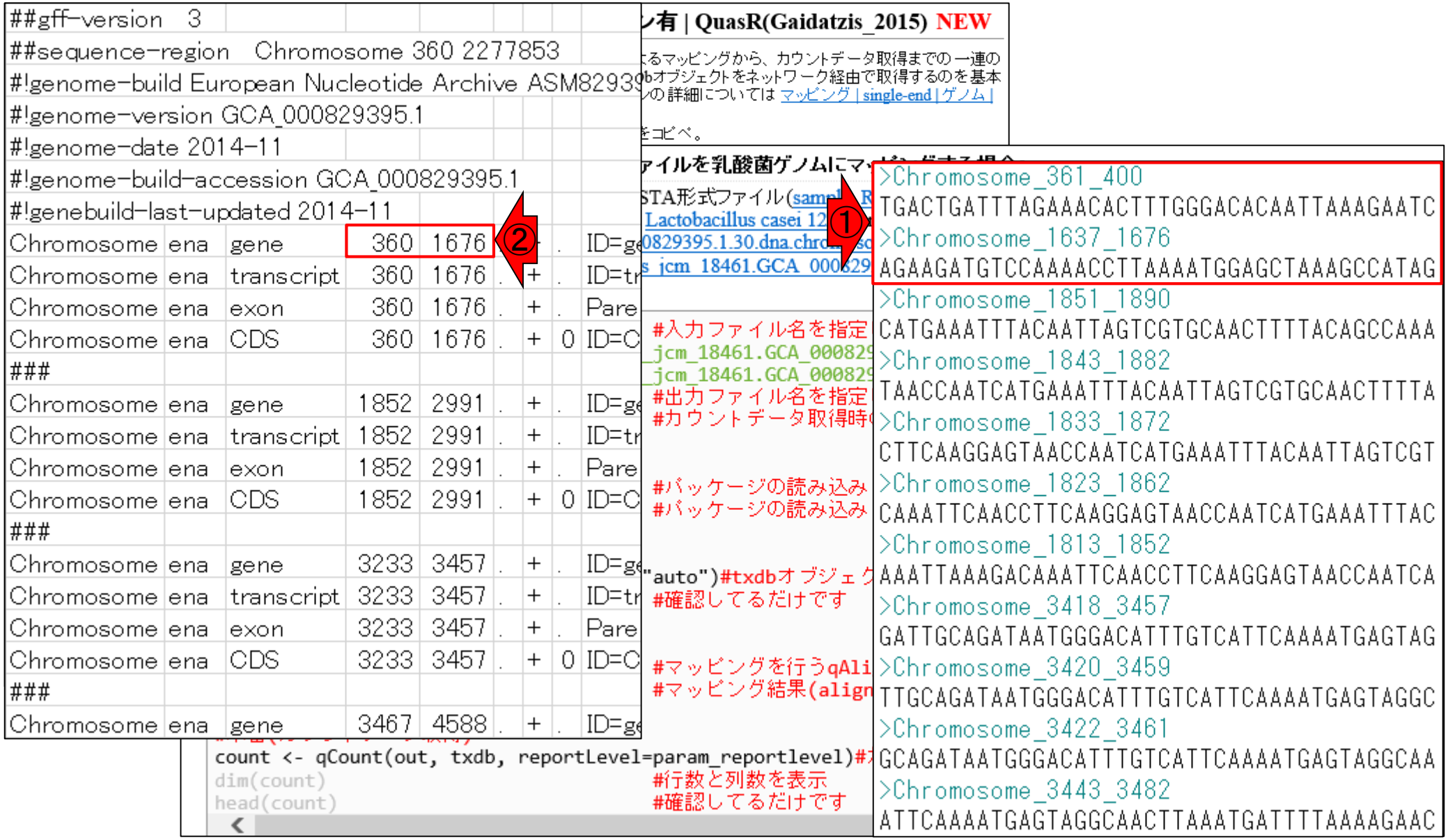
```
#入力ファイル名を指定
_jcm_18461.GCA_000829
_jcm_18461.GCA_000829
#出力ファイル名を指定
#カウントデータ取得時
#パッケージの読み込み
#パッケージの読み込み
"auto")#txdbオブジェク
#確認してるだけです
#マッピングを行うqAli
#マッピング結果(align
count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
#行数と列数を表示
head(count)
#確認してるだけです
```

①

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

マップする側のファイル

①最初の2リードは、②の領域内にマップされるように設計。



```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=ge
Chromosome ena transcript 360 1676 . + . ID=tr
Chromosome ena exon 360 1676 . + . Pare
Chromosome ena CDS 360 1676 . + 0 ID=C
###
Chromosome ena gene 1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon 1852 2991 . + . Pare
Chromosome ena CDS 1852 2991 . + 0 ID=C
###
Chromosome ena gene 3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon 3233 3457 . + . Pare
Chromosome ena CDS 3233 3457 . + 0 ID=C
###
Chromosome ena gene 3467 4588 . + . ID=ge
```

②

ン有 | QuasR(Gaidatzis_2015) NEW

るマッピングから、カウントデータ取得までの一連のbオブジェクトをネットワーク経由で取得するのを基本の詳細については マッピング | single-end | ゲノム |

をコピペ。

ァイルを乳酸菌ゲノムにマ

STA形式ファイル(sampl

Lactobacillus casei 12

0829395.1.30.dna.chro

s_jcm_18461.GCA_000829

①

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

#入力ファイル名を指定
_jcm_18461.GCA_000829
_jcm_18461.GCA_000829
#出力ファイル名を指定
#カウントデータ取得時

#パッケージの読み込み
#パッケージの読み込み

"auto")#txdbオブジェク
#確認してるだけです

#マッピングを行うqAli
#マッピング結果(align

```
count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
head(count)
```

#行数と列数を表示
#確認してるだけです

マップする側のファイル

①3-7番目の5リードは、②の領域に一部がかかるように設計。領域内ではない。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=ge
Chromosome ena transcript 360 1676 . + . ID=tr
Chromosome ena exon 360 1676 . + . Pare
Chromosome ena CDS 360 1676 . + 0 ID=C
###
Chromosome ena gene 1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon 1852 2991 . + . Pare
Chromosome ena CDS 1852 2991 . + 0 ID=C
###
Chromosome ena gene 3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon 3233 3457 . + . Pare
Chromosome ena CDS 3233 3457 . + 0 ID=C
###
Chromosome ena gene 3467 4588 . + . ID=ge
```

②

①

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

```
count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
head(count)
```

マップする側のファイル

①8番目のリードは、②の領域内にマップされるように設計。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=ge
Chromosome ena transcript 360 1676 . + . ID=tr
Chromosome ena exon 360 1676 . + . Pare
Chromosome ena CDS 360 1676 . + 0 ID=C
###
Chromosome ena gene 1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon 1852 2991 . + . Pare
Chromosome ena CDS 1852 2991 . + 0 ID=C
###
Chromosome ena gene 3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon 3233 3457 . + . Pare
Chromosome ena CDS 3233 3457 . + 0 ID=C
###
Chromosome ena gene 3467 4588 . + . ID=ge
```

②

ン有 | QuasR(Gaidatzis_2015) NEW

るマッピングから、カウントデータ取得までの一連のbオブジェクトをネットワーク経由で取得するのを基本ンの詳細については マッピング | single-end | ゲノム |

をコピペ。

ァイルを乳酸菌ゲノムにマッ

STA形式ファイル(sample_R

Lactobacillus casei 12Aの

0829395.1.30.dna.chromoso

s_jcm_18461.GCA_000829

```
#入力ファイル名を指定
_jcm_18461.GCA_00082
_jcm_18461.GCA_00082
#出力ファイル名を指定
#カウントデータ取得時

#パッケージの読み込み
#パッケージの読み込み

"auto")#txdbオブジェク
#確認してるだけで

#マッピングを行うqAli
#マッピング結果(align

count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
head(count)
#行数と列数を表示
#確認してるだけです
```

①

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

マップする側のファイル

①9-10番目のリードは、②の領域に一部がかかるように設計。領域内ではない。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=ge
Chromosome ena transcript 360 1676 . + . ID=tr
Chromosome ena exon 360 1676 . + . Pare
Chromosome ena CDS 360 1676 . + 0 ID=C
###
Chromosome ena gene 1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon 1852 2991 . + . Pare
Chromosome ena CDS 1852 2991 . + 0 ID=C
###
Chromosome ena gene 3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon 3233 3457 . + . Pare
Chromosome ena CDS 3233 3457 . + 0 ID=C
###
Chromosome ena gene 3467 4588 . + . ID=ge
```

②

ン有 | QuasR(Gaidatzis_2015) NEW

るマッピングから、カウントデータ取得までの一連の
bオブジェクトをネットワーク経由で取得するのを基本
ンの詳細については マッピング | single-end | ゲノム |

をコピペ。

ァイルを乳酸菌ゲノムにマッ

STA形式ファイル(sample_R
Lactobacillus casei 12Aの
0829395.1.30.dna.chromoso
s_jcm_18461.GCA_000829

```
#入力ファイル名を指定
_jcm_18461.GCA_00082
_jcm_18461.GCA_00082
#出力ファイル名を指定
#カウントデータ取得時

#パッケージの読み込み
#パッケージの読み込み

"auto")#txdbオブジェク
#確認してるだけです

#マッピングを行うqAli
#マッピング結果(a
```

①

```
count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
head(count)
```

#行数と列数を表示
#確認してるだけです

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

マップする側のファイル

①11番目のリードは、②と③の領域にまたがってマップされるように設計。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=ge
Chromosome ena transcript 360 1676 . + . ID=tr
Chromosome ena exon 360 1676 . + . Pare
Chromosome ena CDS 360 1676 . + 0 ID=C
###
Chromosome ena gene 1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon 1852 2991 . + . Pare
Chromosome ena CDS 1852 2991 . + 0 ID=C
###
Chromosome ena gene 3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon 3233 3457 . + . Pare
Chromosome ena CDS 3233 3457 . + 0 ID=C
###
Chromosome ena gene 3467 4588 . + . ID=ge
```

②

③

ン有 | QuasR(Gaidatzis_2015) NEW

るマッピングから、カウントデータ取得までの一連のbオブジェクトをネットワーク経由で取得するのを基本ンの詳細については マッピング | single-end | ゲノム |

をコピペ。

ァイルを乳酸菌ゲノムにマッ

STA形式ファイル(sample_R

Lactobacillus casei 12Aの

0829395.1.30.dna.chromoso

s_jcm_18461.GCA_000829

```
#入力ファイル名を指定
_jcm_18461.GCA_00082
_jcm_18461.GCA_00082
#出力ファイル名を指定
#カウントデータ取得時

#パッケージの読み込み
#パッケージの読み込み

"auto")#txdbオブジェク
#確認してるだけです

#マッピングを行うqAli
#マッピング結果(align

count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
head(count)
#行数と列数を表示
#確認してるだけで
```

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

①

最低限3リードは…

全11リード中①これらの3リードは、②の領域内にマップされる。従って、得られるカウントの総和は、最低でも3はあるはず。という予想を立てて実際にマッピングを行う。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene       360  1676 . + . ID=ge
Chromosome ena transcript 360  1676 . + . ID=tr
Chromosome ena exon       360  1676 . + . Pare
Chromosome ena CDS        360  1676 . + 0 ID=C
###
Chromosome ena gene       1852 2991 . + . ID=ge
Chromosome ena transcript 1852 2991 . + . ID=tr
Chromosome ena exon       1852 2991 . + . Pare
Chromosome ena CDS        1852 2991 . + 0 ID=C
###
Chromosome ena gene       3233 3457 . + . ID=ge
Chromosome ena transcript 3233 3457 . + . ID=tr
Chromosome ena exon       3233 3457 . + . Pare
Chromosome ena CDS        3233 3457 . + 0 ID=C
###
Chromosome ena gene       3467 4588 . + . ID=ge
```

②

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

①

```
#入力ファイル名を指定
#出力ファイル名を指定
#カウントデータ取得時
#パッケージの読み込み
#パッケージの読み込み
"auto")#txdbオブジェク
#確認してるだけで
#マッピングを行うqAli
#マッピング結果(align
count <- qCount(out, txdb, reportLevel=param_reportlevel)#
dim(count)
head(count)
#行数と列数を表示
#確認してるだけです
```

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

①例題10をやってみましょう。

マッピング実行

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ7

mapping_single_gen
のサンプル名」(例:hum
す。hg19にマップした結
で、Entrez Gene IDに対

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

①

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa) です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3) を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

マッピング実行

①例題10をやってみましょう。②「デスクトップ – hoge – mapping_kiso3」フォルダを作成し、③必要な入力ファイルを揃えてコピペ実行

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeaturesパッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ7

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

①

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"
param_reportlevel <- "gene"

#必要なパッケージをロード
library(QuasR)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, f
txdb

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)
alignmentStats(out)

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                              #行数と列数を表示
head(count)                             #確認してるだけです
```

