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/

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

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

iii ref_genome.fa	
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)	
<pre>>chr1 CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG >chr2 AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA</pre>	<pre>ファイル(F) 編集(E) 書気(O) 表示(V) ヘルブ(H) >chr1_11_45 CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT >chr2_16_50 TATCTATGGCCTAAAAACATAGACACCTTGAGGAG >chr2_1_35 AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT</pre>
TGACGCCCTG >chr3 GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT >chr4 CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC ACCAGACTTTACCGTACAAGATA >chr5 GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	<pre>>chr3_11_45 TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA >chr3_15_49 CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG >chr3_3_37 GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG >chr3_1_35 GGGGGGACTATTTCCCCCGCTTGCAGGAATCGTGTC >chr5_1_35 GCGCGGTCTATTTCCCCCGCTTGCAGGAATCGTGTC</pre>
۲	

①ref_genome.faのおさらい。②chr3と③chr5の違いは…

マップされる側

🧊 ref_genome.fa - 子子帳
ファイル(F) 編集 書式(O) 表示(V) ヘルプ(H)
>chr1
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG
>chr2
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA
TGACC 2 CTG
>chr3
G <u>G</u> GGGG <u>A</u> CTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
>chr4
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT
CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTG
ACCAG
>chr5
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT
* • • • •

①ref_genome.faのおさらい。②chr3と③ chr5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



マップされる側

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

	🦷 sample_RNAseq1.fa (1) 帳 💷 💷 🔤
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)	ファイル(F) 編集(E) 書い(O) 表示(V) ヘルプ(H)
>chr1	Schol 11 45
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG	
>chr2	Schol 16 50
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG	
ΔCGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG	Schr2 1 35
TEACECCCTE	schr3 11 45
TGACGCCCTG	
	schr3 15 49
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	schr3 3 37
>chr4	
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT	schr3 1 35
CGTCTGTATTGGGGGCCTATGCTCGACAAGAGATTGTGTGTG	
	schr5 1 35
>chr5	
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	
	▼
4	- 11

マップする側

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 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)

①使用したオプション。計8リードのうち、マップ

🦳 ref_genome.fa - メモ帳			
ファイル(F) 編	+	1	sample_RNAseq1.fa - 义モ帳
chrl	11 45		ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1	1 25		>chr1 11 45
CGAGGAGG/		FGGATGCTGAGTGGG	CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
≻chr2 chrZ	16 50		>chr2 16 50
AGGGAGGG	1 35	ATAGACACCTTGAGGAG	TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGT/	0 07	GCATCATGAAGGGGCTG	>chr2 1 35
CTCGGGTA1 Chr3	3 37	CACGACGCTGCGGTTCA	AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
TGACGCCCTG			>chr3_11_45
>chr3			TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
GGGGGGGACTATTTCCCCGC	TTGCAGGAATCGTC	TCAGTTGGTATACAGGC	>chr3 15 49
	ΔGGGTGTΔGTCΔGC	<u>CTΔΤΔGTΤΔΔCTΔGTTT</u>	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
Nchr4			>chr3_3_37
	COTCACTCAATCC	TACCANAACAATCTTCT	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
CGAGACGAGCAAGTTATTC	GUTUAGTGAATGGG	TAGCAAAAGAATGTTGT	>chr3_1_35
CGICIGIAIIGGGGGCCIAI	GCTCGACAAGAGAT	IIGIGIGIAGIAIGAGCC	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
ACCAGACTTTACCGTACAA	GATA		>chr5 1 35
>chr5			GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
GCGGGGTCTATTTCCCCGC	TTGCAGGAATCGTC	GTCAGTTGGTATACAGGC	
AGCATCTAGTCGCATCAGA	AGGGTGTAGTCAGC	CTATAGTTAACTAGTTT	
•		•	
CGAGACGAGCAAGTTATTC CGTCTGTATTGGGGCCTAT ACCAGACTTTACCGTACAA >chr5 GCGGGGTCTATTTCCCCGC AGCATCTAGTCGCATCAGA	GCTCAGTGAATGGO GCTCGACAAGAGAT GATA TTGCAGGAATCGTO AGGGTGTAGTCAGO	GTAGCAAAAGAATGTTGT TGTGTGTAGTATGAGCC GTCAGTTGGTATACAGGC CTATAGTTAACTAGTTT	<pre>>chr3_1_35 GGGGGGGGGGGGACTATTTCCCCCGCTTGCAGGAATCGTGTCAG >chr5_1_35 GCGCGGTCTATTTCCCCCGCTTGCAGGAATCGTGTC</pre>

QuasRのマッピング結果

"-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 - 又干帳	
	🔄 sample_RNAseq1.fa - 乄モ帳
ファイル(F) 福美(E) 音丸(O) 衣小(V) / Vレノ(F)	ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
>chr1	>chr1 11 45
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG	
>chr2	>chr2 16 50
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG	
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG	>chr2 1 35
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA	AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
TGACGCCCTG	>chr3_11_45
>chr3	TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	>chr3 15 49
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTT	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr4	>chr3_3_37
	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
COAGACGAGCAAGTTATTCGCTCAGTGAATGGTAGCAAAAGAATGTTGT	>chr3_1_35
CGTCTGTATTGGGGGCCTATGCTCGACAAGAGATTGTGTGTG	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
ACCAGACTTTACCGTACAAGATA	>chr5 1 35
>chr5	GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	-

May 29, 2018

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

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)	sample_RNAseq1.fa - メモ限
>chr1 CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG >chr2 AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA TGACGCCCTG	>chr1_11_45 CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT >chr2_16_50 TATCTATGGCCTAAAAACATAGACACCTTGAGGAG >chr2_1_35 AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT >chr3_11_45
>chr3 GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT >chr4 CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT CGTCTGTATTGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC ACCAGACTTTACCGTACAAGATA >chr5	<pre>>chr3_15_49 CCCGCTTGCAGGAATCGTGTCAGTTGGTATA >chr3_15_49 CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG >chr3_3_37 GGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG >chr3_1_35 GGGGGGACTATTTCCCCCGCTTGCAGGAATCGTGTC >chr5_1_35 GCGCGGTCTATTTCCCCCGCTTGCAGGAATCGTGTC</pre>
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	-

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



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%

iu@bielinux[~/Desktop/mac_share/mapping_kiso1] /home/iu/Desktop/mac share/mapping kisol iu@bielinux[mapping kiso1] ls -l [3:08午後] total 8197 -rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2 140 5月 14 16:39 pigya.2.bt2 -rwxrwxrwx 1 iu iu -rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2 -rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2 -rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2 -rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2 2013 ref genome.fa -rwxrwxrwx 1 iu iu 590 9月 29 396 10月 1 2013 sample RNAseq1.fa -rwxrwxrwx 1 iu iu 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 (2) 100.00% overall alignment rate iu@bielinux[mapping kiso1] [3:08午後]

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

Bowtie2実行結果

															sample_RNAseq1.fa - 义モ帳
															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
@HD		VN:1.0	SO:uns	orte	d										
0sQ		SN:chr1	LN:48												>cnr2_1_35
0sQ		SN:chr2	LN:160												AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
@sg		SN:chr3	LN:100												>chr3 11 45
@sg		SN:chr4	LN:123								_				TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
0 SQ		SN:chr5	LN:100											-	s = h = 2 1 E 40
@PG		ID:bowt	PN:bow	VN:2	CL:'	'/usi	:/bi	.n/	/	lib	/b	owtie2	/bin/lowt	ie2-ali	>CHI.2_12_49
chr1_1	1_45	0	chr1	11	42	35м	*	0	0	C¢ I	: 1 A	s:i:0	XN:i:0	XM:i:0	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
chr2_1	6_50	0	chr2	16	42	35м	*	0	0	TZ I	: 1 A	s:i:0	XN:i:0	XM:i:0	>chr3 3 37
chr2_1	_35	0	chr2	1	42	35M	*	0	0	AC I	I A	AS:i:0	XN:i:0	XM:i:0	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
chr3_1	1_45	0	chr5	11	1	35м	*	0	0	TI I	: 1 A	As:i:0	xs:i:0	XN:i:0	scho2 1 2E
chr3_1	5_49	0	chr3	15	1	35м	*	0	0	cdı	I A	As:i:0	xs:i:0	XN:i:0	scur.2_1_22
chr3_3	_37	0	chr3	3	31	35м	*	0	0	GCI	:] A	s:i:0	XS:i:-6	XN:i:0	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
chr3_1	_35	0	chr3	1	35	35м	*	0	0	GCI	: 1 A	s:i:0	xs:i:-12	XN:i:0	>chr5_1_35
chr5_1	_35	0	chr5	1	16	35M	*	0	0	GCI	I A	AS:i:-0	5 xs:i:-18	XN:i:0	GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC

Bowtie2実行結果

															🦳 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
₫HD -		VN:1.0	SO:unso	orte	d					_	_				Achr2 1 35
0 SQ		SN:chr1	LN:48								_				
0sQ		SN:chr2	LN:160												AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
0 SQ		SN:chr3	LN:100												>chr3 11 45
0sQ		SN:chr4	LN:123								_				TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
0sQ		SN:chr5	LN:100										-2-		Schol 15 10
0PG		ID:bowt:	PN:bow	VN:2	CL:	'/usr	/bi	.n/.	/	lib	/b	owtie2/	bin, owt:	ie2-alio	VUII'5_15_49
chr1_11	_45	0	chr1	11	42	35м	*	0	0	CGI	:) A	s:i:0	XN:i:0	XM:i:0	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
chr2_16	_50	0	chr2	16	42	35м	*	0	0	TZ I	ΊA	s:i:0	XN:i:0	XM:i:0	>chr3 3 37
chr2_1	35	0	chr2	1	42	35м	*	0	0	AC I	ΊA	s:i:0	XN:i:0	XM:i:0	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
chr3_11	_45	0	chr5	11	1	35M	*	0	0	TI I	Π	s:i:0	XS:i:0	XN:i:0	scho2 1 2E
chr3_15	49	0	chr3	15	1	35м	*	0	0	cdı	ΞA	s:i:0	xs:i:0	XN:i:0	SCUL2_1_22
chr3_3	37	0	chr3	3	31	35м	*	0	0	GGI	ΞA	s:i:0	XS:i:-6	XN:i:0	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
chr3_1	35	0	chr3	1	35	35м	*	0	0	GCI	I A	s:i:0	xs:i:-12	XN:i:0	>chr5 1 35
chr5_1	35	0	chr5	1	16	35м	*	0	0	GCI	ΞA	s:i:-6	XS:i:-18	XN:i:0	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

背景:①Bowtie2では、予期せず 全リードがマップされてしまった

bowtie2実行結果

u@bieli	nux[~/Desktop/ma	ac_shar	e/m	apping_kiso1	1]			📬 Ja 🕻	* •))	15:08	₽
0	/home/iu/Des iu@bielinux[sktop [mapp	/ma ing	ac_share, [_kiso1]	/mapp: ls -	ing_ l	kiso1	1	[3:08	午後]	
	- rwx rwx rwx 1 - rwx rwx rwx 1	l iu l iu l iu l iu l iu l iu l iu l iu	iu iu iu iu iu iu iu	4194746 140 53 133 4194746 140 590 396	5月 5月 5月 5月 5月 9月 10月	14 14 14 14 14 14 29 1	16:39 16:39 16:39 16:39 16:39 16:39 2013 2013	pigya.1.bt pigya.2.bt pigya.3.bt pigya.4.bt pigya.rev.1 pigya.rev.1 ref_genome sample_RNAs	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	a	
	iu@bielinux[a -S sample_	RNAs	ing eq]	[_kisol] sam	bowt:	ie2	-x pig	gya -f samp	Le_RNA	seq1.	f
	8 reads; of 8 (100.00% 0 (0.00% 3 (37.50 5 (62.50 100 00% over	thes b) we b) al c) a c) a	e: re igr lig lig	unpaired ed 0 tim ned exac ned >1 t	d; of nes ctly times	the 1 ti	ese: Lme				
	iu@bielinux[[mapp	ing	j_kisol]				j	[3:08	8午後]	5

I

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

9HD			VN:1.0	SO:unso	orte	d														
0sQ			SN:chr1	LN:48																
0sQ			SN:chr2	LN:160																
0sQ			SN:chr3	LN:100																
0 SQ			SN:chr4	LN:123																
0 SQ			SN:chr5	LN:100																
0PG			ID:bowt	PN:bow	VN:2	CL:	'/usi	c/b	in/	/	'lib	/bowtie2	/bin/bowt	ie2-ali	gn-s	wrappe	r basic	-0 -x pi	gya -f sam	nple_RNAs
chr1	_11	_45	0	chr1	11	42	35м	*	0	0	CGI	lAS: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	35м	*	0	0	TZ I	lAS: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	35м	*	0	0	AC I	lAS: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	35м	*	0	0	TI I	IAS: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	35м	*	0	0	CCI	IAS: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	35м	*	0	0	GGI	IAS: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	35м	*	0	0	GGI	IAS: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	35м	*	0	0	GCI	1AS: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
chr5	_1_	35	0	chr5	1	16	35M	*	0	0	GCI]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



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

①basenameをagriにして、 ②bowtie-buildを実行

リファレンス配列の前処理

iu@biel	inux[~/Desktop/r	nac_sha	re/m	apping_kiso [.]	1]			🏚 Ja 📧 🕪 21:54 🔱
0	iu@bielinux /home/iu/De	k[map	ping p/ma	<pre>_kiso1] ac share</pre>	pwd /mapp:	ing	kiso1	[9:51午後]
	iu@bielinux total 8199	k[map	ping	g_ <mark>kiso1</mark>]	ls -	ι _		[9:51午後]
	- rwx rwx rwx	1 iu	iu	4194746	5月	14	16:39	pigya.1.bt2
	- rwx rwx rwx	1 iu	iu	140	5月	14	16:39	pigya.2.bt2
	- rwx rwx rwx	1 iu	iu	53	5月	14	16:39	pigya.3.bt2
$\underline{}$	- rwx rwx rwx	1 iu	iu	133	5月	14	16:39	pigya.4.bt2
	-rwxrwxrwx	1 iu	iu	4194746	5月	14	16:39	pigya.rev.1.bt2
\times	-rwxrwxrwx	1 iu	iu	140	5月	14	16:39	pigya.rev.2.bt2
	- rwx rwx rwx	1 iu	iu	590	9月	29	2013	ref_genome.fa
	- rwx rwx rwx	1 iu	iu	396	10月	1	2013	sample_RNAseq1.fa
	- rwx rwx rwx	1 iu	iu	1616	5月	15	15:08	sample_RNAseq1.sam
	iu@bielinu>	k[map	ping]_kiso1]	bowt:	ie-k	ouild I	ref_genome.fa agri

bowtie-build完了

Bowtie-build実行が無事完
了したようだ。①clearして…

iu@bielinux[~/Desktop/mac_share/mapping_kiso1] 1	Ja 📧	▶ ◀))	21:59	垛
eftabSz: 80				
ftabLen: 1048577				
ftabSz: 4194308				
offsLen: 17				
offsSz: 68				
isaLen: 0				
isaSz: 0				
lineSz: 64				
sideSz: 64				
sideBwtSz: 56				
sideBwtLen: 224				
numSidePairs: 2				
numSides: 4				
numLines: 4				
ebwtTotLen: 256				
ebwtTotSz: 256				
reverse: 0	0000000			1000
Total time for backward call to driver() for mir	ror i	ndex	: 00	:0
0:00	-			
<pre>iu@bielinux[mapping_kiso1] clear</pre>	[9:54	午後]

①Isで確認。作成されたインデックス ②bowtie1ではebwtとなっており、③ とが分かります。このことから、bow	スファイルの拡張子部分が bowtie2のbt2とは異なるこ tie2で作成したインデックス
lu@bletinux[~/Desktop/mac_share/mapping_kiso1] ファイルはbowtielでは利用できない	のたろうと思ったりします
(home (iu/Desktop (mac_share (mapping_kisol)) (home (iu/Desktop (mac_share (mapping_kisol)))	
1 jughielinux[mapping_kisol]]s]	
total 16394	
= -rwxrwxrwx 1 iu iu 4194810 5 17 21.54 anri 1 ebwt	
-rwxrwxrwx 1 iu iu 72 5月 17 21:54 agri 2.ebwt	
-rwxrwxrwx 1 iu iu 53 5月 17 21:54 agri.3.ebwt	
-rwxrwxrwx 1 iu iu 133 5月 17 21:54 agri.4.ebwt	
- rwxrwxrwx 1 iu iu 4194810 5月 17 21:54 agri.rev.l.ebwt	
-rwxrwxrwx 1 iu iu 72 5月 17 21:54 agri.rev.2.ebwt	
-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2	
email - rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.2.bt2	
L=rwxrwxrwx 1 iu iu 53 5月 14 16:39 pigya.3.bt2	
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2	
- rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2	
-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2	
-rwxrwxrwx 1 iu iu 590 9月 29 2013 ref_genome.fa	
LCL - rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa	
-rwxrwxrwx 1 1u 1u 1616 5月 15 15:08 sample_RNAseq1.sam	5 m .

Bowtie実行

iu@bielinux[~/Desktop/mac_share/mapping_kiso1]

iu@bielinux[mapping kiso1] ls -l

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 /home/iu/Desktop/mac share/mapping k カSAMファイル名は③bowtie1 default.samとしています

total 16394 -rwxrwxrwx 1 iu iu 4194810 5月 17 21:54 agri.1.ebwt 17 21:54 agri.2.ebwt -rwxrwxrwx 1 iu iu 72 5月 -rwxrwxrwx 1 iu iu 53 5月 17 21:54 agri.3.ebwt -rwxrwxrwx 1 iu iu 133 5月 17 21:54 agri.4.ebwt -rwxrwxrwx 1 iu iu 4194810 17 21:54 agri.rev.1.ebwt 5月 -rwxrwxrwx 1 iu iu 17 21:54 agri.rev.2.ebwt 72 5月 -rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.1.bt2 5月 14 16:39 pigya.2.bt2 -rwxrwxrwx 1 iu iu 140 -rwxrwxrwx 1 iu iu 53 14 16:39 pigya.3.bt2 5月 133 14 16:39 pigya.4.bt2 -rwxrwxrwx 1 iu iu 5月 14 16:39 pigya.rev.1.bt2 -rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.2.bt2 -rwxrwxrwx 1 iu iu 140 5月 590 9月 2013 ref genome.fa -rwxrwxrwx 1 iu iu 29 2013 sample RNAseq1.fa -rwxrwxrwx 1 iu iu 396 10月 -rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample RNAseq1.sam iu@bielinux[mapping kiso1] bowtie agri -f sample RNAseg1.fa -S bowtiel default.sam

