

NGS 基本フォーマットとツール 復習と補足

NGS Basic Formats and Tools
reviewing the last lecture + supplement

NIBB

2021's Genome Informatics Training Course

Hiroyo NISHIDE  (hiroyo@nibb.ac.jp, @piroyon  )

Remote login & file transfer

Remote login to bias.nibb-gitc.link

- log in

```
$ ssh courseXX@bias.nibb-gitc.link
```

- log out

```
$ exit
```

「courseXX」 is
your account

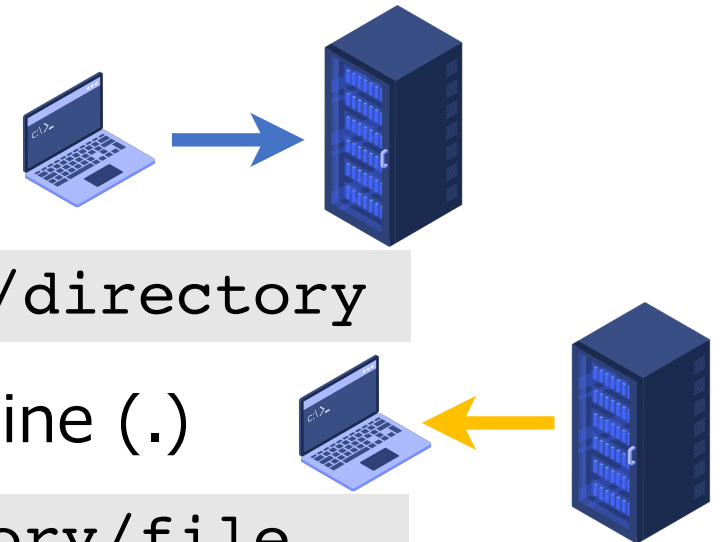
File transfer (optional)

- Your machine to bias.nibb-gitc.link (~ / directory)

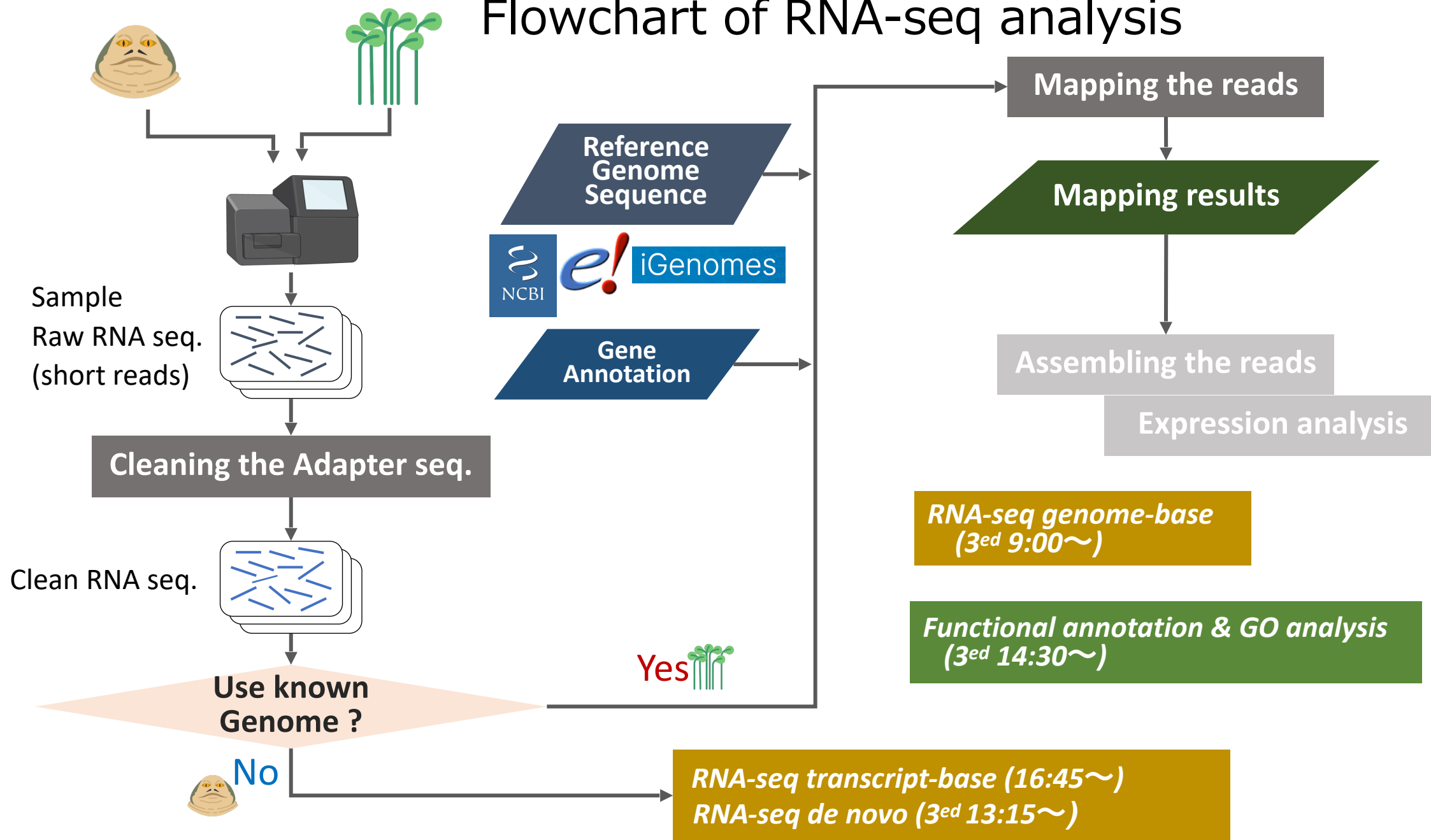
```
$ scp file courseXX@bias.nibb-gitc.link:~/directory
```

- bias.nibb-gitc.link (~ / directory / file) to your machine (.)

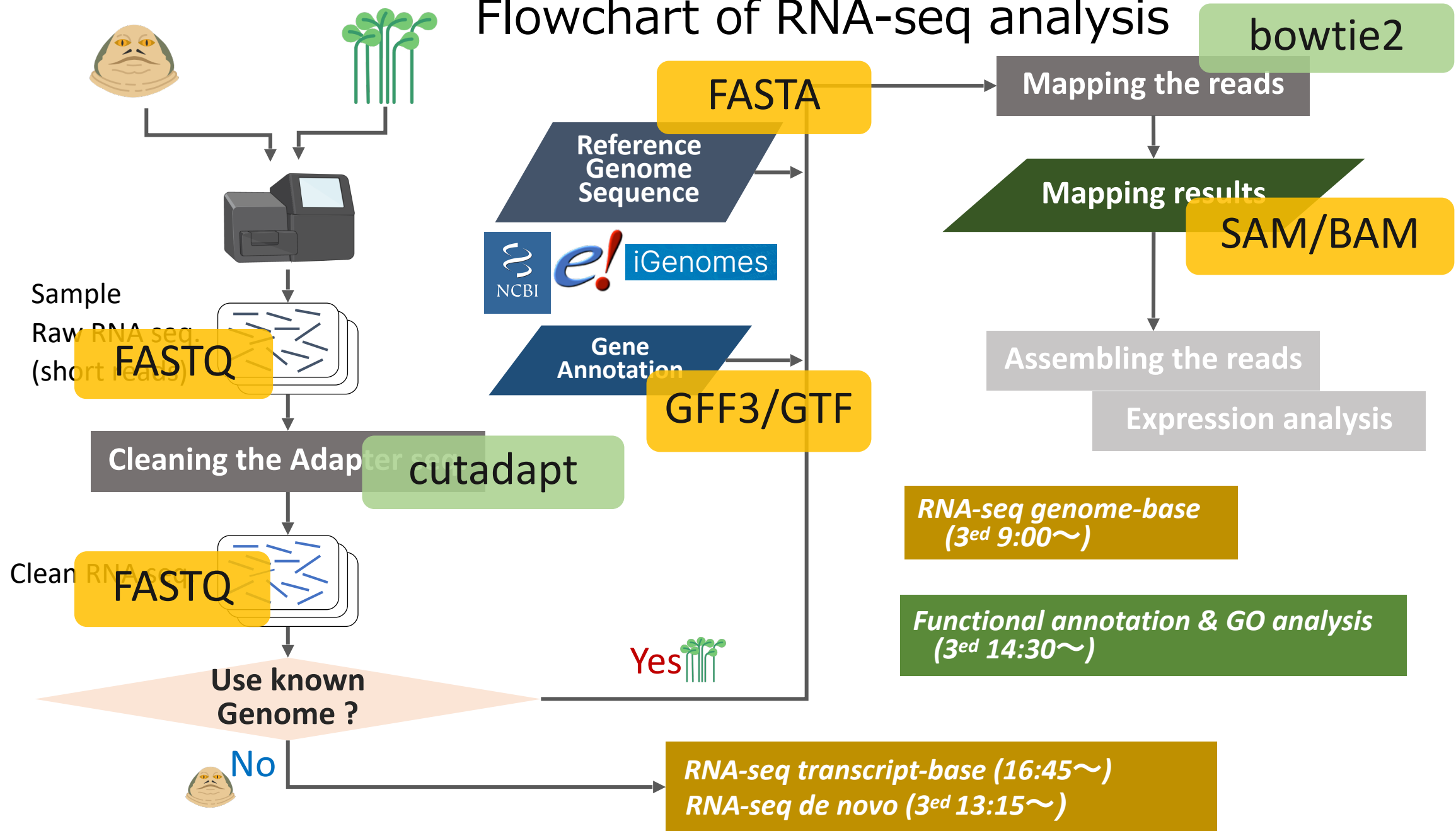
```
$ scp courseXX@bias.nibb-gitc.link:directory/file .
```