③

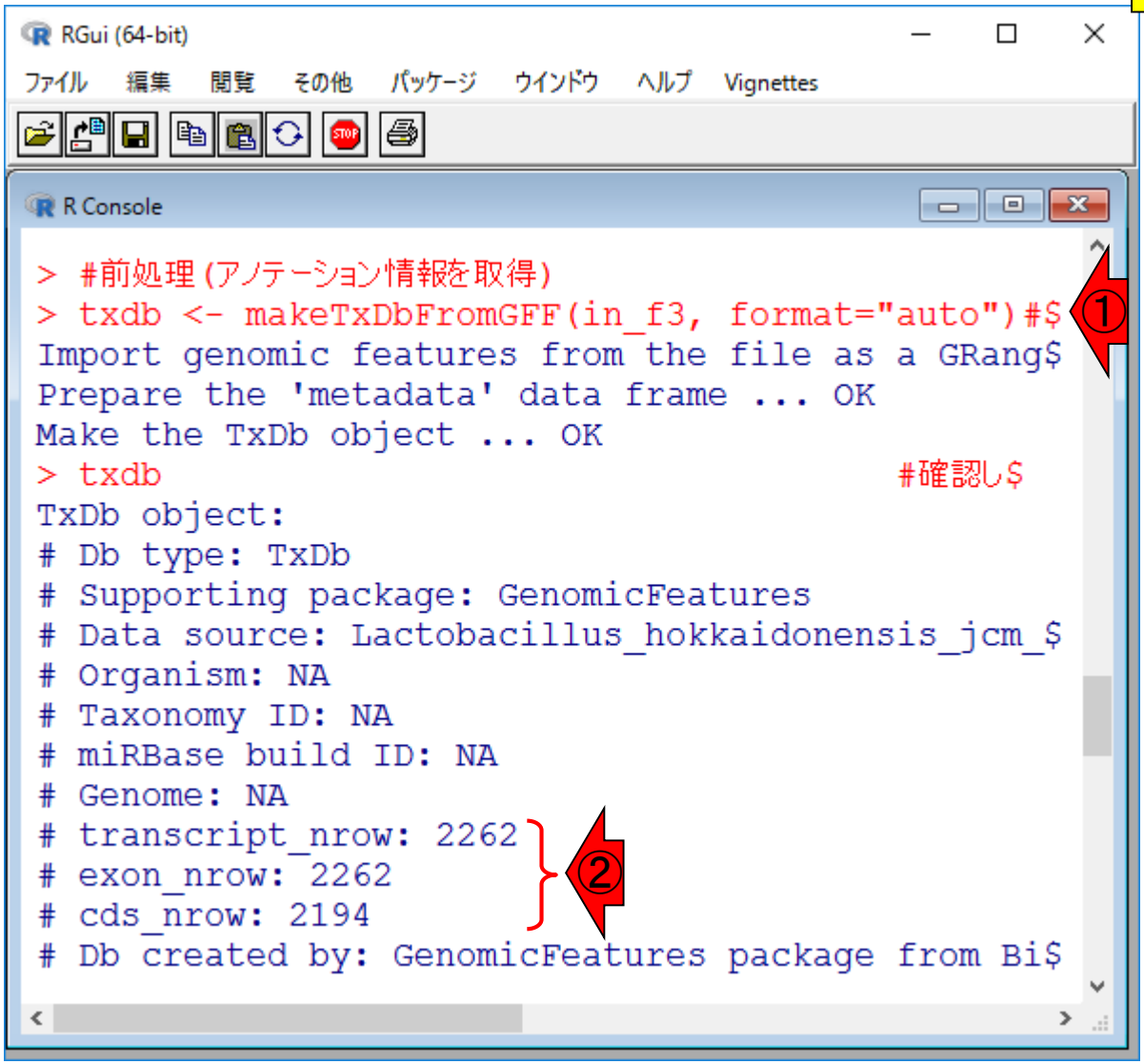
R Console

```
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso3"
> list.files()
[1] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
[2] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
[3] "mapping_single_genome7.txt"                $
[4] "sample_RNAseq4.fa"                         $
>
```