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

u@bieli	nux[~/Desktop/n	nac	_sha	re/m	apping_kiso ⁺	1]			🄃 🗔 📧 🕪 22:31 🕸
	- rwx rwx rwx	1	iu	iu	53	5月	17	21:54	agri.3.ebwt
0	- rwx rwx rwx	1	iu	iu	133	5月	17	21:54	agri.4.ebwt
	- rwx rwx rwx	1	iu	iu	4194810	5月	17	21:54	agri.rev.l.ebwt
	- rwx rwx rwx	1	iu	iu	72	5月	17	21:54	agri.rev.2.ebwt
	- rwx rwx rwx	1	iu	iu	4194746	5月	14	16:39	pigya.1.bt2
	- rwx rwx rwx	1	iu	iu	140	5月	14	16:39	pigya.2.bt2
9)	- rwx rwx rwx	1	iu	iu	53	5月	14	16:39	pigya.3.bt2
-	- rwx rwx rwx	1	iu	iu	133	5月	14	16:39	pigya.4.bt2
	- rwx rwx rwx	1	iu	iu	4194746	5月	14	16:39	pigya.rev.1.bt2
\succ	- rwx rwx rwx	1	iu	iu	140	5月	14	16:39	pigya.rev.2.bt2
	- rwx rwx rwx	1	iu	iu	590	9月	29	2013	ref_genome.fa
	- rwx rwx rwx	1	iu	iu	396	10月	1	2013	<pre>sample_RNAseq1.fa</pre>
	- rwx rwx rwx	1	iu	iu	1616	5月	15	15:08	<pre>sample_RNAseq1.sam</pre>
	iu@bielinu>	([n	napp	ping	<pre>_kiso1]</pre>	bowt:	ie a	agri -1	f sample_RNAseq1.fa -S
	bowtiel_def	้อเ	ilt.	san	n				
	# reads pro	oce	esse	ed:	8				
	<pre># reads wit</pre>	h	at	lea	ast one i	repor	ted	align	nent: 8 (100.00%)
~	<pre># reads that</pre>	at	fai	llea	to alig	gn: 0	(0.	.00%)	
1	Reported 8	al	.igr	nmer	nts to 1	outp	ut s	stream	(s)
	iu@bielinu>	([n	napp	ping	_kiso1]				[10:30午後]
and the second s									

it

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…

Bowtie実行

u@bieli	pielinux[~/Desktop/mac_share/mapping_kiso1] 11 12:31												
	- rwx rwx rwx	1	iu	iu	53	5月	17	21:54	agri.3.ebwt				
Q	- rwx rwx rwx	1	iu	iu	133	5月	17	21:54	agri.4.ebwt				
	- rwx rwx rwx	1	iu	iu	4194810	5月	17	21:54	agri.rev.l.ebwt				
	- rwx rwx rwx	1	iu	iu	72	5月	17	21:54	agri.rev.2.ebwt				
	- rwx rwx rwx	1	iu	iu	4194746	5月	14	16:39	pigya.1.bt2				
	- rwxrwxrwx	1	iu	iu	140	5月	14	16:39	pigya.2.bt2				
	-rwxrwxrwx	1	iu	iu	53	5月	14	16:39	pigya.3.bt2				
$\underline{}$	-rwxrwxrwx	1	iu	iu	133	5月	14	16:39	pigya.4.bt2				
	- rwx rwx rwx	1	iu	iu	4194746	5月	14	16:39	pigya.rev.1.bt2				
\times	- rwx rwx rwx	1	iu	iu	140	5月	14	16:39	pigya.rev.2.bt2				
	- rwx rwx rwx	1	iu	iu	590	9月	29	2013	ref_genome.fa				
	- rwx rwx rwx	1	iu	iu	396	10月	1	2013	sample_RNAseq1.fa				
	- rwx rwx rwx	1	iu	iu	1616	5月	15	15:08	<pre>sample_RNAseq1.sam</pre>				
	iu@bielinux	<[n	napp	ping	[_kiso1]	bowt:	ie a	agri -1	f sample_RNAseq1.fa -S				
围	bowtiel_det	fau	lt.	san	١								
	# reads pro	oce	esse	ed:	8								
	<pre># reads wit</pre>	th	at	lea	ist one i	repor	ted	alignm	nent: 8 (100.00%)				
	<pre># reads that</pre>	at	fag	ilea	to alig	gn: 0	(0)	.00%)					
1	Reported 8	al	Lig	nmer	nts to 1	outp	ut s	stream	(s)				
(COX)	iu@bielinux	<[n	napp	oing	[kiso1]	0.10			[10:30午後]				
and the second	A578		05 5	a 107	9 1040				Agata Sector Sector A				

bowtie1_default.sam

@HD	VN:1.0	SO:uns	orte	d													
0sg	SN:chr1	LN:48															
0 SQ	SN:chr2	LN:160															
0 SQ	SN:chr3	LN:100															
0 SQ	SN:chr4	LN:123															
0sQ	SN:chr5	LN:100															
@PG	ID:Bowt:	VN:1.1	CL:"	'bowt	ie -	-w	rap	pei	: k	as	sic-0 a	gri -f sam	ple_RNA	seq1.fa	-S bowti	e1_defau	lt.sam"
chr1_11_45	0	chr1	11	255	35м	*	0	0	C	I	XA:i:0	MD:2:35	NM:i:0				
chr2_16_50	0	chr2	16	255	35м	*	0	0	ТZ	I	XA:i:0	MD:2:35	NM:i:0				
chr2_1_35	0	chr2	1	255	35м	*	0	0	A	I	XA:i:0	MD:Z:35	NM:i:0				
chr3_11_45	0	chr5	11	255	35м	*	0	0	TI	I	XA:i:0	MD:Z:35	NM:i:0				
chr3_15_49	0	chr5	15	255	35м	*	0	0	co	I	XA:i:0	MD:2:35	NM:i:0				
chr3_3_37	0	chr3	3	255	35м	*	0	0	GG	I	XA:i:0	MD:2:35	NM:i:0				
chr3_1_35	0	chr3	1	255	35м	*	0	0	GG	I	XA:i:0	MD:2:35	NM:i:0				
chr5_1_35	0	chr5	1	255	35м	*	0	0	GC	I	XA:i:1	MD:Z:3G31	NM:i:1				

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)

①5列目はマッピングクオリティ(MAPQ)だが、、 **bowtie1_default.sam** 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:uns	orte	d													
0sQ	SN:chr1	LN:48															
0 SQ	SN:chr2	LN:160															
0sQ	SN:chr3	LN:100															
0sQ	SN:chr4	LN:123															
0sQ	SN:chr5	LN:100															
@PG	ID:Bowt:	VN:1.1	CL:"	bwt	cie -	w	rap	per	r k	oasi	ic-0 a	gri -f sam	nple_RNA	seq1.fa	-S bowti	.e1_defau	lt.sam"
chr1_11_45	0	chr1	11	255	35м	*	0	0	C	112	<pre>KA:i:0</pre>	MD:2:35	NM:i:0				
chr2_16_50	0	chr2	16	255	35м	*	0	0	ТZ	112	KA:i:0	MD:2:35	NM:i:0				
chr2_1_35	0	chr2	1	255	35м	*	0	0	A	112	KA:i:0	MD:2:35	NM:i:0				
chr3_11_45	0	chr5	11	255	35м	*	0	0	тı	11:2	KA:i:0	MD:2:35	NM:i:0				
chr3_15_49	0	chr5	15	255	35м	*	0	0	co	I: 2	KA:i:0	MD:2:35	NM:i:0				
chr3_3_37	0	chr3	3	255	35м	*	0	0	GG	112	KA:i:0	MD:2:35	NM:i:0				
chr3_1_35	0	chr3	1	255	35м	*	0	0	GG	112	KA:i:0	MD:2:35	NM:i:0				
chr5 1 35	0	chr5	1	255	35м	*	0	0	GC	112	(A:i:1	MD:Z:3G31	NM:i:1				

おさらい。①Bowtie2をデフォルトオプションで Sample_RNAseq1.san)。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

0HD			VN:1.0	SO:unso	orted	d															
0sQ			SN:chr1	LN:48																	
0sQ			SN:chr2	LN:160																	
0sQ			SN:chr3	LN:100																	
0sQ			SN:chr4	LN:123																	
0sQ			SN:chr5	LN:100																	
0PG			ID:bowt	PN:bow	VN:2		/usr	c/b	in/	/	lik	6/J	bowtie2/	/bin/bowt:	ie2-ali	gn-s	wrappe	r basic	-0 -x pi	gya -f sam	ple_RNAs
chr1	11	45	0	chr1	11	42	35м	*	0	0	CC I	11	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	35м	*	0	0	TZ I	I];	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	35м	*	0	0	AC :	I];	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	35м	*	0	0	T 1 :	I] /	AS:i:0	xs:i:0	7 1:0	XM:i:0	xo:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3	15	49	30	chr3	15	1	35м	*	0	0	cc :	I];	AS:i:0	xs:i:0	x i:0	XM:i:0	xo:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU
chr3	3	37	7 0	chr3	3	31	35м	*	0	0	GC I	I];	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	35м	*	0	0	GC I	I];	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	35м	*	0	0	GC I	I],	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最	後の	ノードo	hr5_1_	<mark>351</mark>	は、	(2)b	ow	tie、	3bo	wtie2
お	ま	ナ:	톼	是行	发	0	D	l	J			- -	*	の結 いる 列上	課とも ことが	に、 わかる 其ミス	らくの音 る。この マッチ	『分)リ・ ティー	で 一ト 許変	他の に 冬して)リ- ,リ: ない	ード。ファ	と異な レンフ マップ	よって 、配 され、
@HD	VN:1.0	SO:uns	orte	d															- 1 ° -			. +¬ı		
@sg	SN:chr1	LN:48												(よし)	ようし	アサイ	シシャ	ιC	L 17		21-	-起	凶 9 1	して
@so	SN:chr2	LN:160												思わ	れる。	その。	ようなネ	見上	で	XA:	МГ): N	M: A	AS:
@sg	SN:chr3	LN:100													າເ ເມັດ:				ムフ	иц И	-+-1	、	4	,
@sg	SN:chr4	LN:123													よどの	況明を	読むる	2理	. 円午	L7J	96	, 1 7)	·	
@ S Q	SN:chr5	LN:100																						
@PG	ID:Bowt:	VN:1.1	CL:'	"bowt	tie –	w:	rap	per	c b	as	ic-	0 a	gri	i -f sam	ple_RN	Aseq1.f	a -S bo	wtie	∍1_d	lefau	lt.s	am"		
chr1_11_45	0	chr1	11	255	35M	*	0	0	CG	11	XA:	i:0	MI):Z:35	NM:i:0									
chr2_16_50	0	chr2	16	255	35M	*	0	0	ТZ	11	XA:	i:0	MI):Z:35	NM:i:0									
chr2_1_35	0	chr2	1	255	35M	*	0	0	AG	11	XA:	i:0	MI):Z:35	NM:i:0									
chr3_11_45	0	chr5	11	255	35M	*	0	0	тı	11	XA:	i:0	MI):Z:35	NM:i:0									
chr3_15_49	0	chr5	15	255	35M	*	0	0	CC	11	XA:	i:0	MI):Z:35	NM:i:0									
chr3_3_37	0	chr3	3	255	35M	*	0	0	GG	11	XA:	i:0	MI):Z:35	NM:i:0									
chr3_1_35	0	chr3	1	255	35M	*	0	0	GG	11	XA:	i:0	MI):Z:35	NM:i:0									
chr5_1_35	0	chr5	1	255	35M	*	0	0	GC	1:	XA:	i:1	MI	:Z:3G31	NM:i:1	(2)								
0 B C	Thebowt	LN:100	VN • 2	CT.		c/h	in/	_	141		how	+ie'	2/2	in/howt	102-01-		-wrappe:	r he	aia	-0 -1	r ni		-f asw	DIA DNA
chr1 11 45	10.00%0.	chr1	11	42	25M	*	0		Cd	т 1	78.	i • 0	27.5	ZN.::.0	vm·i·(vouiu	vc·i·0	NM.	1.0	MD - 2	. PI	yya vm.,	2.101	IDIC_KNA
chr2 16 50	0	chr2	16	42	35M	*	0	0	πz	 	AD.	1.0	- 1	(N+i+0	XM·i·C	xo·i·(xc·i·0	NM ·	1.0	MD · 2	1.35	VT - 1	Z.100	
chr2 1 35	0	chr2	1	42	35M	*	0	0	ΔC	T1	AS:	i:0	2	N:i:0	XM:i:0	xo:i:(XG:1:0	NM :	i:0	MD: 2	: 35	YT - 2	Z:UU	
chr3 11 45	0	chr5	11	1	35M	*	0	0	TI	I]	AS:	i:0	2	<pre>ks:i:0</pre>	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	35м	*	0	0	cc	11	AS:	i:0	X	ks:i:0	XN:i:0	хм:і: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	GG	11	AS:	i:0	2	(S: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	35м	*	0	0	GG	11	AS:	i:0	2	(s:i:-12	XN:i:0	XM:i:0	xo:i:0	XG:	i:0	NM:i	:0	MD:	z:35	Y -2:UU
chr5_1_35	1 0	chr5	1	16	35м	*	0	0	GC	11	AS:	i:-	6 X	(s:i:-18	XN:i:0	XM:i:1	xo:i:0	XG:	i:0	NM:i	:1	MD:	z:3G31	3:00

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



May 29, 2018

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

u@bieli	nux[~/Desktop/r	nac	_sha	re/m	apping_kiso1]			🄃 🗔 📧 🕪 15:28 🕸
	- rwx rwx rwx	1	iu	iu	133	5月	17	21:54	agri.4.ebwt
0	- rwx rwx rwx	1	iu	iu	4194810	5月	17	21:54	agri.rev.1.ebwt
	- rwx rwx rwx	1	iu	iu	72	5月	17	21:54	agri.rev.2.ebwt
	- rwx rwx rwx	1	iu	iu	4194746	5月	14	16:39	pigya.1.bt2
	- rwx rwx rwx	1	iu	iu	140	5月	14	16:39	pigya.2.bt2
	- rwx rwx rwx	1	iu	iu	53	5月	14	16:39	pigya.3.bt2
9)	- rwx rwx rwx	1	iu	iu	133	5月	14	16:39	pigya.4.bt2
<u> </u>	- rwx rwx rwx	1	iu	iu	4194746	5月	14	16:39	pigya.rev.1.bt2
	- rwx rwx rwx	1	iu	iu	140	5月	14	16:39	pigya.rev.2.bt2
\succ	- rwx rwx rwx	1	iu	iu	590	9月	29	2013	ref_genome.fa
	- rwx rwx rwx	1	iu	iu	396	10月	1	2013	<pre>sample_RNAseq1.fa</pre>
	- rwx rwx rwx	1	iu	iu	1616	5月	15	15:08	<pre>sample_RNAseq1.sam</pre>
	iu@bielinu>	<[n	napp	ping	_kisol]	bowt:	ie a	agri -1	f sample_RNAseq1.fa -S
	bowtie1_det	fal	ilt.	san	n				
国	# reads pro	DCe	esse	ed:	8				
	<pre># reads wit</pre>	th	at	lea	ast one i	report	ted	alignm	nent: 8 (100.00%)
	<pre># reads that</pre>	at	fai	llea	d to alig	gn: 0	(0.	.00%)	
	Reported 8	al	igr	nmer	nts to 1	outpu	ut s	stream	(s)
	iu@bielinu>	<[n	napp	oing	g_kisol]	bowt:	ie -	·m 1 -·	-beststrata -v 0 agr
	i -f sample	E_F	RNAS	seq	L.fa -S k	powtie	e1_()uasR.s	sam (2)

it

実行結果

iu

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

@bieli	nux[~/Desktop/mac_share/mapping_kiso1] 15:31 🔱 Ja 📧 4)) 15:31 🔱
	-rwxrwxrwx l iu iu 133 5月 14 16:39 pigya.4.bt2
0	-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
	-rwxrwxrwx 1 iu iu 140 5月 14 16:39 pigya.rev.2.bt2
	-rwxrwxrwx l iu iu 590 9月 29 2013 ref_genome.fa
	-rwxrwxrwx l iu iu 396 10月 1 2013 <mark>sample_RNAseq1.fa</mark>
	-rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample_RNAseq1.sam
9)	<pre>iu@bielinux[mapping_kiso1] bowtie agri -f sample_RNAseq1.fa -S</pre>
\leq	bowtie1_default.sam
	<pre># reads processed: 8</pre>
\succ	<pre># reads with at least one reported alignment: 8 (100.00%)</pre>
	<pre># reads that failed to align: 0 (0.00%)</pre>
	Reported 8 alignments to 1 output stream(s)
	<pre>iu@bielinux[mapping_kiso1] bowtie -m 1beststrata -v 0 agr</pre>
	i -f sample_RNAseq1.fa -S bowtie1_QuasR.sam
臣	# reads processed: 8
	<pre># reads with at least one reported alignment: 5 (62.50%)</pre>
	# reads that failed to align: 1 (12.50%)
-	<pre># reads with alignments suppressed due to -m: 2 (25.00%)</pre>
5	Reported 5 alignments to 1 output stream(s)
(Arel-	iu@bielinux[mapping_kisol] [3:29午後]

実行結果

iu

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…

\$
*
ŝ
r
1

複数個所にマップされるこ の2つのリードなのだろう

QuasRのマッピング結果

■ "-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)	sample_RNAseq1.fa - 义モ帳
<pre>>chr1 CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG >chr2 AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA</pre>	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
TGACGCCCTG >chr3	<pre>>chr3_11_45 TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA</pre>
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT >chr4	<pre>>chr3 15 49 CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG >chr3_3_37 CCCCACTATTTCCCCCCCTTCCACCAATCCTCTCACCACC</pre>
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT CGTCTGTATTGGGGGCCTATGCTCGACAAGAGATTGTGTGTAGTATGAGCC ACCAGACTTTACCGTACAAGATA	<pre>>chr3_1_35 GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG >chr3_1_35 GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC >chr5_1_35 CCCCCCCTTCCACCAATCCTCCCCCCTTCCACCAATCCTCTCCCCCC</pre>
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	

実行結果

iu

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…

@biel	inux[~/Desktop/mac_share/mapping_kiso1] 15:31 🔱 Ja 📧 🕬 15:31 🔱
	-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
Q.	-rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
	-rwxrwxrwx l iu iu 140 5月 14 16:39 pigya.rev.2.bt2
	-rwxrwxrwx l iu iu 590 9月 29 2013 ref_genome.fa
	-rwxrwxrwx 1 iu iu 396 10月 1 2013 sample_RNAseq1.fa
	-rwxrwxrwx 1 iu iu 1616 5月 15 15:08 sample_RNAseq1.sam
9)	<pre>iu@bielinux[mapping_kiso1] bowtie agri -f sample_RNAseq1.fa -S</pre>
\leq	<pre>bowtiel_default.sam</pre>
	# reads processed: 8
\times	<pre># reads with at least one reported alignment: 8 (100.00%)</pre>
	<pre># reads that failed to align: 0 (0.00%)</pre>
	Reported 8 alignments to 1 output stream(s)
	iu@bielinux[mapping_kiso1] bowtie -m 1beststrata -v 0 agr
(THE)	i -f sample_RNAseq1.fa -S bowtie1_QuasR.sam
臣	# 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%)
5	Reported 5 alignments to 1 output stream(s)
Same and	

QuasRのマッピング結果

■ "-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 - 又干帳	
	🔄 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	>chr1 11 45
CGAGGAGGAACGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGTGGG	CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>chr2	>chr2 16 50
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG	TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG	>chr2_1_35
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA	AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
TGACGCCCTG	>chr3_11_45
>chr3	TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	>chr3_15_49
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr4	>chr3_3_37
CGAGACGAGCAAGTTATTCGCTCAGTGAATGGGTAGCAAAAGAATGTTGT	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
CETCTETATTEGEGCCTATECTCGACAAGAGATTETETETATEAGCC	>chr3_1_35
	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
ACCAGACTITACCOTACAAGATA	>chr5_1_35
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	-

(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
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…

a@bielinux[~/Desktop/mac_share/mapping_kiso1]
-rwxrwxrwx 1 iu iu 133 5月 14 16:39 pigya.4.bt2
② -rwxrwxrwx 1 iu iu 4194746 5月 14 16:39 pigya.rev.1.bt2
-rwxrwxrwx l iu iu 140 5月 14 16:39 pigya.rev.2.bt2
rwxrwxrwx l iu iu 396 10月 1 2013 sample_RNAseq1.fa
-rwxrwxrwx l iu iu 1616 5月 15 15:08 sample_RNAseql.sam
<pre>jiu@bielinux[mapping_kiso1] bowtie agri -f sample_RNAseq1.fa -S</pre>
bowtie1_default.sam
// # reads processed: 8
<pre># reads with at least one reported alignment: 8 (100.00%)</pre>
<pre># reads that failed to align: 0 (0.00%)</pre>
Reported 8 alignments to 1 output stream(s)
<pre>iu@bielinux[mapping_kiso1] bowtie -m 1beststrata -v 0 agr</pre>
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)
iu@bielinux[mapping_kiso1] [3:29午後]

こ	h-	で	す
	Ŭ	-	1

bowtie1_QuasR.sam

@HD	VN:1.0	SO:uns	orte	d													
0sQ	SN:chr1	LN:48															
0 SQ	SN:chr2	LN:160															
0sQ	SN:chr3	LN:100															
0sQ	SN:chr4	LN:123															
0sQ	SN:chr5	LN:100															
@PG	ID:Bowt:	VN:1.1	CL:	'bowt	ie -	w	rap	per	: k	a	віс-0 -г	n 1best	stra	ta -v 0	agri -f	sample_P	NAseq1.
chr1_11_45	0	chr1	11	255	35м	*	0	0	CG	I	XA:i:0	MD:2:35	NM:i:0				
chr2_16_50	0	chr2	16	255	35м	*	0	0	ТĮ	I	XA:i:0	MD:2:35	NM:i:0				
chr2_1_35	0	chr2	1	255	35м	*	0	0	AG	I	XA:i:0	MD:2:35	NM:i:0				
chr3_11_45	4	*	0	0	*	*	0	0	тı	I	XM:i:1						
chr3_15_49	4	*	0	0	*	*	0	0	cc	I	XM:i:1						
chr3_3_37	0	chr3	3	255	35м	*	0	0	GG	I	XA:i:0	MD:2:35	NM:i:0				
chr3_1_35	0	chr3	1	255	35м	*	0	0	GG	I	XA:i:0	MD:2:35	NM:i:0				
chr5_1_35	4	*	0	0	*	*	0	0	GC	I	XM:i:0						

bowtie1_QuasR.