Flowchart of RNA-seq analysis



Flowchart of RNA-seq analysis



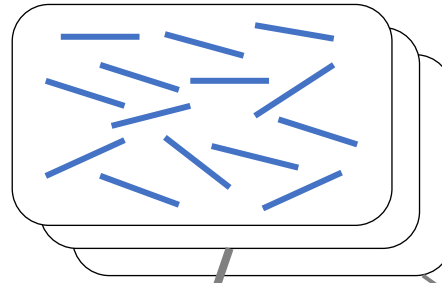
Short Read Mapping & Data Format

Genome (Reference) sequence
format (Sequence)

```
>chr
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACCTGGTTACCTGCCGTGAGTAAATTA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAG
AGTACACAACATCCATGAAACGCATTAGCACCACC
```

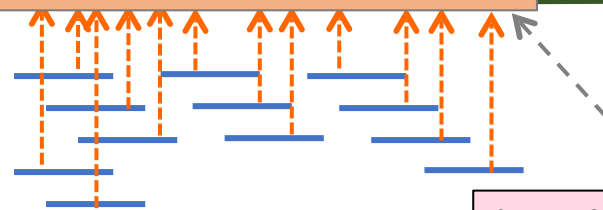
Short read sequence
format (Sequence + Quality)

Sample



```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGGCTGGCGCACCGACCTATGTTCGGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFDFDFHDFHIIIEGIHJJJGFGHGGHGGHGGIJDGIJHHGGGHHI
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CAGGACATCGCCTTTGATCGGTTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CCCFDFDFHDFHIIJGHIJJJJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

Mapping to a reference seq.



Quality control
Adapter removal

format (Result of mapping)

chr	RefSeq	start_codon	190	192	1.000	+	.	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	CDS	190	252	1.000	+	0	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	stop_codon	253	255	1.000	+	.	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	exon	190	255	1.000	+	.	gene_id "b0001"; transcript_id "b0001";

format
(Gene annotation)

```
@HD      VN:1.0      SO:unsorted
@SQ      SN:chr      LN:4639675
@PG      ID:bowtie2  PN:bowtie2  VN:2.2.4    CL:"/bio/bin/bowtie2-align
SRR1515276.40 0 chr 4423609 42 51M * 0 0 GGAATTCCTCACTGCCA
SRR1515276.158 16 chr 501700 42 51M * 0 0 ACGCACCGAGTGCAAAG
SRR1515276.212 4 * 0 0 * * 0 0 GGCCGCTTTCAGCGTGT
SRR1515276.319 0 chr 2922768 42 51M * 0 0 GCTTAAGTTGATTAAGG
SRR1515276.367 16 chr 2753873 42 51M * 0 0 GCGTGTCCGTCCGCAGC
SRR1515276.411 0 chr 3440721 42 51M * 0 0 ACGGCATAATTCTTGA
SRR1515276.434 0 chr 4198737 42 51M * 0 0 GCGCGGTACGCATCTGG
```

Binary
digitization

format
(Result of mapping)

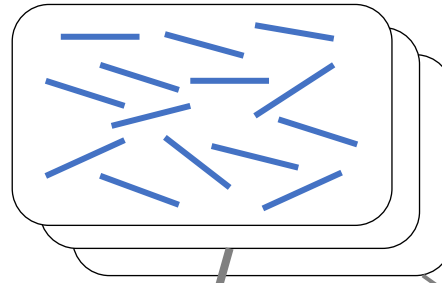
Short Read Mapping & Data Format

Genome (Reference) sequence
FASTA format (Sequence)

```
>chr
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACCTGGTTACCTGCCGTGAGTAAATTA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAG
AGTACACAACATCCATGAAACGCATTAGCACCACC
```

Short read sequence

Sample



FASTQ format (Sequence + Quality)

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGGCTGGCGCACCGACCTATGTTCGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTCCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFDFDFHDFHIIIEGIHJJJGFHGGHGGHGGIJDGIJHHGGGHHI
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CAGGACATCGCCTTTGATCGGTTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CCCFDFDFHDFHIIJGHIJJJJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

Mapping to a reference seq.

Quality control
Adapter removal

SAM format (Result of mapping)

chr	RefSeq	start_codon	190	192	1.000	+	.	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	CDS	190	252	1.000	+	0	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	stop_codon	253	255	1.000	+	.	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	exon	190	255	1.000	+	.	gene_id "b0001"; transcript_id "b0001";