②

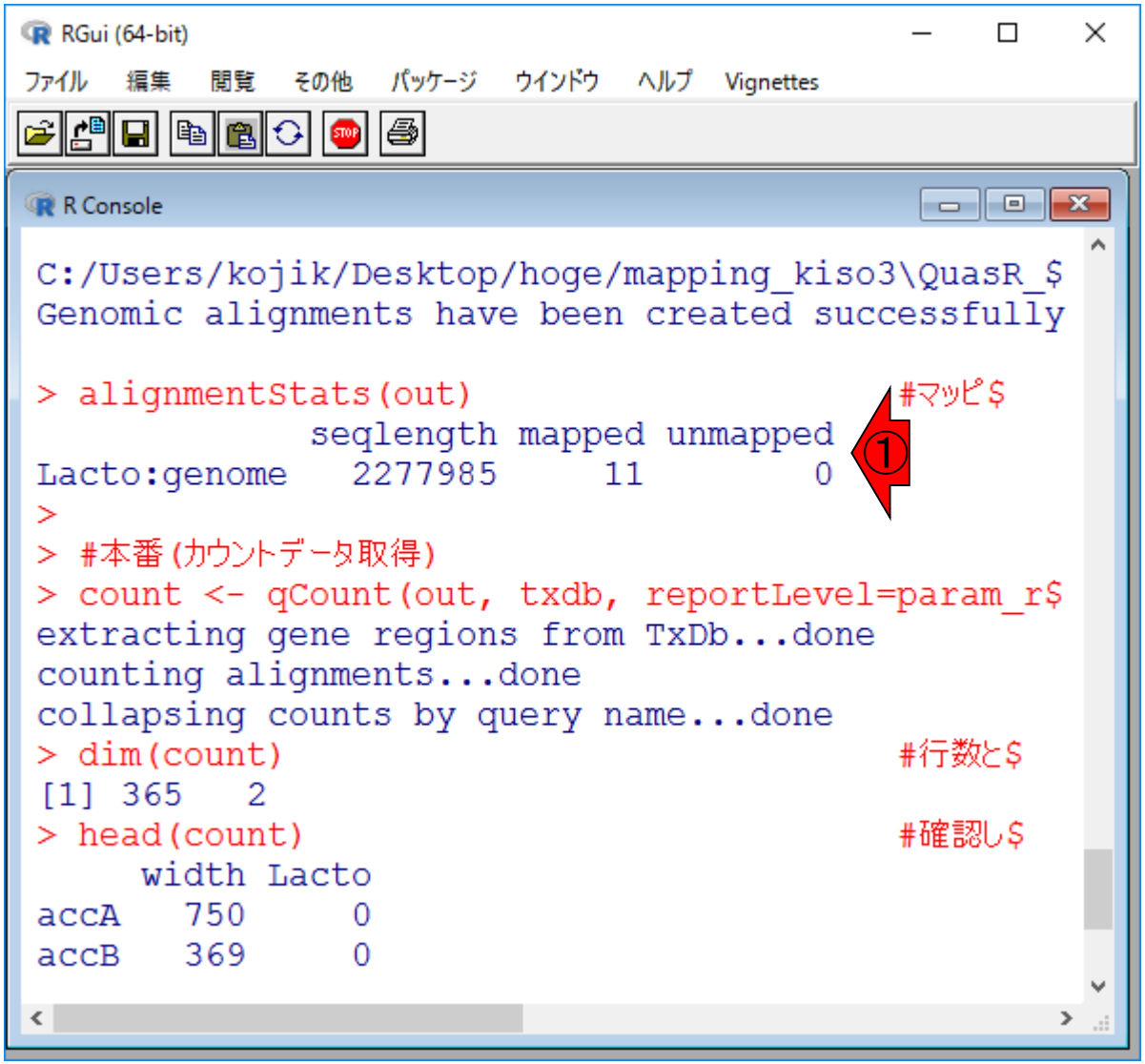
途中経過

①アノテーションファイル(.gff3)読み込みのあたり。得られるカウントデータは、②2,000行以上からなるのだろう、などと妄想



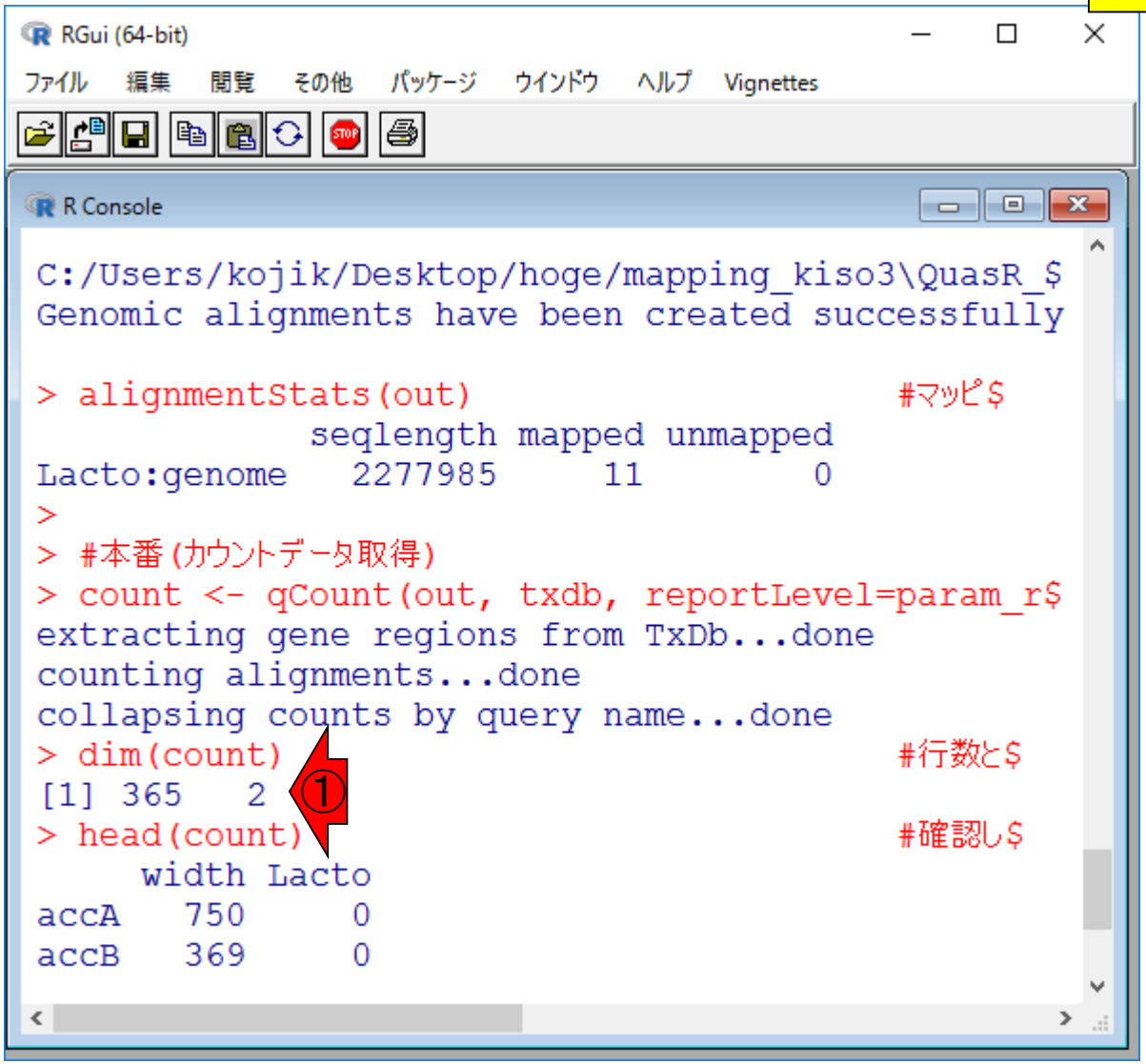
途中経過

①全11リードがマップされたのは妥当。そのように設計しているから



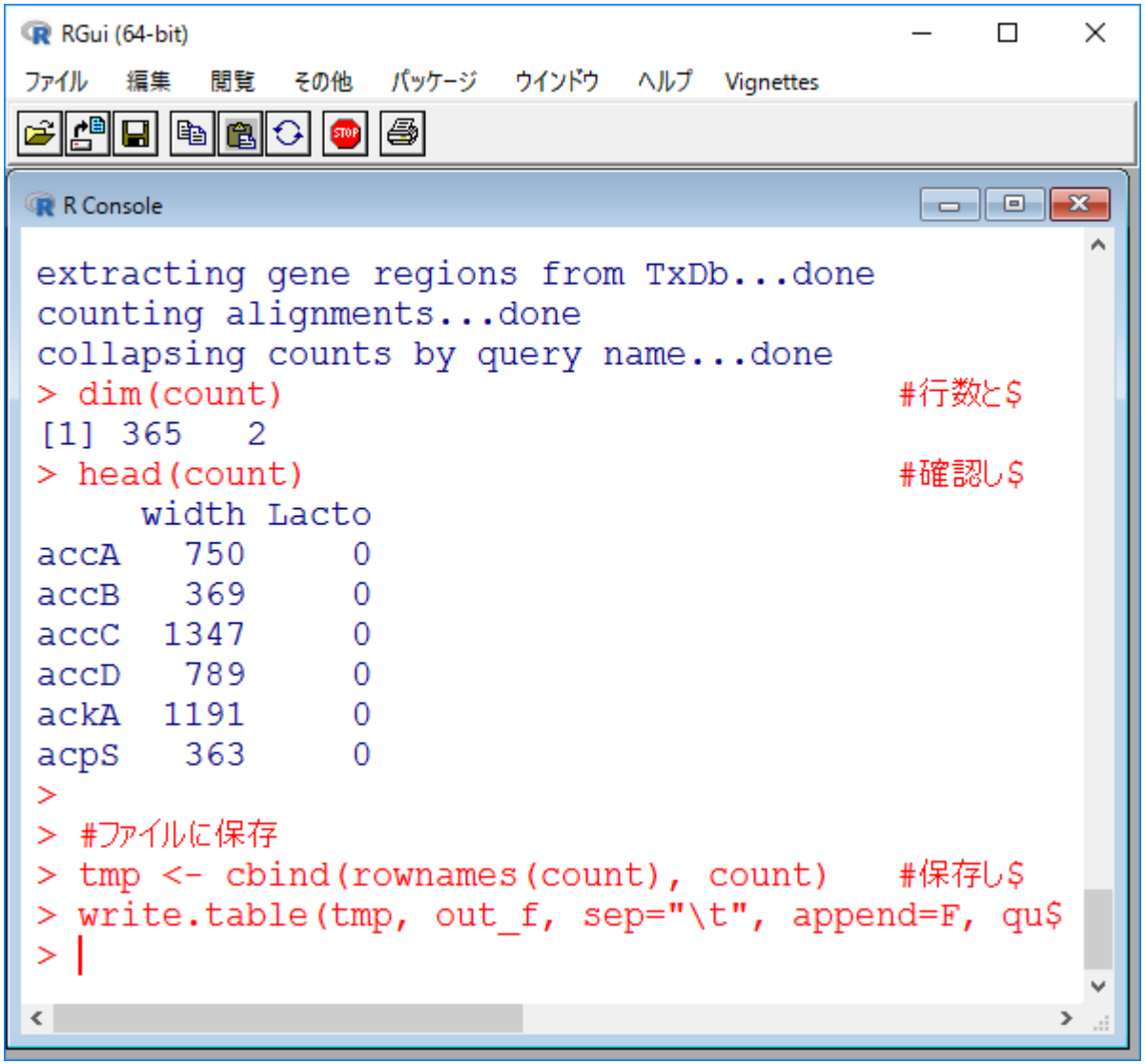
ん?!

①列数はともかく、なんで365行しかないのだろう?!2000行以上ないとオカシイはずなのに…と疑問を持ちつつも、最後まで眺める。



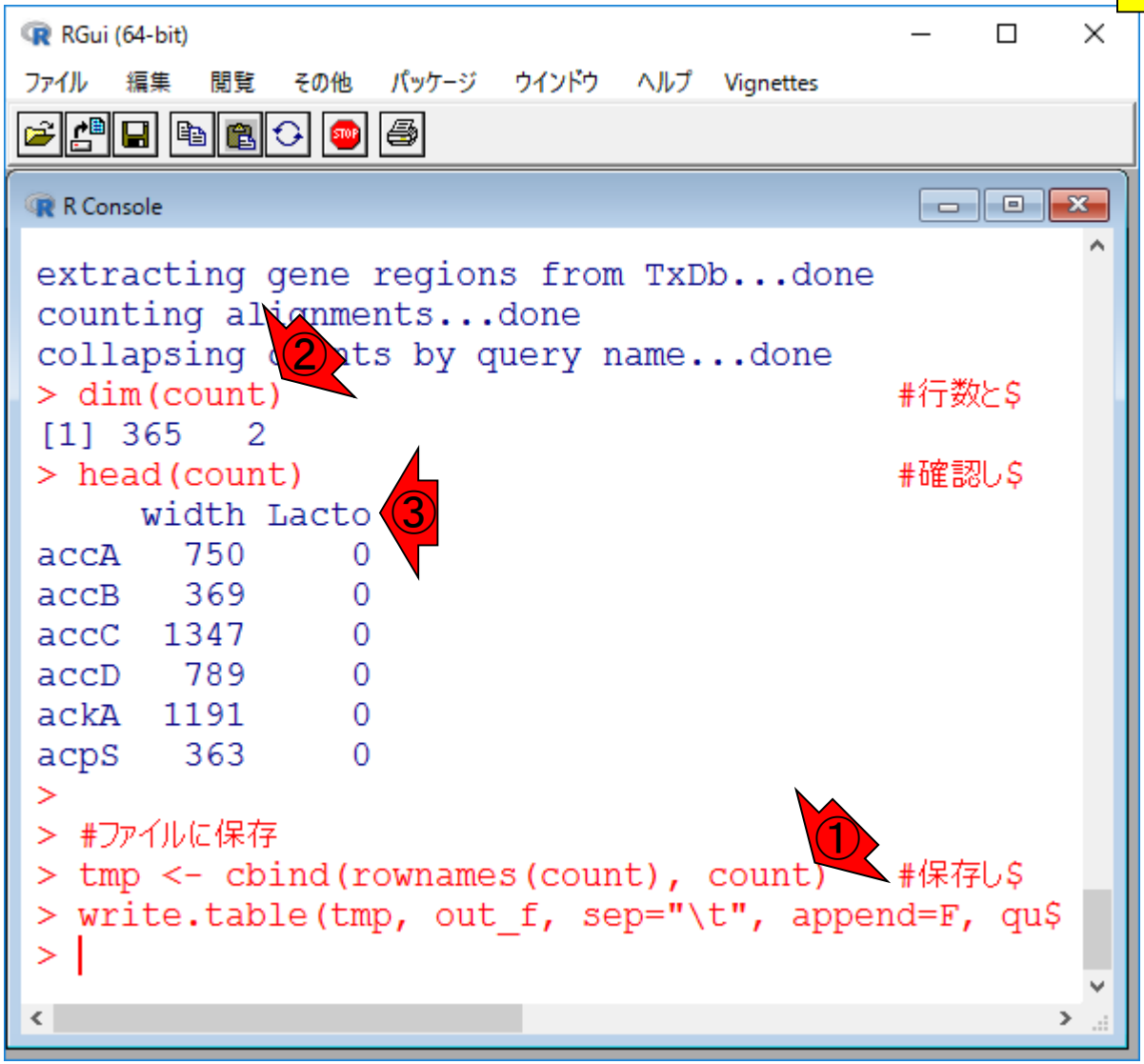
最後まで完了

最後まで無事完了したら、こんな感じになります



カウントデータの概要

①countオブジェクトが、得たいカウントデータ情報。②（ヘッダー部分を除く）行数は365で、列数は2。③Lactoという列名が…



カウントデータの列名

①のリストファイル（mapping_single_genome7.txt）中に、②列名として記述したLactoです。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  ①入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "juncti

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
                                       #パッケージの読み込み

                                       o")#txdbオブジェクトの作成
                                       認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

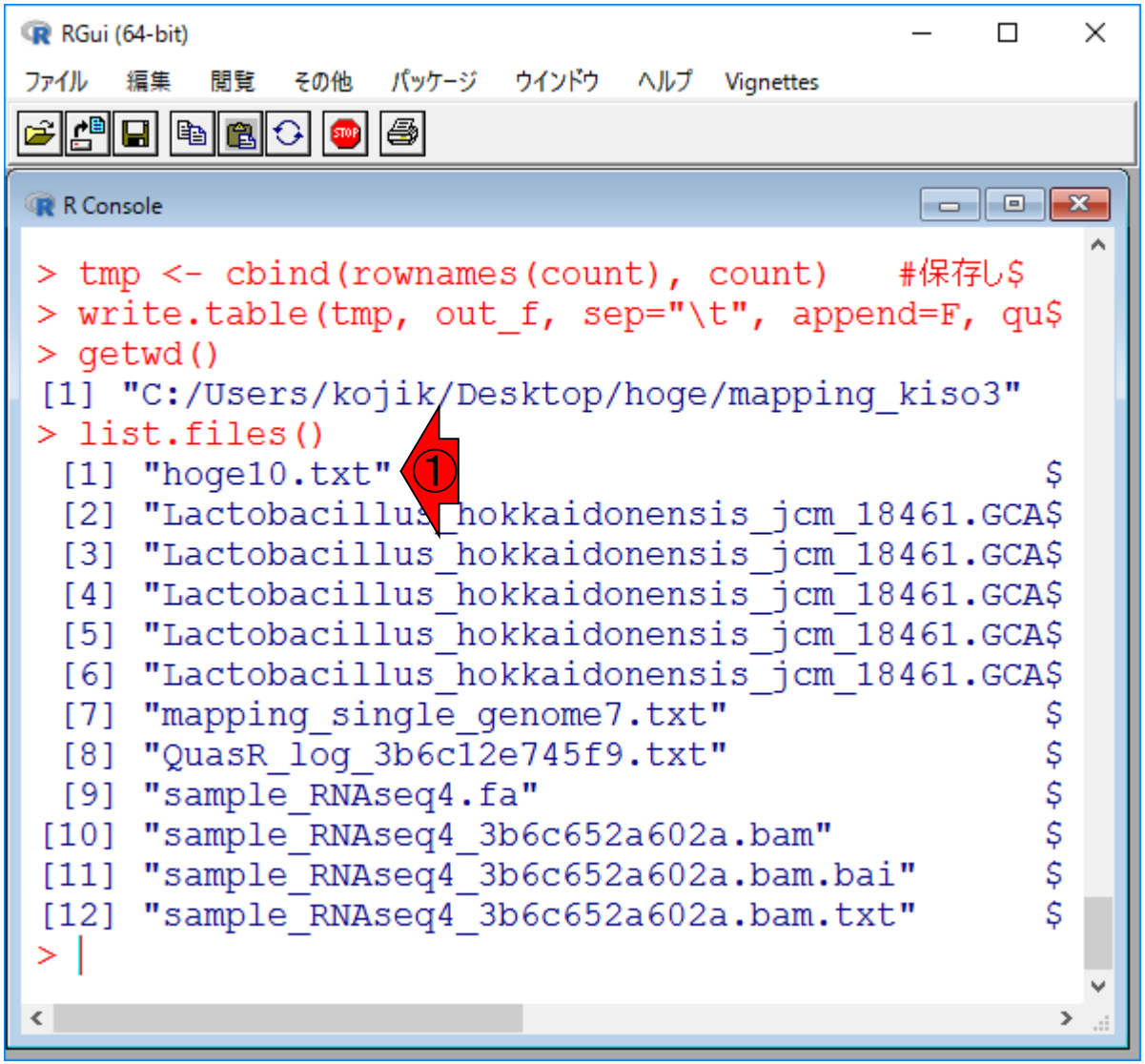
FileName	SampleName
sample_RNAseq4.fa	Lacto ②

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - **結果の解釈**
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

