

NGS 基本フォーマットとツール 復習と補足

基礎生物學研究所
ゲノムインフォマティクストレーニングコース
内山 郁夫 (uchiyaama@nibb.ac.jp)

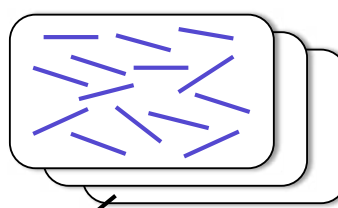
ショートリードのマッピング

ゲノム配列
(リファレンス reference 配列)

形式 (配列)

```
>chr
AGCTTTTCATTTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACCTGGTTACCTGCCGTGAGTAAATTTAAA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAATTACAG
AGTACACAACATCCATGAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGG
```

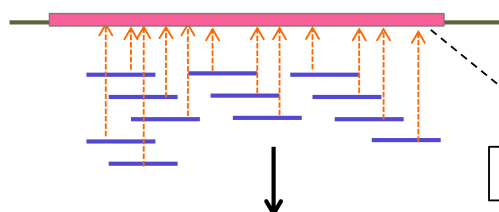
サンプル (ゲノムDNA/RNA)
(リード read 配列)



形式
(配列 + クオリティ値)

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:
ATCCGGCTGGCGCACCACCTATGTTCCGGGCGAATACAAGCTGC
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDCA?>A?
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:
CCCFDFDFHDFHIIIEGIIHJJJGFGHGHGGHGGIJDGIJHE
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:
CAGGACATCGCCTTTGATCGGTTGAGACTCGGACCAACCTGCAT
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:
CCCFDFDFAFHFIJGHIIJJJJJHEHIIJGHIFEHIIA@FIFE
```

リファレンス配列へのマッピング



クオリティチェック
アダプター除去

形式 (遺伝子アノテーション)

chr	RefSeq	start_codon	190	192	1.000	+	.	gene_id	"b0001"
chr	RefSeq	CDS	190	252	1.000	+	0	gene_id	"b0001"
chr	RefSeq	stop_codon	253	255	1.000	+ <td>.<td>gene_id</td><td>"b0001"</td></td>	. <td>gene_id</td> <td>"b0001"</td>	gene_id	"b0001"
chr	RefSeq	exon	190	255	1.000	+ <td>.<td>gene_id</td><td>"b0001"</td></td>	. <td>gene_id</td> <td>"b0001"</td>	gene_id	"b0001"

形式 (マッピング結果)

```
@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 ACGACCCGAGTGCAAG
SRR1515276.212 4 * 0 0 * * 0 0 GGCCGCTTTGAGCGTGT
SRR1515276.319 0 chr 2922768 42 51M * 0 0 GCTTAAGTTGATTAAGG
SRR1515276.367 16 chr 2753873 42 51M * 0 0 GCGTGTCCGTCGCGAGC
SRR1515276.411 0 chr 3440721 42 51M * 0 0 ACGGCATAAATTCCTTCA
```

復習: cutadaptによるアダプターの除去

実習用ディレクトリ ~/data/IU

入力

- リード配列 (FASTQ 形式; 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

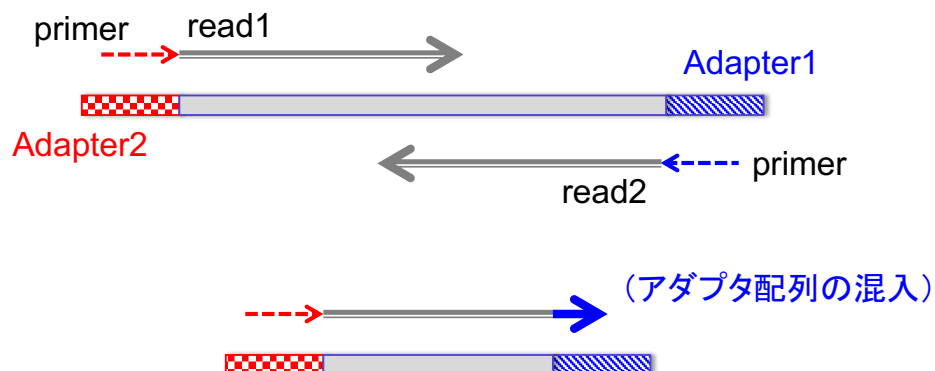
Adapter2: AGATCGGAAGAGCGTCG

◆ アダプター配列除去の実行

除去後のデータ (FASTQ形式) は etec_1.cut.fq, etec_2.cut.fq とする。