```
@HD      VN:1.0      SO:unsorted
@SQ      SN:chr      LN:4639675
@PG      ID:bowtie2  PN:bowtie2  VN:2.2.4    CL:"/bio/bin/bowtie2-align
SRR1515276.40 0 chr 4423609 42 51M * 0 0 GGAATTCCTCACTGCCA
SRR1515276.158 16 chr 501700 42 51M * 0 0 ACGCACCGAGTGCAAAG
SRR1515276.212 4 * 0 0 * * 0 0 GGCCGCTTTCAGCGTGT
SRR1515276.319 0 chr 2922768 42 51M * 0 0 GCTTAAGTTGATTAAGG
SRR1515276.367 16 chr 2753873 42 51M * 0 0 GCGTGTCCGTCCGCAGC
SRR1515276.411 0 chr 3440721 42 51M * 0 0 ACGGCATAATTCTTGA
SRR1515276.434 0 chr 4198737 42 51M * 0 0 GCGCGGTACGCATCTGG
```

GTF (GFF3) format
(Gene annotation)

Binary
digitization

BAM format
(Result of mapping)

復習：cutadaptによる アダプター配列の除去

実習用ディレクトリ ~/gitc/data/HN

入力

- ショートリード配列 (FASTQ フォーマット, paired-end)

`etec_1.fq`

`etec_2.fq`

- アダプター配列 (それぞれを3'端から除去)

Adapter1: `AGATCGGAAGAGCGGTT`

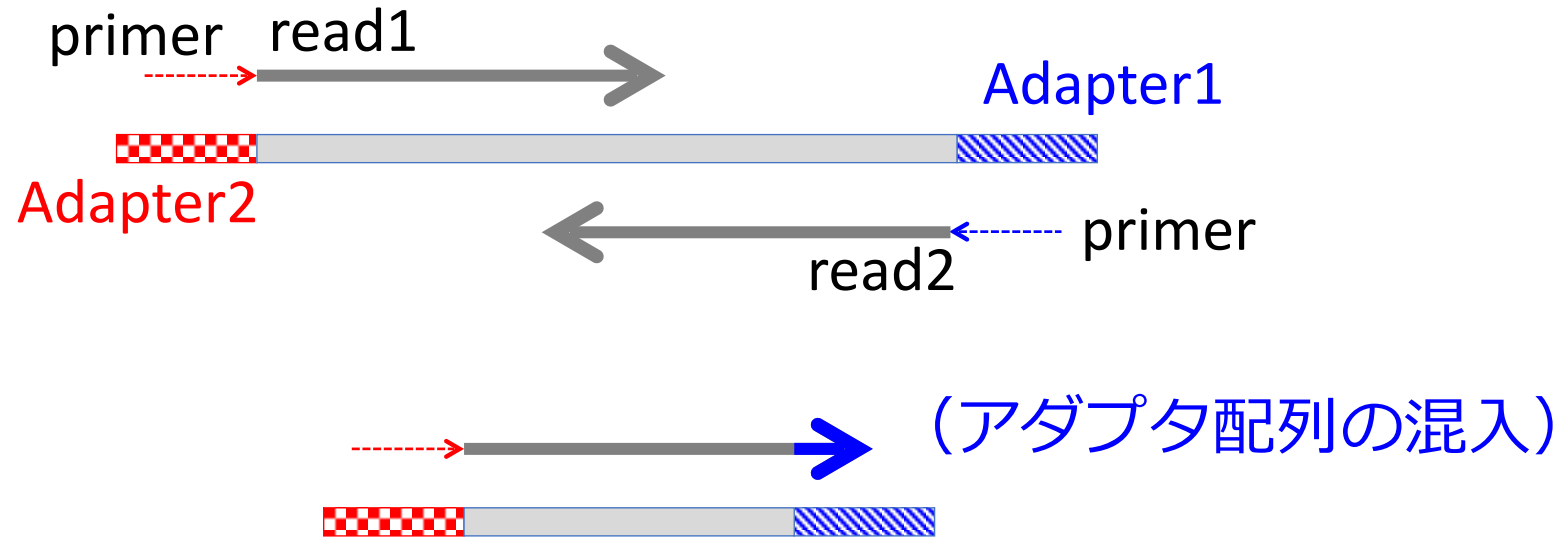
Adapter2: `AGATCGGAAGAGCGTCG`

◆ アダプター配列除去の実行

除去後のデータ (FASTQフォーマット) は `etec_1.cut.fq`、`etec_2.cut.fq`とする。

```
$ cutadapt -a AGATCGGAAGAGCGGTT -A AGATCGGAAGAGCGTCG  
           -o etec_1.cut.fq -p etec_2.cut.fq  
           etec_1.fq etec_2.fq
```

Illuminaにおけるアダプター配列



Adapter1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Adapter2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA

cutadapt -a (-A) オプションでは、指定した配列とマッチした箇所以降の3'側を切り捨てるので、アダプタ配列は全長を指定しなくてもよい。

cutadapt その他のオプション

- **-q** [5' *cutoff*,] 3' *cutoff* (例: **-q 20**)
 - クオリティ値が指定したカットオフより低い塩基を3'端から除く (カンマ区切りでカットオフを2つ指定した場合は5'端からも除く)
- **-m** *min length* (例: **-m 30**)
 - アダプター除去後の配列長が指定した長さ以下になったら配列全体を捨てる。
 - ペアエンドの場合、ペアのどちらかが捨てられる場合は両方を捨てる。
→ 2つのファイルで対応する配列の出現順が揃うようにする。
- **-o** *overlap length* (例: **-o 5**)
 - アダプターとリードとの間で、マッチしたと見なす最低のオーバーラップ長を指定。デフォルトは3。



復習：bowtie2 用インデックスの作成

実習用ディレクトリ ~/gitc/data/HN

bowtie2でマッピングをするには、リファレンスゲノム配列にインデックスが必要

入力

- ゲノム配列（FASTAフォーマット）

eco_o139.fa 腸管毒素原性大腸菌(ETEC) O139:H28のゲノム配列

- ◆ bowtie2用インデックスの作成（インデックス名は **etec** とする）

```
$ bowtie2-build eco_o139.fa etec
```

復習：bowtie2の実行 (paired-end)

実習用ディレクトリ ~/gitc/data/HN

入力

- ショートリード配列 (FASTQフォーマット, paired-end, アダプター除去済)

etec_1.cut.fq

etec_2.cut.fq

- リファレンスゲノム配列のインデックス名 (先ほど作ったもの)

etec

- ◆ bowtie2によるマッピングの実行 (結果ファイル: **etec_bowtie2.sam**)

```
$ bowtie2 -x etec -1 etec_1.cut.fq -2 etec_2.cut.fq  
-S etec_bowtie2.sam
```

マッピング結果ファイル (SAMフォーマット)

ヘッダ (@で始まる)

リファレンス配列に関する情報																																
@HD	VN:1.0	SO:unsorted																														
@SQ	SN:ETEC_chr	LN:4979619																														
@SQ	SN:pETEC_80	LN:79237																														
@SQ	SN:pETEC_35	LN:34367																														
@SQ	SN:pETEC_73	LN:70609																														
@SQ	SN:pETEC_6	LN:6199																														
@SQ	SN:pETEC_74	LN:74224																														
@SQ	SN:pETEC_5	LN:5033																														
@PG	ID:bowtie2	PN:bowtie2	VN:2.3.0	CL:"/bio/bin/bowtie2-align-s --wrapper basic-0 -x etec -S etec_bowtie2.sam -l etec_1.cut.fq -2 etec_2.cut.fq"																												
SRR345261.25	89	ETEC_chr	3758170	1	49M	=	3758170	0	ACACGCGGCATGGCTG...	###?ED>EBDBDDE,E...	AS:i:-1	XS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:24T24	YT:Z:UP													
SRR345261.25	133	ETEC_chr	3758170	0	*	=	3758170	0	NNNNNNNNNNNNNNNN...	#####...	YT:Z:UP	YF:Z:NS																				
SRR345261.50	73	ETEC_chr	4361458	1	49M	=	4361458	0	CAAGCGTTAATCGGAA...	:HEGDFHHHH@BGG=B...	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YT:Z:UP													
SRR345261.50	133	ETEC_chr	4361458	0	*	=	4361458	0	NNNNNNNATNNNNNN...	#####...	YT:Z:UP	YF:Z:NS																				
SRR345261.75	73	ETEC_chr	4362922	1	49M	=	4362922	0	CGGTGGATGCCCTGGC...	DDDDBD6DB>BB>1>>	AS:i:-2	XS:i:-2	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:37T11	YT:Z:UP													
SRR345261.75	133	ETEC_chr	4362922	0	*	=	4362922	0	NNNNNNTTNNNTCGG...	#####...	YT:Z:UP	YF:Z:NS																				
SRR345261.100	73	ETEC_chr	679991	42	49M	=	679991	0	GTGGTTTAATGAGTCC...	GGGGGGGGB=ED=EEG...	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YT:Z:UP														
SRR345261.100	133	ETEC_chr	679991	0	*	=	679991	0	NNNNNNCACCGNTAGT...	#####...	YT:Z:UP	YF:Z:NS																				
SRR345261.125	73	ETEC_chr	4376280	42	49M	=	4376280	0	CTCAGGATGAGGGTCA...	EEEE=B<<@BDEEDE:...	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YT:Z:UP														
SRR345261.125	133	ETEC_chr	4376280	0	*	=	4376280	0	NNNNNTTTCNCNTTAG...	#####...	YT:Z:UP	YF:Z:NS																				
SRR345261.150	89	ETEC_chr	779844	42	49M	=	779844	0	TTCAGGAAACCTGAA...	B@8D>ECC?BG@ECC>...	AS:i:-5	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:43T5	YT:Z:UP														
SRR345261.150	133	ETEC_chr	779844	0	*	=	779844	0	CNCNGGAGTACNTTGA...	#####...	YT:Z:UP	YF:Z:NS																				
SRR345261.175	83	ETEC_chr	3605306	42	49M	=	3605113	-242	CCGCTTGC CGCGGGCCA...	EDE<8??;?@DGGDDE...	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YS:i:-3	YT:Z:CP													
SRR345261.175	163	ETEC_chr	3605113	42	49M	=	3605306	242	CCGGGTCTGTCGTGG...	DGGDGFDDGGGGGEGD...	AS:i:-3	XN:i:0	XM:i:3	XO:i:0	XG:i:0	NM:i:3	MD:Z:38A6C2T0	YS:i:0	YT:Z:CP													
SRR345261.200	77	*	0	0	*	*	0	0	AAAAAAAAAAAAAAAA...	#####...	YT:Z:UP																					
SRR345261.200	141	*	0	0	*	*	0	0	AAAAAAAAAAAAAAAA...	8@#####...	YT:Z:UP																					
SRR345261.225	83	ETEC_chr	2879707	1	49M	=	2879600	-156	CACAACACGAGCTGAC...	8?D8BEBGD@GG8GCE...	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YS:i:-1	YT:Z:CP												
SRR345261.225	163	ETEC_chr	2879600	1	49M	=	2879707	156	CCCACCTTCCTCCAGT...	GGGBGDEGGG@GG<G8...	AS:i:-1	XS:i:-1	XN:i:0	XM:i:1	XO:i:0	XG:i:0	NM:i:1	MD:Z:45G3	YS:i:0													
YT:Z:CP																																
SRR345261.250	99	ETEC_chr	4361346	1	49M	=	4361525	228	GTACTTTTCAGCGGGGA...	ECE=>EC?FDG<EGDA...	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YS:i:0	YT:Z:CP												
SRR345261.250	147	ETEC_chr	4361525	1	49M	=	4361346	-228	CCGGGCTCAACCTGGG...	#####BC...	AS:i:0	XS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:49	YS:i:0	YT:Z:CP												

FLAG

マップされた
染色体と位置
(* はマップ
されなかつ
た)

MAPQ

CIGAR

ペアの相手がマッ
プされた染色体
(同じなら=) と
位置、フラグメン
トの長さ (右側の
リードは負値)

リード配列

配列クオリティ
値

オプション

AS アライメントスコア
XS 他の位置でのベストスコア
YF リードがfiltering out され
た理由

同じ名前のリード
=ペアエンドの
リード対

復習：SAMからBAMへの変換

実習用ディレクトリ ~/gitc/data/HN

人が読めるテキストデータのSAMから、コンピュータが扱い易いBAM（圧縮したバイナリデータ）へ変換する

入力

- SAMフォーマットファイル（さきほどbowtie2によって作成されたもの）

`etec_bowtie2.sam`

- ◆SAMからBAMへ変換する（結果ファイル：`etec_bowtie2.bam`）