カウントデータ

マッピング実行後は、こんな感じになります。①hoge10.txtが目的のカウントデータです

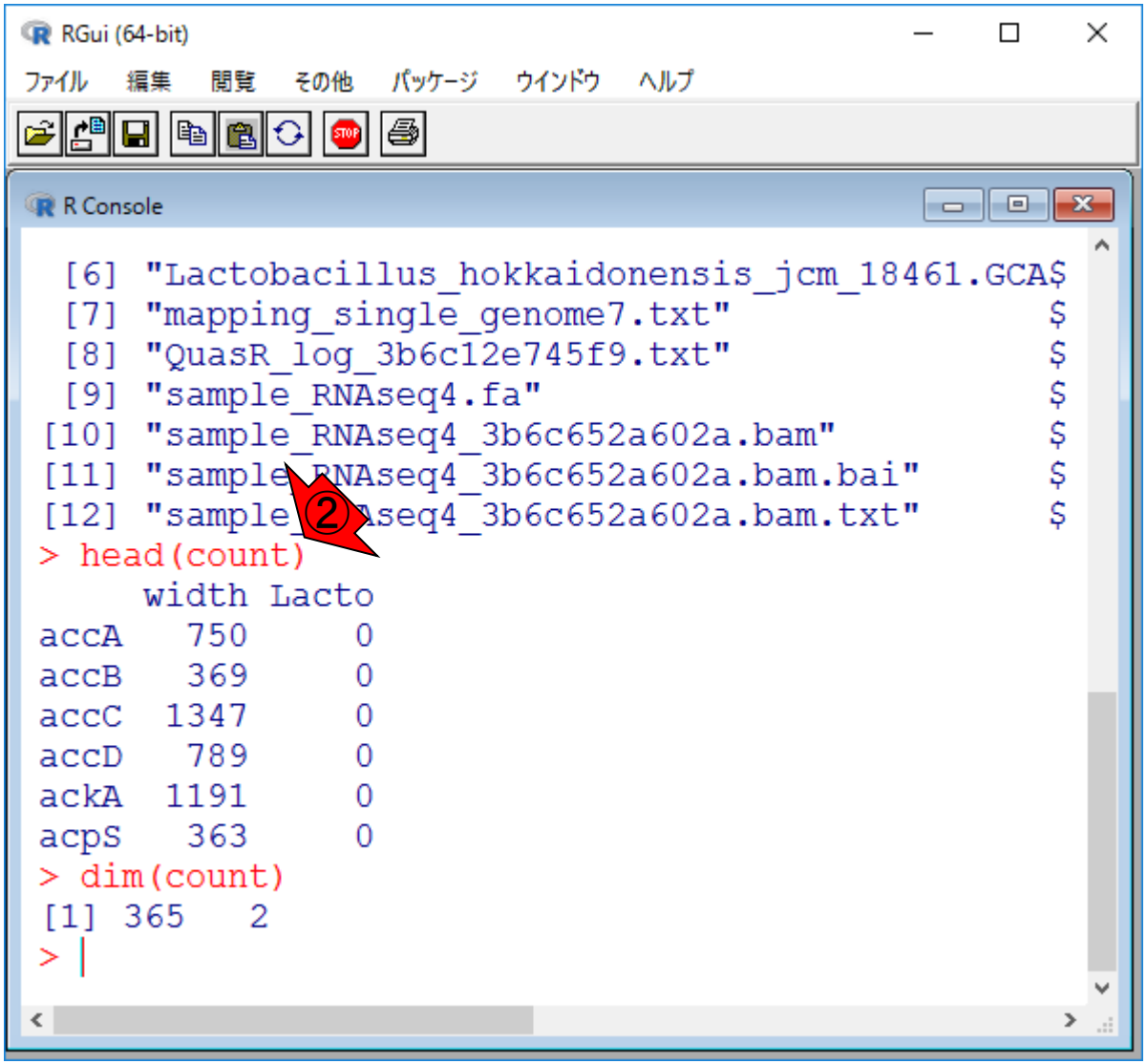


①

	width	Lacto
accA	750	0
accB	369	0
accC	1347	0
accD	789	0
ackA	1191	0
acpS	363	0
addA	3711	0
adh	1056	0
adhE	2607	0
adk	660	0
alaS	2655	0
aldA	1464	0
aldB	714	0
amt	1323	0

カウントデータ

①hoge10.txtの中身は、②countオブジェクトと同じです



①

	width	Lacto
accA	750	0
accB	369	0
accC	1347	0
accD	789	0
ackA	1191	0
acpS	363	0
addA	3711	0
adh	1056	0
adhE	2607	0
adk	660	0
alaS	2655	0
aldA	1464	0
aldB	714	0
amt	1323	0

1行目のカウントが0

①最初におやっ?!と思うのは、1行目のカウントが0であるという点です。この理由は…

RGui (64-bit)

ファイル　編集　閲覧　その他　パッケージ　ウインドウ　ヘルプ

R Console

```
 [6] "Lactobacillus_hokkaidonensis_jcm_18461.GCA$
 [7] "mapping_single_genome7.txt"                $
 [8] "QuasR_log_3b6c12e745f9.txt"                $
 [9] "sample_RNAseq4.fa"                         $
[10] "sample_RNAseq4_3b6c652a602a.bam"           $
[11] "sample_RNAseq4_3b6c652a602a.bam.bai"       $
[12] "sample_RNAseq4_3b6c652a602a.bam.txt"       $
> head(count)
      width Lacto
accA    750     0
accB    369     0
accC   1347     0
accD    789     0
ackA   1191     0
acpS    363     0
> dim(count)
[1] 365   2
>
```

①

	width	Lacto
accA	750	0
accB	369	0
accC	1347	0
accD	789	0
ackA	1191	0
acpS	363	0
addA	3711	0
adh	1056	0
adhE	2607	0
adk	660	0
alaS	2655	0
aldA	1464	0
aldB	714	0
amt	1323	0

1行目のカウントが0

①GFF3ファイル内に出現する最初のgeneのカウントは、②2だと思っていたからです

```
##gff-version  3
##sequence-region   Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome  ena  gene        360   1676  .  +  .  ID=ge
Chromosome  ena  transcript  360   1676  .  +  .  ID=tr
Chromosome  ena  exon        360   1676  .  +  .  Pare
Chromosome  ena  CDS         360   1676  .  +  0  ID=C
###
Chromosome  ena  gene        1852  2991  .  +  .  ID=ge
Chromosome  ena  transcript  1852  2991  .  +  .  ID=tr
Chromosome  ena  exon        1852  2991  .  +  .  Pare
Chromosome  ena  CDS         1852  2991  .  +  0  ID=C
###
Chromosome  ena  gene        3233  3457  .  +  .  ID=ge
Chromosome  ena  transcript  3233  3457  .  +  .  ID=tr
Chromosome  ena  exon        3233  3457  .  +  .  Pare
Chromosome  ena  CDS         3233  3457  .  +  0  ID=C
###
Chromosome  ena  gene        3467  4588  .  +  .  ID=ge
```

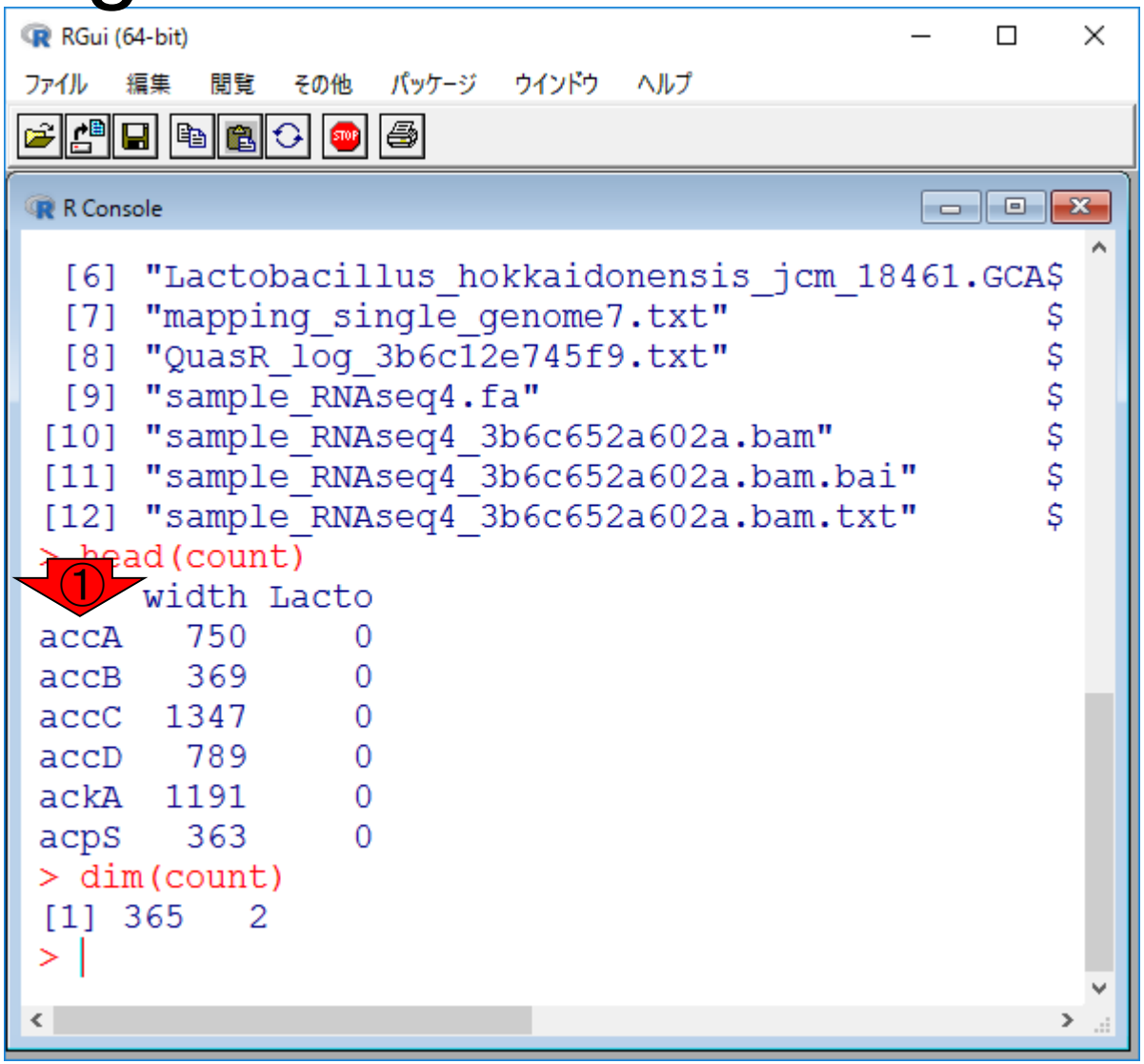
①

②

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

geneレベルのカウントデータ

geneレベルのカウントデータを得ているので、①が遺伝子名(genename)のアルファベット順にソートされているのだろう。



①	width	Lacto
accA	750	0
accB	369	0
accC	1347	0
accD	789	0
ackA	1191	0
acpS	363	0
addA	3711	0
adh	1056	0
adhE	2607	0
adk	660	0
alaS	2655	0
aldA	1464	0
aldB	714	0
amt	1323	0

GFFファイルをよく見る

GFFファイル内にgenenameの情報があるはずだという確信のもと、①のあたりをよく見る

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ①ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + . ①ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + . ①ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + . ①ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

GFFファイルをよく

①赤枠内のName=の右側の文字がgenenameのようですね。②この遺伝子領域にはName=*genename*がないこともわかる。これらが原因で2000行超にはならずに365行となってしまったのかも…と考える。そして、②の領域[3233, 3457]内には、確か少なくとも1リード完璧にマップされたような…といううろ覚えの記憶をたどる。

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + ① ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + ① ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + ② ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + ① ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

dnaAのカウント数

少なくとも2リードは完全に領域内にマップされているはずの、①dnaAのカウント数を調べる。

①

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + . ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + . ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + . ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

dnaAのカウント数

少なくとも2リードは完全に領域内にマップされているはずの、①dnaAのカウント数を調べる。②カウント数は2ですね。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene       360  1676 . + . ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360  1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon       360  1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS        360  1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene       1852 2991 . + . ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon       1852 2991 . + . P
Chromosome ena CDS        1852 2991 . + 0 ID
###
Chromosome ena gene       3233 3457 . + . ID
Chromosome ena transcript 3233 3457 . + . ID
Chromosome ena exon       3233 3457 . + . P
Chromosome ena CDS        3233 3457 . + 0 ID
###
Chromosome ena gene       3467 4588 . + . ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