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:unso	orte	d														
0 SQ	SN:chr1	LN:48																
0 SQ	SN:chr2	LN:160																
@SQ	SN:chr3	LN:100																
@sg	SN:chr4	LN:123																
@sg	SN:chr5	LN:100																
@PG	ID:Bowt:	VN:1.1	CL:	'bowt	ie -	w	rap	per	r ba	sic	:-0 -:	m 1	best	tstra	ta -v O	agri -f	sample_R	NAseq1.
chr1_11_45	0	chr1	11	255	35м	*	0	0	CGI	1 XA	:::0	MD	:Z:35	NM:i:0				
chr2_16_50	0	chr2	16	255	35м	*	0	0	TZ I	JXA	:::0	MD	:Z:35	NM:i:0				
chr2_1_35	0	chr2	1	255	35м	*	0	0	AGI	JXA	:::0	MΓ	:Z:35	NM:i:0				
chr3_11_45	4	*	0	0	*	*	0	0	ΤΊΙ	1 XM	1:i:1							
chr3 15 49	4	*	0	0	*	*	0	0	CCI	1 XM	I:i:1							
chr3_3_37	0	chr3	3	255	35м	*	0	0	GGI	1 XA	:i:0	ML	:Z:35	NM:i:0				
chr3 1 35	0	chr3	1	255	35м	*	0	0	GGI	1 XA	:::0	M	Z:35	NM:i:0				
chr5_1_35	4	*	0	0	*	*	0	0	GCI	1 XM	1:i:0	(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)

a																		
@HD	VN:1.0	SO:unso	orte	d						_								
0 SQ	SN:chr1	LN:48																
0sQ	SN:chr2	LN:160																
0 SQ	SN:chr3	LN:100																
0 SQ	SN:chr4	LN:123																
0 SQ	SN:chr5	LN:100																
@PG	ID:Bowt:	VN:1.1	CL:"	'bowt	ie -	-w	rap	per	r ba	si	- 0-	m 1	best	tstra	ata -v O	agri -f	sample_F	NAseq1.
chr1_11_45	0	chr1	11	255	35м	*	0	0	COI	1 X2	A:i:0	MD	:Z:35	NM:i:0				
chr2_16_50	0	chr2	16	255	35м	*	0	0	TZ I	1 X2	A:i:0	MD	:Z:35	NM:i:0				
chr2_1_35	0	chr2	1	255	35м	*	0	0	AGI	1 X2	A:i:0	МΓ	:Z:35	NM:i:0				
chr3_11_45	4	*	0	0	*	*	0	0	TI I	: X1	M:i:1	C						
chr3_15_49	4	*	0	0	*	*	0	0	CCI	: XI	4:i:1							
chr3_3_37	0	chr3	3	255	35м	*	0	0	GGI	1 X2	A:i:0	ML	:Z:35	NM:i:0				
chr3_1_35	0	chr3	1	255	35м	*	0	0	GGI	1 X2	A:i:0	М	z:35	NM:i:0				
chr5_1_35	4	*	0	0	*	*	0	0	GCI	: XI	M:i:0	KG	<u>}</u>					

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

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 - のでの実行結果)

Bowtieマニュアル中のSAM bowtie output

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

SAM bowtie output

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

- 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
 - The read is one of a pair and has no reported alignments

Bowtie (Langmead et al., Genome Biol., 10: R25, 2009) erse reference strand

8

SAM形式の解説

Optional fields. Fields are tab-separated. For descriptions of all ssible 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.

Aligned read belongs to stratum <N>. See Strata for definition.

XM:i:
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).

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-m1(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

XM:i:<N>



For a read with no reported alignments, $\langle N \rangle$ 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).

12.

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。

	bo. unbe	Drce	d										reau na	u no an	gnmei	its.	11 -m W	as specif	led and the
SN:chr1	LN:48												read's a	lianmer	nts we	res	supress	ed becau	se the -m
SN:chr2	LN:160																		
SN:chr3	LN:100												ceiling v	vas exc	eeded	$, < \mathbb{N}$	I> equa	ls the -m	ceiling 1, t
SN:chr4	LN:123																		
SN:chr5	LN:100																		
D:Bowt:	VN:1.1	CL:"	'bowt	ie -	wr	app	er	ba	asic-	-0 -r	n 1	best	tstrat	ta -v 0	agri	-f	sample	RNAseq1.	
0	chr1	11	255	35м	*	0	0	C()	IIXA	:i:0	MD:	z:35	NM:i:0						
0	chr2	16	255	35м	*	0	0 !	FZ 1	IIXA	i 🖌	MD:	z:35	NM:i:0						
0	chr2	1	255	35м	*	0	0 2	AG 1	IJXA	<u>, (2</u>		z:35	NM:i:0						
4	*	0	0	*	*	0	0 !	C 1 1	I: XM	:i:1									
4	*	0	0	*	*	0	0	cc i	I: XM	:i:1									
0	chr3	3	255	35м	*	0	0 0	GQ 1	IIXA	:i:0	ML:	z:35	NM:i:0						
0	chr3	1	255	35м	*	0	0 0	GQ 1	IIXA	:i:0	M	Z:35	NM:i:0						
4	*	0	0	*	*	0	0 0	GC 1	I XM	:i:0	(3)								
	N:chr1 N:chr3 N:chr3 N:chr5 D:Bowt: 0 0 0 4 4 0 0 4 4 0 0	N:chr1 LN:40 N:chr2 LN:160 N:chr3 LN:100 N:chr4 LN:123 N:chr5 LN:100 D:Bowt: VN:1.1 0 chr1 0 chr2 0 chr2 4 * 4 * 0 chr3 0 chr3 4 *	N:chr1 LN:40 N:chr2 LN:160 N:chr3 LN:100 N:chr4 LN:123 N:chr5 LN:100 D:Bowt: VN:1.1 CL: 0 chr1 11 0 chr2 16 0 chr2 1 4 * 0 4 * 0 0 chr3 3 0 chr3 1 4 * 0	N:chr1 LN:40 N:chr2 LN:160 N:chr3 LN:100 N:chr4 LN:123 D:Bowt: VN:1.1 CL: "bowt 0 chr1 11 255 0 chr2 16 255 0 chr2 1 255 4 * 0 0 4 * 0 0 0 chr3 3 255 0 chr3 1 255 4 * 0 0	N:chr1 LN:48 N:chr2 LN:160 N:chr3 LN:100 N:chr4 LN:123 D:Bowt: VN:1.1 CL: "bowtie 0 chr1 11 255 35M 0 chr2 16 255 35M 0 chr2 1 255 35M 4 * 0 0 * 4 * 0 0 * 0 chr3 3 255 35M 0 chr3 1 255 35M 0 chr3 1 255 35M	N:chr1 LN:40 N:chr2 LN:160 N:chr3 LN:100 N:chr4 LN:123 N:chr5 LN:100 D:Bowt: VN:1.1 CL: "bowtiewr 0 chr1 11 255 35M * 0 chr2 16 255 35M * 0 chr2 1 255 35M * 4 * 0 0 * * 4 * 0 0 * * 0 chr3 3 255 35M *	N:chr1 LN:40 Image: Constraint of the second s	N:chr1 LN:48 Image: Constraint of the second s	N:chr1 LN:48 Image: Constraint of the second s	N:chr1 LN:48 Image: Second Secon	N:chr1 LN:40 N:chr2 LN:160 N:chr3 LN:100 N:chr4 LN:123 N:chr5 LN:100 D:Bowt: VN:1.1 CL: "bowtiewrapper basic-0 -r 0 chr1 11 255 35M * 0 0 CC IIXA:i:0 0 chr2 16 255 35M * 0 0 TZ IIXA:i:0 0 chr2 1 255 35M * 0 0 AC IIXA:i:0 4 * 0 0 * * 0 0 CC IIXA:i:1 4 * 0 0 * * 0 0 CC IIXA:i:1 0 chr3 3 255 35M * 0 0 GC IIXA:i:0 0 chr3 1 255 35M * 0 0 GC IIXA:i:0 4 * 0 0 * * 0 0 GC IIXA:i:0 1 255 35M * 0 0 GC IIXA:i:0	N:chr1 LN:48 Image: Second secon	N:chr1 LN:48 Image: Constraint of the second s	N:chr1 LN:43 read's a N:chr2 LN:160 ceiling v N:chr3 LN:100 ceiling v N:chr4 LN:123 ceiling v N:chr5 LN:100 ceiling v N:chr5 LN:100 ceiling v D:Bowt: VN:1.1 CL: "bowtiewrapper basic-0 -m 1beststrated 0 chr1 11 255 0 chr2 16 255 35M * 0 0 TI XA:::0 MD:2:35 NM:i:0 0 chr2 1 255 35M * 0 0 TI XA:::0 MD:2:35 NM:i:0 0 chr2 1 255 35M * 0 0 TI XA:::0 MD:2:35 NM:i:0 4 0 0 * 0 0 CI XA:::0 MD:2:35 NM:i:0 0 chr3 3 255 35M * 0 0 CI XA:::0 MD:2:35 NM:i:0 0 chr3 1 255 35M * 0 0 GC II XA:::0 MD:2:35 NM:i:0	N:chr1 LN:48 Image: Constraint of the second s	N:chr1 LN:40 read's alignments we N:chr2 LN:160 ceiling was exceeded N:chr3 LN:100 ceiling was exceeded N:chr4 LN:123 ceiling was exceeded N:chr5 LN:100 ceiling was exceeded D:Bowt: VN:1.1 CL:"bowtiewrapper basic-0 -m 1beststrata -v 0 agri 0 chr1 11 255 0 chr2 16 255 0 chr2 1255 4 0 0 ci lixA:::0 0 chr3 3 255 0 chr3 1 255 0 chr3 1 255 4 0 0 cc lixA:::0 0 chr3 1 255 4 0 0 cc lixA:::0 0 chr3 1 255 4 0 0 cc lixA:::0 0 chr3 1 cc lixA:::0 0 chr3 1 cc lixA:::0 0 chr3 1 cc lixA:::0	N:chr1 LN:48 Image: Constraint of the second s	N:chr1 LN:43 read's alignments were supress N:chr2 LN:160 read's alignments were supress N:chr3 LN:100 read's alignments were supress N:chr4 LN:123 read's alignments were supress N:chr5 LN:100 read's alignments were supress D:Bowt VN:1.1 CL: "bowtiewrapper basic-0 -m 1beststrata -v 0 agri -f sample 0 chr1 11 255 35M * 0 0 CC IJ XA:i:0 MD:2:35 NM:i:0 read's alignments were supress 0 chr1 11 255 35M * 0 0 CC IJ XA:i:0 MD:2:35 NM:i:0 read's alignments were supress 0 chr1 11 255 35M * 0 0 CC IJ XA:i:0 MD:2:35 NM:i:0 read's alignments were supress 0 chr2 16 255 35M * 0 0 CI J XA:i:0 MD:2:35 NM:i:0 read's alignments were supress 4 * 0 0 cc IJ XA:i:0 MD:2:35 NM:i:0 read's alignments were supress 0 chr2 1255	Nichrif LN:48 Image: Constraint of the second decimal second deci

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

Bio-Linux環境で行ったマッピングは、① bowtie2のデフォルト、2bowtieのデフォ マッピング結果の違い 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 iu@bielinux[mapping kisol] bowtie agri -f sample RNAseq1.fa -S bowtiel 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) iu@bielinux[mapping kiso1] bowtie -m 1 --best --strata -v 0 agr i -f sample RNAseq1.fa -S bowtiel 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)









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 RNAs	seal.f
a -S sample RNAseq1.sam	🔄 sample_RNAseq1.fa - メモ帳
8 reads; of these:	ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)
8 (100.00%) were unpaired; of these:	>chr1 11 45
0 (0.00%) aligned 0 times	CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
3 (37.50%) aligned exactly 1 time	>chr2_16_50
5 (62.50%) aligned >1 times (2)	TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
100.00% overall alignment rate	->chr2_1_35
	AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
	>chr3_11_45
	TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
	7chr3_15_49
	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
	>chr3_3_37
	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
<pre>iu@bielinux[mapping_kiso1] bowtie -m 1beststrata -v</pre>	>chr3_1_35
i -f sample_RNAseq1.fa -S bowtie1_QuasR.sam	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
# reads processed: 8	>chr5_1_35
# reads with at least one reported alignment: 5 (32.50%)	GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
# reads that failed to aligh: 1 (12.50%)	
# redus with alignments suppressed due to -M: 2 (25.00%) Reported 5 alignments to 1 output stream(s)	
Reported 5 artgiments to 1 output stream(s)	



Δ

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	
ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)	sample_RNAseq1.1a - X-Ek
Nchr1	ファイル(F) 編集(E) 書式(O) 表示(V) ヘルフ(H)
	>chr1_11_45
	CGCTTACGAGATCAGGCTAAGAGTGGATGCTGAGT
>cnr2	>chr2_16_50
AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACATAGACACCTTGAGGAG	TATCTATGGCCTAAAAACATAGACACCTTGAGGAG
ACGCAGGTAGGCTGAGGATAAAGCCGTTTGCACGCATCATGAAGGGGCTG	>chr2_1_35
CTCGGGTATGGTTAGTCTTTGCCTCTAGATTTTCACGACGCTGCGGTTCA	AGGGAGGGGGTCCAGTATCTATGGCCTAAAAACAT
TGACGCCCTG	>chr3_11_45
>chr3	TTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATA
GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	>chr3_15_49
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	CCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGG
>chr4	>chr3_3_37
	GGGGACTATTTCCCCGCTTGCAGGAATCGTGTCAG
	>chr3_1_35
	GGGGGGACTATTTCCCCGCTTGCAGGAATCGTGTC
	>chr5_1_35
>chr5	GCGCGGTCTATTTCCCCGCTTGCAGGAATCGTGTC
GCGGGGTCTATTTCCCCGCTTGCAGGAATCGTGTCAGTTGGTATACAGGC	
AGCATCTAGTCGCATCAGAAGGGTGTAGTCAGCCTATAGTTAACTAGTTT	-

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

Whith Solution and Solutio	
9SQ SN:chr1 LN:48 Image: SN:chr2 LN:160 Image: SN:chr2 LN:160 Image: SN:chr2 LN:100 Image: SN:chr2 LN:100 Image: SN:chr2 LN:123 Image: SN:chr2 Image: SN:ch	
ASQ SN:chr2 LN:160 Image: Constraint of the second s	
SQ SN:chr3 LN:100 Image: SQ SN:chr4 LN:123 Image: SQ SN:chr4 LN:123 Image: SQ	
SQ SN:chr4 LN:123 SQ SN:chr4 LN:123 SQ SN:chr4 LN:123 SQ SQ SN:chr4 LN:123 SQ SQ SN:chr4 LN:123 SQ SQ SN:chr4 LN:123 SQ SQ SQ SQ SN:chr4 LN:123 SQ	
SQ SN:chr5 LN:100	
PG ID:bowt: PN:bow VN:2CL:"/usr/bin//lib/bowtie2/bin/bowtie2-align-swrapper basic-0 -x pigya	sample_RNAs
chr1_11_45 0 chr1 11 42 35M * 0 0 cc IlAs:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:	συ
chr2 16 50 0 chr2 16 42 35M * 0 0 TZ IJAS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:	συ
chr2 1 35 0 chr2 1 42 35M * 0 0 ACIJAS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:35 YT:	σσ
chr3 11 45 0 chr5 11 1 35M * 0 0 TILLAS: 10 XS: 10 XN: 10 XM: 10 XO: 10 XG: 10 NM: 10 MD:	35 YT:Z:UU
chr3 15 49 0 chr3 15 1 35M * 0 0 cc IlAs:i:0 XS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:	35 YT:Z:UU
chr3 3 37 0 chr3 3 31 35M * 0 0 GC IIAS:i:0 XS:i:-6 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:	35 YT:Z:UU
chr3_1_35 0 chr3 1 35 35M * 0 0 GCIJAS:i:0 XS:i:-12 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:	35 YT:Z:UU
chr5_1_35 0 chr5 1 16 35M * 0 0 GC I]AS:i:-6 XS:i:-18 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:	3G31 YT:Z:UU



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

9HD			VN:1.0	SO:uns	orted	d														
9sQ			SN:chr1	LN:48																
9sQ			SN:chr2	LN:160																
sõ			SN:chr3	LN:100																
sQ			SN:chr4	LN:123																
sõ			SN:chr5	LN:100																
PG			ID:bowt	PN:bow	VN:2	CL:'	"/usi	r/b	oin/	/	11	b/bowtie2	/bin/bowt	ie2-ali	gn-s	wrappe	basic	-0 -x pi	gya -f sam	ple_RNAs
chr1	11	45	0	chr1	11	42	35м	*	0	0	CC	IJAS: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	35м	*	0	0	ТZ	IIAS:i:0	XN:i:0	XM:i:0	xo:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
hr2	1 3	35	0	chr2	1	42	35м	*	0	0	AC	IIAS:i:0	XN:i:0	хм:і:0	xo:i:0	XG:i:0	NM:i:0	MD:Z:35	YT:Z:UU	
chr3	11	5	0	chr5	11	1	35м	*	0	0	т	IIAS: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	(3)	35м	*	0	0	cc	IJAS: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	35м	*	0	0	GG	IJAS: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	35м	*	0	0	GC	IJAS: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		35м	*	0	0	GC	I]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
						\checkmark			1										<u>.</u>	

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)
 - □ 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。

乳酸菌学会誌NGS連載の、①第6回がDDBJ 他のBowtie2実行手段。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) このウェブベージのR関連部分は、<u>インストール IIについて</u>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) 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) What's new? 書籍|日本乳酸菌学会誌|について (last modified 2018/05/10) NEW Silhouetteスコアの新たな使い道提唱論文(Zhao et) 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) Silhouetteスコアの新たな使い道提唱論文(Zhao et) |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) 「平成29年度NGSハンズオン講習会」の動画が公開 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、マッビーグ、そしてQC (last modified 2017/06/25) 門田からメール返信をもらえない場合は (last modif 書籍|日本乳酸菌学会誌|<u>第6回ゲノムアセンブリ(</u>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アレ(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 p/) fied 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 (Non-modified 2018/03/23)

 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)(#ast 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)|

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)

• <u>はじめに</u> (last modified 2015/03/31)

 参考資料 | 書籍、学会誌 (last modified 2017/11/13) 講演資料 (last modified) 参考咨料 講習会 講美

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

全体像のおさらい

RNA-Seq data analysis

_F 最近の総説 (Lowe et al., PLoS Comput. Biol., **13**: e1005457, 2017) ₉

processed to yield useful information. Data analysis usually requires a combination of bioinformatics software tools that vary according to the experime (1) designed (2) als. The process can be broken down into the following four stages: <u>quality control</u>, <u>alignment</u>, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a corr (2) d-line interface, eith (2) a Unix environment or within the R/Bioconductor statistical



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

単ーサンプル

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
 - コ 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



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)

ーサンプル

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

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 <u>基本的な利用法(Windows2015.04.03版</u> と
的にまとめた <u>書籍</u> もあります。(2015/04/03)	• マッピング <u>基礎</u> (last modified 2013/06/19)
	・ マッピング single-end ゲノム basic aligner(基礎) <u>QuasR(Gaidatzis 2015)</u> (last modified 2014/06/21)
	・ マッピング single-end ゲノム basic aligner(応用) <u>QuasR(Gaidatzis_2015)</u> (last modified 2015/06/28)
What's new?	・ マッピング single-end ゲノム splice-aware aligner <u>QuasR(Gaidatzis_2015)</u> (last modified 2014/06/21)
 Silhouetteスコアの新たな使い道提唱論文(2) 	・ マッピング paired-end ゲノム basic aligner(基礎) <u>QuasR(Gaidatzis 2015)</u> (last modified 2016/02/11)
 Silhouetteスコアの新たな使い道提唱論文(Z 	・ マッピング paired-end ゲノム basic aligner(応用) <u>QuasR(Gaidatzis_2015)</u> (last modified 2016/02/11)
 「平成29年度NGSハンズオン講習会」の動画 	・ マッピング paired-end トランスクリプトーム basic aligner(基礎) <u>QuasR(Gaidatzis 2015)</u> (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) <u>QuasR(Gaidatzis 2015)</u> (last modified 2016/02/10)
	・ <u>マップ後目について</u> (last modified 2013/06/19)
• 問用からマール 返信をもらえたい 堪合け (1ag	• <u>マップ後 出力ファイル形式について</u> (last modified 2013/11/05)
• はじめに (last modified 2015/03/31)	• <u>マップ後 出力ファイルの読み込み BAM</u> 形式 について (last modified 2016/09/14)
• 参考资料 書籍 学会註 (last modified 2012)	• マップ後 出力ファイルの読み込み BAM形式 <u>rbamtools(Kaisers 2015)</u> (last modified 2016/09/14)
• 参考資料 <u> </u>	• マップ後 出力ファイルの読み込み BAM形式 <u>GenomicAlignments(Lawrence 2013)</u> (last modified 2016/09/14)
	• マッブ後 出力ファイルの読み込み <u>Bowtie形式</u> (last modified 2013/06/18)
	• マップ後 出力ファイルの読み込み <u>SOAP形式</u> (last modified 2013/06/19)
	• マッブ後 出力ファイルの読み込み <u>htSeqTools(Planet_2012)</u> (last modified 2013/06/19)
	• <u>マッブ後 カワント 情報取得 について</u> (last modified 2017/01/11)
	 マッブ後 カウント 情報取得 single-end ゲノム アノテーション有 QuasR(Gaidatzis 2015) (/ tmodified 2015/02/26)
	 マッフ後 カワント情報取得 single-end ケノム アノテーション無 QuasR(Gaidatzis 2015) (1) modified 2014/06/22)
	 マッフ後 カワント情報取得 paired-end ケノム アノテーション有 QuasR(Gaidatzis 2015)(Ast modified 2016/02/13)
	• マッブ後 カワント 情報取得 paired-end ゲノム アノテーション無 <u>QuasR(Gaidatzis 2015)</u> (last modified 2015/07/02)
	• マップ後 カワント 情報取得 paired-end トランスクリプトーム <u>QuasR(Gaidatzis 2015)</u> (last modified 2016/02/12)
	• マッフ後 カワント 情報取得 トランスクリフト ーム <u>BEDファイルから</u> (last modified 2014/06/21)
	 マッフ後 <u>町列長とカワント数の関係</u> (last modified 2015/07/03)
	• <u>IEXTENSE</u> (last modified 2014/06/22) TH (IEE TWO CONTINUES (LAST IN EXAMPLE 1) (IEE TO A CONTINUES)
	• 止現11 基礎 KPK or CPK (肥列 長 補止) (last modified 2015/07/04)