```
$ samtools view -b etec_bowtie2.sam -o etec_bowtie2.bam
```

- ◆作成したBAMファイルをヘッダ付きでSAMに変換してlessで表示する

```
$ samtools view -h etec_bowtie2.bam | less
```

復習：BAMのインデックス作成と検索

実習用ディレクトリ ~/gitc/data/HN

BAMフォーマットファイルを扱いやすくするためにソートし、インデックスを作成する

入力

- BAMフォーマットファイル（さきほどSAMからの変換によって作成されたもの）

etec_bowtie2.bam

- ◆リファレンス配列上の位置の順にソートする

（結果ファイル：**etec_bowtie2_sorted.bam**）

```
$ samtools sort etec_bowtie2.bam -o etec_bowtie2_sorted.bam
```

- ◆ソートされたBAMファイルに対してインデックスを作成する（.baiファイルができる）

```
$ samtools index etec_bowtie2_sorted.bam
```

- ◆インデックスを使って、リファレンスの染色体配列(染色体名：ETEC_chr)の10000-12000 の範囲にマッピングされた結果のみを表示する

```
$ samtools view etec_bowtie2_sorted.bam ETEC_chr:10000-12000
```

SAM/BAM フォーマット補足

- Bowtie2のデフォルトオプションでマッピングした結果のSAM/BAMファイルは、元のFASTQファイルに含まれている各リードの配列とクオリティデータをすべて含んでいる。以下のコマンドでSAM/BAMファイルからFASTQファイルを作成できる。

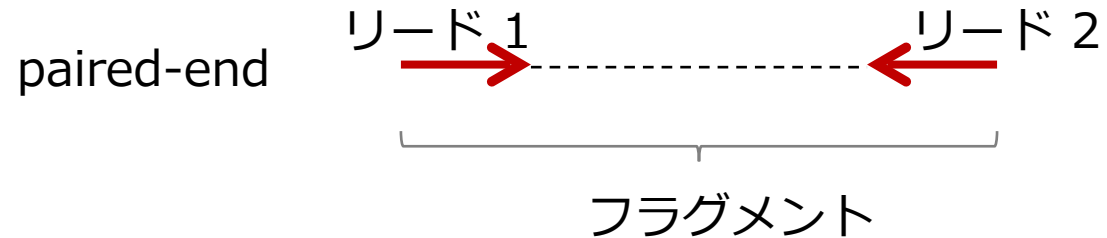
```
$ samtools fastq etec_bowtie2.bam -1 r1.fq -2 r2.fq
```

- 個々のリード配列を記録する代わりに、リファレンス配列を参照して、各リードのリファレンス上の位置とアライメント情報のみを記録することによって、さらに圧縮率を高めたバイナリ形式としてCRAM形式がある。

```
$ samtools view -C etec_bowtie2.sam -T eco_o139.fa  
-o etec_bowtie2.cram
```

Bowtie2のオプション1

ペアエンドリード対の検索



- **-I** *min_length* フラグメント長の最小値 (default: 0)
- **-X** *max_length* フラグメント長の最大値 (default: 500)
- **--fr** / **--rf** / **--ff** リード1とリード2の相対的な向き (default: fr)



- 条件を満たさない(discordant)リード対もデフォルトでは出力される。その際、2カラム目(FLAG)の2ビット目（ペアが正しくアラインされたか？）に0がセットされる。

マッピング結果のフラグ (FLAG)

- True/Falseの2状態を1/0で表した変数。複数のフラグをまとめて、2進数の数値で表現される。
- フラグ値は10進数で表示されるが、2進数に変換することで解釈される。

FLAG値

10進数

83

2進数

01010011

解釈

ペアリードである

各リードが適切にアラインされている

逆鎖にマップされている

1番目のリードである

unix コマンドによる 10進数→2進数の変換

```
$ echo 'obase=2;83' | bc
1010011
```

samtools を使ったフラグ値についての確認

```
$ samtools flags 83
0x53    83      PAIRED,PROPER_PAIR,REVERSE,READ1
```

各フラグの説明を表示

```
$ samtools flags
```

Paired end readでのFLAG値



ペアリードがある 両方適切にマッピングされている 自分がマッピングされていない ペア相手がマッピングされていない 逆順にマッピングされた ペア相手は逆順にマッピングされた Read1の配列である Read2の配列である	2進数表記	10進数表記
1 1 1 1 1 1 1 1	11111111	255

通常のpaired end seqで consistentにアラインしていれば この4通りになる	0 1 0 1 0 0 1 1	01010011	83
	0 1 1 0 0 0 1 1	01100011	99
	1 0 0 1 0 0 1 1	10010011	147
	1 0 1 0 0 0 1 1	10100011	163
片方しかアラインしていない場合	0 1 0 0 1 0 0 1	01001001	73
	0 1 0 1 1 0 0 1	01011001	89
	0 1 0 0 0 1 0 1	01000101	69
	0 1 1 0 0 1 0 1	01100101	101
	1 0 0 0 1 0 0 1	10001001	137
	1 0 0 1 1 0 0 1	10011001	153
	1 0 0 0 0 1 0 1	10000101	133
	1 0 1 0 0 1 0 1	10100101	165
どっちもアラインしていない場合	0 1 0 0 1 1 0 1	01001101	77
	1 0 0 0 1 1 0 1	10001101	141

Samtoolsを用いたフラグによるフィルタリング

- `samtools view -f フラグ値 BAMファイル`

指定したフラグ値中で1であるフラグが、BAMファイル中のフラグ値でもすべて1になっている行のみを抜き出す。

例) ペアリードでかつ両方が適切にアラインされている行のみを抜き出す