①

R Console

②

```
> count["dnaA",]
width Lacto
 1317      2
> 1676 - 360 + 1
[1] 1317
>
```

dnaAのカウント数

少なくとも2リードは完全に領域内にマップされているはずの、①dnaAのカウント数を調べる。②カウント数は2ですね。ちなみに、③1317は、dnaAの領域の長さです。

```
##gff-version  3
##sequence-region  Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene       360  1676 . + . ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360  1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon       360  1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS        360  1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene       1852 2991 . + . ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon       1852 2991 . + . P
Chromosome ena CDS        1852 2991 . + 0 ID
###
Chromosome ena gene       3233 3457 . + .
Chromosome ena transcript 3233 3457 . + . I
Chromosome ena exon       3233 3457 . + . P
Chromosome ena CDS        3233 3457 . + 0 ID
###
Chromosome ena gene       3467 4588 . + . ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

①

②

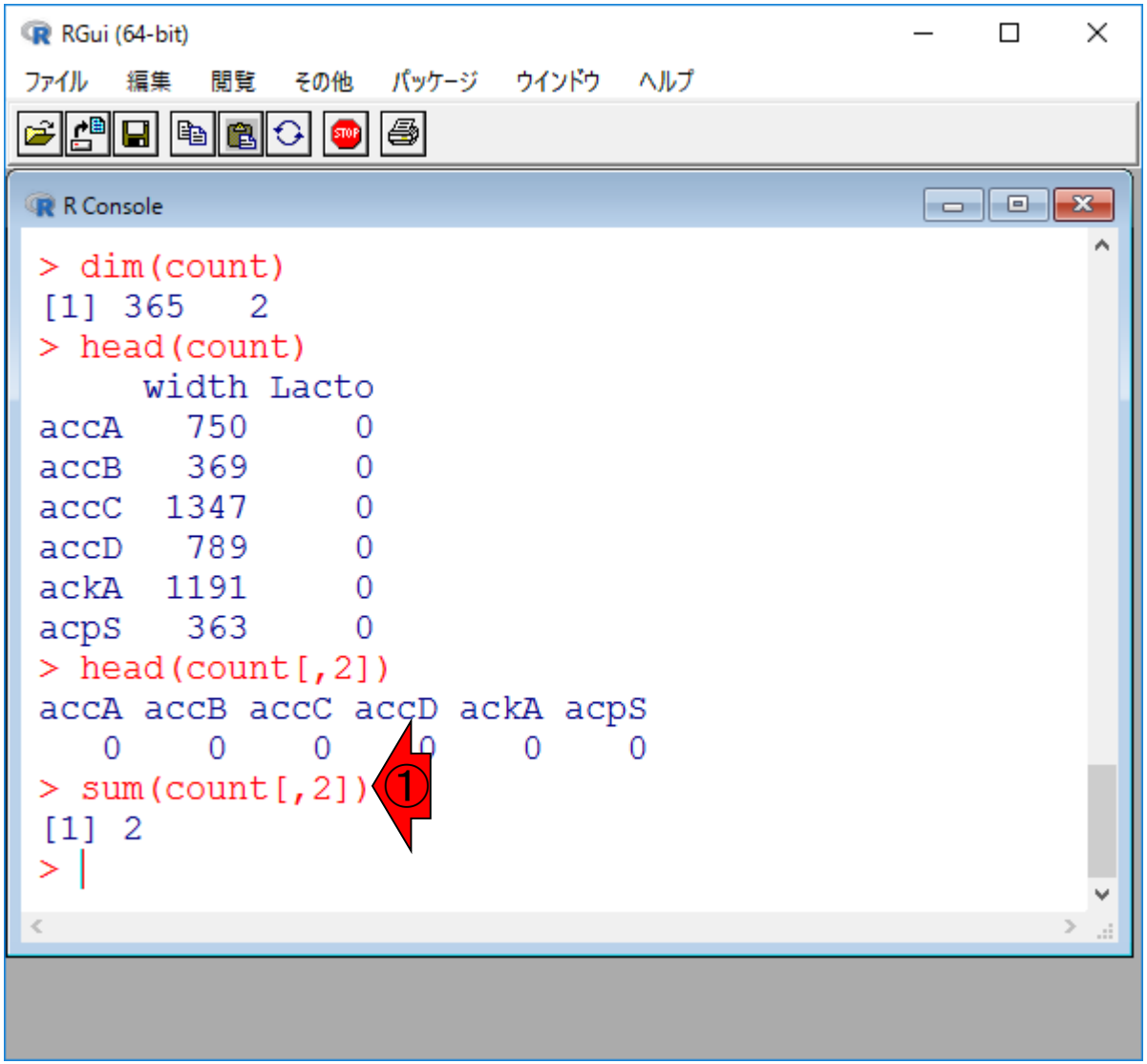
③

R Console

```
> count["dnaA",]
width Lacto
 1317      2
> 1676 - 360 + 1
[1] 1317
>
```

カウント総数の確認

全365遺伝子の領域にマップされたリードの総数は、①2でした



Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - **カウント情報取得時のオプション**
    - grepでgenenameの個数を確認

?qCount

カウントデータ取得部分では、qCount関数を利用しているので、他にどのようなオプションがあるのかを調べてみる。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

10.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

マップする側のファイルは、サンプルデータ47のFASTA形式ファイル(sample_RNAseq4.fa)です。マップされる側のファイルは、Ensembl (Zerbino et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FASTA形式ゲノム配列ファイル (Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式のアノテーションファイル(Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウント情報を取得しています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル
out_f <- "hoge10.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junct

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel)#カウントデータ行列を取得してcountに格納
dim(count)                             #行数と列数を表示
head(count)                            #確認してるだけです
```

①

?qCount

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ

R Console

```
> dim(count)
[1] 365   2
> head(count)
     width Lacto
accA   750     0
accB   369     0
accC  1347     0
accD   789     0
ackA  1191     0
acpS   363     0
> head(count[,2])
accA accB accC accD ackA acpS
   0    0    0    0    0    0
> sum(count[,2])
[1] 2
> ?qCount
```

①

qCount {QuasR} R Documentation

Quantify alignments

Description

Quantify alignments from sequencing data.

Usage

```
qCount(proj,
       query,
       reportLevel=c(NULL, "gene", "exon","promoter","junction"),
       selectReadPosition=c("start", "end"),
       shift=0L,
       orientation=c("any", "same", "opposite"),
       useRead=c("any", "first", "last"),
       auxiliaryName=NULL,
       mask=NULL,
       collapseBySample=TRUE,
       includeSpliced=TRUE,
       includeSecondary=TRUE,
       mapqMin=0L,
       mapqMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```

orientation

①orientation部分の説明。②デフォルトはanyのようだ。これは、昔のRNA-seqが二本鎖の状態でsequenceされていたためである。つまりリードの方向性（orientation）はわからなかったということ。リファレンス配列上の領域はわかるが、+鎖側(same strand)か−鎖側(opposite strand)かは不明だった、ということ。

qCount {QuasR}

Quantify alignments

Description

Quantify alignments from sequencing dat

Usage

```
qCount(proj,
       query,
       reportLevel=c(NULL, "gene",
       selectReadPosition=c("start
       shift=0L,
       orientation=c("any", "same", "opposite"),
       useRead=c("any", "first", "last"),
       auxiliaryName=NULL,
       mask=NULL,
       collapseBySample=TRUE,
       includeSpliced=TRUE,
       includeSecondary=TRUE,
       mapqMin=0L,
       mapqMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```

① orientation

sets the required orientation of the alignments relative to the query region in order to be counted, one of:

②
- `any` (default): count alignment on the same and opposite strand
- `same` : count only alignment on the same strand
- `opposite` : count only alignment on the opposite strand

orientation

①今回のマップする側のリードは、+鎖側(same strand)で設計したので、おそらくanyでもsameでも同じ結果になる。しかし、おそらくoppositeを指定したら(orientation = “opposite”)、カウントの総和は0になるだろう。未確認

qCount {QuasR}

Quantify alignments

Description

Quantify alignments from sequencing dat

Usage

```
qCount(proj,
       query,
       reportLevel=c(NULL, "gene",
       selectReadPosition=c("start
       shift=0L,
       orientation=c("any", "same", "opposite"),
       useRead=c("any", "first", "last"),
       auxiliaryName=NULL,
       mask=NULL,
       collapseBySample=TRUE,
       includeSpliced=TRUE,
       includeSecondary=TRUE,
       mapqMin=0L,
       mapqMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```

`orientation`

sets the required orientation of the alignments relative to the query region in order to be counted, one of:

① • `any` (default): count alignment on the same and opposite strand

② • `same` : count only alignment on the same strand

③ • `opposite` : count only alignment on the opposite strand

少しずれたリード

①1塩基くらいずれていても、領域内の大部分にマップされたリードということでカウント情報として加えるにはどうすればよいのか？という視点でオプション名を眺める。①shiftとかのオプションをshift = 1などとすればいいのかな…などと妄想しながら説明文を読む。

qCount {QuasR}

Quantify alignments

Description

Quantify alignments from sequencing data.

Usage

```
qCount(proj,
       query,
       reportLevel=c(NULL, "gene", "exon","promoter","junction"),
       selectReadPosition=c("start", "end"),
①      shift=0L,
       orientation=c("any", "same", "opposite"),
       useRead=c("any", "first", "last"),
       auxiliaryName=NULL,
       mask=NULL,
       collapseBySample=TRUE,
       includeSpliced=TRUE,
       includeSecondary=TRUE,
       mapqMin=0L,
       mapqMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```

少しずれたリード

①shiftの説明部分。なんで②3’-endに限定するのか不明だが、もしかしたら③のところのデフォルトが④startということと関連するのかな?!などと想像する

qCount {QuasR} R Documentation

Quantify alignments

Description

Quantify alignments from sequencing data

Usage

```
qCount(proj,
       query,
       reportLevel=c(NULL, "gene",
       selectReadPosition=c("start
       shift=0L,
       orientation=c("any", "same"
       useRead=c("any", "first", "
       auxiliaryName=NULL,
       mask=NULL,
       collapseBySample=TRUE,
       includeSpliced=TRUE,
       includeSecondary=TRUE,
       mapqMin=0L,
       mapqMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```

③ selectReadPosition

defines the part of the alignment that has to be contained within a query region to produce an overlap (see Details). Possible values are:

④
- `start` (default): start of the alignment
- `end`: end of the alignment

① shift

② controls the shifting alignments towards their 3'-end before quantification. `shift` can be one of:

- an “integer” vector of the same length as the number of alignment files
- a single “integer” value
- the character string `"halfInsert"` (only available for paired-end experiments)

The default of `0` will not shift any alignments.

少しずれたリード

ただ、こういう場合は、プログラムのバグの可能性など様々な要因が考えられるので、まずは③selectReadPositionには手をつけず、①shift = 1を付けてどのような結果になるかを見てみる。

qCount {QuasR}

Quantify alignments

Description

Quantify alignments from sequencing data

Usage

```
qCount(proj,
       query,
       reportLevel=c(NULL, "gene",
       selectReadPosition=c("start
       shift=0L,
       orientation=c("any", "same"
       useRead=c("any", "first", "
       auxiliaryName=NULL,
       mask=NULL,
       collapseBySample=TRUE,
       includeSpliced=TRUE,
       includeSecondary=TRUE,
       mapqMin=0L,
       mapqMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```

③ selectReadPosition

defines the part of the alignment that has to be contained within a query region to produce an overlap (see Details). Possible values are:

④
- `start` (default): start of the alignment
- `end`: end of the alignment

① shift

② controls the shifting alignments towards their 3'-end before quantification. `shift` can be one of:

- an "integer" vector of the same length as the number of alignment files
- a single "integer" value
- the character string `"halfInsert"` (only available for paired-end experiments)

The default of `0` will not shift any alignments.

少しずれたリード

1塩基のずれのみでカウントされるかもしれないのは、①のリード。②の遺伝子領域が対応します。

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=gene:LOOC260_10
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP845
Chromosome ena exon 360 1676 . + . Parent=transcript:
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Pa
###
Chromosome ena gene 1852 2991 . + . ID=gene:LOOC260_10
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP845
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Pa
###
Chromosome ena gene 3233 3457 . + . ID=gene:LOOC260_10
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP845
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Pa
###
Chromosome ena gene 3467 4588 . + . ID=gene:LOOC260_10
```

①

②

```
>Chromosome_361_400
TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
>Chromosome_1637_1676
AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
>Chromosome_1851_1890
CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
>Chromosome_1843_1882
TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
>Chromosome_1833_1872
CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
>Chromosome_1823_1862
CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
>Chromosome_1813_1852
AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
>Chromosome_3418_3457
GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
>Chromosome_3420_3459
TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
>Chromosome_3422_3461
GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
>Chromosome_3443_3482
ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
```

少しずれたリード

1塩基のずれのみでカウントされるかもしれないのは、①のリード。②の遺伝子領域が対応します。③その遺伝子名はdnaN。

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + . ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + . ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + . ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + . ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

例題11

①例題11をやってみましょう。qCount関数実行時のオプションとして、②shift = 1となるように、③で指定しています。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。
「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動し以下をコピペ。

1.サンプルデータ7[illegible]STQ形式ファイル(SRR037439.fastq)の BSgenome.Hsapiens.UCSC.hg19 へのマッピング結果の場合:

mapping_single_g
のサンプル名」(例:
す。hg19にマップした結
で、Entrez Gene IDに対

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "
param_txdb2 <-
param_reportlev

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

①

②

11.mapping_single_genome7.txt中のFASTA形式ファイルを乳酸菌ゲノムにマッピングする場合:

例題10と基本的に同じですが、カウント情報を取得するqCount関数の実行時に、shift = 1オプションを付けています。

```
in_f1 <- "mapping_single_genome7.txt"  #入力ファイル名を指定してin_f1に格納(RNA-seqファイル)
in_f2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファイ
in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル名
out_f <- "hoge11.txt"                  #出力ファイル名を指定してout_fに格納
param_reportlevel <- "gene"            #カウントデータ取得時のレベルを指定："gene", "exon", "promoter", "junctio
param_shift <- 1                       #カウントデータ取得時の領域からの許容するずれの度合い?!を指定

#必要なパッケージをロード
library(QuasR)                         #パッケージの読み込み
library(GenomicFeatures)               #パッケージの読み込み

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, format="auto")#txdbオブジェクトの作成
txdb                                   #確認してるだけです