	こんな感じで、エラーメッセージが	<mark>出てなければOK</mark>
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バッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム 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(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 「ファイル」ー「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	軸列への <mark>Bowtie</mark> によるマッピングから、 カウントデータ取得まで ので、 <u>GenomicAlignments</u> バッケージを利用して、 マップされた -ド 数をカウントしています。 ·クトリに移動し以下をコビベ。	
1. <u>サンブルデータ</u> 18,19のRNA-seqデータ(<u>sample_RNAseq1.fa</u>)の <u>ref_ge</u>	<u>nome.fa</u> へのマッピングの場合(<u>mapping single genomel.txt</u>):	
オプションを"-m 1beststrata -v 0"とした例です。		_
in_f1 <- "mapping_single_genome1.txt" #入力ファイル名 in_f2 <- "ref_genome.fa" #入力ファイル param_mapping <- "-m 1beststrata -v 0"#マッビ	を指定してin_f1に格納(RNA-seqファイル) RConsole	
#必要なパッケージをロード library(QuasR)#パッケージの #パッケージの #パッケージのlibrary(GenomicAlignments)#パッケージの #パッケージの	- } > m <- reduce(granges(k)) >	#GRanges\$
<pre>#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 +</pre>	<pre>* #本留(DDDF1前報取得) > tmp <- as.data.frame(m) > for(i in 1:length(tmpfname)){ - tmpcount <- summarizeOverlaps(m, - count <- assays(tmpcount)\$counts - colnames(count) <- tmpsname[i] - tmp <- cbind(tmp, count)</pre>	#出力ファ\$ #サンプル\$ tmpfname[i])#\$ #Summariz\$ #行列coun\$ #保存した\$
<pre>#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(i > tmpsname <- out@alignments[,2] #サンプル名(i for(i in 1:length(tmpfname)){ #サンプル数(> if(i == 1){ k <- readGAlignments(tmpfname[i]) #BAM形式ファ > </pre>	- } > #ファイルに保存 > out_f <- sub(".bam", "_range.txt", > write.table(tmp, out_f, sep="\t", a >	tmpfname[i])#\$ ppend=F, quot\$
<		>

*_range.txt

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バッケージを用いた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 バ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。

1	<u>サンブルデータ</u> 18,19のRNA-seqデータ(<u>sample_R</u>)	<u> Aseq1.fa</u>)の <u>ref</u>	R Con	sole	
	オプションを"-m 1beststrata -v 0"とした例です。			f < -sub("bam" "range tyt" tmpfname	_[i])#\$ [^]
	<pre>in_f1 <- "mapping_single_genome1.txt"</pre>	#入力ファイル	> wri	ite.table(tmp, out_f, sep="\t", append=F	, quot\$
	param_mapping <- "-m 1beststrata	#八月ファイル -v 0"#マッピ	> get	twd ()	
			[1]	"C:/Users/kojik/Desktop/hoge/mapping_kis	o1"
	#必安はハックーンでロート libnany(QuasR)	#13 ss ケージの	> 113	st.files()	
	library(GenomicAlignments)	#バッケージの	[1]	"mapping_single_genome1.txt"	
			[2]	"QuasR_log_146c2f3d4e64.txt"	
	#前処理(マッピング)		[3]	"ref_genome.fa"	
	<pre>time_s <- proc.time()</pre>	#計算時間を計	[4]	"ref genome.fa.fai"	
	time e <- proc.time()	arameter=para #計算時間を計	[5]	"ref_genome.fa.md5"	
	time_e - time_s	#計算時間を表	[6]	"ref_genome.fa.Rbowtie"	
	out	#マッピングに	[7]	"sample RNAseq1.fa"	
	alignmentStats(out)	#マッピンク結	[8]	"sample_RNAseq1_146c6c6d54aa.bam"	
	#本番(マップされたリードの和集合領域同定)		[9]	"sample_RNAseq1_146c6c6d54aa.bam.bai"	
	<pre>tmpfname <- out@alignments[,1]</pre>	#ファイル名(i	[10]	"sample_RNAseq1_146c6c6d54aa.bam.txt"	
	<pre>tmpsname <- out@allgnments[,2] for(i in 1:length(tmpfname)){</pre>	#サンブル冶(1)#サンブル粉(1)	[11]	"sample RNAseq1 146c6c6d54aa.bed"	
	if(i = 1)	#)))//gx([12]	"sample RNAseq1 146c6c6d54aa QC.pdf"	
	k <- readGAlignments(tmpfname[i])	#BAM 形式ファ	[13]	"sample RNAseq1 146c6c6d54aa range.txt"	
	<		>		
					• · · ·
			<		>

	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 ませんが、①*_range.txtという名前のファイルが作成されま
fange.txt	。これは、②.bamという名前のファイルを内部的に入力とし
■ マップ後 カウント情報取得 single-end くて	読み込み、その文字列中の.bamを_range.txtに置換したもの
QuasRバッケージを用いたsingle-end RNA-seqデータのリファレンを	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
の一連の流れを示します。アノテーション情報かない場合を想定してい 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	いるので、 <mark>GenomicAlignments</mark> バッケージを利用して、マッフされた 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ビベ。
1. <u>サンブルデータ</u> 18,19のRNA-seqデータ(<u>sample_RNAseq1.fa</u>)の <u>ref</u>	R Console
<pre>A ノンヨノを"-m1-beststrata -v 0"としたり]です。 for(i in 1:length(tmpfname)){ #サンブル数(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)){ #サンブル数(tmpcount <- summarizeOverlaps(m, tmpfname[i])#GRanges count <- assays(tmpcount)\$counts #SummarizedI colnames(count) <- tmpsname[i] #行列countの tmp <- cbind(tmp, count) #R存したい情 #ファイルに保存</pre>	<pre>> 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" [5] "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.ba" [10] "sample_RNAseq1_146c6c6d54aa.bam.txt"</pre>
out_f <- sub(".bam", "_range.txt", tmpfname[i])#変 write.table(tmp, out_f, sep="\t", append=F, quote=	[12] "sample_RNAseq1_146c6c6d54aa_QC.pdf" [13] "sample RNAseq1 146c6c6d54aa 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 <mark>無</mark> Q	uas R (G	aidatz	is_2015)	
						ि ^{दे} *.bec	b	
[] $[]$ $[]$ $[]$ $[]$ $[]$ $[]$ $[]$						11	45	
1. <u>サンブルデータ</u> 18,19のRNA-seqデータ(<u>sample_RNAseq1.fa</u>)の <u>ref_genome.fa</u> へのマッピングの場合(<u>mapping_single_ge</u> _Chr2							1	35
オブションを"-m1beststrata -v 0"とした例です。 chr2						16	50	
in f1 <- ["mapping single genome1.txt"] #入力ファイル名を指定してin f1に格納(RNA-segファイル) chr3					1	35		
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 1beststrata -v 0"#マッピング時のオプションを指定 Chr3						3	37	
<pre>#必要なバッケージをロード library(QuasR) library(GenomicAlignments) #前処理(マッピング) time s <- proc time()</pre>	#パッケー #パッケー #計算時間	- ジの読み込み - ジの読み込み 見を計測するため		÷	_ran	ge.txt		
out <- qAlign(in_f1, in_f2, a	lignmentParameter=	=param_mapping)#	seanames	start	end	width	strand	namae
A	В	を表示(一番右側の	chr1	11	45	35	+	1
1 FileName	SampleName	グに用いたバラメ〜 グ結果(alignment	chr2	1	50	50	+	2
2 sample RNAseg1.fa	namae		chr3	. 1	37	37	+	2
<pre>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] #サンブル名(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)){ #サンブル数(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 </pre>								

*

range.txt

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バッケージを用いた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、<u>GenomicAlignments</u> バ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。

1.サンプルデータ18	,19のRNA-segデータ(sample	<u>NAseal.fa)のref_genome.faへのマッピングの場合(m</u> ;	apping single genomel.txt):	
オプションを"-m1	beststrata -v 0"とした例で1	5. <u>サンプルデータ</u> 18-20の複数のRNA-seqデータ(sau	mple RNAseq1.fa	<u>Aseq2.fa)をref_genome.fa</u>
		「「「「「」」「「」」「「」」「」」「「」」「」」「「」」「」」「」」「「」」」「」」「」」」」	b :	
in_f1 <- "map	ping_single_genome1.txt	全部のマッピング結果をまとめて和集合領域を定め	2.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
in_t2 <- "ret	_genome.ta" 	行列の形式(2列目以降がカウント情報)にし、配列日	長情報と別々のファイルにして保	存するやり方です。
hau.am_mahhtuR	< III I Dest strata	in f1 <- "mapping single genomed tyt"	#入力ファイルタを指定して	in f1に移納/RNA_segフィ
#必要なバッケー	-ジをロード	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)
library(QuasR)	out f1 <- "hoge5 count.txt"	#出力ファイル名を指定して	out f1に格納
library(Genom	icAlignments)	<pre>out_f2 <- "hoge5_genelength.txt"</pre>	#出力ファイル名を指定して	but_f2に格納
#益加速/ラッピ	6 - N	param_mapping <- "-m 1beststrata	a -v 1 #マッビング時のオブS	ノョンを指定
#刑処理(イツC	()	#心華たじったこのを口にじ		
out <- aAlign	(in f1, in f2, alignmentF	library(QuasR)	#バッケージの読み込み	
time e <- pro	c.time()	library(GenomicAlignments)	#バッケージの読み込み	
time_e - time	_s			
out		#前処理(マッピング)		
alignmentStat	s(out)	time_s <- proc.time()	#計算時間を計測するため	ー、ビンガネ(にき_***
#本悉(マップさ)	れたリードの和集会領域同定)	out <- dAlign(in_fi, in_f2, alignmentF	'arameter=param_mapping)# #計質時間を計測するため	マッピングを1TO qAligni
tmpfname <- o	ut@alignments[.1]	time_e - time s	#計算時間を表示(一番右側の	数字。単位はsecond)
tmpsname <- o	ut@alignments[,2]	out	#マッピングに用いたバラメ·	ータや入力ファイルの情報
for(i in 1:le	ngth(tmpfname)){	alignmentStats(out)	#マッピング結果(alignment	: statistics)の表示。s
if(i == 1){				
k <- read	GAlignments(tmp+name[i])	#本番(マッフされたリートの和集合領域可定)	#ファイルタ(きゃ 41の17月日	- 相坐)を+mnfnomoとして
		tmpsname <- out@alignments[,1]	#ファイル石(III_TIの191日) #サンブル名(in_f1の2列日)	に相当)をtmp=nameとして、 こ相当)をtmp=nameとして、
		for(i in 1:length(tmpfname)){	#サンブル数(ファイル数)分	だけループを回す
		<		>

	①例題5をやってみましょう。「デスクトップ
複数サンブル	- noge - mapping_kiso2」フォルタを作成し 、必要な入力ファイルを揃えてコピペ実行
5. <u>サンブルデータ</u> 18-20の複数のRNA-seqデータ(<u>sample RNAseq1.fa</u> と にマッビングする場合(<u>mapping single genome4.txt</u>):	sample RNAseq2.fa)をref genome.fa
全部のマッピング結果をまとめて和果音領域を定め、カウンド情報を得 行列の形式(2列目以降がカウンド情報)にし、配列長情報と別々のファイ	a149万です。一般的なカランドチーダ イルにして保存するやり方です。
<pre>in_f1 <- "mapping_single_genome4.txt" #入力ファイル名 in_f2 <- "ref_genome.fa" #入力ファイル名 out_f1 <- "hoge5_count.txt" #出力ファイル名 out_f2 <- "hoge5_genelength.txt" #出力ファイル名 papare mapping < " m 1 best structs x 1"#マッドン기</pre>	を指定してin_f1に格納(RNA-seqフ を指定してin_f2に格納(リファレン) を指定してout_f1に格納 を指定してout_f2に格納
param_mapping <m #="" -v="" i="" ibeststrata="" v="" ッピン<br="">#必要なパッケージをロード</m>	R Console
#必要なパックージをロート library(QuasR) #パッケージの読 library(GenomicAlignments) #パッケージの読	 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。
<pre>#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</pre>	'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。
<pre>#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: </pre>	<pre>> getwd() [1] "C:/Users/kadota/Desktop/hoge/mapping_kiso2" > list.files() [1] "mapping_single_genome4.txt" [2] "ref_genome.fa" [3] "sample_RNAseg1.fa"</pre>
	[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
--	---
YEQ SX り し り 5.サンブルデータ18-20の複数の RNA-seqデータ(sample RNAseq1.faとsaIC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 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in_f1 <- "mapping_single_genome4.txt"#人力ファイル名を #人力ファイル名を #人力ファイル名を #出力ファイル名を #コファイル名 #コファイル名 #ロファイジの読み #ニカファイジの読み #ニカファイジの読み #ニカファージの読み #ニカファージの読み #ニカファージの読み #ニカファージの読み #ニカファージの読み #ニカファージの読み #ニカファージの読み #ニカファージの読み #ニカングレジの読み #ニカングレジク #計算時間を計測す #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニト #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレジク #ニカングレ	イル(hoge5_count.txt)のほうです ample RNAseq2.fa)をref genome.fa をやり方です。一般的なカウントデータ 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" [3] "mapping_single_genome4.txt" [5] "ref_genome.fa.fai" [6] "ref_genome.fa.fai" [7] "ref_genome.fa.Rbowtie" [8] "ref_genome.fa.Rbowtie"
<pre>#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 </pre>	<pre>[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" ></pre>



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

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

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

(Rで)塩基配列解析

アノテーション情報あり

このウェブベージのR関連部分は、インストール目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) • 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) What's new? ・ 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) Silhouetteスコアの新たな使い道提唱論文(2) • マッピング | paired-end | ゲノム | basic aligner(応用) | QuasR(Gaidatzis 2015) (last modified 2016/02/11) Silhouetteスコアの新たな使い道提唱論文(Z ・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) 「平成29年度NGSハンズオン講習会」の動画 • 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) マップ後目について(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) 門田からメール返信をもらえない場合は (las) マッブ後|出力ファイルの読み込み|BAM形式|について (last modified 2016/09/14) • はじめに (last modified 2015/03/31) ・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) 参考資料 | 書籍、学会誌 (last modified 201) ・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) 参老咨料 I 講習会 講義 講演 溶料 (last m マッブ後|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) <u>マッブ後|カウント情報取得||こついて</u> (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) (Normodified 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)

・ マッブ後 | 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)

正規化日こついて (last modified 2014/06/22)

• 正規化 | 基礎 | <u>RPK or CPK (配列長補正)</u> (last modified 2015/07/04)



①GFF3ファイルの中身は、こんな感じでした。

おさらい

##gff-version	3							ノ有 QuasR(Gaidatzis_2015) NEW
##sequence-	regio	n Chromo	some 3	860 227	785	3		
#!genome-bui	ld Eu	ropean Nuc	leotide	Archiv	e Aŝ	SM	82939	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 (イノム)
#!genome−ver	rsion	GCA_00082	29395.1					
#!genome-dat	e 20	14-11						
#!genome-bui	ld-ac	cession GC	A 000	829395	.1			アイルを乳酸菌ケノムにマッピンクする場合:
#!genebuild-la	ast-u	pdated 2014	4-11					STA形式ファイル(<u>sample_RNAseq4.fa</u>)です。マップされる側のファイルは、 <u>Ensembl</u> (<u>Zerbino</u> Lactobacillus casei 12Aの_multi-FASTA形式ゲノム配列ファイル
Chromosome	ena	gene	360	1676.	+		ID=ge	0829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式の
Chromosome	ena	transcript	360	1676.	+		ID=tr	<u>s_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3</u>)を読み込んでカウン
Chromosome	ena	exon	360	1676.	+		Pare	
Chromosome	ena	CDS	360	1676.	+	0	ID=C	#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.cnromosome.Cnromosome.ta #ハリファー jcm 18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"世社 ファイル
Chromosome	ena	gene	1852	2991.	+		ID=ge	#出力ファイル名を指定してout_fに格納
Chromosome	ena	transcript	1852	2991.	+		ID=tr	#バワンドナーダ取得時のレベルを指定: gene , exon , prom zr , juncti
Chromosome	ena	exon	1852	2991.	+		Pare	
Chromosome	ena	CDS	1852	2991.	+	0	ID=C	#バッケージの読み込み #パッケージの読み込み
###								
Chromosome	ena	gene	3233	3457.	+		ID=ge	"auta")#tydbオゴミリークトの作成
Chromosome	ena	transcript	3233	3457.	+		ID=tr	auto J#txdbs シシェシトのTFM。 #確認してるだけです
Chromosome	ena	exon	3233	3457.	+		Pare	
Chromosome	ena	CDS	3233	3457.	+	0	ID=C	#マッピングを行うoAlign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
Chromosome	ena	gene	3467	4588 .	+		ID=ge	
		count <- qCo	ount(out	t, txdb,	rep	ort	Level	- =param_reportlevel)#カウントデータ行列を取得してcountに格納
		dim(count)						#FT
		<						
May 29, 2018								78

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

QuasRパッケージを用いたsingle-end RNA-segデータのリファレンスゲノム配列へのBowtiel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 バ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バ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ブ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。 「ファイルレー「ディレクト」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ビベ。 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: 1.サンブルデータ70 <u>mapping single gen</u> のサンプル名」(例:hum 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 す。hg19にマップした結 (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 で、Entrez Gene IDに対 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 in f1 <- "mapping</pre> ト情報を取得しています。 in f2 <- "BSgenom out f <- "hoge1.t 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) param mapping <in f2 <- "Lactobacillus hokkaidonensis jcm 18461.GCA 000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファー param txdb1 <- "h param txdb2 <- "k in f3 <- "Lactobacillus hokkajdonensis jcm 18461.GCA 000829395.1.30.chromosome.Chromosome.gff3"#シンファイル: param reportlevel out f <- "hoge10.txt" #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 "iuncti param reportlevel <- "gene" #必要なバッケージを library(QuasR) library(GenomicFe #必要なバッケージをロード library(QuasR) #バッケージの読み込み #前処理(マッピング library(GenomicFeatures) #バッケージの読み込み out <- qAlign(in alignmentStats(ou #前処理(アノテーション情報を取得) #前処理(TxDbオブジ txdb <- makeTxDbFromGFF(in f3, format="auto")#txdbオブジェクトの作成 txdb <- makeTxDbF #確認してるだけです txdb txdb #本番(カウントデー #前処理(マッピング) count <- qCount(o #マッビングを行うgAlign関数を実行した結果をoutに格納 out <- qAlign(in f1, in f2)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 #行数と列数を表示 dim(count) 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は、①のリストファイル マップする側のファイル^{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-segデータのリファレンスゲノム配列へのBowtiel こよるマッビングから、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 バ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バ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ブ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ビベ。 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: 1.サンブルデータ7のFAS mapping single genom 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 のサンブル名」(例:hum et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FAS (2) 式ゲノム配列ファイル す。hg19にマップした結 (Lactobacillus hokkaidonensis jcm_18461.GCA_000829395.1.30.dna.chromosome.Chron 、 me.fa) です。マッピング結果に対して、GFF3形式の で、Entrez Gene IDに対 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 in f1 <- "mapping</pre> ト情報を取得しています。 in f2 <- "BSgenom out f <- "hoge1.t カファイル名を指定してin_f1に格納(RNA-seqファイル) in f1 <- "mapping single genome7.txt"</pre> param mapping <-🖬 18461.GCA 000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファー in f2 <- "Lactobacillus hokkaidonensis param txdb1 <- "h param_txdb2 <- "k in f3 <- "Lactobacillus hokkaidonensis icm 18461.GCA 000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル param reportlevel out f <- "hoge10.txt" #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", "juncti param reportlevel <- "gene" #必要なバッケージを library(QuasR) library(GenomicFe #必要なバッケージをロード library(QuasR) #バッケージの読み込み #前処理(マッピング ⊷ケージの読み込み out <- qAlig SampleName FileName alignmentSt #前処理(TxDb o")#txdbオブジェクトの作成 sample_RNAseq4.fa Lacto txdb <- make 認してるだけです txdb #本番(カウントデー #前処理(マッピング) count <- qCount(o out <- qAlign(in f1, in f2) #マッビングを行うgAlign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 #確認してるだけです head(count) May 29, 2018