```
$ samtools view -f 3 etec_bowtie2_sorted.bam
```

3は2進数で 11 だから、1番目と2番目のフラグが1である行を抜き出す（それ以外のフラグは無視）

- `samtools view -F フラグ値 BAMファイル`

指定したフラグ値中で1であるフラグが、BAMファイル中のフラグ値ではすべて0になっている行のみを抜き出す。

例) ペアリードの両方が適切にアラインされていない行のみを抜き出す

```
$ samtools view -F 2 etec_bowtie2_sorted.bam
```

2番目のフラグが0である行を抜き出す。

Bowtie2のオプション2

アライメント出力のモード

Bowtie2におけるスコア =
ミスマッチに対するペナルティ

- 一般に、1つのリードは複数の箇所にマップされる。



- `default` (`best one mode`)

条件を満たすアライメントを検索し、最高スコアのものを1つ出力
(ただし、検索は完全でないので、最高スコアを取りこぼす可能性はある)
上記の例では、BまたはD (どちらかがランダムに選ばれる)

- `-a` 条件を満たすアライメントをすべて出力 上記の例では、A,B,C,D,E

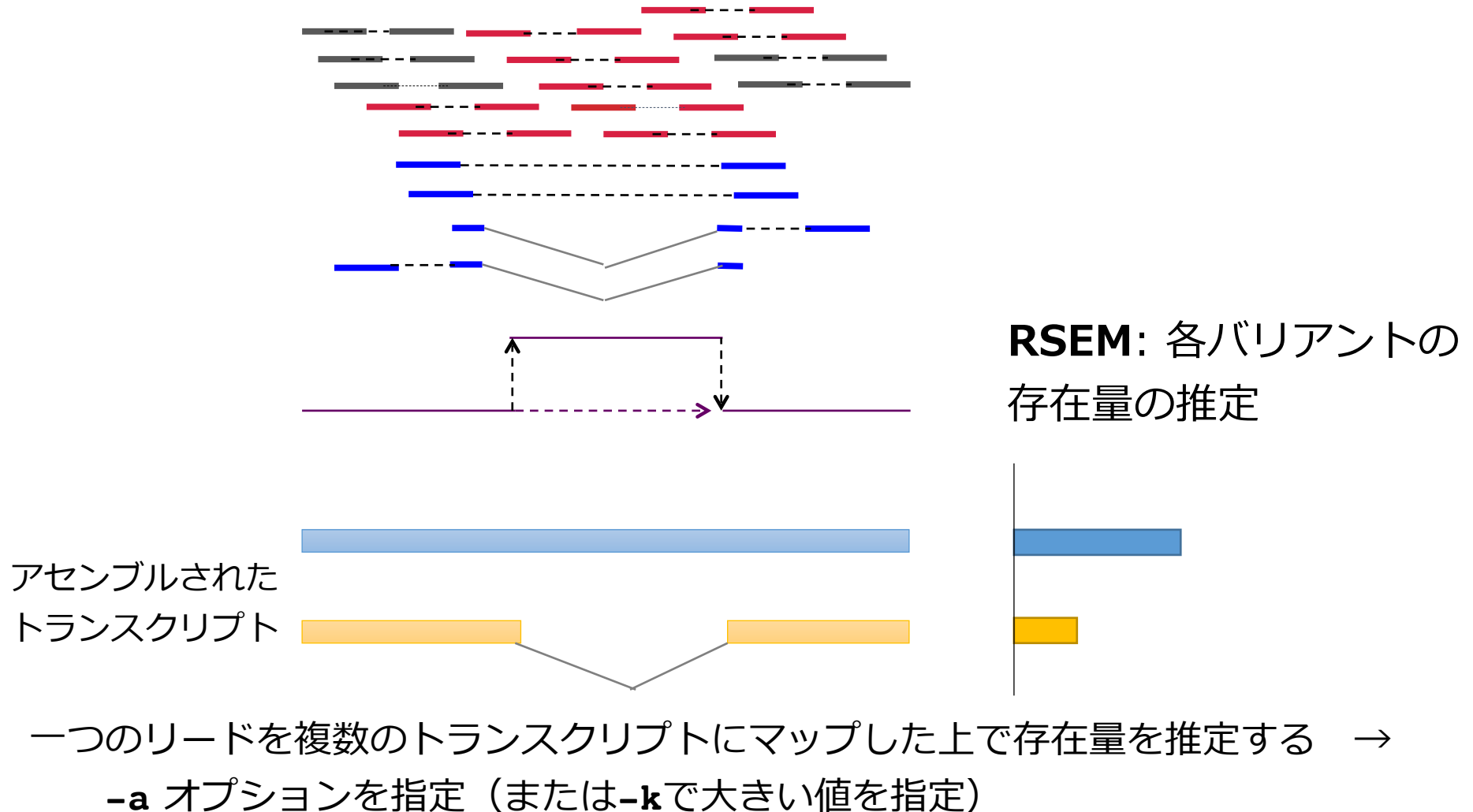
- `-k` `num_of_alignment`

条件を満たすアライメントを、見つかった順に指定した数だけ出力
上記の例で、`-k 2` のとき、左から順に見つかるとうすると、AとBが出力される
(実際には位置の順に見つかるわけではない)

- `-a` や `-k` を指定したとき、最高スコアでないアライメントには9番目のフラグ
(`secondary alignment`)に1がセットされる

(参考) *De novo* Assembly によるRNA-Seq解析

デノボ・アセンブルによる転写配列の構築



マッピングクオリティ (MAPQ)

- マッピングクオリティ (MAPQ) 値は以下の式で計算される。

$$\text{MAPQ} = -10\log_{10}(P_e)$$

- ただし、 P_e はリードが間違った位置にマップされている確率の推定値。
- MAPQ は、リードがその位置にどの程度ユニークにマップされたかを示す指標であり、その位置でのアライメントスコアが、他のすべての位置におけるスコアよりずっと大きいときに大きくなる。
- Bowtie2 のデフォルトでは同じスコアのアライメントが複数の位置で得られた場合、ランダムに一つの位置を出力し、MAPQ に低い値を設定する。
- MAPQ が低いアライメントの位置は信用できないので、下流の解析の際には捨てた方が良い場合もある。

Samtoolsを用いたMAPQによるフィルタリング

- `samtools view -q 閾値 BAMファイル`

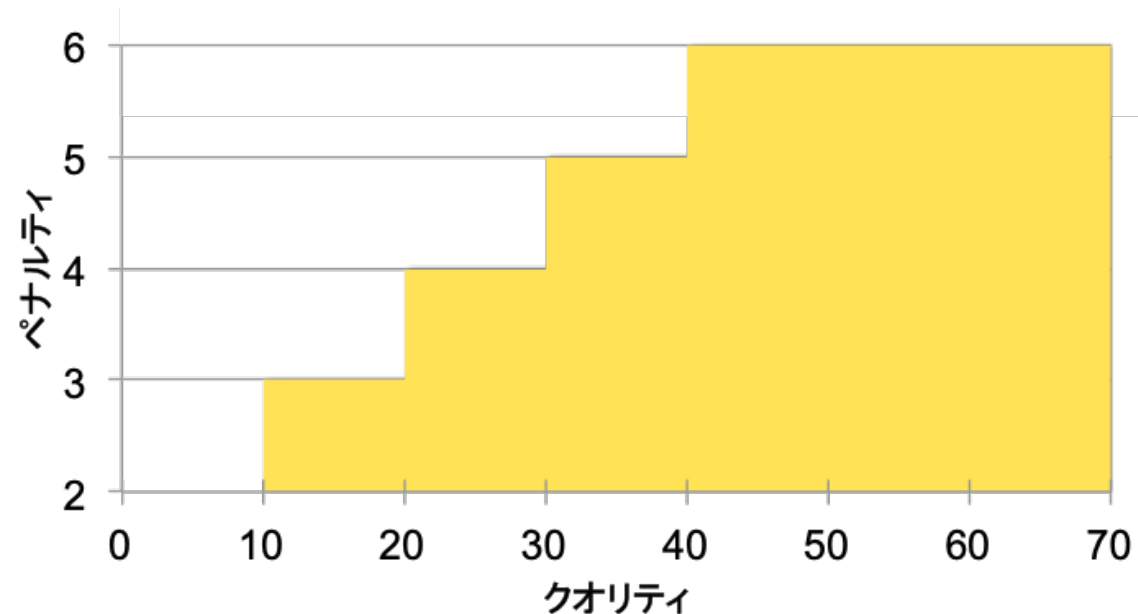
MAPQの値が閾値より小さい行を除く

例) MAPQが20以上の行のみを出力