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)            #マッピングを行うqAlign関数を実行した結果をoutに格納
alignmentStats(out)                    #マッピング結果(alignment statistics)の表示。seqlength：リファレンス配列の

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLevel=param_reportlevel,#カウントデータ行列を取得してcountに格納
                shift=param_shift)     #カウントデータ行列を取得してcountに格納
```

③

例題11

①例題11をやってみましょう。qCount関数実行時のオプションとして、②shift = 1となるように、③で指定しています。④こんな感じの状態でコピペ実行して構いません。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeaturesパッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動

1.サンプルデータ7

mapping_single_g
のサンプル名」(例:
す。hg19にマップした結
で、Entrez Gene IDに対

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "
param_txdb2 <-
param_reportlev

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

①

11.mapping_single_genome7.txt中のFASTA形

例題10と基本的に同じですが、カウント情報を

②

```
in_f1 <- "mapping_single_genome7.t
in_f2 <- "Lactobacillus_hokkaidon
in_f3 <- "Lactobacillus_hokkaidone
out_f <- "hoge11.txt"
param_reportlevel <- "gene"
param_shift <- 1

#必要なパッケージをロード
library(QuasR)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, forr
txdb

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)
alignmentStats(out)

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLe
                shift=param_shift)
```

③

④

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
> getwd()
[1] "C:/Users/kojik/Desktop/hoge/mapping_kiso3"
> list.files()
 [1] "hoge10.txt"                                          $
 [2] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
 [3] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
 [4] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
 [5] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
 [6] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_$
 [7] "mapping_single_genome7.txt"                          $
 [8] "QuasR_log_3b6c12e745f9.txt"                          $
 [9] "sample_RNAseq4.fa"                                   $
[10] "sample_RNAseq4_3b6c652a602a.bam"                     $
[11] "sample_RNAseq4_3b6c652a602a.bam.bai"                 $
[12] "sample_RNAseq4_3b6c652a602a.bam.txt"                 $
>
```

コピペ実行後

①例題11のコピペ実行後は、②のような感じになります。

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeaturesパッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動

1.サンプルデータ7

① 11.mapping_single_genome7.txt中のFASTA形

例題10と基本的に同じですが、カウント情報を

```
in_f1 <- "mapping_single_genome7.t
in_f2 <- "Lactobacillus_hokkaidon
in_f3 <- "Lactobacillus_hokkaidone
out_f <- "hoge11.txt"
param_reportlevel <- "gene"
param_shift <- 1

#必要なパッケージをロード
library(QuasR)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, forr
txdb

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)
alignmentStats(out)

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportLe
                shift=param_shift)
```

② RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
call dbDisconnect() when finished working with a $
> dim(count)                                    #行数と$
[1] 365   2
> head(count)                                   #確認し$
      width Lacto
accA    750     0
accB    369     0
accC   1347     0
accD    789     0
ackA   1191     0
acpS    363     0
>
> #ファイルに保存
> tmp <- cbind(rownames(count), count)    #保存し$
> write.table(tmp, out_f, sep="\t", append=F, quo$
>
```

カウント総数の確認

全365遺伝子の領域にマップされたリードの総数は、①3になりました!

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeaturesパッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動

1.サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の BSgenome.Hsapiens

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

11.mapping_single_genome7.txt中のFASTA形

例題10と基本的に同じですが、カウント情報を

```
in_f1 <- "mapping_single_genome7.t
in_f2 <- "Lactobacillus_hokkaidone
in_f3 <- "Lactobacillus_hokkaidone
out_f <- "hoge11.txt"
param_reportlevel <- "gene"
param_shift <- 1

#必要なパッケージをロード
library(QuasR)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, for
txdb

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)
alignmentStats(out)

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportL
                shift=param_shift)
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
[1] 365   2
> head(count)                                    #確認し$
     width Lacto
accA   750      0
accB   369      0
accC  1347      0
accD   789      0
ackA  1191      0
acpS   363      0
>
> #ファイルに保存
> tmp <- cbind(rownames(count), count)    #保存し$
> write.table(tmp, out_f, sep="\t", append=F, quo$
> sum(count[,2])
[1] 3
>
```

①

①dnaNのカウント数は1。

dnaNのカウント数

マップ後 | カウント情報取得 | single-end | ゲノム | アノテーション有 | QuasR(Gaidatzis_2015) NEW

QuasRパッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッピングから、カウントデータ取得までの一連の流れを示します。アノテーション情報は、GenomicFeatures パッケージ中の関数を利用してTxDbオブジェクトをネットワーク経由で取得するのを基本としつつ、TxDbパッケージを読み込むやり方も示しています。マッピングのやり方やオプションの詳細については マッピング | single-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis_2015)などを参考にしてください。

「ファイル」-「ディレクトリの変更」で解析したいファイルを置いてあるディレクトリに移動

1.サンプルデータ7のFASTQ形式ファイル(SRR037439.fastq)の BSgenome.Hsapiens

```
in_f1 <- "mapping
in_f2 <- "BSgenom
out_f <- "hoge1.t
param_mapping <-
param_txdb1 <- "h
param_txdb2 <- "k
param_reportlevel

#必要なパッケージを
library(QuasR)
library(GenomicFe

#前処理(マッピング)
out <- qAlign(in_
alignmentStats(ou

#前処理(TxDbオブジ
txdb <- makeTxDbF
txdb

#本番(カウントデー
count <- qCount(o
```

11.mapping_single_genome7.txt中のFASTA形

例題10と基本的に同じですが、カウント情報を

```
in_f1 <- "mapping_single_genome7.t
in_f2 <- "Lactobacillus_hokkaidone
in_f3 <- "Lactobacillus_hokkaidone
out_f <- "hoge11.txt"
param_reportlevel <- "gene"
param_shift <- 1

#必要なパッケージをロード
library(QuasR)
library(GenomicFeatures)

#前処理(アノテーション情報を取得)
txdb <- makeTxDbFromGFF(in_f3, for
txdb

#前処理(マッピング)
out <- qAlign(in_f1, in_f2)
alignmentStats(out)

#本番(カウントデータ取得)
count <- qCount(out, txdb, reportL
                shift=param_shift)
```

RGui (64-bit)

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

R Console

```
accA    750      0
accB    369      0
accC   1347      0
accD    789      0
ackA   1191      0
acpS    363      0
>
> #ファイルに保存
> tmp <- cbind(rownames(count), count)    #保存し$
> write.table(tmp, out_f, sep="\t", append=F, quo$
> sum(count[,2])
[1] 3
> count["dnaN",]
width Lacto
 1140     1
>
```

①

Contents

- マッピング（アラインメント）の続き
  - おさらい：入力ファイル（マップする側、される側）、QuasRの結果、Bowtie2の結果
  - マップされなかったリード：Bowtie（デフォルト）、Bowtie（QuasRと同じオプション）
  - SAM形式の解説、マッピング結果の違い、課題
  - Linux環境以外でのBowtie2実行手段
- カウント情報取得
  - アノテーション情報がない場合：単一サンプル、複数サンプル
  - アノテーション情報がある場合
    - 概要
    - マップする側のファイルの説明
    - マッピング実行
    - 結果の解釈
    - カウント情報取得時のオプション
    - grepでgenenameの個数を確認

残る問題

（他にも沢山あるが）なぜ得られたカウントデータの行数が2000行超にならずに365行となってしまったのか、について考える

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + ① ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + ① ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + ② ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + ① ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

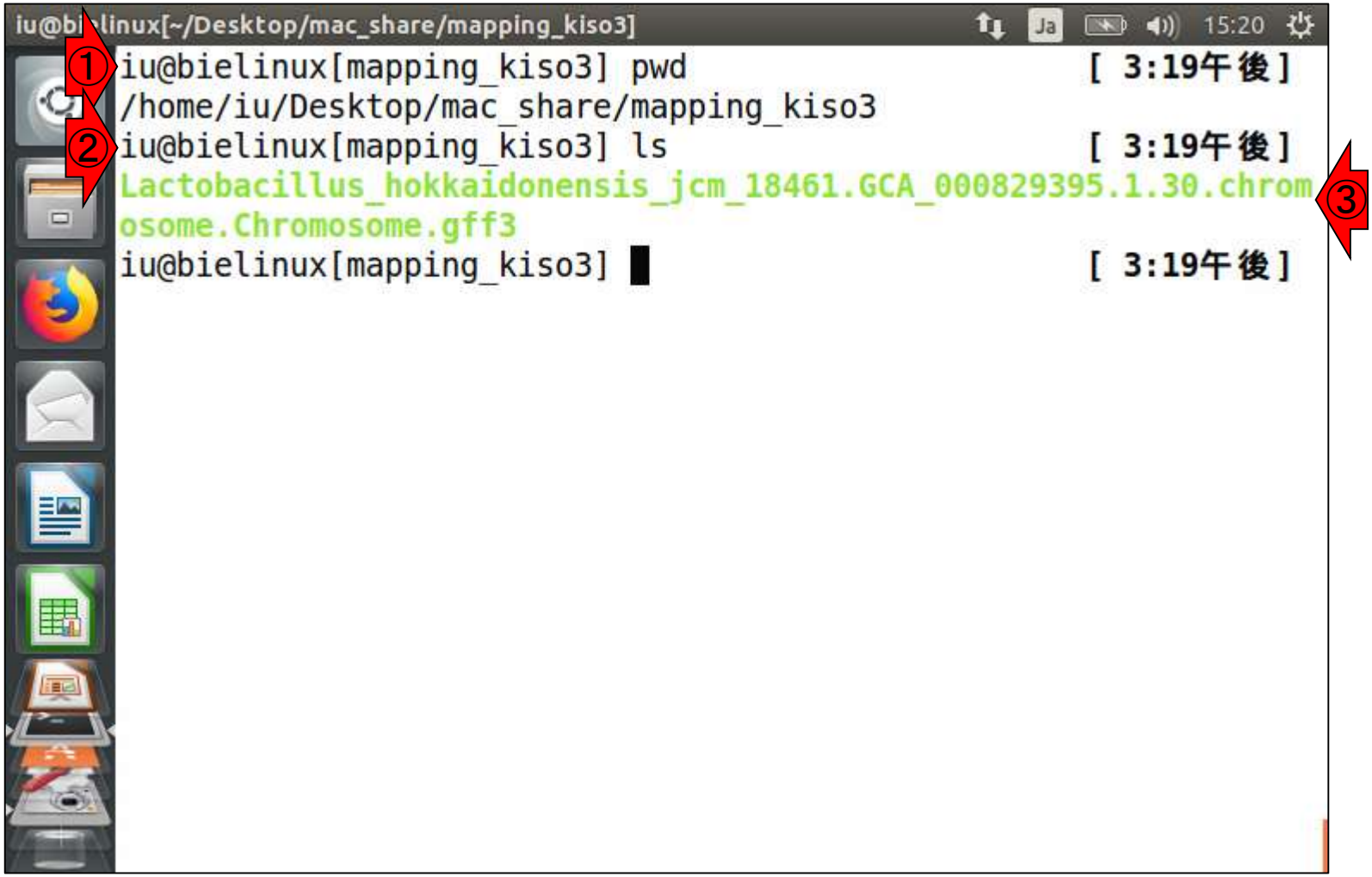
残る問題

Linux環境で、genenameの個数が本当に365個だったのかを検証する。

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + ① ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + ① ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + . ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + ① ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

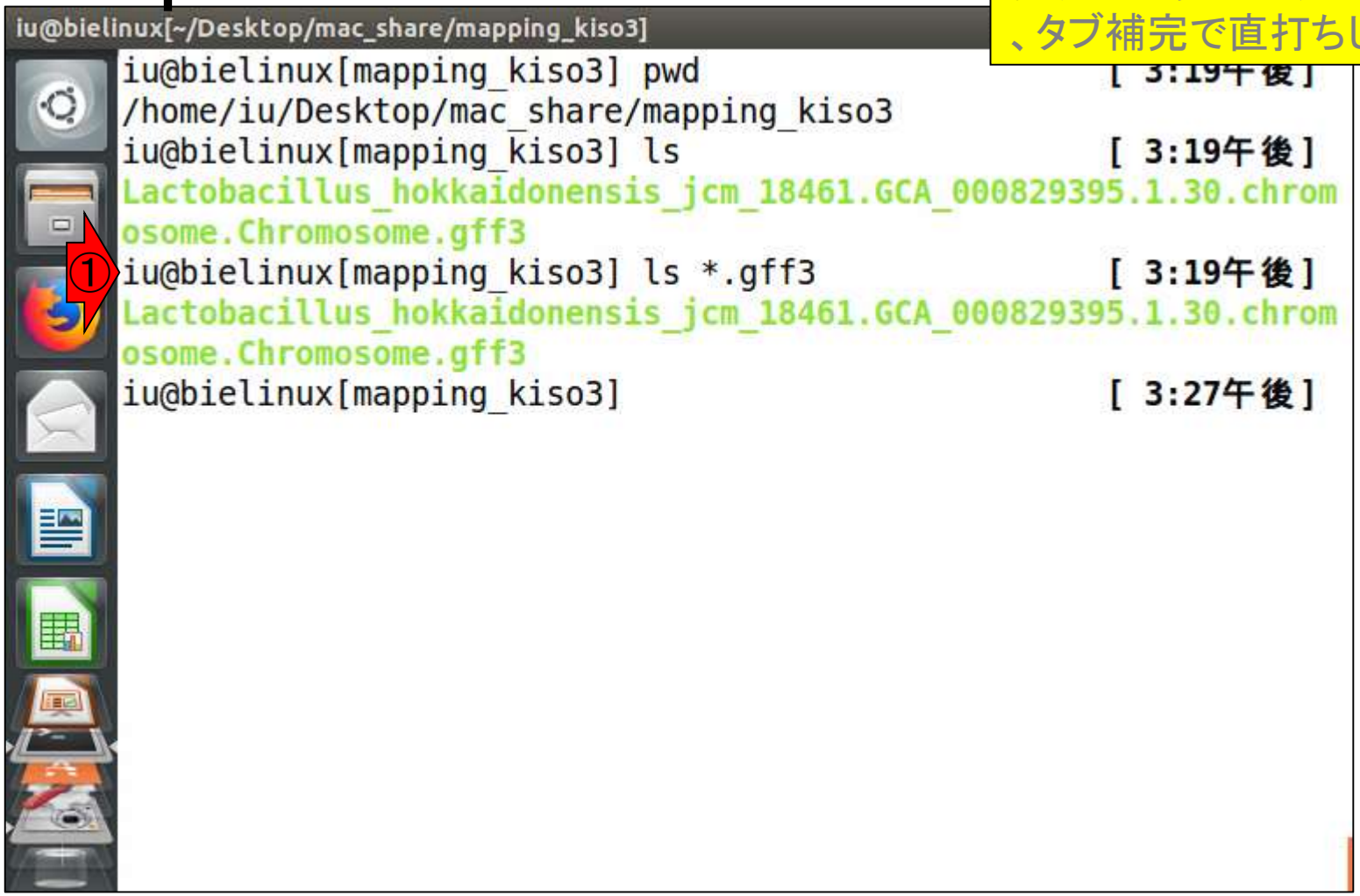
Bio-Linuxのターミナル画面

お約束の①pwdと②ls。このような状況で③gff3ファイルのみを取り扱う



Tips：ワイルドカード

①*.gff3と書くことで、.gff3で終わる全てのファイルのみをリストアップすることができます。今回の場合は、ファイルが1つしかないので、タブ補完で直打ちしてもよいといえばよい。



残る問題

（他にも沢山あるが）なぜ得られたカウントデータの行数が2000行超にならずに365行となってしまったのか、について考える