```
$ cutadapt ☐ AGATCGGAAGAGCGGTT ☐ AGATCGGAAGAGCGTCG  
           ☐ etec_1.cut.fq ☐ 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: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

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 ~/data/IU

入力

- ゲノムデータ (FASTA形式)
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)

\$ bowtie2-build

復習: bowtie2の実行 (paired-end)

実習用ディレクトリ ~/data/IU

入力

- リード配列 (FASTQ 形式; paired-end; アダプターを除去したもの)
etec_1.cut.fq
etec_2.cut.fq
- リファレンス配列のインデックス名
etec (先ほど作ったもの)

◆ bowtie2によるマッピングの実行 (出力: etec_bowtie2.sam)

```
$ bowtie2 [ ] etec [ ] etec_1.cut.fq [ ] etec_2.cut.fq
           [ ] etec_bowtie2.sam
```

マッピング結果ファイル (SAMファイル)

ヘッダ (@で始まる)

```
@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 -1 etec_1.cut.fq -2 etec_2.cut.fq
SRR345261.25 89 ETec chr 3758170 1 49M = 3758170 0 ACACCGCGCATGGCTG... ##?ED>EBDBDE,E... AS:i:-1 XS:i:-1 XN:i:0
SRR345261.25 133 ETec chr 3758170 0 * = 3758170 0 NNNNNNNNNNNNNNNNN... #####... YT:Z:UP YF:Z:NS
SRR345261.50 73 ETec chr 4361458 1 49M = 4361458 0 CAAGCCTTAATCGGAA... :HEGD?HHHHH@BGG=B... AS:i:0 XS:i:0 XN:i:0
SRR345261.50 133 ETec chr 4361458 0 * = 4361458 0 NNNNNNNNNNNNNNNNN... #####... YT:Z:UP YF:Z:NS
SRR345261.75 73 ETec chr 4362922 1 49M = 4362922 0 CGGTGGATGCCCTGGC... DDDDBD6<B>DB>BB>1> AS:i:-2 XS:i:-2 XN:i:0
SRR345261.75 133 ETec chr 4362922 0 * = 4362922 0 NNNNNNTTNNNTTGGC... #####... YT:Z:UP YF:Z:NS
SRR345261.100 73 ETec chr 679991 42 49M = 679991 0 GTGGTTTAATGAGTGC... GGGGGGGGB=ED=EEG... AS:i:0 XN:i:0 XM:i:0
SRR345261.100 133 ETec chr 679991 0 * = 679991 0 NNNNNNACCAGTAGT... #####... YT:Z:UP YF:Z:NS
SRR345261.125 73 ETec chr 4376280 42 49M = 4376280 0 CTCAGGAACCTGAA... EEEE=B<<@BDEEDE... AS:i:0 XN:i:0 XM:i:0
SRR345261.125 133 ETec chr 4376280 0 * = 4376280 0 NNNNNNTTCCNTTAG... #####... YT:Z:UP YF:Z:NS
SRR345261.150 89 ETec chr 779844 42 49M = 779844 0 TTCAGGAACCTGAA... B@8D>ECC?BG@ECC... AS:i:-5 XN:i:0 XM:i:1
SRR345261.150 133 ETec chr 779844 0 * = 779844 0 CCGGTTGCGCGGCCA... #####... YT:Z:UP YF:Z:NS
SRR345261.175 83 ETec chr 3605306 42 49M = 3605113 -242 CCGGTTGCGCGGCCA... EDE<8?7;?@DGGDD... AS:i:0 XN:i:0 XM:i:0
SRR345261.175 163 ETec chr 3605113 42 49M = 3605306 242 CCGGTTGCGCGGCCA... DGGDGFDDGGGGGEGD... AS:i:-3 XN:i:0 XM:i:3
SRR345261.200 77 * 0 0 * * 0 0 AAAAAAAAAAAAAAAAA... #####... YT:Z:UP
SRR345261.200 141 * 0 0 * * 0 0 AAAAAAAAAAAAAAAAA... 8@#####... YT:Z:UP
SRR345261.225 83 ETec chr 2879707 1 49M = 2879600 -156 CACACACGAGCTGAC... 87D8BEBGD@GG8GCE... AS:i:0 XS:i:0 XN:i:0
SRR345261.225 163 ETec chr 2879600 1 49M = 2879707 156 CCCACCTTCCTCCAGT... GGGBGDEGGG@GG<G@... AS:i:-1 XS:i:-1 XN:i:0
SRR345261.250 99 ETec chr 4361346 1 49M = 4361525 228 GTACTTTTCAGCGGGA... ECE=>EC?FDG<EGDA... AS:i:0 XS:i:0 XN:i:0
SRR345261.250 147 ETec chr 4361525 1 49M = 4361346 -228 CCGGCTCAACCTGGG... #####B... AS:i:0 XS:i:0 XN:i:0
```