```
$ samtools view -q 20 etec_bowtie2.bam
```

(参考) Bowtie2におけるアライメントスコア

- マッチは0で、ミスマッチにマイナスのペナルティ（最高スコアは0点）
- ミスマッチペナルティは、クオリティ値に応じて -2 から -6 の値をとる（下図）
- あいまい塩基（N）のペナルティは -1
- ギャップペナルティは、ギャップの長さ n に対して $-(5 + 3n)$
- スコアのカットオフは、長さ L に対して $-0.6(L+1)$



Bowtie2のオプション3

アライメントのモード

- **--end-to-end** リード配列全長に渡るアライメント
(default)

```
Read:      GACTGGGCGATCTCGACTTCG
           ||||| ||||| ||||| |||
Reference:  GACTG--CGATCTCGACATCG
```

- **--local** リード配列のうち、類似度の高い一部の領域のみ
を抜き出してアラインしたもの

```
Read:      ACGGTTGCGTTAA-TCCGCCACG
           ||||| ||||| |||||
Reference:  TAACTTGCGTTAAATCCGCCTGG
```

CIGAR文字列

- リードとリファレンス配列とのアライメントの詳細を表す。
- ギャップなしでアラインされている場合 nM (n はリード配列の長さ) となる。
- ギャップが入っている場合、 nD (欠失) または nI (挿入) (n は欠失/挿入長さ) が入る。

5M2D4M1I5M

```
ref  AGACGAGATTA-GCATG
      ⋮ ⋮⋮ ⋮⋮⋮ ⋮⋮⋮ ⋮⋮ ⋮⋮
read ACACG--ATTAGGCTTG
```

- ローカルアライメントのとき、両端の除かれる部分は nS で、またTopHatなどのスプライシングを考慮するアライメントにおいて、イントロンとしてスキップされるリファレンス配列上の領域は nN で表される。

5S4M1I5M