```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
Chromosome ena gene 360 1676 . + ① ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome ena transcript 360 1676 . + . ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome ena exon 360 1676 . + . Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome ena CDS 360 1676 . + 0 ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###
Chromosome ena gene 1852 2991 . + ① ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome ena transcript 1852 2991 . + . ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome ena exon 1852 2991 . + . Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome ena CDS 1852 2991 . + 0 ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###
Chromosome ena gene 3233 3457 . + ② ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome ena transcript 3233 3457 . + . ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome ena exon 3233 3457 . + . Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome ena CDS 3233 3457 . + 0 ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###
Chromosome ena gene 3467 4588 . + ① ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de
```

ID=geneを含む行数

Geneの領域数をカウントする。全体をざっと眺めて、①ID=geneを含む行数をカウントすればよいだろう

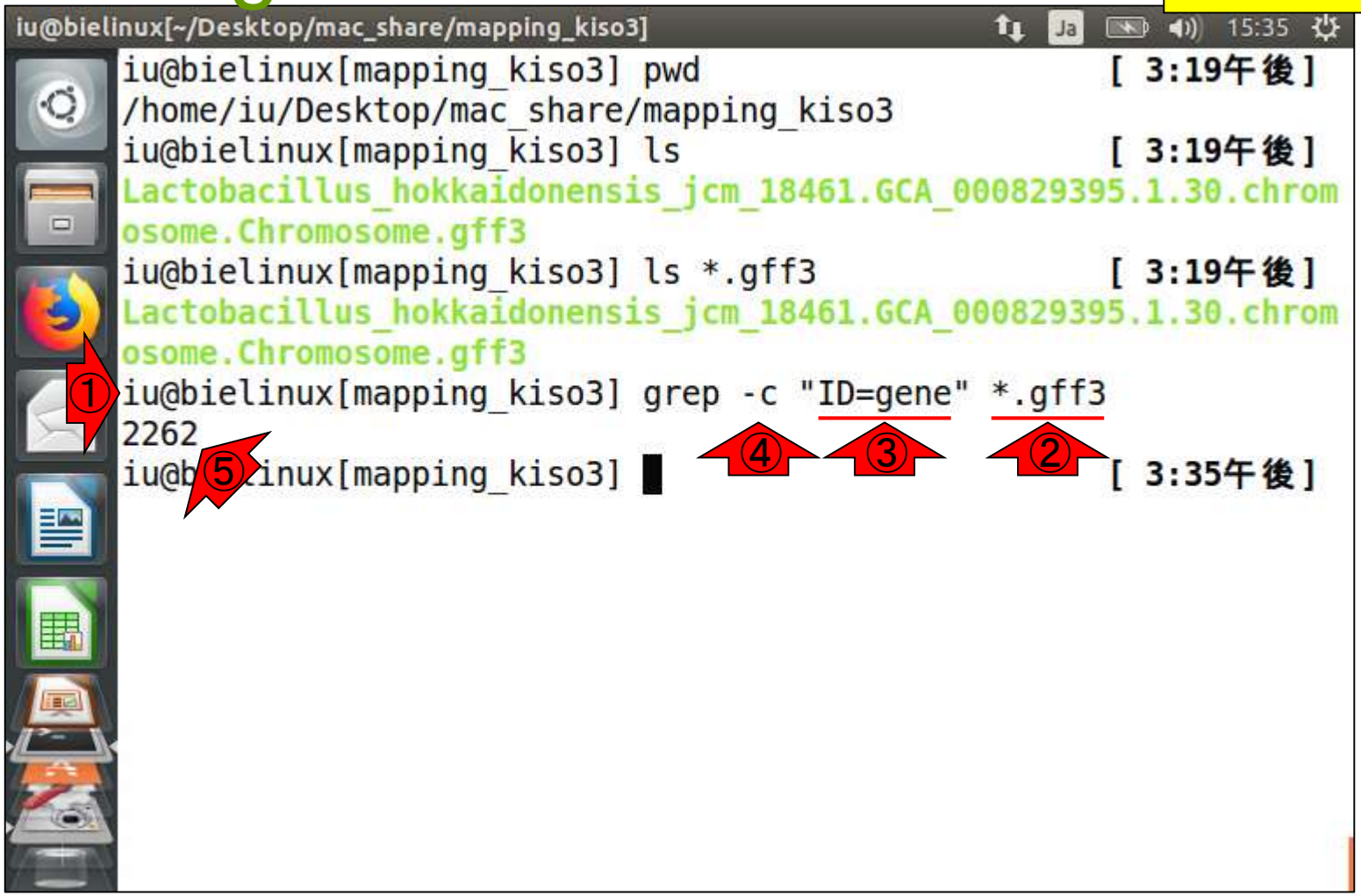
```
##gff-version 3
##sequence-region Chromosome 360 2277853
#!genome-build European Nucleotide Archive ASM82939v1
#!genome-version GCA_000829395.1
#!genome-date 2014-11
#!genome-build-accession GCA_000829395.1
#!genebuild-last-updated 2014-11
```

①

Chromosome	ena	gene	360	1676	.	+	.	ID=gene:LOOC260_100010;Name=dnaA;biotype=protein_coding;d
Chromosome	ena	transcript	360	1676	.	+	.	ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=d