82

N 1997			マップマ	する側のフ	<u>ァイルは、①のリストファイル</u>
マッ	プする側	のファイ	、 ル の に記 の が 中	載されてし 身	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sin 流れを示します。アノテーショ としつつ、TxDbパッケージを basic aligner(応用) QuasR(C	ngle-end RNA-seqデータのリファレンス aン情報は、 <u>GenomicFeatures</u> バッケー :読み込むやり方も示しています。マッ <u>Gaidatzis 2015</u> などを参考にしてくださ 5 再しで経知したいファイルを発いてまる	ゲノム配列への <mark>Bowtie</mark> によるマッピ ジ中の関数を利用してTxDbオブジョ ピングのやり方やオプションの詳細 い。	ングから、カウントデータ -クトをネットワーク経由で こついては <u>マッピング si</u>	取得までの 一連の 取得するのを基本 ngle-end ゲノム	
	0 mapping single genome7 to		3 酸菌ゲ ノノ		
<pre>I. <u>サンプルテータ</u>70 FAS II mapping single genom のサンプル名J(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)</pre>	<pre>0.mapping single genome/.tx マップする側のファイルは、サ et al., Nucleic Acids Res., 2018 (Lactobacillus hokkaidonensis アノテーションファイル(Lactoba ト 情報を取得しています。 in_f1 <- "mapping_sing in_f2 <- "Lactobacillu out_f <- "hoge10.txt" param_reportlevel <- " #必要なパッケージをロート library(QuasR)</pre>	▲ 中のFASIA定丸ファイルマ ンプルデータ47のFASTA形式)から提供されている Lactob jcm 18461.GCA 00082939 acillus hokkaidonensis jcm ■ Lactob Lact	オファイル(<u>sample R</u> acillus casei 12Aの 1 <u>5.1.30.dna.chromosc</u> 18461.GCA 000829 ファイル名を指定 18461.GCA_000829 ファイル名を指定 ントデータ取得時 ントデータ取得時	>Chromoson TGACTGATT1 >Chromoson AGAAGATGTC >Chromoson TAACCAATC4 >Chromoson CTTCAAGGAC >Chromoson	ne_361_400 FAGAAACACTTTGGGACACAATTAAAGAATC ne_1637_1676 CCAAAACCTTAAAAATGGAGCTAAAGCCATAG ne_1851_1890 FACAATTAGTCGTGCAACTTTTACAGCCAAA ne_1843_1882 ATGAAATTTACAATTAGTCGTGCAACTTTTA ne_1833_1872 GTAACCAATCATGAAATTTACAATTAGTCGT ne_1823_1862
out <- qAlig alignmentSta	Name	SampleName	-) ーンの記み込み	>Chromoson	DUTTUAAGGAGTAAUUAATUATGAAATTTAU ne 1813-1852
#前処理(TxDb: txdb <- make txdb	ple_RNAseq4.fa	lacto 🐹	')#txdb オブジェク してるだけです	AAATTAAAGA >Chromoson	ACAAATTCAACCTTCAAGGAGTAACCAATCA ne_3418_3457
#本番(カウントデー・ count <- qCount(o く	#前処理(マッピング) out <- qAlign(in_f1, i alignmentStats(out)	■ n_f2) #マッ #マッ	ピングを行うqAli ビング結果(aligr	GATTGCAGA1 >Chromoson TTGCAGATA4	FAATGGGACATTTGTCATTCAAAATGAGTAG ne_3420_3459 ATGGGACATTTGTCATTCAAAATGAGTAGGC
	#本番(カウントデータ取得 count <- qCount(out, t dim(count) head(count)) xdb, reportLevel=parar #行数 #確認	m_reportlevel)#: と列数を表示 してるだけです	>Chromoson GCAGATAATO >Chromoson	ne_3422_3461 GGGACATTTGTCATTCAAAATGAGTAGGCAA ne_3443_3482 CACTACCCAACTTAAAATCATTTTAAAAACAAC
May 29, 2018	<			ATTOAAAATU	83

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



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 マップする側のファイル 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-segデータのリファレンスゲノム配列へのBowtiel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 バ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バ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ブ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」ー「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ビベ。 1.サンブルデータ7のFAs 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: mapping single genome 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 のサンブル名」(例:hum et al., Nucleic Acids Res., 2018)から提供されている Lactobacillus casei 12Aの multi-FAS7(1) 式ゲノム配列ファイル す。hg19にマップした結 (Lactobacillus hokkaidonensis jcm 18461.GCA 000829395.1.30.dna.chromosome.Chron 、 me.fa) です。マッピング結果に対して、GFF3形式の で、Entrez Gene IDに対 アノテーションファイル(Lactobacillus hokkaidonensis jcm 18461.GCA 000829395.1.30.dna.chromosome.Chromosome.gff3(2) み込んでカウン in f1 <- "mapping ト情報を取得しています。 in f2 <- "BSgenom out f <- "hoge1.t 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) param mapping <in_t2 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa_#入力ファー in_f3 <- "Lactobacillus_hokkaidonensis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.gff3"(2)コファイル param txdb1 <- "h param txdb2 <- "k param reportlevel out f <- "hoge10.txt"</pre> #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 #必要なバッケージを library(QuasR) library(GenomicFe #必要なバッケージをロード library(QuasR) #バッケージの読み込み #前処理(マッピング library(GenomicFeatures) #バッケージの読み込み out <- qAlign(in alignmentStats(ou #前処理(アノテーション情報を取得) #前処理(TxDbオブジ txdb <- makeTxDbFromGFF(in f3, format="auto")#txdbオブジェクトの作成 txdb <- makeTxDbF #確認してるだけです txdb txdb #本番(カウントデー #前処理(マッピング) count <- qCount(o #マッビングを行うgAlign関数を実行した結果をoutに格納 out <- qAlign(in f1, in f2)</pre> 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 #行数と列数を表示 dim(count) #確認してるだけです head(count) < May 29, 2018 85

N. 6							<u> </u> 1 マ	'ップする側(sample_RNAseq4.fa)のリー
マット	プオン	ス作) -	7 7		ドは	、②アノテーションファイル(.gff3)中の③
	/ 7 '(2 IS	₹ '] ∨.				gene	領域を参考にしながら作成しています。
##gff=version 3						ン有 QuasR(Ga	全て	完全一致でマッフされるように設計。
##sequence-regio	n Chromo	some 3	360 2277	853		:るマッピングから、カウ 	リントデータ	マ取得までの 一連の
#!genome-build Eu	uropean Nuc	leotide	Archive) ASI	M82939	10オフシェクトをネットワ Vの詳細については <u>マ</u> 、	ーク経田 C <u>ゾビング si</u>	『取得するの』を基本 ingle-end ゲノム
#!genome-version	GCA_00082	29395.1				ドードベ		
#!genome-date 20	14-11					``''	$\overline{1}$	
#!genome-build-a	ccession GC	000_A	829395.	1		ドイルで 孔岐困ソン		>Chromosome_361_400
#!genebuild-last-u	pdated 2014	4-11				SIAボエレアイル(S Lactobacillus casei	12AD 1	TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
Chromosome ena	gene	360	1676.	+	3D=ge	0829395.1.30.dna.c	hromose	>Chromosome_1637_1676
Chromosome ena	transcript	360	1676.	+ .	ID=tr	s jcm 18461.GCA	000829	AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
Chromosome ena	exon	360	1676.	+ .	Pare			>Chromosome_1851_1890
Chromosome ena	CDS	360	1676 .	+	0 ID=C	#入力ファイル名	るを指定	CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
###						_jcm_18461.GCA _jcm_18461_GCA	000829	>Chromosome_1843_1882
Chromosome ena	gene	1852	2991	+	3 D=g	#出力ファイル名	るを指定	TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA
Chromosome ena	transcript	1852	2991	+		#カウントデータ	?取得時(>Chromosome_1833_1872
Chromosome ena	evon	1852	2991	+	Pare			CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
Chromosome ena		1852	2001	+		#バッケージの読	「み込み	>Chromosome_1823_1862
###		1002	2001.			#バッケージの訝	きみ込み	CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC
		0000	2/67	_	20			>Chromosome_1813_1852
Chromosome ena	gene	0200	2457			"auto")#txdb才	ブジェク	AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA
Chromosome ena	transcript	3233	0457	т.	ID-tr	#帷認してるたけ	гса	>Chromosome_3418_3457
Chromosome ena	exon	3233	3437.	+.	Pare			GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG
Chromosome ena	CDS	3233	3457.	+		#マッピングを行	j∋qAli	>Chromosome_3420_3459
###			4500			#マッピング結开	e(align	TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
Chromosomelena	gene	3467	4588 .	+	o D=a			>Chromosome_3422_3461
	count <- qCo	ount(out	t, txdb,	repo	rLevel	=param_report1	evel)#	GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
	head(count)					#1丁致⊂列致で塗 #確認してるだけ	v⊼ ⊦です	>Chromosome_3443_3482
	<							ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC

マップする側のファイル

##gff-version	3							ン有 QuasR(Gaidatzis_	2015) NEW
##sequence-	regio	n Chromo:	some 3	360 227	785	3		、 るマッピングから、カウントデータ	
#!genome-bui	ld Eu	ropean Nuc	leotide	 Archiv 	e Aŝ	βM	8293	bオブジェクトをネットワーク経由で レの詳細については マッピング [si	『取得するのを基本 ingle-endlゲノム」
#!genome−ver	rsion	GCA_00082	9395.1	1					
#!genome-dat	e 20	14-11						건가 소의 秘화년 ㅋㅋㅋㅋ	
#!genome-bui	ld-ac	cession GC:	A_000	829395	.1			アイルぞ孔酸医ケノムにイ	>Chromosome_361_400
#!genebuild-la	ist-u	pdated 2014	4—11					STAポシュ、ファイル(same) R Lactobacillus casei 12(1)	TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
Chromosome	ena	gene	360	1676	2		ID=g	0829395.1.30.dna.chro	>Chromosome_1637_1676
Chromosome	ena	transcript	360	1676.	+		ID=tr	s jcm 18461.GCA 000829	AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
Chromosome	ena	exon	360	1676.	+		Pare		>Chromosome_1851_1890
Chromosome	ena	CDS	360	1676.	+	0	ID=C	- #入力ファイル名を指定 form 18461 CCA 000820	CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
###									>Chromosome_1843_1882
Chromosome	ena	gene	1852	2991.	+		ID=g	#出力ファイル名を指定	
Chromosome	ena	transcript	1852	2991.	+		ID=ti	#カワントナーダ取得時! 	>Uhromosome_1833_1872
Chromosome	ena	exon	1852	2991.	+		Pare		
Chromosome	ena	CDS	1852	2991.	+	0	ID=C	#バッケージの読み込み #バッケージの読み込み	>Uhromosome_1823_1862
###								「#ハッツリーン V20L0天)(20)	UAAATILAAUUTILAAGGAGTAAULAATLATGAAATITAU
Chromosome	ena	gene	3233	3457.	+		ID=g		
Chromosome	ena	transcript	3233	3457.	+		ID=tr	「auto")#txdbオフンエク 1#確認してるだけです	AAATTAAAGACAAATTCAACCIICAAGGAGTAACCAATCA
Chromosome	ena	exon .	3233	3457.	+		Pare		[//TTTCCACATAATCCCACATTTCTCATTCAAAAATCACTAC
Chromosome	ena	CDS	3233	3457.	+	0	ID=C	#フッピングを行う aAli	Chromosomo 3420 3459
###								#マッビング結果(align	
Chromosome	ena	gene	3467	4588 .	+		ID=g		Chromosome 3422 3461
		count <- qCo	unt(out	t, txdb,	rep	ort	tLevel	.=param reportlevel)#フ	
	-	dim(count)						#行数と列数を表示	>Chromosome 3443 3482
		head(count)						#帷認してるたけです	ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
May 29, 2018		*							87

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。

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。

7.1	,-	パオン	Z /E		יה			▲ Ⅱ 、 かかるように設計。領域内ではない。
X)		94	<u>۹ ک</u>	₹IJ V	ノ.		<u> </u>	
##gff-version	3							ン有 QuasR(Gaidatzis_2015) NEW
##sequence-i	regior	n Chromos	some 3	360 227	7785	3		ころマッピングから、カウントデータ取得までの 一連の
#!genome-bui	ld Eu	ropean Nuc	leotide	Archiv	ve AS	SM	32939	∲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 との詳細についてはマッピング!single-end!ゲノム!
#!genome=ver	rsion	GCA_00082	9395.1					
#!genome-dat	e 201	14-11						ᅚᅸᅸᄿᇲ ᇦᇧᇿᅕᅈᆁᅍᄷᅕᄰᇧᇿᆝᇊᆋᅸᇲᄰᆂᅎᄺᄉ
#!genome-bui	ld-ac	cession GC	A_000	829395	5.1			アイルを乳酸風リノムにく >Chromosome_361_400
#!genebuild-la	st-u	odated 2014	1-11					FTA形式ファイル(<u>sample R</u> Lactobacillus casei 12Aの d Lactobacillus casei 12Aの d
Chromosome	ena	gene	360	1676	. +		ID=ge	0829395.1.30.dna.chromosc >Chromosome_1637_1676
Chromosome	ena	transcript	360	1676	. +		ID=tr	<u>s jcm 18461.GCA 000829 AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG</u>
Chromosome	ena	exon	360	1676	. +		Pare	>Chromosome_1851_1890
Chromosome	ena	CDS	360	1676	. +	0	ID=C	#入力ファイル名を指定 CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA
###								_jcm_18461.GCA_000829>Chromosome_1843_1882
Chromosome	ena	gene	1852	2991	2		ID=ge	#出力ファイル名を打定 TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTA
Chromosome	ena	transcript	1852	2991	\ +		ID=tr	#カワントテータ収入す >Chromosome_1833_1872
Chromosome	ena	exon .	1852	2991	+		Pare	
Chromosome	ena	CDS	1852	2991	+	0	ID=0	#バッケージの読み込み >Uhromosome_1823_1862
###	0			2001				
Chromosome	ena	gene	3233	3457	+		ID=gr	>Uhromosome_1813_1852
Chromosome	ena	transcript	3233	3457	+			"auto")#txdbオ ブジェ (AAATTAAAGAUAAATTUAAUUTTUAAGGAGTAAUUAATUA # #確認してるだけです。
Chromosome	ena	evon	3233	3457		•	Pare	
Chromosome	ena		3233	3457				
###	Chu	020	0200	0.07	· ·			1 #マッピングを行うqA11 /UNFOMOSOME_34ZU_3459 #マッピング結果(align TTCCACATAATCCCACATTTCTCATTCAAAAATCACTACCC
Chromosome	ona	gene	3467	4588	+		ID= a	
Onionosome				t tydh	non	ont		
	0	dim(count)	unc(ou	c, crub	, rep	ont	Level	
	ł	nead(count)						
		<						

マップする側のファイル

##gff-version 3	3						ン有 QuasR(Gaidatzis_	2015) NEW	
##sequence-reg	gion Chrom	osome	360 227	785	3		:るマッピングから、カウントデータ	?取得までの一連の	
#!genome-build	European Ni	cleotide	e Archive	e As	ŞΜξ	3293	Ibオブジェクトをネットワーク経由で Vの詳細については マッピング Isi	*取得するのを基本	
#!genome-version	on GCA_000	329395.	1						
#!genome-date :	2014-11						건가 수 있 자유 부 년 기기는	122 H-12 H A	
#!genome-build-	accession C	GA_000	829395.	1			アイルぞ孔酸風ケノムにマッ	>Chromosome	e_361_400
#!genebuild-last	-updated 20	14-11					STAポジェ、ノアイル(<u>sample R</u> Lactobacillus casei 12Aの t	TGACTGATTT/	AGAAACACTTTGGGACACAATTAAAGAATO
Chromosome er	na gene	360	1676.	+		ID=ge	0829395.1.30.dna.chromoso	>Chromosom@	e_1637_1676
Chromosome er	na transcrip	t 360	1676.	+		ID=tr	s jcm 18461.GCA 000829	AGAAGATGTC	CAAAACCTTAAAATGGAGCTAAAGCCATAG
Chromosome er	na exon	360	1676.	+		Pare		>Chromosom	e_1851_1890
Chromosome er	na CDS	360	1676.	+	0	ID=C	#人力ファイル名を指定 icm 18461_CCA_000820	CAIGAAAIII/	ACAATTAGICGIGCAACITITACAGCCAAA
###							_jcm 18461.GCA 000829	>Chromosome	e_1843_1882
Chromosome er	na gene	1852	2991.	+		ID=ge	#出力ファイル名を指定	TAACCAATCA	IGAAATITACAATTAGICGIGCAACITITA
Chromosome er	na transcrip	t 1852	2991.	+		ID=tr	#カワントナーダ収得時(>Uhromosome	
Chromosome er	na exon	1852	2991.	+		Pare			TAACCAATCATGAAATTTACAATTAGTCGT
Chromosome er	na CDS	1852	2991.	+	0	ID=C	#バッケージの読み込み #バッケージの読み込み	CANATTONAC	8_1823_1802 0TT0440040T440044T04T04444TT44
###									UTTUAAGGAGTAAUUAATUATGAAATITAU - 1919 1959
Chromosome er	na gene	3233	3457	2		ID=ge	╵╸╷┾╺╹╲╫┾╍┨┝╶┽╶╝╝╎╲╴╱		9_1010_1002 0&&&TTC&&CCTTC&&CC&CT&&CC&&TC&
Chromosome er	na transcrip	t 3233	3457.	+		ID=tr	#確認してるだけで <mark>ゴー</mark>	AAATTAAAGAG	$\sim 3/18 - 3/17$
Chromosome er	na exon	3233	3457.	+		Pare			8_0410_0407 AATGGGACATTTGTCATTCAAAATGAGTAG
Chromosome er	na CDS	3233	3457.	+	0	ID=C	#マッピングを行うoAli		\sim 3/20 3/59
###							#マッビング結果(align	TTGCAGATAA	
Chromosome er	na gene	3467	4588 .	+		ID=ge		>Chromosom	e 3422 3461
	count <- q	Count(ou	t, txdb,	rep	ort	Level		GCAGATAATG	GGACATTIGICATICAAAATGAGTAGGCAA
	dim(count)						#行数と列数を表示	>Chromosom	e 3443 3482
	head(count)					#0睡話としてる/こけです	ATTCAAAATGA	AGTAGGCAACTTAAATGATTTTAAAAGAAC
May 29, 2018	-							1	89

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。

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							ン有 QuasR(Gaidatzis_2015) NEW
##sequence-regio	on Chromo	some 3	360 227	785	3		
#!genome-build Ei	uropean Nuc	leotide	Archiv	e As	SM	82939	ゆオブジェクトをネットワーク経由で取得するのを基本 Pの詳細についてはマッピング」single-endlゲノムト
#!genome-versior	1 GCA_00082	29395.1					
#!genome-date 20)14-11						드는 ^
#!genome-build-a	ccession GC	A_000	829395	.1			アイルを乳酸菌ワンムにく アイ形式デコスイル(comple P) >Chromosome_361_400
#!genebuild=last=u	updated 2014	1-11					Lactobacillus casei 12AO I
Chromosome ena	gene	360	1676.	+		ID=ge	<u>0829395.1.30.dna.chromoso</u> >Chromosome_1637_1676
Chromosome ena	transcript	360	1676.	+		ID=tr	<u>s_1cm_18461.GCA_000829</u> AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG
Chromosome ena	exon	360	1676.	+		Pare	>Chromosome_1851_1890
Chromosome ena	CDS	360	1676.	+	0	ID=C	#人力ファイル名を指定 CAIGAAAIIIACAAIIAGICGIGCAACIIIIACAGCCAAA jcm 18461_GCA_000829、ol
###							_jcm_18461.GCA_000829 >Uhromosome_1843_1882
Chromosome ena	gene	1852	2991.	+		ID=ge	#出力ファイル名を指定 AAUUAA UA GAAA AUAA AG UG GUAAU A
Chromosome ena	transcript	1852	2991.	+		ID=tr	
Chromosome ena	exon	1852	2991.	+		Pare	CITCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT
Chromosome ena	CDS	1852	2991.	+	0	ID=C	#ハッケージの読み込み 2011 0110 SOME_1023_1002 #バッケージの読み込み CAAATTCAACCTTCAACCACTAACCAATCATCAAATTTAC
###							Chromosomo 1813 1852
Chromosome ena	gene	3233	3457	2		ID=ge	
Chromosome ena	transcript	3233	3457.	+		ID=tr	# m
Chromosome ena	exon	3233	3457.	+		Pare	
Chromosome ena	CDS	3233	3457.	+	0	ID=C	#マッピングを行うgAli>Chromosome 3420 3459
###							#マッビング結果(at TIGCAGATAATGGGACATTIGTCATTCAAAATGAGTAGGC
Chromosome ena	gene	3467	4588.	+		ID=ge	>Chromosome 3422 3461
	count <- qCo	unt(ou	t, txdb,	rep	ort	Level	=param_reportlevel #7GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA
	dim(count)						#行数と列数を表示 #確認してるだけです >Chromosome_3443_3482
	<						ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC
May 29, 2018							90