```
ref  ACGGCTGATTA-GCATG
      ⋮⋮⋮ ⋮⋮ ⋮⋮
read  taaccATTAGGCTTG
```

インデックスを使った高速検索

ハッシュテーブル

ゲノム配列

ACACGTTACGGT.....

リード配列

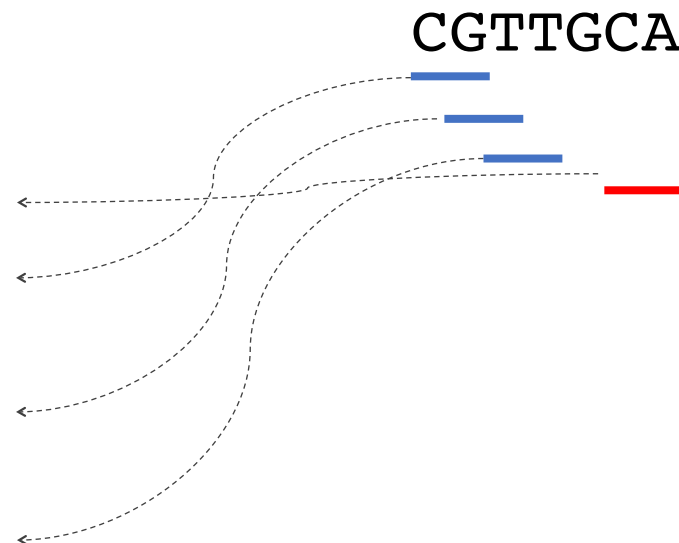
CGTTGCA

① インデックス作成

ハッシュテーブル
各2-merの出現位置を記録

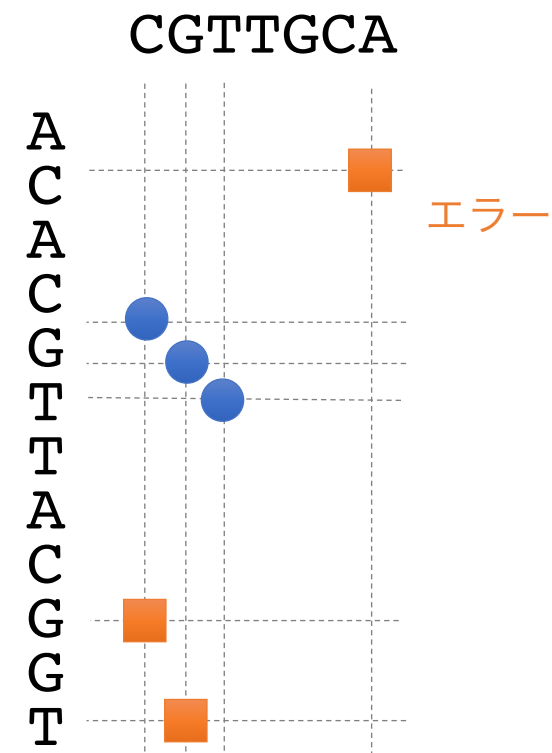
2-mer	positions
AC	1, 3, 8
CA	2
CG	4, 9
GG	10
GT	5, 11
TA	7
TT	6

② インデックスを使った初期検索(seed検索)



③ 見つかったseedを延長してアライメント

ACACGTTACGGT.....
CGTT**G**CA



インデックスを使った高速検索

接尾辞配列 (suffix array)

ACACGTTACGGT

接尾辞

1	ACACGTTACGGT
2	CACGTTACGGT
3	ACGTTACGGT
4	CGTTACGGT
5	GTTACGGT
6	TTACGGT
7	TACGGT
8	ACGGT
9	CGGT
10	GGT
11	GT
12	T

辞書順で
ソート

1	ACACGTTACGGT
8	ACGGT
3	ACGTTACGGT
2	CACGTTACGGT
9	CGGT
4	CGTTACGGT
10	GGT
11	GT
5	GTTACGGT
12	T
7	TACGGT
6	TTACGGT

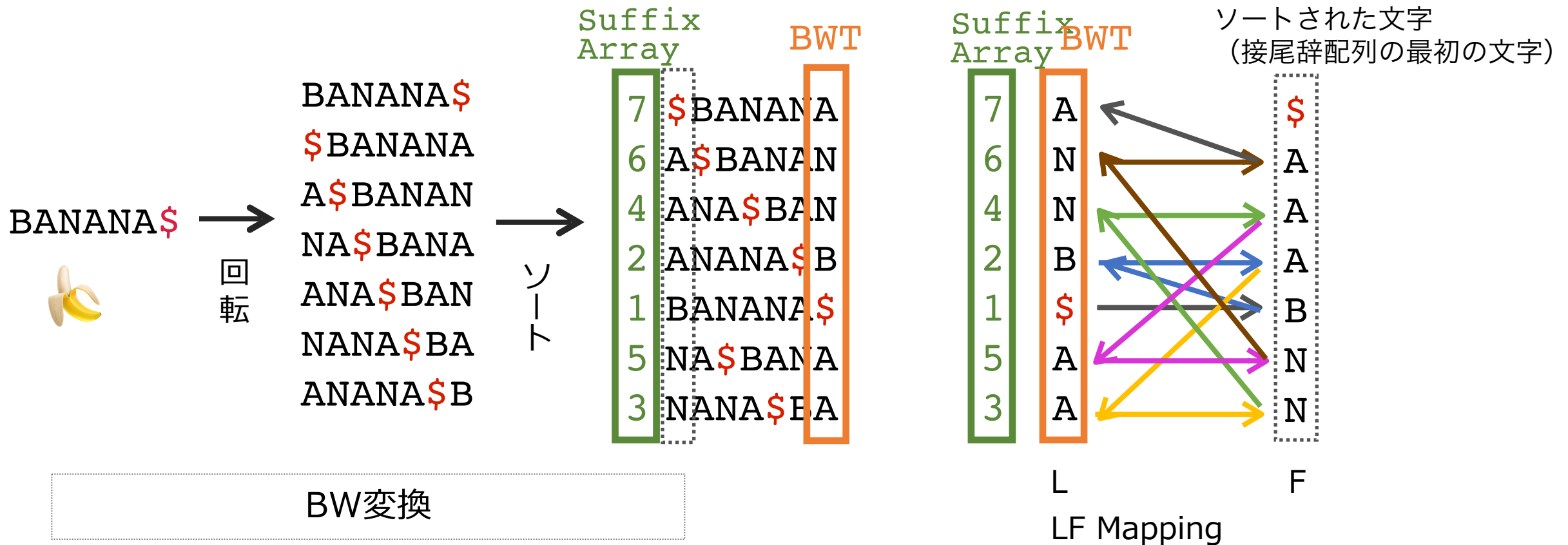
接尾辞配列

ACG 3,8

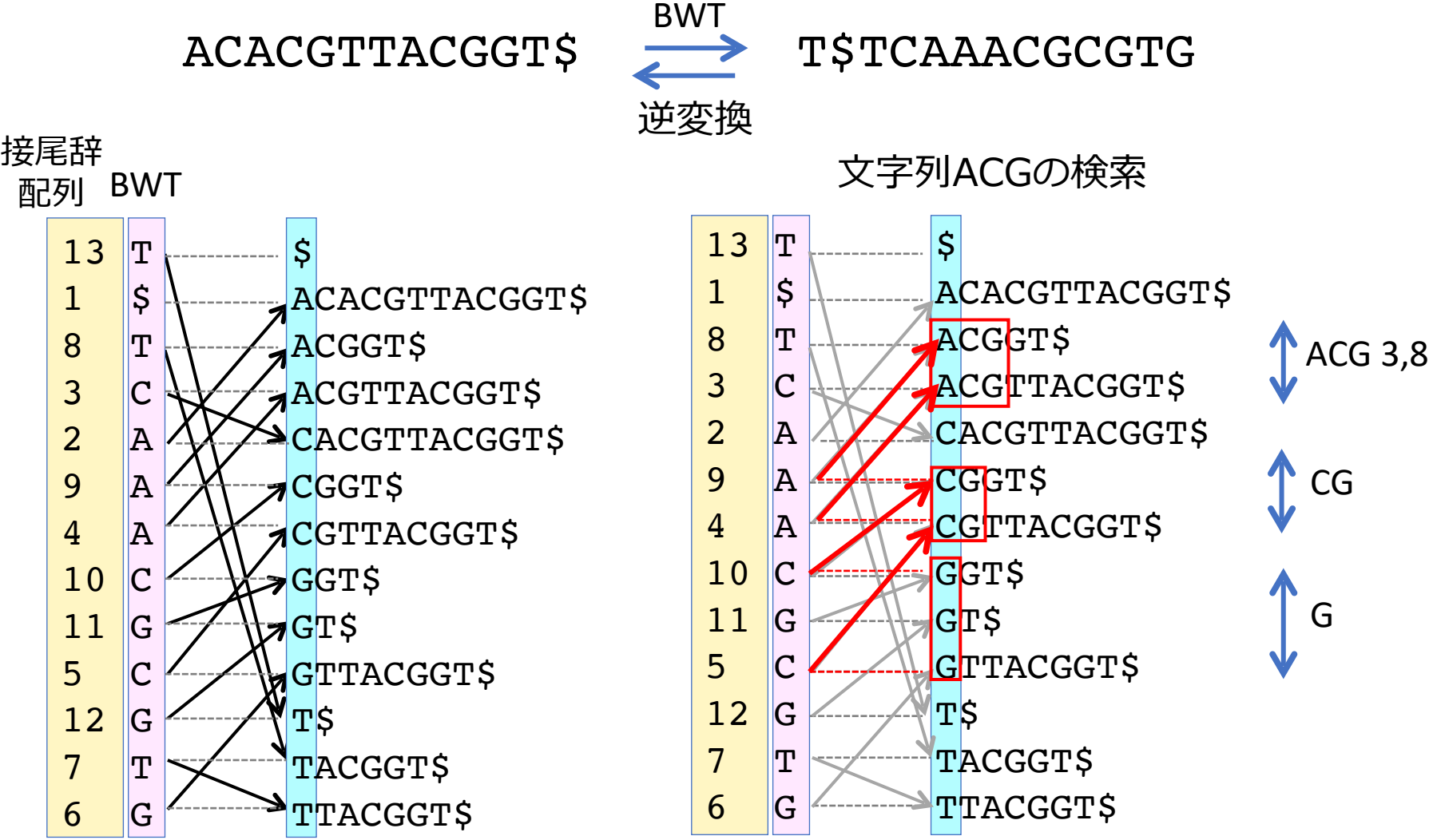
CG 4,9

Burrows-Wheeler 変換 (BWT)

- **BANANA\$** の BWT は、**ANNB\$AA**
- BWTから元の **BANANA\$** に戻ることができる (逆変換)



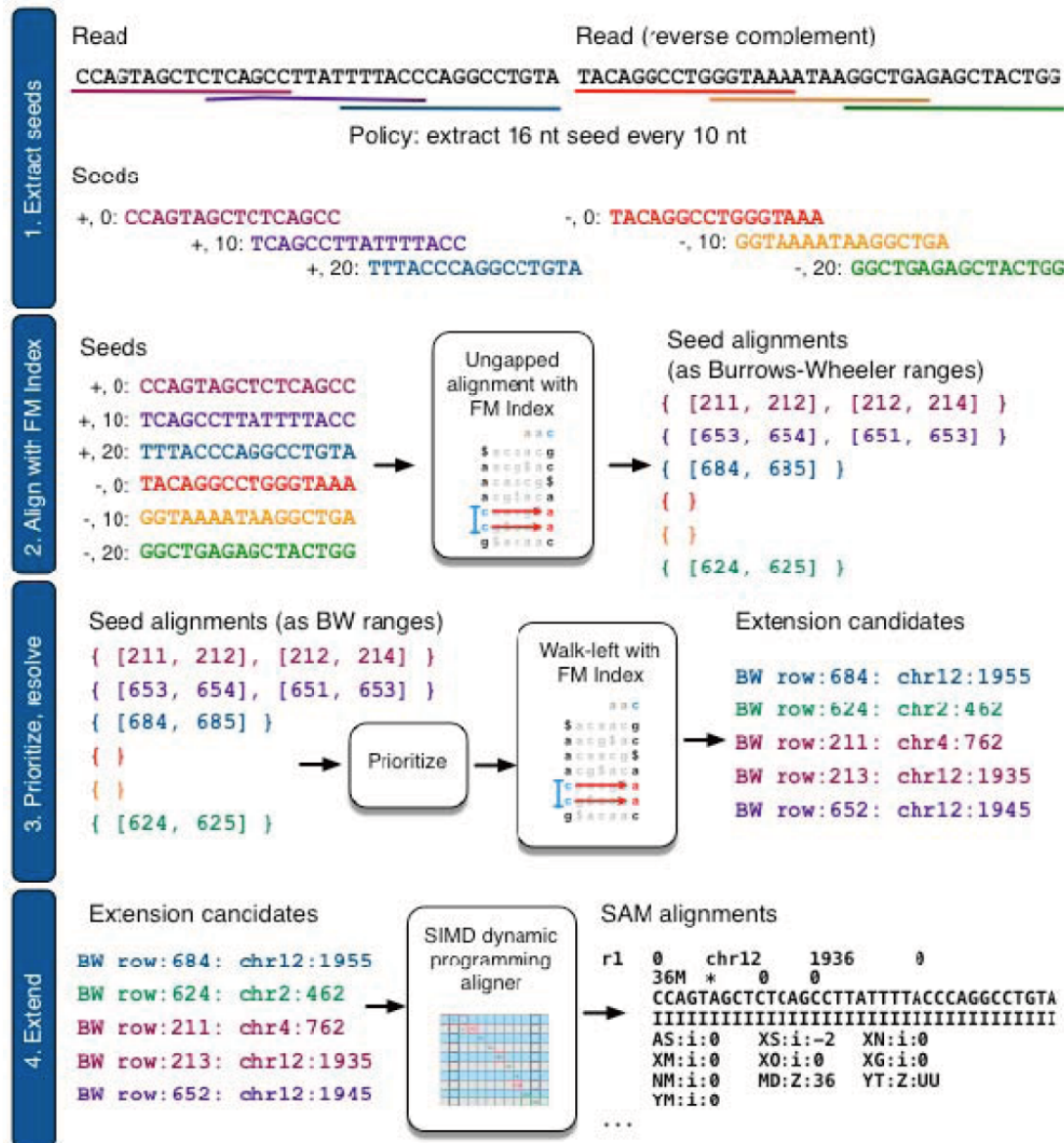
Burrows-Wheeler 変換 (BWT)に基づくインデックス (FM-Index)



矢印 (LF mapping) を辿って元の配列を再構築できる (逆変換)。

矢印を辿って接尾辞配列上での効率の良い文字列検索の実現

Bowtie2 アルゴリズムの詳細



1. Seed 配列の抽出

各リード配列およびその相補配列から i 塩基ごとに L 塩基の配列を抽出してseed配列とする（図では $i=10$, $L=16$ ）。

2. FM index を用いた検索

各seed配列がゲノム上に出現する位置がBW rangeとして得られる。最大1つのミスマッチを考慮した検索が可能。

3. ヒットの優先付け、位置の取得