同じ名前のリード = ペアエンドのリード対

FLAG

マップされた染色体と位置 (* はマップされなかった)

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

復習: **SAM**から**BAM**への変換

実習用ディレクトリ ~/data/IU

入力

- SAM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 (出力ファイル名: etec_bowtie2.bam)

```
$ samtools   etec_bowtie2.sam   
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   etec_bowtie2.bam  less
```

復習: **BAM**ファイルのインデックスづけと検索

実習用ディレクトリ ~/data/IU

入力

- BAM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
(出力ファイル: etec_bowtie2_sorted.bam)

```
$ samtools  etec_bowtie2.bam  
 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

```
$ samtools  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  etec_bowtie2_sorted.bam  

```

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、もとの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
```

- リファレンス配列を参照して、配列を記録する代わりに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
```

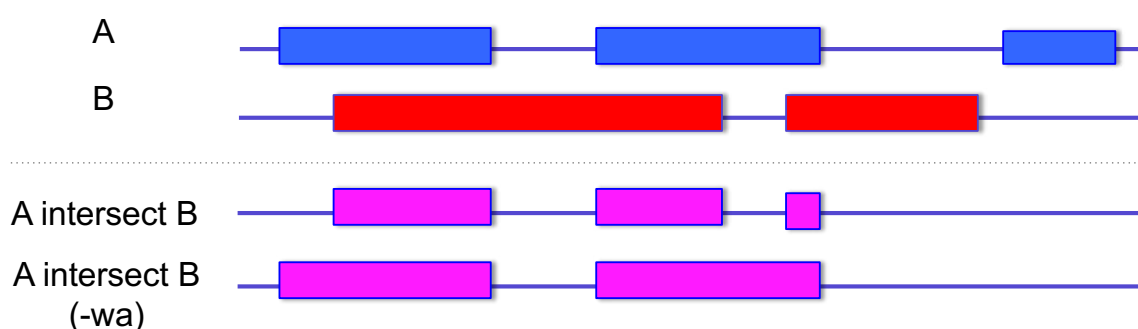
Bedtools

- BED, GFF(GTF), BAM 形式で記述された「ゲノム上の領域の集合」に対して様々な操作を行うツールを集めたもの。

例) intersection

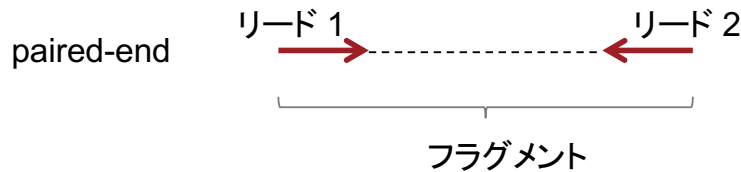
BAM形式のマッピング結果(etec_bowtie2.bam)の中で、コード領域(eco_o130_cds.gff) とオーバーラップしているリードを抽出して、BED形式で表示する。

```
$ bedtools intersect -abam etec_bowtie2.bam  
-b eco_o139_cds.gff -wa -bed | less
```



Bowtie2のオプション1

ペアエンドリード対の検索



- `-I int` フラグメント長の最小値(default: 0)
- `-X int` 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。

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	2進数	解釈
83	01010011	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 を使ったフラグの解釈

```
% samtools flags 83
```

```
0x53    83    PAIRED, PROPER_PAIR, REVERSE, READ1
```

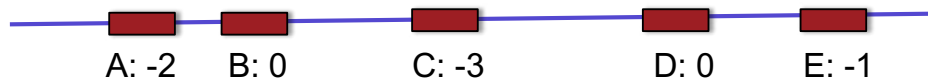
各フラグの説明を表示