マップする側のファイル
^{11番目のリードは、②と③の領域にまた}

##gff-version	3							ン有 QuasR(Gaidatzis_2(015) NEW			
##sequence-	regioi	n Chromo	some (360 221	7785	53			 得までの 一連の			
#!genome-bui	ld Eu	ropean Nuc	leotide	Archi	ve A	SM	8293	ibオブジェクトをネットワーク経由で取 2の詳細については マッピング Leing	得するのを基本 le-endlゲノム			
#!genome−vei	rsion	GCA_00082	29395.1	1								
#!genome-dat	e 201	14-11							•- /"			
#!genome-bui	ld-ac	cession GC	DA 000	829395	5.1			アイルを乳酸菌ケノムにマイン	Chromosome_3	361_400		
#!genebuild-la	st-u	pdated 2014	4-11					STA形式ファイル(<u>sample R</u> 丁 Lactobacillus casei 12Aの」	GACTGATTTAGA	AACACTTTG	GGACACAATT	AAAGAATC
Chromosome	ena	gene	360	1676	. +		ID=g	0829395.1.30.dna.chromoso	Chromosome_1	637_1676		
Chromosome	ena	transcript	360	1676	. +		ID=tr	s jcm 18461.GCA 000829 A	GAAGATGTCCAA	AACCTTAAA,	ATGGAGCTAA	AGCCATAG
Chromosome	ena	exon	360	1676	. +		Pare	>	Chromosome_1	851_1890		
Chromosome	ena	CDS	360	1676	. +	0	ID=C	#入力ファイル名を指定 C	ATGAAATTTACA	ATTAGTOGT	GCAACTTTTA	CAGCCAAA
###								_jcm_18461.GCA_000829	Chromosome_1	843_1882		
Chromosome	ena	gene	1852	2991	+		ID=g	#出力ファイル名を指定│	AACCAATCATGA	AATTTACAA	TTAGTCGTGC	AACTTTTA
Chromosome	ena	transcript	1852	2991	. +		ID=tr	#カウントデータ取得時(>	Chromosome_1	833_1872		
Chromosome	ena	exon	1852	2991	. +		Pare	C	TTCAAGGAGTAA	ACCAATCATG	AAATTTACAA'	TTAGTCGT
Chromosome	ena	CDS	1852	2991	. +		ID=C	#バッケージの読み込み >	Chromosome_1	823_1862		
###	ona		1002	2001				#ハッケーシの読み込み [[AAATICAACCII	CAAGGAGTA	ACCAAICAIG.	AAAIIIAC
Chromosome	ena	gene	3233	3457	2		ID= a		Chromosome_1	813_1852		
Chromosome	ona	transcript	3233	3457				"auto")#txdbオブジェクA	AATTAAAGACAA	ALICAACCI	ICAAGGAGIA	ACCAAICA
Chromosome	ona	avon	3233	3457		·	Dara	#0進認してる/こけで9 >	Uhromosome_a	3418_3457		
Chromosome	ona		3233	2457	. '				ALIGUAGATAAT	IGGGACATIT	JICATICAAA.	ATGAGTAG
###	ena		0200	3437	. '	- 0	ID-C	#マッビンクを行うqAli > #マッビング結果(alignim	Uhromosome_3	3420_3459		
### Olavea wala a a wala			2467	4500	2	_		# 3 O C D D #DVK (arrgu	TGUAGATAATGU	GACATITGI	CALICAAAAD	GAGTAGGU
Chromosome	ena	igene	3407	4000	Y		ID-gi El avial	>	Uhromosome_3	3422_3461		
		dim(count)	ount(ou	τ, τχαρ	, re	por	crevel	=param_reportievel)#/G #行数と列数を表示		ALATITUTUA	IICAAAAIGA	GIAGGUAA
	ł	head(count)						#確認してるだけで	UNITOMOSOME_3	0443_348Z	* * * * * * * * * * * * *	
May 20, 2019		<							TICAAAAIGAGI	AGGUAAUTT	RAAIGATITI	AAAAGAAC
iviay 29, 2010								,				J

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 最低限3リードは. ントの総和は、最低でも3はあるはず。という ##gff-version 3 予想を立てて実際にマッピングを行う。 レ有 | QuasR(Ga ##sequence-region Chromosome 360 2277853 、るマッビングから、 カウントデータ取得までの 一連の #!genome-build European Nucleotide Archive ASM8293(なぜのデ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 | #!genome-version GCA 000829395.1 を⊐ビべ。 #!genome-date 2014-11 アイルを乳酸菌ゲノムにマッ >Chromosome_361_400 #!genome-build-accession GCA 000829395.1 STA形式ファイル(sam<mark>el</mark> 📲 TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC #!genebuild-last-updated 2014-11 Lactobacillus casei 12 Chromosome_1637 1676 ID=ge0829395.1.30.dna.chro 360 1676 |Chromosome|ena lgene jcm 18461.GCA 000829 AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG 1676 ID=+ 3601 |Chromosome|ena +ltranscript. >Chromosome 1851 1890 1676 |Chromosome|ena 360 Pare +lexon CATGAAATTTACAATTAGTCGTGCAACTTTTACAGCCAAA #入力ファイル名を指定 0 ID=C ICDS 360 1676 +|Chromosome|ena jcm 18461.GCA 000829 >Chromosome 1843 1882 ### icm 18461.GCA 000829 TAACCAATCATGAAATTTACAATTAGTCGTGCAACTTTTA #出力ファイル名を指定 ID=ge 1852 2991 + lChromosome∣ena gene #カウントデータ取得時 >Chromosome 1833 1872 2991 1852 ID=ti |Chromosome|ena + transcript CTTCAAGGAGTAACCAATCATGAAATTTACAATTAGTCGT |Chromosome|ena 1852 2991 Pare +exon >Chromosome 1823 1862 #バッケージの読み込み 1852 2991 0 ID=C CDS + |Chromosome|ena #バッケージの読み込み CAAATTCAACCTTCAAGGAGTAACCAATCATGAAATTTAC ### >Chromosome 1813 1852 3233 3457 |Chromosome|ena ID=gd"auto")#txdbオ ブジ 🖕 ク AAATTAAAGACAAATTCAACCTTCAAGGAGTAACCAATCA gene 3233 3457 |Chromosome|ena +ID=tr #確認してるだけで transcript >Chromosome 3418 3457 3457 |Chromosome|ena 3233 Pare +lexon GATTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAG 3233 3457 CDS +0 ID=C Chromosomelena #マッピングを行うqAli >Chromosome 3420 3459 #マッピング結果(align TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC ### 3467 4588 + ID=g Chromosome ena lgene >Chromosome 3422 3461 count <- qCount(out, txdb, reportLevel=param_reportLevel)# GCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGCAA dim(count) #行数と列数を表示 >Chromosome 3443 3482 #確認してるだけです head(count) ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAAGAAC

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-segデータのリファレンスゲノム配列へのBowtiel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 バ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バ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ブ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」ー「ディレクト」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ビベ。 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: 1.サンブルデータ70 <u>mapping single gen</u> のサンプル名」(例:hum 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 す。hg19にマップした結 (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 で、Entrez Gene IDに対 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 in_f1 <- "mapping</pre> ト情報を取得しています。 in f2 <- "BSgenom out f <- "hoge1.t 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) param mapping <in f2 <- "Lactobacillus hokkaidonensis jcm 18461.GCA 000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファー param txdb1 <- "h param txdb2 <- "k in f3 <- "Lactobacillus hokkaidonensis jcm 18461.GCA 000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル param reportlevel 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 #必要なバッケージを library(QuasR) library(GenomicFe #必要なバッケージをロード library(QuasR) #バッケージの読み込み #前処理(マッピング library(GenomicFeatures) #バッケージの読み込み out <- qAlign(in alignmentStats(ou #前処理(アノテーション情報を取得) #前処理(TxDbオブジ txdb <- makeTxDbFromGFF(in f3, format="auto")#txdbオブジェクトの作成 txdb <- makeTxDbF #確認してるだけです txdb txdb #本番(カウントデー #前処理(マッピング) count <- qCount(o #マッビングを行うgAlign関数を実行した結果をoutに格納 out <- qAlign(in f1, in f2)</pre> 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 #行数と列数を表示 dim(count) #確認してるだけです head(count) <

		①例題10をやってみましょう。②「デスクトッ
	い、ぶりに	プ - hoge - mapping_kiso3」フォルダを作成し
<u> </u>	ヒノソ夫仃	、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	ν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。 流れを示します。アノテーシ としつつ、 <u>TxDb</u> パッケージ basic aligner(応用) QuasR 「ファイル」-「ディレク <mark>ト、</mark> の	single-end RNA-seqデータのリファレンスゲノム配列へのBoy ョン情報は、 <u>GenomicFeatures</u> バッケージ中の関数を利用し 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 <u>(Gaidatzis 2015)</u> などを参考にしてください。 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	<u>wtie</u> 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 で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 プションの詳細については <u>マッピング single-end ゲノム </u> 」以下をコビベ。
1. <u>サンブルデータ</u> 70	10. <u>mapping single genome7.txt</u> 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:
mapping single gen A のサンブル名」(例: hum す。hg19にマップした結 で、Entrez Gene IDに対 in_f1 <- "mapping in_f2 <- "BSgenom	マップする側のファイルは、 <u>サンブルデータ</u> 470 et al., Nucleic Acids Res., 2018)から提供されて (Lactobacillus hokkaidonensis jcm 18461.GC アノテーションファイル(Lactobacillus hokkaido ト情報を取得しています。	DFASTA形式ファイル(<u>sample_RNAseq4.fa</u>)です。マップされる側のファイルは、 <u>Ensembl</u> (Zerbino いる <u>Lactobacillus casei 12A</u> の multi-FASTA形式ゲノム配列ファイル A 000829395.1.30.dna.chromosome.Chromosome.fa)です。マッピング結果に対して、GFF3形式の mensis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.gff3)を読み込んでカウン
<pre>out_f <- "hoge1.t param_mapping <- param_txdb1 <- "h param_txdb2 <- "k param_reportlevel</pre>	<pre>in_f1 <- "mapping_single_genome7.t in_f2 <- "Lactobacillus_hokkaidone in_f3 <- "Lactobacillus_hokkaidone out_f <- "hoge10.txt"</pre>	xt" #入力ファイル名を指定してin_f1に格納(RNA-seqファイル) nsis_jcm_18461.GCA_000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファイ nsis_jcm_18461.GCA_000829395.1.30.chromosome.Chromosome.fa" #入力ファイ の R Console
#必要なバッケージを library(QuasR) library(GenomicFe #前処理(マッピング)	param_reportievel <- gene #必要なパッケージをロード library(QuasR)	<pre>> getwd() [1] "C:/Users/kojik/Desktop/hoge/mapping_kiso3" > list.files()</pre>
out <- qAlign(in_ alignmentStats(ou #前処理(TxDbオブジ	library(GenomicFeatures) #前処理(アノテーション情報を取得) txdb <- makeTxDbEromGEE(in f3.	[1] "Lactobacillus_hokkaidonensis_jcm_18461.GA_\$ [2] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_\$
txdb <- makeTxDbF txdb #本番(カウントデー)	txdb #前処理(マッビング)	[4] "sample_RNAseq4.fa" \$
count <- qCount(o	<pre>out <- qAlign(in_f1, in_f2) alignmentStats(out)</pre>	
	<pre>#本番(カウントデータ取得) count <- qCount(out, txdb, reportL dim(count) head(count)</pre>	evel=param_reportlevel)#カウントデータ行列を取得してcountに格納 #行数と列数を表示 #確認してるだけです



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。
RGui (64-bit) ー □ 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 ご言語 記 ② ●	×
R Console C:/Users/kojik/Desktop/hoge/mapping_kiso3\QuasR_\$ Genomic alignments have been created successfully > alignmentStats(out) #マッピ\$ seqlength mapped unmapped Lacto:genome 2277985 11 0 > #本番(カウントデータ取得) > count <- qCount(out, txdb, reportLevel=param_r\$ extracting gene regions from TxDbdone counting alignmentsdone	
<pre>collapsing counts by query namedone > dim(count) #行数と\$ [1] 365 2 1 > head(count) #確認し\$ width Lacto accA 750 0 accB 369 0 <</pre>	

最後まで完了

🙀 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					×
extracting gene regions counting alignmentsde	from TxD one	bdone			^
collapsing counts by que	ery name.	done	#/寻数	ı۶c	
[1] 365 2			#1JØ>	ιCφ	
> head(count)			#確認	3U\$	
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	€UŞ	
> write.table(tmp, out_	E, sep="∖	t", apper	nd=F,	qu	Ş
>					
<					>

最後まで無事完了したら 、こんな感じになります



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。 カウントデータの列名 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-seg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バ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ブ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ベ。 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: 1.サンブルデータ7のFAS mapping single genome 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 のサンブル名」(例:hum 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 す。hg19にマップした結 (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 で、Entrez Gene IDに対 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 in f1 <- "mapping</pre> ト情報を取得しています。 in f2 <- "BSgenom out f <- "hoge1.t (1)カファイル名を指定してin_f1に格納(RNA-seqファイル) in f1 <- "mapping single genome7.txt"</pre> param mapping <-📩 18461.GCA 000829395.1.30.dna.chromosome.Chromosome.fa" #入力ファーヘ in f2 <- "Lactobacillus hokkaidonensis param txdb1 <- "h param txdb2 <- "k in f3 <- "Lactobacillus hokkaidonensis icm 18461.GCA 000829395.1.30.chromosome.Chromosome.gff3"#入力ファイル param reportlevel out f <- "hoge10.txt" #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", "juncti param reportlevel <- "gene" #必要なバッケージを library(QuasR) library(GenomicFe #必要なバッケージをロード library(QuasR) #バッケージの読み込み #前処理(マッピング マケージの読み込み out <- qAlig FileName SampleName alignmentSt #前処理(TxDb o")#txdbオブジェクトの作成 txdb <- make sample_RNAseq4.fa Lacto 認してるだけです txdb #本番(カウントデー #前処理(マッビング) count <- qCount(o #マッビングを行うgAlign関数を実行した結果をoutに格納 out <- qAlign(in f1, in f2) 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 #行数と列数を表示 dim(count) #確認してるだけです head(count)

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

	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

あし、トギーク
<u> </u>
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
<pre>[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_2 Aseq4_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 > </pre>

①hoge10.txtの中身は、② countオブジェクトと同じです

	width	Lacto
accA	750	0
ассВ	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 RGui (64-bit) × \Box パッケージ ウインドウ ヘルプ ファイル その他 編集 閲覧 🛎 🖆 🖬 🖻 🔁 😔 🥌 x R Console "Lactobacillus hokkaidonensis jcm 18461.GCA\$ [6] "mapping single genome7.txt" [7] "QuasR log 3b6c12e745f9.txt" \$ [8] "sample RNAseq4.fa" \$ [9] ֆ Տ "sample RNAseq4 3b6c652a602a.bam" [10] [11] "sample RNAseq4 3b6c652a602a.bam.bai" [12] "sample RNAseq4 3b6c652a602a.bam.txt" > head(count) width Lacto 750 accA 0 369 accB 0 accC 1347 0 accD 789 0 ackA 1191 0 363 acpS 0 > dim(count) [1] 365 2 > ٠

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…

		width	Lacto
1	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	<mark>660</mark>	0
	alaS	2655	0
	aldA	1464	0
	aldB	714	0
	amt	1323	0

<u>1行目のカウントが0</u>

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-	regior	n Chromo:	some 3	360 22	77	850	3	
#!genome-bui	ld Eui	ropean Nuc	leotide	Archi	ve	AS	SM	82939
#!genome-ver	rsion	GCA_00082	29395.1					
#!genome-dat	e 201	4-11						
#!genome-bui	ld–ac	cession GC	A 000	82939!	5.1			
#!genebuild-la	ist-up	dated 2014	4-11					
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



geneレヘルのカワントナ-
ファイル 編集 閲覧 その他 バッケージ ウインドウ ヘルブ
R Console
<pre>[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" \$ [13] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [14] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [15] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [16] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [17] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [18] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [19] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [10] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [10] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [11] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [12] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [12] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [12] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [12] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [13] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [14] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [15] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [15] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [16] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [17] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [18] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [19] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [19] "sample_RNAseq4_3b6c652a602a.bam.txt" \$ [10] "sam</pre>
[1] 365 2
< >

geneレベルのカウントデータを そているので、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。

	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ファイルをよく見る

					-															
##gff-version	3																			
##sequence-	regio	n Chromo	some	360 2	277	85	3													
#!genome-bu	ild Eı	uropean Nu	ucleoti	de Arc	hiv	еA	SM8293	9v1												
#!genome-ver	rsion	GCA_0008	329395	5.1																
#!genome-da	te 20	14-11																		
#!genome-bu	ild-a	ccession G	CA_00	08293	395.	1														
#!genebuild-la	ast-u	pdated 20	14-11																	
Chromosome	ena	gene	360	1676		(1)	ID=gen	e:LOOC2	260_100	010;Nan	ne=dna <i>l</i>	\;biotype	=proteir	n_codin	g;d					
Chromosome	ena	transcript	360	1676		- / -	ID=tran	script:B	AP8458	1;Parent	=gene:L	.00C260	_100010);Name⊧	=dı					
Chromosome	ena	exon	360	1676			Parent=	transcri	pt:BAP8	84581;Na	ame=BA	P84581-	1;consti	tutive=1	1;e					
Chromosome	ena	CDS	360	1676		- 0	ID=CDS	S:BAP84	581;Par	ent=trar	nscript:B	AP84583	l;protein	_id=BA	P8					
###																				
Chromosome	ena	gene	1852	2991		(1)	ID=gen	e:LOOC2	260_100	020;Nan	ne=dnaN	N;biotype	e=proteii	n_codin	g;d					
Chromosome	ena	transcript	1852	2991		- /-	ID=tran	script:B	AP84582	2;Parent	=gene:L	.00C260	_100020);Name⊧	=dı					
Chromosome	ena	exon	1852	2991			Parent=	transcri=	pt:BAP8	84582;Na	ame=BA	P84582-	1;consti	tutive=1	1;e					
Chromosome	ena	CDS	1852	2991		- 0	ID=CDS	S:BAP84	582;Pare	ent=trar	nscript:B	AP84582	2;protein	_id=BA	P8					
###																				
Chromosome	ena	gene	3233	3457		(1)	ID=gen	e:LOOC2	260_100	030;biot	ype=pro	tein_coo	ling;desc	cription=	=S4					
Chromosome	ena	transcript	3233	3457		Ξ.	ID=tran	script:B	AP8458	3;Parent	=gene:L	.00C260	_100030);biotyp	e=					
Chromosome	ena	exon	3233	3457	. +		Parent=	transcri	pt:BAP8	34583;Na	ame=BA	P84583-	1;consti	tutive=:	1;e					
Chromosome	ena	CDS	3233	3457	. +	- 0	ID=CDS	S:BAP84	583;Par	ent=trar	nscript:B	AP84583	3;protein	_id=BA	P8					
###																				
Chromosome	ena	gene	3467	4588		(1)	ID=gen	e:LOOC2	260_100	040;Nan	ne=recF	;biotype	=protein	_coding	;;de					
						7														
								17	赤枠内]のN	lame	e=の	右	側の	文 ⁻	字がg	gen	ename	ው መ	ようで
----------------	------------	------------	---------	----------	------	--------------	-------	------------------	----------------------	-------	------	----------	------------------	--------	----------------	--------	----------	---------------------	--------------	------------------
GF	— -	フマノ	7 川	、ち		\mathbf{F}	2	すれ	a. 2	ະຫຼ	遺伝	云子令	<mark>湏</mark> 垣	支にに	tNa	ame=	ge	nenam	eが	ない
	2		/	<u> </u>	. C	ト		こと	もわた	いる。	、これ	nst	が原	うして	<u> </u>	200行	一起	にはな	167	ずに
##gtt-version	3			000.0	077			<mark>365</mark>	行とな	いて		まった	<u>=0</u> ,	かも	••••	と考え	える	っそし	て、	$(2)\mathcal{O}$
##sequence-	regio	n Chromo	some	360 2	211	85	3	領北	<mark>或[323</mark>	3. 34	457	内に	[は	、 確:	かり	少なく	[と=	も1リー	-ド 字	呈壁に
#!genome-bu	ild Ei	uropean Nu	icleoti	de Arc	chiv	e A	SM82	7	, , , + +	n + -	トンフ	+-	4	い ふろう	て正	ゴラク	 ∖ ≣=	しま <i>、</i> 1倍た+	شا ب	Ζ.
#!genome-vei	rsion	GCA_0008	329395	5.1				X D	// 01	し/この	トノ	<u>с</u>		·))	25			記名	<u> </u>	ຈັ
#!genome-da	te 20	14-11																		
#!genome-bu	ild-a	ccession G	CA_00	08293	395	.1														
#!genebuild-la	ast-u	pdated 20	14-11																	
Chromosome	ena	gene	360	1676		1	ID=ge	ene	:LOOC	260_	1000	010 N	lam	ne=dn	ıаА	bioty	pe=	=protei	n_co	ding;d
Chromosome	ena	transcript	360	1676	. •	ŦŢ.	ID=tr	ans	cript:B	AP84	4581	.;Pare	ent∘	=gene	e:L(00C2	60_	10001	D;Na	me=di
Chromosome	ena	exon	360	1676		+ .	Parer	nt=t	ranscr	ipt:B	AP8	4581	;Na	me=l	BAF	P8458	31-1	;const	tuti	/e=1;e
Chromosome	ena	CDS	360	1676		+ 0	ID=C	DS:	BAP84	581;	Pare	ent=t	ran	script	:BA	\P845	581;	proteir	ı_id=	BAP8=
###																				
Chromosome	ena	gene	1852	2991		1	ID=ge	ene	:L00C	260_	1000	020 N	lam	ne=dr	ιaΝ	;bioty	pe=	=protei	n_co	oding;d
Chromosome	ena	transcript	1852	2991		Ŧ .	ID=tr	ans	cript:B	AP84	4582	Pare:	ent	=gene	e:L(00C2	60_	10002	0;Na	me=di
Chromosome	ena	exon	1852	2991		+ .	Parer	nt=t	ranscr	ipt:B	AP8	4582	;Na	me=l	BAF	P8458	82-1	;const	tuti	/e=1;e
Chromosome	ena	CDS	1852	2991		+ 0	ID=C	DS:	BAP84	582;	Pare	ent=t	ran	script	:B/	\P845	82;	proteir;	ı_id=	BAP8=
###																				
Chromosome	ena	gene	3233	3457		2	ID=ge	ene	:LOOC	260_	1000	030;b	ioty	/pe=p	prot	tein_c	odi	ng;des	cript	ion=S4
Chromosome	ena	transcript	3233	3457		+7.	ID=tr	ans	cript:B	AP84	4583	;Pare	ent	=gene	e:L(00C2	60_	10003);bio	type=
Chromosome	ena	exon	3233	3457		+ .	Parer	nt=t	ranscr	ipt:B	AP8	4583	;Na	me=l	BAF	P8458	3-1	;const	tuti	/e=1;e
Chromosome	ena	CDS	3233	3457		+ 0	ID=C	DS:	BAP84	583;	Pare	ent=t	ran	script	:BA	AP845	583;	proteir	_id=	BAP8
###																				
Chromosome	ena	gene	3467	4588		1	ID=ge	ene	:L00C	260_	1000	040 N	lam	ie=re	cF;	biotyp	be=	proteir	_co	ding;de
						7														