Chromosome	ena	exon	360	1676	.	+	.	Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;e
Chromosome	ena	CDS	360	1676	.	+	0	ID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP8
###								
Chromosome	ena	gene	1852	2991	.	+	.	ID=gene:LOOC260_100020;Name=dnaN;biotype=protein_coding;d
Chromosome	ena	transcript	1852	2991	.	+	.	ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=d
Chromosome	ena	exon	1852	2991	.	+	.	Parent=transcript:BAP84582;Name=BAP84582-1;constitutive=1;e
Chromosome	ena	CDS	1852	2991	.	+	0	ID=CDS:BAP84582;Parent=transcript:BAP84582;protein_id=BAP8
###								
Chromosome	ena	gene	3233	3457	.	+	.	ID=gene:LOOC260_100030;biotype=protein_coding;description=S
Chromosome	ena	transcript	3233	3457	.	+	.	ID=transcript:BAP84583;Parent=gene:LOOC260_100030;biotype=
Chromosome	ena	exon	3233	3457	.	+	.	Parent=transcript:BAP84583;Name=BAP84583-1;constitutive=1;e
Chromosome	ena	CDS	3233	3457	.	+	0	ID=CDS:BAP84583;Parent=transcript:BAP84583;protein_id=BAP8
###								
Chromosome	ena	gene	3467	4588	.	+	.	ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;de

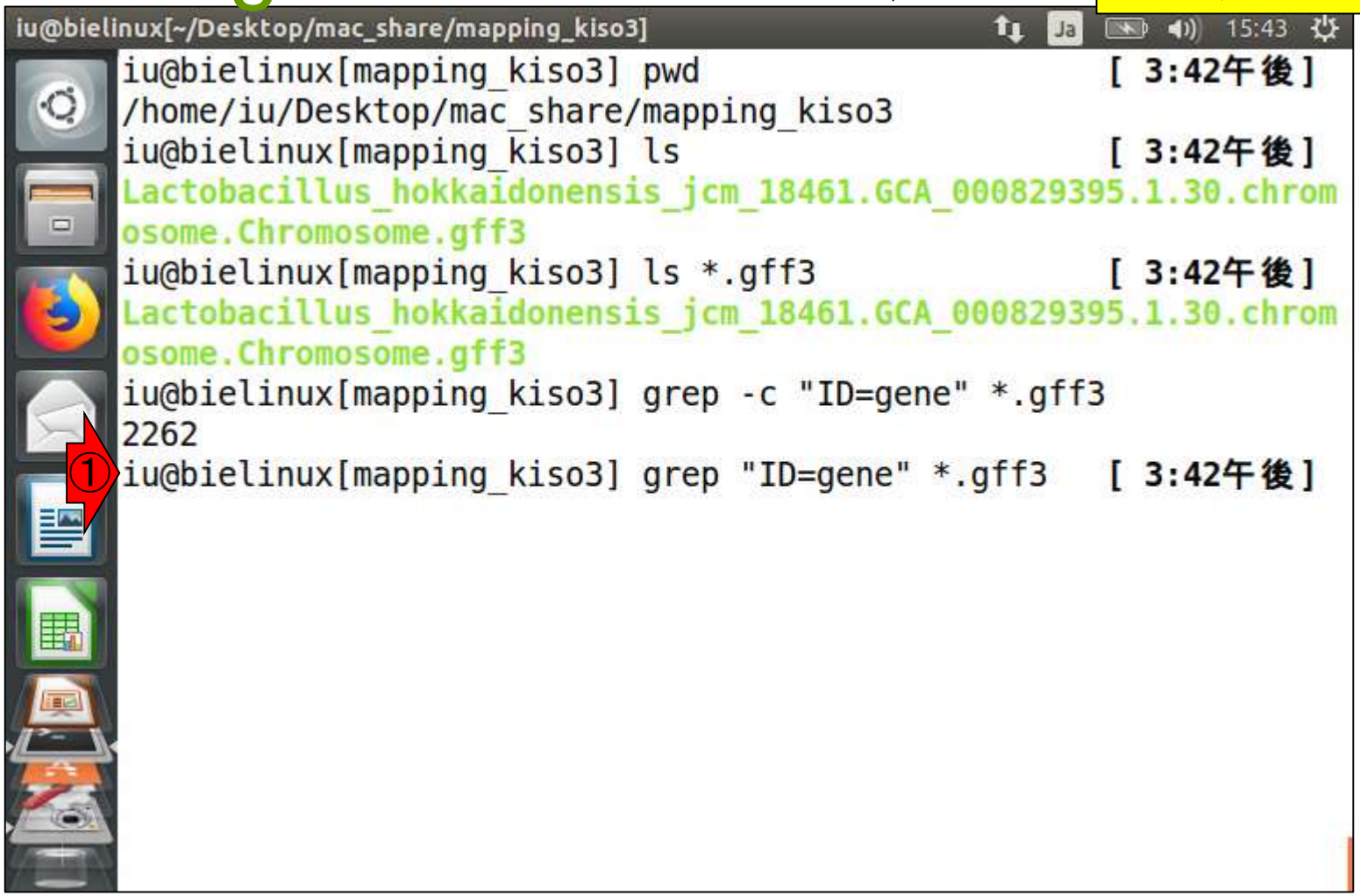
ID=geneを含む行数

①grepコマンドで、②*.gff3というファイルに対して、③ID=geneという文字列を含む、④行数は、⑤2262行



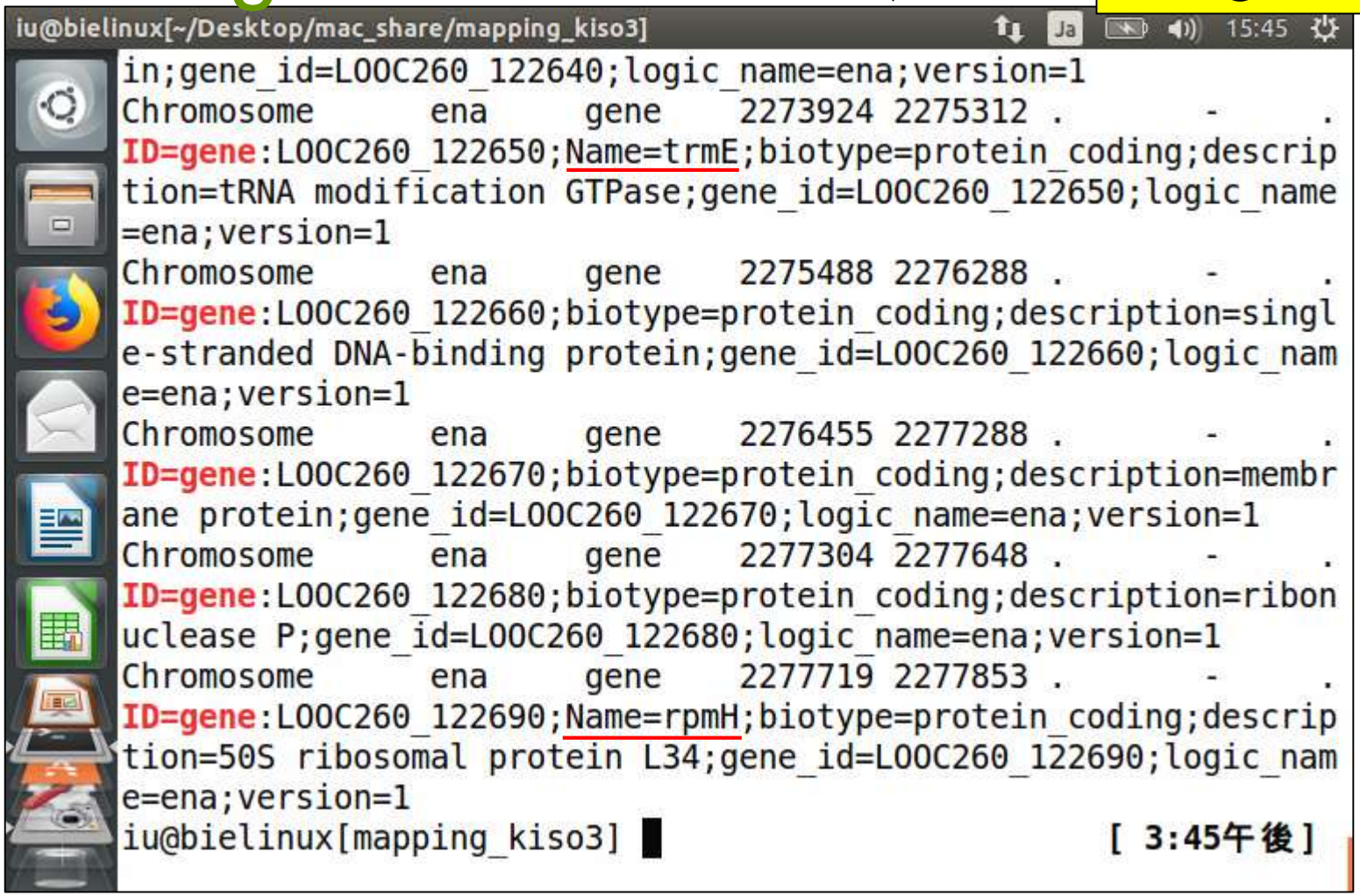
ID=geneを含む行を表示

①を実行すると、ID=geneという文字列を含む行がそのまま表示される。2262行分あるので、画面がざっと流れる。



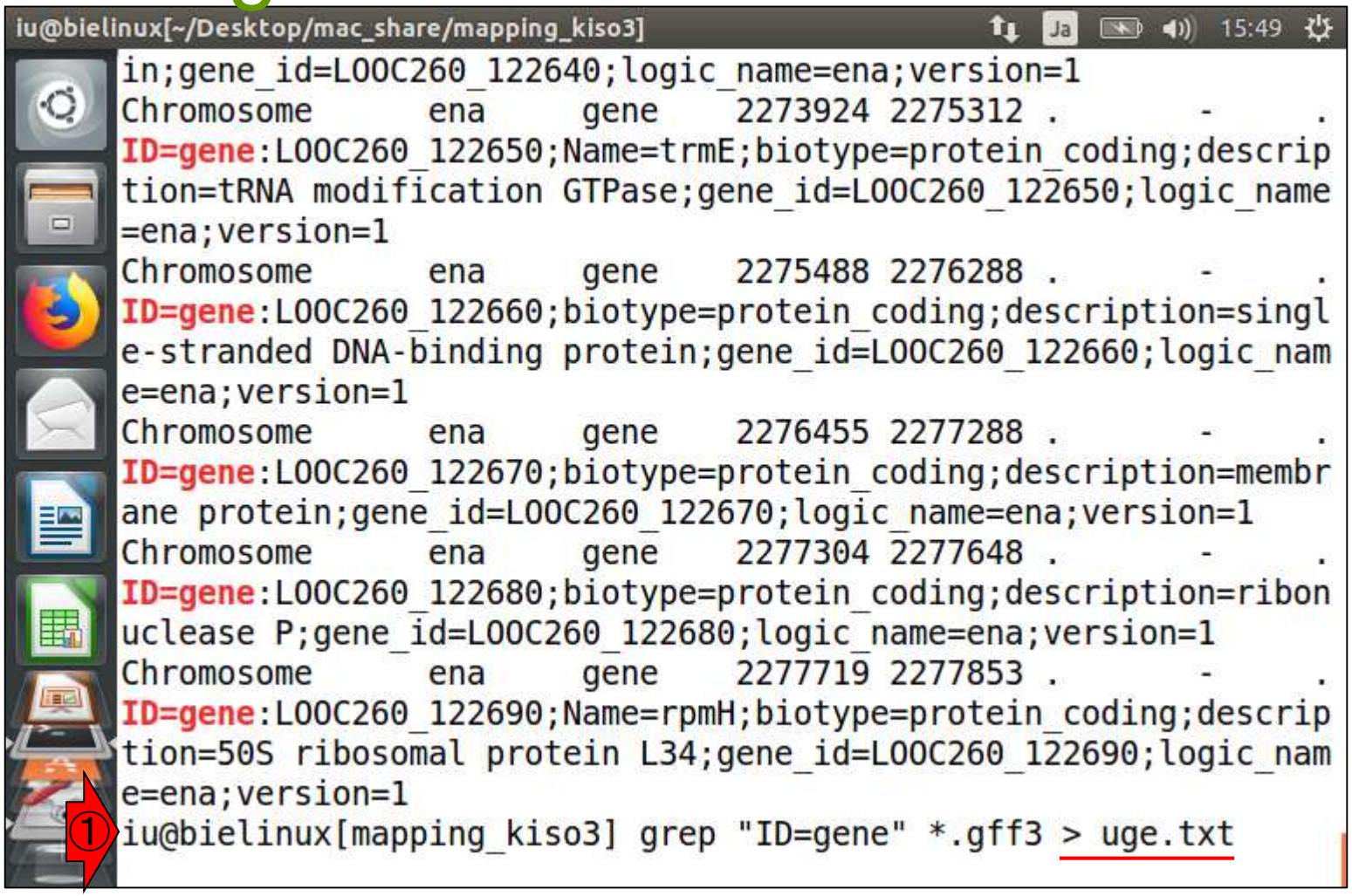
ID=geneを含む行を表示

画面がざっと流れた結果。この結果をざっと眺めても、genenameがあるのは5個中①2個と少ないですね。



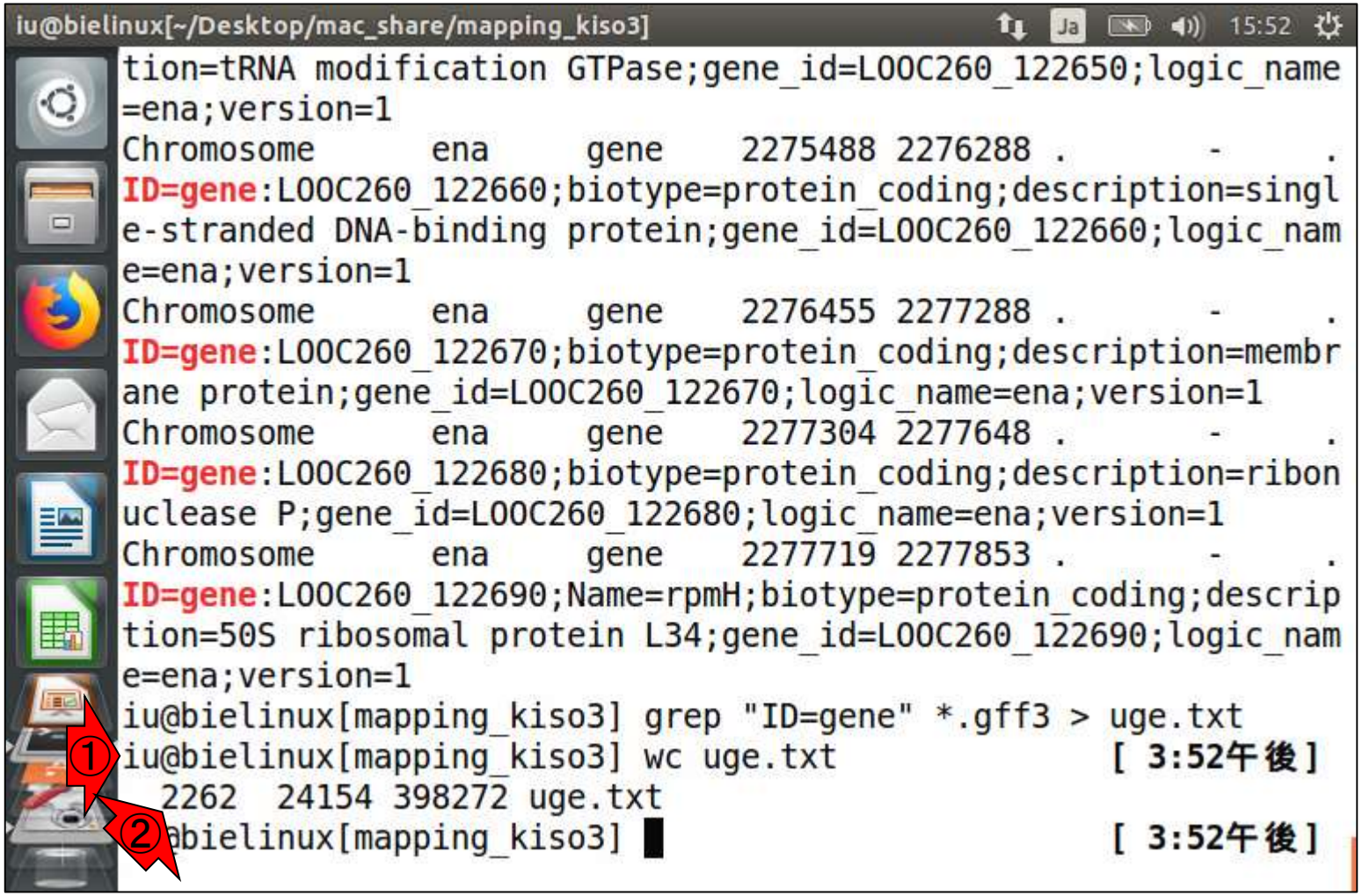
ID=geneを含む行をファイル出力

①のようにするとターミナル画面上に表示するのではなく、uge.txtに保存できます



行数を確認

①wcでuge.txtの行数を確認。確かに2262行ですね。



Name=を含む行数を表示

①grepコマンドで、②uge.txtというファイルに対して、③Name=という文字列を含む、④行数は、⑤457行。大分365個に近づいてきました。本当はここで365行になって一件落着のつもりでしたが…。Name=はあるがgenename部分が空っぽとか、同じ遺伝子名のものがあるとか…