BW rangeの幅が小さいヒットに高い優先度をつけて、ランダムに候補をピックアップし、ゲノム上の位置を取得。

4. アライメントの計算

得られた位置の周辺で、ギャップ入りのアライメントスコアを計算。これを各候補位置について繰り返して、最高スコアを与えるゲノム上の位置を出力。

Bowtie2のオプション4

検索の精度と速度に関するオプション

- **-N** int seed 検索時にミスマッチを許す数 (0 or 1)
- **-L** int seed の長さ
- **-i** func seed をとる間隔 (リード長を基に決める式を指定)
- **-D** int 最高スコアが更新されないときアライメント計算を打ち切るまでの回数
- **-R** int リードが高反復のseedをもつときにre-seedを行う最大回数

上記のオプションを同時に設定する preset optionがある。

高速（低感度）→ 高感度（低速）の順に4段階のオプションが用意されている。

- end-to-endモードの場合 (default: sensitive)

--very-fast / --fast / --sensitive / --very-sensitive

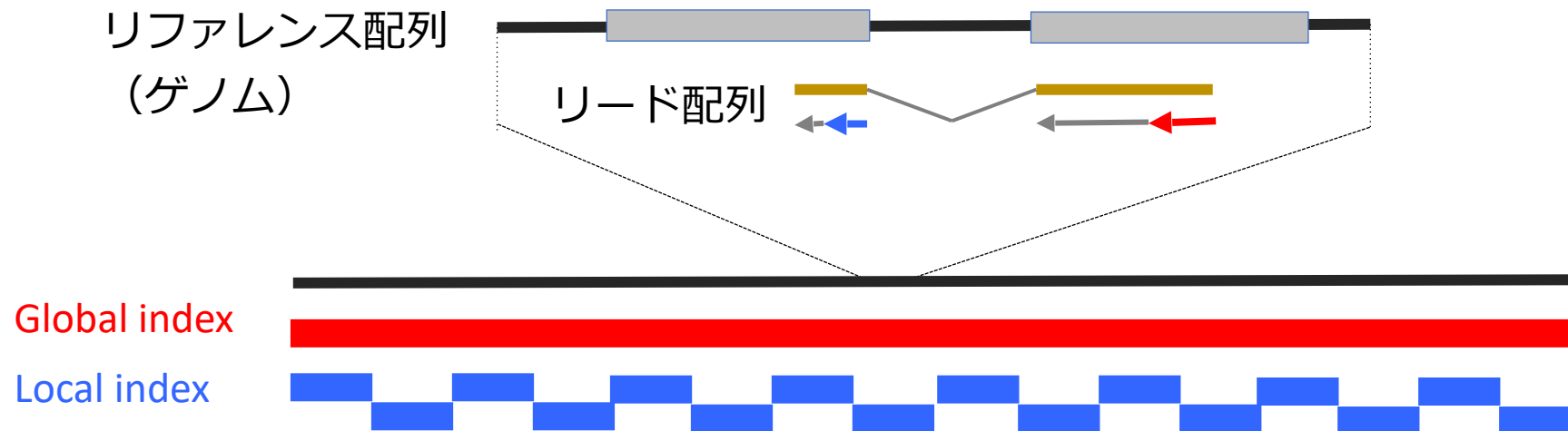
- localモードの場合 (default: sensitive-local)

--very-fast-local / --fast-local / --sensitive-local / --very-sensitive-local

(参考) HISAT2

スプライシングを考慮した高速マッピングツール

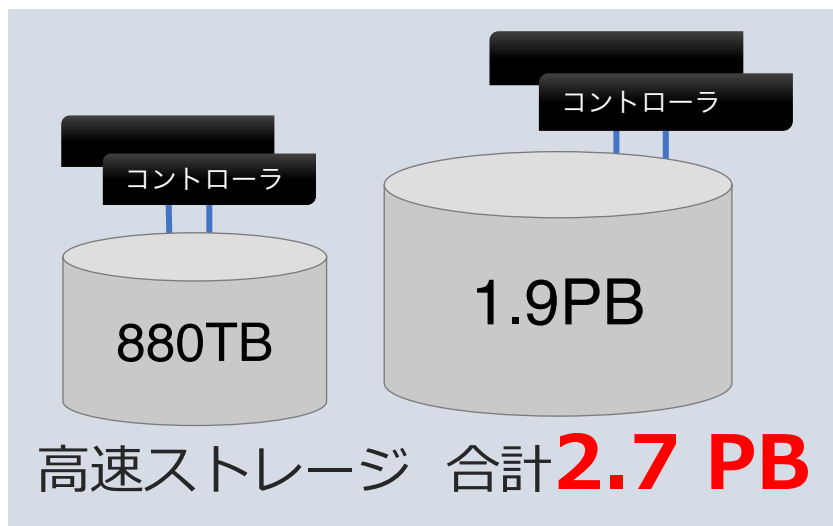
- スプライシングを考慮して、一つのリードをゲノム上で離れた箇所にもたがってマッピングする。
- global とlocalの2つのインデックスを2段階で用いることにより、高速かつ正確にスプライスされたアライメントを実現。



生物情報解析システム (bias5)

共有メモリ型計算サーバ
4TB Mem 80 cores
ldas-smp.nibb.ac.jp

大容量
ストレージ
600TB



FDR インフィニバンド
スイッチ

共有メモリ型計算サーバ
527GB Mem 64 cores
diaf-smp1.nibb.ac.jp

共有メモリ型計算サーバ
527GB Mem 64 cores
diaf-smp2.nibb.ac.jp

分散処理計算機クラスタ
800cores 4.8GB Mem/core
bias5-node[01-20].nibb.ac.jp

共有メモリ型計算サーバ
3TB Mem 72 cores
bias5-smp.nibb.ac.jp

ログインノード
768GB Mem
bias5.nibb.ac.jp

データベースサーバ
512GB Mem