Γ

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。

dnaAのカウント数

							-		
##gff-version	3								
##sequence-	regio	n Chromo	some	360 2	27	78	53	3	
#!genome-bu	ild Eu	uropean Nu	icleoti	de Arc	chi	ve	A	\SM82939v1	
#!genome-vei	rsion	GCA_0008	329395	5.1					
#!genome-da	te 20	14-11							
#!genome-build-accession GCA_000829395.1									
#!genebuild-l	ast-u	pdated 20	14-11						
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;c	
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								Τ
##sequence-	regio	n Chromo	some	360 22	27	78	53	3	
#!genome-bu	ild Ει	uropean Nu	icleoti	de Arc	chi	ve	AS	ASM82939v1	
#!genome-vei	rsion	GCA_0008	29395	5.1					
#!genome-da	te 20	14-11							
#!genome-bu	ild-a	ccession G	CA_00	08293	395	5.1			
#!genebuild-l	ast-u	pdated 20	14-11						
Chromosome	ena	gene	360	1676		+		ID=gene:LOOC260_100010;Name= <u>dnaA</u> ;biotype=protein_coding;	d
Chromosome	ena	transcript	360	1676		+		ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=	dı
Chromosome	ena	exon	360	1676		+		Parent=transcript:BAP84581;Name=BAP84581-1;constitutive=1;	е
Chromosome	ena	CDS	360	1676		+	0	DID=CDS:BAP84581;Parent=transcript:BAP84581;protein_id=BAP	8
###									
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=	dı
Chromosome	ena	exon	1852	2991		+			
Chromosome	ena	CDS	1852	2991		+	0	D C > count ["dnaA",] (2)	
###								width Lacto	
Chromosome	ena	gene	3233	3457		+		IC 1317 2	
Chromosome	ena	transcript	3233	3457		+		> 1676 - 360 + 1	
Chromosome	ena	exon	3233	3457		+		P; [1] 1317	
Chromosome	ena	CDS	3233	3457		+	0		~
###								< > >	.:
Chromosome	ena	gene	3467	4588		+		ID=gene:LOOC260_100040;Name=recF;biotype=protein_coding;c	Je

- -									少なく	ともな	<u>2リート</u>	[、] は完	全	に領地	或内	ミマ	<mark>ップ</mark>	され
dpc	ν	M +		` ,	L	米	έh		ている	はす	゛の、	1)dna	A Ø)カウ	ント	数を	<mark>調べ</mark>	る。
SID	1A	UJJJ	・ノ			す	议	, (2 カウ	ント	<u>数は2</u>	です	ね。	ちな	みに	2、3	<mark>1317</mark>	は、
##gff-version	3								dnaA 🗸) 領	域の長	長さで	す。					
##sequence-	regio	n Chromo	some	360 2	277	85	3								-			
#!genome-bu	ild Eı	uropean Nu	icleoti	de Arc	hiv	e A	SN	182939v1										
#!genome-ver	rsion	GCA_0008	29395	5.1														
#!genome-da	te 20)14-11																
#!genome-bu	ild-a	ccession G	CA_00	08293	395	.1												
#!genebuild-la	ast-i	pdated 20	14-11															
Chromosome	ena	gene	360	1676		٢.	ID	=gene:LOO	C260_1	000	10;Nar	ne= <u>dr</u>	naA;	biotyp	e=p	oroteir		ding;d
Chromosome	ena	transcript	360	1676		⊦.	ID	=transcript:	BAP84	581;	Parent	=gen	e:LC	DOC26	0_1	00010);Nar	ne=dı
Chromosome	ena	exon	360	1676		⊦.	Pa	arent=transc	ript:BA	P84	581;Na	ame=	BAP	84581	l-1;o	consti	tutiv	e=1;e
Chromosome	ena	CDS	360	1676		+ () ID	=CDS:BAP8	4581;P	arer	nt=trar	nscrip	t:BA	P8458	31;p	rotein	_id=	BAP8
###																		
Chromosome	ena	gene	1852	2991		⊦.	ID	=gene:LOO	C260_1	0002	20;Nar	ne=dr	naN	;biotyp	oe=p	orotei	n_co	ding;d
Chromosome	ena	transcript	1852	2991		⊦.	ID	=transcript:	RAP84	582:	Parent	=gen	e:I (00026	i <mark>0 1</mark>	0002):Nar	ne <u>=d</u> i
Chromosome	ena	exon	1852	2991		⊦.	Pa	🥨 R Console								l		
Chromosome	ena	CDS	1852	2991		+ () ID	> count["	'dnaA"	,] 🕻	2							^
###								width Lac	to									
Chromosome	ena	gene	3233	3457		⊦.	3	1317	2		1							
Chromosome	ena	transcript	3233	3457		⊦.	Π	> 1676 -	360 +	1								
Chromosome	ena	exon	3233	3457		⊦.	Pa	[1] 1317										- 1
Chromosome	ena	CDS	3233	3457		+ () ID	>										~
###								<										> .:
Chromosome	ena	gene	3467	4588		+.	ID	=gene:LOO	C260_1	0004	40;Nar	ne=re	cF;ł	biotype	e=p	rotein	_cod	ing;de

カウント総数の確認
(R 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
> 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
> sum(count[,2])
· · · · · · · · · · · · · · · · · · ·
<u>ار خ</u>

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

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。

?qCount

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によるマッビングから、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ブ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ビベ。





?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 イ	qCount {QuasR} R Documentation
R Console	Quantify alignments
<pre>> dim(count) [1] 365 2 > head(count) width Lacto</pre>	Quantify alignments from sequencing data.
accA 750 0 accB 369 0	Usage
<pre>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 0 > sum(count[,2]) [1] 2 > ?qCount[]</pre>	<pre>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, includeSpliced=TRUE, includeSecondary=TRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeedTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, mapqMin=0L, MarcelleeEdTRUE, marcelleeE</pre>
	absIsizeMin=NULL, absIsizeMax=NULL, maxInsertSize=500L, clObj=NULL)



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 orientation る。しかし、おそらくoppositeを指定したら(orientation = fopposite")、カウントの総和は0になるだろう。未確認 qCount {QuasR} Quantify alignments Description orientation sets the required orientation of the alignments relative to the query region in order to be counted, one of: Quantify alignments from sequencing dat any (default): count alignment on the same and Usage opposite strand qCount (proj, same : count only alignment on the same strand query, reportLevel=c(NULL, "gene" selectReadPosition=c("start • opposite : count only alignment on the opposite strand 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

qCount {QuasR}

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。

Quantify alignment_

```
Description
Quantify alignments from sequencing data.
Usage
gCount (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,
       mapgMax=255L,
       absIsizeMin=NULL,
       absIsizeMax=NULL,
       maxInsertSize=500L,
       clObj=NULL)
```





<u> 1塩基のずれのみでカウントされるかもしれないの いしずれたリード。②の遺伝子領域が対応します。 </u>

						-			
##gff-version	3								
##sequence-	regio	n Chromo	some	360 2	27	78	53		
#!genome-bu	ild Ei	uropean Nu	ucleoti	de Arc	chi	ve	AS	SM82939v1	
#!genome-vei	rsion	GCA_0008	329395	5.1					
#!genome-da	te 20)14-11							
#!genome-build-accession GCA_000829395.1									>Chromosome_361_400
#!genebuild-l	ast-i	pdated 20	14-11						TGACTGATTTAGAAACACTTTGGGACACAATTAAAGAATC
Chromosome	ena	gene	360	1676		+		ID=gene:LOOC260_1	<pre>Q>Chromosome_1637_1676</pre>
Chromosome	ena	transcript	360	1676		+		ID=transcript:BAP845	<u>AGAAGATGTCCAAAACCTTAAAATGGAGCTAAAGCCATAG</u>
Chromosome	ena	exon	360	1676		+		Parent=transcript:	;>Chromosome_1851_1890
Chromosome	ena	CDS	360	1676		+	0	ID=CDS:BAP84581	
###]>Uhromosome_1843_1882
Chromosome	ena	gene	1852	2991		2)-		ID=gene:LOOC260_1	
Chromosome	ena	transcript	1852	2991		+		ID=transcript:BAP84	
Chromosome	ena	exon	1852	2991		+		Parent=transcript:BA	Chromocomo 1922 1962
Chromosome	ena	CDS	1852	2991		+	0	ID=CDS:BAP84582;P	ηνοπη οποδοπιθ_το23_του2 Θελλαττελλεεττελλεελλεελατελτελλαττιλε
###									Character addition add and a constrained and an antitac
Chromosome	ena	gene	3233	3457		+		ID=gene:LOOC260_1	
Chromosome	ena	transcript	3233	3457		+		ID=transcript:BAP845	\sim Chromosome 3418 3457
Chromosome	ena	exon	3233	3457		+		Parent=transcript:BA	
Chromosome	ena	CDS	3233	3457		+	0	ID=CDS:BAP84583;P	4>Chromosome 3420 3459
###									TTGCAGATAATGGGACATTTGTCATTCAAAATGAGTAGGC
Chromosome	ena	gene	3467	4588		+		ID=gene:LOOC260_1	Q>Chromosome 3422 3461
]gcagataatgggacatTtgtcattcaaaatgagtaggcaa
									>Chromosome 3443 3482

ATTCAAAATGAGTAGGCAACTTAAATGATTTTAAAAGAAC

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 遣 伝 子 タ は dna N

少し		げれ7	1-1	ノー	_				は、 (3)そ	<mark>1の</mark> の遺	リー 伝	<mark>-ド。②</mark> 子名は	の遺 dnal	伝- N。	子領	域な	が対応	iL	ます。	
##gff-version	3																			<u> </u>
##sequence-	regio	n Chromo	some	360 2	27	78	53	}												
#!genome-bu	ild Ei	uropean Ni	icleoti	de Arc	hi	ve	AS	SM82939)v1											
#!genome-ver	rsion	GCA_0008	329395	5.1																
#!genome-da	te 20)14-11																		
#!genome-bu	ild-a	ccession G	CA_00	08293	395	5.1														
#!genebuild-l	ast-i	updated 20	14-11																	
Chromosome	ena	gene	360	1676		+		ID=gene	e:LOOC	260_1	100	010;Nai	me=d	InaA	\;biot	ype	=prote	in_	_coding	g;d
Chromosome	ena	transcript	360	1676		+		ID=tran	script:B	AP84	581	l;Paren	t=gei	ne:L	000	260	_10003	10;	Name-	=dı
Chromosome	ena	exon	360	1676		+		Parent=	transcr	ipt:B/	AP8	4581;N	ame=	=BA	P845	581-	1;cons	titı	utive=1	1;e
Chromosome	ena	CDS	360	1676		+	0	ID=CDS	:BAP84	581;F	Pare	ent=tra	nscri	ot B	AP84	1581	;prote	in_	id=BA	P8
###														3						
Chromosome	ena	gene	1852	2991				ID=gene	e:LOOC	260_1	100	020;Nai	me=c	lnaN	l;bio	type	=prote	in_	_codin	g;d
Chromosome	ena	transcript	1852	2991		+	•	ID=tran	script:B	AP84	582	2;Paren	t=gei	ne:L	000	260	_10002	20;	Name-	=dı
Chromosome	ena	exon	1852	2991		+	•	Parent=	transcr	ipt:B/	AP8	4582;N	ame=	=BA	P845	82-	1;cons	titı	utive=1	l;e
Chromosome	ena	CDS	1852	2991		+	0	ID=CDS	:BAP84	582;F	Pare	ent=tra	nscri	ot:B	AP84	1582	;prote	in_	id=BA	P8
###																				
Chromosome	ena	gene	3233	3457		+		ID=gene	e:LOOC	260_1	100	030;bio	type=	=pro	tein_	_cod	ing;de	scr	iption=	=S4
Chromosome	ena	transcript	3233	3457		+		ID=tran	script:B	AP84	583	3;Paren	t=gei	ne:L	.000	260	_10003	30;	biotyp	e=
Chromosome	ena	exon	3233	3457		+		Parent=	transcr	ipt:B/	AP8	4583;N	ame=	=BA	P845	583-	1;cons	titı	utive=1	l;e
Chromosome	ena	CDS	3233	3457		+	0	ID=CDS	:BAP84	583;F	Pare	ent=tra	nscrij	ot:B	AP84	1583	;prote	in_	id=BA	P8
###																				
Chromosome	ena	gene	3467	4588		+		ID=gene	e:LOOC	260_1	100	040;Nai	me=r	ecF	;biot	ype=	=protei	in_	coding	;;de



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Bowtiel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、<u>GenomicFeatures</u>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」ー「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ビベ。



N 199		①例題11をやってみましょう。qCount関数実行時のオコ	プシ
	11	ョンとして、②shift = 1となるように、③で指定しています	;
的起		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バッケージを用いたs 流れを示します。アノテーシ としつつ、 <u>TxDb</u> パッケージ ³ basic aligner(応用) QuasR	ingle-end RNA-seqデータのリファレンスゲノム配列へのBov ョン情報は、 <u>GenomicFeatures</u> バッケージ中の関数を利用し を読み込むやり方も示しています。 マッピングの <i>や</i> り方やオ (Gaidatzis 2015)などを参考にしてください。	<u>vtie</u> lこよるマッピングから、カウントデータ取得までの <i>一</i> 連の で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 ブションの詳細については <u>マッピング single-end ゲノム </u>	
「ファイル」ー「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	□ RGui (64-bit) — □	×
1. <u>サンブルデータ</u> 76 SI	の形式ファイル(SRR037430 fasto)の RSgenome Hsaniens		<u> </u>
mapping single g のサンプル名」(例 m m す hg19にマップレンジ	1. <u>mapping single genome7.txt</u> 中のFASTAだ 例題10と基本的に同じですが、カウント情報をl		
で、Entrez Gene IDに対	<pre>in_f1 <- "mapping_single_genome7.t</pre>		
in_f2 <- "BSgenom	in_f3 <- "Lactobacillus_hokkaidore		^
param_mapping <-	<pre>out_f <- "hoge11.txt"</pre>	<pre>> getWd() [1] "Ge(Userng(beright)) / Degister (begg(menning)) / ige2"</pre>	
param_txdb1 <- 2	param_reportievel <- gene param_shift <- 1	<pre>> list files()</pre>	
param_reportlev		[1] "hoge10.txt"	s
#必要なバッケージを library(QuasR)	#必要なハックニンをロート library(QuasR)	[2] "Lactobacillus hokkaidonensis jcm 18461.GCA	ŝ
library(ĞenomicFe	library(GenomicFeatures)	[3] "Lactobacillus hokkaidonensis jcm 18461.GCA	ŝ
#前処理(マッピング)	#前処理(アノテーション情報を取得)	[4] "Lactobacillus hokkaidonensis jcm 18461.GCA	Ş
alignmentStats(ou	txdb <- makeTxDbFromGFF(in_f3, for	[5] "Lactobacillus_hokkaidonensis_jcm_18461.GCA	Ş
#前処理(TxDbオブジ	txdb	[6] "Lactobacillus_hokkaidonensis_jcm_18461.GCA_	Ş
txdb <- makeTxDbF	#前処理(マッピング)	<pre>[7] "mapping_single_genome7.txt"</pre>	Ş
	<pre>out <- qAlign(in_f1, in_f2)</pre>	<pre>[8] "QuasR_log_3b6c12e745f9.txt"</pre>	Ş
#本畨(カワントテー) count <- qCount(o	alignmentStats(out)	[9] "sample_RNAseq4.fa"	Ş
<	#本番(カウントデータ取得)	[10] "sample_RNAseq4_3b6c652a602a.bam"	Ş
	<pre>count <- qCount(out, txdb, reportL</pre>	[11] "sample_RNAseq4_3b6c652a602a.bam.bai"	Ş
	shift=param_shift)	[12] "sample_RNAseq4_3b6C652a602a.bam.txt"	ې ۲
L	3		~
		<	>

N 199			①例題11のコピペ実行後は
	ペ宝行後		、②のような感じになります。
マップ後 カウン マップ後 カウン Quas N/ッケージを用いたs 流れを示します。アノテーシ としつつ、TxDb//ッケージ basic aligner(応用) Quas R 「ファイル」ー「ディレクトリの 1.サンプルデータ700 ST mapping single g 10 m す。hg191こマップしよ結 で、Entrez Gene IDIこ対 in_f1 <- "mapping in_f2 <- "BSgenom out_f <- "hoge1.t param_mapping <- param_txdb1 <- "h param_txdb1 <- "h param_txdb2 <- "k param_txdb2 <- "k param_txdb1 <- "h param_txdb2 <- "k param_txdb2 <- "k param_txdb1 <- "h param_txdb2 <- "k param_txdb2 <- "k param_	への実に行後の 小情報取得 single-end ゲノム アノテー single-end RNA-seqデータのリファレンスゲノム配列へのBov (a) 「情報は、GenomicFeatures バッケージ中の関数を利用し 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 (Gaidatzis 2015)などを参考にしてください。 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 (DE式ファイルを置いてあるディレクトリに移動) co形式ファイルを置いてあるディレクトリに移動) co形式ファイル(SER037430 fastorの BSgenome Hsaniens 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1 in_f1 <- "mapping_single_genome7.tx in_f2 <- "Lactobacillus_hokkaidon" in_f3 <- "Lactobacillus_hokkaidon" in_f3 <- "Lactobacillus_hokkaidon" in_f3 <- "Lactobacillus_hokkaidon" 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, form 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, reportled out <- qCount(out, txdb, reportled)	ション有 QuasR(Gaidatzis_2015) NEW mielctaマッピングから、カウントデータ取得までの一連の 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 プションの詳細についてはマッピング single-end ゲノム] RGui (64-bit) アイル 編集 閲覧 その他 パッケージ ウイント アイルに保存 > #ファイルに保存 > tmp <- cbind (rownames (come) Numite table (tmp out for the second secon	<pre>(1)例題11のコピペ実行後は 、②のような感じになります。</pre> - □ × ジ ヘルブ Vignettes finished working with a \$ #行数と\$ #確認し\$
	<	> <	

- -			全365遺伝子の領域に	マップされた
╶╶┼┐┍┶┑	いな参うる	· [刃	リードの総数は、①3に	なりました!
ノリーノ	✓ 「 小心 女人 ∨ノ 1 圧 小情報取得 single-end ゲノム アノテー	. 口心 ション有 QuasR(Gaidatzis 201)	5) NEW	E.
QuasR バッケージを用いた 流れを示します。アノテーシ としつつ、 <u>TxDb</u> パッケージ basic aligner(応用) QuasR [ファイル」ー「ディレクトリの	single-end RNA-seqデータのリファレンスゲノム配列へのBoy /ョン情報は、 <u>GenomicFeatures</u> パッケージ中の関数を利用し 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 (<u>Gaidatzis 2015)</u> などを参考にしてください。 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	v <u>tie</u> によるマッピングから、カウントデータ取得 てTxDbオブジェクトをネットワーク経由で取得 プションの詳細については <u>マッピング single -</u>	- までの 一連の するのを基本 end ゲノム	
1. <u>サンブルデータ</u> 7のFAS	CO形式ファイル(SRR037430 fasto)の BSgenome Hsaniens	RGui (64-bit)		- 🗆 ×
<u>mapping single genome</u> のサンブル名J(例:hum す。hg19にマップした結	11. <u>mapping single genome7.txt</u> 中のFASTA形 例題10と基本的に同じですが、カウント情報を1		ィージ ウインドウ ヘルフ Vignettes	
<pre>9 . hgl9/Cマックし/2福 で、Entrez Gene IDIこ対 in_f1 <- "mapping in_f2 <- "BSgenom out_f <- "hoge1.t param_mapping <- param_txdb1 <- "h param_txdb2 <- "k 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</pre>	<pre>in_f1 <- "mapping_single_genome7.to in_f2 <- "Lactobacillus_hokkaidoner in_f3 <- "Lactobacillus_hokkaidoner 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)</pre>	<pre> R Console [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 > </pre>		■■× #確認し\$
#本番(カウントデー・ count <- qCount(o	out <- qAlign(in_f1, in_f2) alignmentStats(out) #本番(カウントデータ取得) count <- qCount(out, txdb, reportLo shift=param_shift)	<pre>> #ファイルに保存 > tmp <- cbind(rown > write.table(tmp > sum(count[,2]) [1] 3 > </pre>	ames(count), count) out_f, sep="\t", appe	#保存し\$ nd=F, quo\$ ♪

Г

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 OuasRバッケージを用いたsingle-end RNA-seqデータのリファレンスゲノム配列へのBowtieによるマッビングから、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 バ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バ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ブ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」ー「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」 RGui (64-bit) \times 1.サンブルデータ7のFASTO形式ファイル(SRR037439 fasto)の RSgenome Heapien mapping single genome 11. mapping single genome 7. txt中のFASTA形 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 のサンブル名」(例:hum 例題10と基本的に同じですが、 カウント 情報を 🖆 💾 🖪 🔁 🚭 👄 9 す。hg19にマップした結 で、Entrez Gene IDに対 in f1 <- "mapping single genome7.tx - - X R Console in f1 <- "mapping in f2 <- "Lactobacillus hokkaidone in f2 <- "BSgenom in f3 <- "Lactobacillus hokkaidone out f <- "hoge1.t 750 accA 0 out f <- "hoge11.txt"</pre> param mapping <-369 param txdb1 <- "h param reportlevel <- "gene" accB 0 param_txdb2 <- "k param shift <- 1 accC 1347 0 param reportlevel 789 accD 0 #必要なバッケージをロード #必要なバッケージを ackA 1191 0 library(QuasR) library(QuasR) library(GenomicFe library(GenomicFeatures) 363 0 acpS #前処理(マッピング > #前処理(アノテーション情報を取得) out <- qAlign(in > #ファイルに保存 txdb <- makeTxDbFromGFF(in f3, form alignmentStats(ou txdb > tmp <- cbind(rownames(count), count) #保存U\$ #前処理(TxDbオブジ > write.table(tmp, out f, sep="\t", append=F, quo\$ txdb <- makeTxDbF #前処理(マッピング) txdb > sum(count[,2])out <- qAlign(in f1, in f2)</pre> #本番(カウントデー [1] 3 alignmentStats(out) count <- qCount(o > count["dnaN",] #本番(カウントデータ取得) width Lacto count <- qCount(out, txdb, reportL</pre> 1140 shift=param shift) > | < >

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

残る	問題

##gff-version	3						
##sequence-	regio	n Chromo	some	360 22	2778	353	3
#!genome-bu	ild Eı	uropean Nu	icleoti	de Arc	hive	A	SM82939v1
#!genome-ver	rsion	GCA_0008	329395	5.1			
#!genome-da	te 20	14-11					
#!genome-bu	ild-a	ccession G	CA_00	08293	395.	1	
#!genebuild-la	ast-i	pdated 20	14-11			V	
Chromosome	ena	gene	360	1676	(1	ID=gene:LOOC260_100010Name=dnaAbiotype=protein_coding;d
Chromosome	ena	transcript	360	1676	. +	•	ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=dr
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_100020Name=dnaN;biotype=protein_coding;d
Chromosome	ena	transcript	1852	2991	. +		ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=dr
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
###						V	
Chromosome	ena	gene	3233	3457		2	ID=gene:LOOC260_100030;biotype=protein_coding;description=S4
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
###						V	
Chromosome	ena	gene	3467	4588	. -	1)	ID=gene:LOOC260_100040Name=recF;biotype=protein_coding;de
						7	

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

「「「「」」	ζ F	旧町										닅	á(こ3	<mark>65個</mark>	国だ	った	<u>-</u> の7	かを	検討	Eする	0
7天~	ון כ	可迟																			
##gff-version	3																				
##sequence-	regio	n Chromo	some	360 2	27	78	53	3													
#!genome-bu	ild Ei	uropean Nu	ucleoti	de Arc	chiv	/e	AS	SM82939)v1												
#!genome-vei	rsion	GCA_0008	329395	5.1																	
#!genome-da	te 20)14-11																			
#!genome-bu	ild-a	ccession G	CA 00	08293	395	5.1															
#!genebuild-la	ast-i	pdated 20	14-11																		
Chromosome	ena	gene	360	1676		-(ID=gene	e:LOOC2	260	100	010	Nam	ne=d	InaA	bio	type	=pr	oteir	o codi	ng;d
Chromosome	ena	transcript	360	1676		+		ID=tran	script:B	AP8	4581	1;Pa	arent	=ger	ne:L	000	260	10	0010	;Name	e=dı
Chromosome	ena	exon	360	1676		+		Parent=	transcr	ipt:B	AP8	8458	31;Na	me=	=BA	P84	581-	1;co	onsti	tutive=	=1;e
Chromosome	ena	CDS	360	1676		+	0	ID=CDS	:BAP84	581;	Pare	ent=	=tran	scrip	ot:B	AP8	4581	L;pro	otein	id=B	AP8
###																		-			
Chromosome	ena	gene	1852	2991		-(ID=gene	e:LOOC2	260	100	020	Nam	ne=d	naN	;bio	type	=pr	oteir	n_codi	ng;d
Chromosome	ena	transcript	1852	2991		+		ID=tran	script:B	AP8	4582	2;Pa	rent	=ger	ne:L	000	260	10	0020	;Name	e=dı
Chromosome	ena	exon	1852	2991		+		Parent=	transcr	ipt:B	AP8	8458	32;Na	me=	=BA	P84	582-	1;co	onsti	tutive=	=1;e
Chromosome	ena	CDS	1852	2991		+	0	ID=CDS	:BAP84	582;	Pare	ent=	=tran	scrip	ot:B	AP8	4582	2;pro	otein	_id=B	AP8
###																					
Chromosome	ena	gene	3233	3457		+		ID=gene	e:LOOC2	260_	100	030	;biot	ype=	pro	tein	_coc	ling	desc	ription	n=S4
Chromosome	ena	transcript	3233	3457		+		ID=tran	script:B	AP8	4583	3;Pa	arent	=ger	ie:L	000	260	_10	0030	;bioty	pe=
Chromosome	ena	exon	3233	3457		+		Parent=	transcr	ipt:B	AP8	8458	33;Na	me=	=BA	P84	583-	1;co	onsti	tutive=	=1;e
Chromosome	ena	CDS	3233	3457		+	0	ID=CDS	:BAP84	583;	Pare	ent=	=tran	scrip	ot:B	AP8	4583	3;pro	otein	id=B	AP8
###																					
Chromosome	ena	gene	3467	4588		-(ID=gene	e:LOOC2	260_	100	040	Nam	ne=r	ecF	biot	ype:	=pro	otein	_codir	ng;de
	-							-													

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

Bio-Linuxのターミナル画面



Tips: ワイルドカード iu@bielinux[~/Desktop/mac_share/mapping_kiso3]

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。



inux[~/Desktop/mac_share/mapping_kiso3]	、タフ補完で直打な
<pre>iu@bielinux[mapping kiso3] pwd</pre>	[3:19十夜]
<pre>/home/iu/Desktop/mac share/mapping kiso3</pre>	
<pre>iu@bielinux[mapping kiso3] ls</pre>	[3:19午後]
Lactobacillus hokkaidonensis jcm 18461.GCA 00	0829395.1.30.chron
osome.Chromosome.gff3	
<pre>iu@bielinux[mapping kiso3] ls *.gff3</pre>	[3:19午後]
Lactobacillus hokkaidonensis jcm 18461.GCA 00	0829395.1.30.chron
osome.Chromosome.gff3	
<pre>iu@bielinux[mapping kiso3]</pre>	[3:27午後]

(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

残る	問題

##gff-version	3						
##sequence-	regio	n Chromo	some	360 22	2778	353	3
#!genome-bu	ild Eı	uropean Nu	icleoti	de Arc	hive	A	SM82939v1
#!genome-ver	rsion	GCA_0008	329395	5.1			
#!genome-da	te 20	14-11					
#!genome-bu	ild-a	ccession G	CA_00	08293	395.	1	
#!genebuild-la	ast-i	pdated 20	14-11			V	
Chromosome	ena	gene	360	1676	(1	ID=gene:LOOC260_100010Name=dnaAbiotype=protein_coding;d
Chromosome	ena	transcript	360	1676	. +	•	ID=transcript:BAP84581;Parent=gene:LOOC260_100010;Name=dr
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_100020Name=dnaN;biotype=protein_coding;d
Chromosome	ena	transcript	1852	2991	. +		ID=transcript:BAP84582;Parent=gene:LOOC260_100020;Name=dr
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
###						V	
Chromosome	ena	gene	3233	3457		2	ID=gene:LOOC260_100030;biotype=protein_coding;description=S4
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
###						V	
Chromosome	ena	gene	3467	4588	. -	1)	ID=gene:LOOC260_100040Name=recF;biotype=protein_coding;de
						7	

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 お行物をわらいよわげといだろう

D=	g	ene	を言	会	ŀ		彳	于数				体を む行	ざっと 数をナ	眺め ⁻ Iウン	て、 トす	<mark>①ID=</mark> g ればよ	seneを言 こいだろ	含う
##gff-version	3																	—
##sequence-	regio	n Chromo	some	360 2	27	78	53	}										
#!genome-bu	ild E	uropean Ni	ucleoti	de Aro	chi	ve	AS	SM82939v	/1									
#!genome-vei	rsion	GCA_0008	329395	5.1														
#!genome-da	te 20)14-11																
#!genome-bu	ild-a	ccession G	CA_00	0829	39	5.1												
#!genebuild-l	ast-i	updated 20	14-11															
Chromosome	ena	gene	360	1676		+		ID=gene:	LOOC2	260_10	00	10;Nam	ne=dna	A;biot	ype=	=protei	n_codin	g;d
Chromosome	ena	transcript	360	1676		+		ID=trans	cript:B	AP845	81;	;Parent	=gene:	LOOC	260_	10001	0;Name	=d
Chromosome	ena	exon	360	1676		+		Parent=t	ranscri	pt:BAF	P84	4581;Na	ame=B/	\P845	81-1	1;const	itutive=	1;e
Chromosome	ena	CDS	360	1676		+	0	ID=CDS:	BAP84	581;Pa	irei	nt=tran	script:E	3AP84	581	;proteir	n_id=BA	P8
###																		
Chromosome	ena	gene	1852	2991		+		ID=gene:	LOOC2	260_10	00	20;Nam	ne=dna	N;biot	ype	=protei	n_codin	g;d
Chromosome	ena	transcript	1852	2991		+		ID=transe	cript:B	AP845	82;	;Parent	=gene:	LOOC	260_	_10002	0;Name	=d
Chromosome	ena	exon	1852	2991		+		Parent=t	ranscri	pt:BAF	284	1582;Na	ame=BA	\P845	82-3	1;const	itutive=	1;e
Chromosome	ena	CDS	1852	2991		+	0	ID=CDS:	BAP84	582;Pa	irei	nt=tran	script:E	3AP84	582	;proteir	n_id=BA	P8
###																		
Chromosome	ena	gene	3233	3457		+		ID=gene:	LOOC2	260_10	00	30;biot	ype=pr	otein_	cod	ing;des	cription	=S
Chromosome	ena	transcript	3233	3457		+		ID=trans	cript:B	AP845	83;	;Parent	=gene:	L00C	260_	_10003	0;biotyp	e=
Chromosome	ena	exon	3233	3457		+		Parent=t	ranscri	pt:BAF	284	1583;Na	ame=BA	\P845	83-3	1;const	itutive=	1;e
Chromosome	ena	CDS	3233	3457		+	0	ID=CDS:	BAP84	583;Pa	irei	nt=tran	script:E	3AP84	583	;proteir	n_id=BA	P8
###																		
Chromosome	ena	gene	3467	4588		+		ID=gene:	LOOC2	260_10	000	40;Nam	ne=recF	;bioty	/pe=	⊧proteir	n_coding	g;de

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



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。

	D=geneを含む行を表示	を含む行がそ 行分あるので
iu@biel	inux[~/Desktop/mac_share/mapping_kiso3] 1 I	💌 🜒) 15:43 🔱
	<pre>iu@bielinux[mapping_kiso3] pwd</pre>	[3:42午後]
Q	<pre>/home/iu/Desktop/mac_share/mapping_kiso3</pre>	-
	iu@bielinux[mapping_kiso3] ls	[3:42午後]
	Lactobacillus_nokkaidonensis_jcm_18461.GCA_00082939	5.1.30.chrom
	jughielinux[mapping_kiso3] ls * aff3	[3:42年後]
	Lactobacillus hokkaidonensis icm 18461.GCA 00082939	5.1.30.chrom
0	osome.Chromosome.gff3	
	<pre>iu@bielinux[mapping kiso3] grep -c "ID=gene" *.gff3</pre>	}
	2262	
1	<pre>iu@bielinux[mapping_kiso3] grep "ID=gene" *.gff3</pre>	[3:42午後]
IIII IIII		
2		
-		

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。

@biel	inux[~/Desktop/mac_sha	re/mapping	_kiso3]		tı 🛛	• 💌 🜒	15:45 🔱
a	in;gene_id=LOOC	2 <mark>60_</mark> 1226	40;logic_	name=ena	;version=	1	
9	Chromosome	ena 122650.	gene	22/3924	22/5312 .	coding.	
	tion=tRNA modif	ication	GTPase · de	ene id-l(1000000000000000000000000000000000000	650·log	ic name
	=ena:version=1	reaction	on ase, ge		000200_122	050, tog	
	Chromosome	ena	gene	2275488	2276288 .		
9)	ID=gene:L00C260	122660;	biotype=p	orotein_c	coding;des	criptio	n=singl
	e-stranded DNA-	binding	protein;	gene_id=l	_00C260_12	2660;lo	<pre>gic_nam</pre>
	e=ena;version=1			2276455	2277200		
	Chromosome	ena	gene	22/0455	22//288 .	crintio	
	ane protein gen	_122070; e_id=1.00	C260 1226	570·logic	name=ena	·versio	n=1
	Chromosome	ena	aene	2277304	2277648	,	
V	ID=gene:L00C260	122680;	biotype=	orotein d	oding;des	criptio	n=ribon
围	uclease P;gene_	id=L00C2	60_122680	0;logic_r	name=ena;v	ersion=	1
	Chromosome	ena	gene	2277719	2277853 .		- -
	ID=gene:L00C260	122690;	Name=rpm	l;biotype	e=protein_	coding;	descrip
	tion=505 riboso	mal prot	ein L34;	gene_1d=L	12 1000260	2690;10	gic_nam
6	iu@bielinux[map	ping_kis	03]			[3:4	5午後]

ID=geneを含む行を表示

iu

D=geneを含む行をファイル出力

iu@bielinux[~/Desktop/mac_share/mapping_kiso3]

🕦 Ja 📧 🜒 15:49 <u>신논</u>

in;gene id=L00C260 122640;logic name=ena;version=1 Chromosome 2273924 2275312 . ena gene ID=gene:L00C260 122650;Name=trmE;biotype=protein coding;descrip tion=tRNA modification GTPase;gene id=L00C260 122650;logic name =ena;version=1 Chromosome 2275488 2276288 . ena aene ID=gene:L00C260 122660;biotype=protein coding;description=singl e-stranded DNA-binding protein;gene id=L00C260 122660;logic nam e=ena;version=1 Chromosome 2276455 2277288 . dene ena ID=gene:L00C260 122670;biotype=protein coding;description=membr ane protein;gene id=L00C260 122670;logic name=ena;version=1 Chromosome 2277304 2277648 . ena gene ID=gene:L00C260 122680;biotype=protein coding;description=ribon uclease P;gene id=L00C260 122680;logic_name=ena;version=1 2277719 2277853 . Chromosome ena gene ID=gene:L00C260 122690;Name=rpmH;biotype=protein_coding;descrip tion=50S ribosomal protein L34;gene id=L00C260 122690;logic nam e=ena;version=1 iu@bielinux[mapping kiso3] grep "ID=gene" *.gff3 > uge.txt

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

①wcでuge.txtの行数を確 認。確かに2262行ですね。

行数を確認

iu@bielinux[~/Desktop/mac_share/mapping_kiso3] 💌 🜒) 15:52 🔱 T1 Ja tion=tRNA modification GTPase;gene id=L00C260 122650;logic name =ena;version=1 Chromosome 2275488 2276288 . ena aene ID=gene:L00C260 122660;biotype=protein coding;description=singl e-stranded DNA-binding protein; gene id=L00C260 122660; logic nam e=ena;version=1 Chromosome 2276455 2277288 . ena gene ID=gene:L00C260 122670;biotype=protein coding;description=membr ane protein; gene id=L00C260 122670; logic name=ena; version=1 ena dene 2277304 2277648 . Chromosome ID=gene:L00C260 122680;biotype=protein coding;description=ribon uclease P;gene id=L00C260 122680;logic name=ena;version=1 Chromosome 2277719 2277853 . ena gene ID=gene:L00C260 122690;Name=rpmH;biotype=protein coding;descrip tion=50S ribosomal protein L34;gene id=L00C260 122690;logic nam e=ena;version=1 iu@bielinux[mapping kiso3] grep "ID=gene" *.gff3 > uge.txt iu@bielinux[mapping kiso3] wc uge.txt [3:52午後] 2262 24154 398272 uge.txt 2)abielinux[mapping kiso3] [3:52午後]

①grepコマンドで、②uge.txtというファイ
 ①grepコマンドで、②uge.txtというファイ
 Name=という文字列を含
 い(Digrep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
 iu@bielinux[~/Desktop/mac_share/mapping_kiso3]
 1
 近づいてきました。本当はここで365行

2275488 2276288 . Chromosome ena dene になって一件落着のつもりでしたが…。 ID=gene:L00C260 122660;biotype=protein coding;desd Name=はあるがgenename部分が空っぽ e-stranded DNA-binding protein; gene id=L00C260 122 とか、同じ遺伝子名のものがあるとか… e=ena;version=1 Chromosome 2276455 2277288 . ena aene ID=gene:L00C260 122670;biotype=protein coding;description=membr ane protein;gene id=L00C260 122670;logic name=ena;version=1 Chromosome 2277304 2277648 . ena gene ID=gene:L00C260 122680;biotype=protein coding;description=ribon uclease P;gene id=L00C260 122680;logic name=ena;version=1 2277719 2277853 . Chromosome ena aene **ID=gene**:L00C260 122690;Name=rpmH;biotype=protein coding;descrip tion=50S ribosomal protein L34;gene id=L00C260 122690;logic nam e=ena;version=1 iu@bielinux[mapping kiso3] grep "ID=gene" *.gff3 > uge.txt iu@bielinux[mapping kiso3] wc uge.txt [3:52午後] 2262 24154 398272 uge.txt iu@bielinux[mapping kiso3] grep -c "Name=" uge.txt [3:52午後] 457 inux[mapping_kiso3] 3:56午後] iuas