```
% samtools flags
```


Bowtie2のオプション2

アライメント出力のモード

- 一般に、1つのリードは複数の箇所にマップされる。

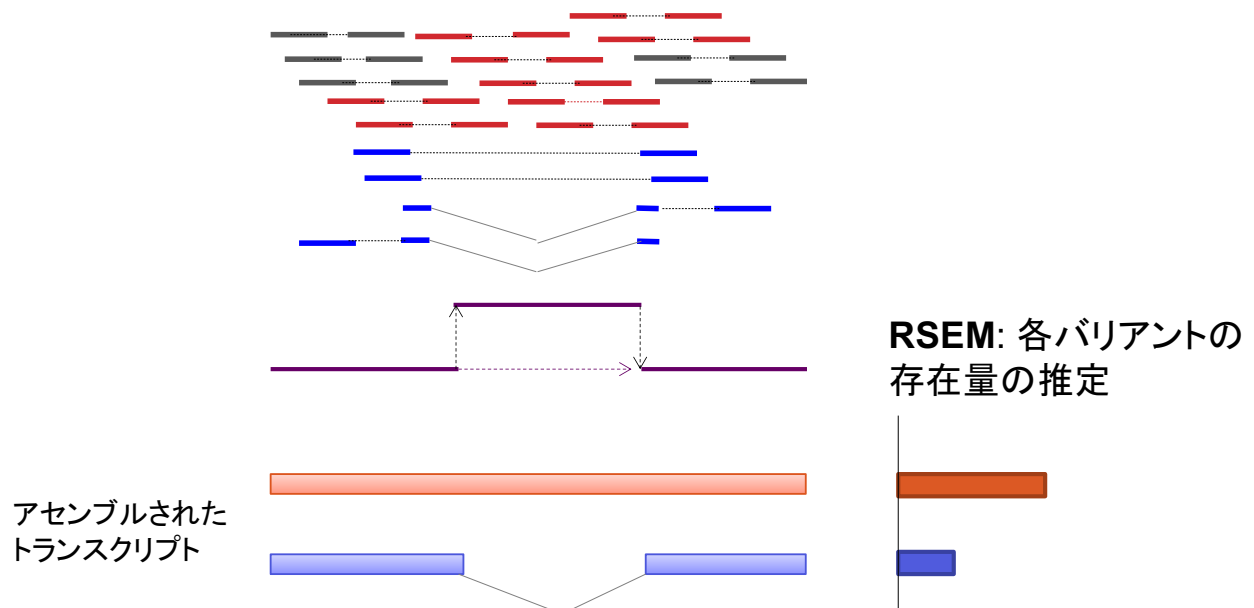


スコア=ミスマッチに
対するペナルティ

- **default (best one mode)**
条件を満たすアライメントを検索し、最高スコアのを1つ出力
(ただし、検索は完全でないため、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)
- **-k <int>**
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と、AとB
(実際には位置の順に見つかるわけではない)
- **-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** や **-a** 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

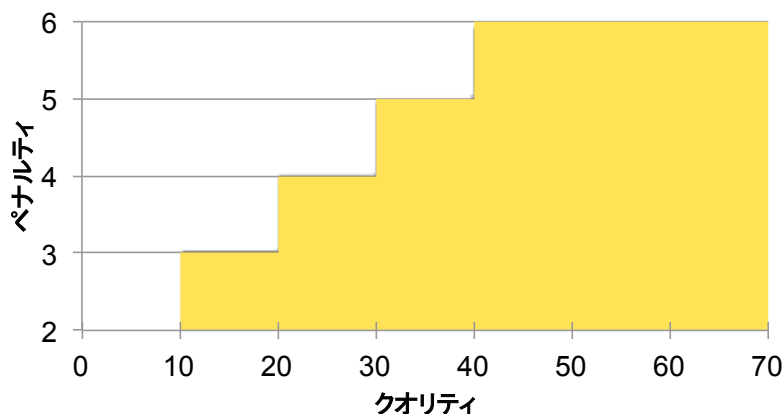
(参考) デノボ・アセンブルによるRNA-Seq解析

デノボ・アセンブルによる転写配列の構築



(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(最高スコアが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 $-0.6(L+1)$



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よい場合もある。

Samtoolsを用いた MAPQによるフィルタリング

- `samtools view -q 閾値 BAMファイル名`

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
```

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: TAACTTGC GTTAAATCCGCCTGG
```

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 は挿入・欠失の長さ)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

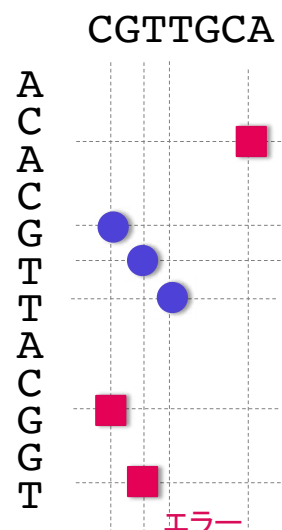
①インデックス作成

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)

CGTTGCA

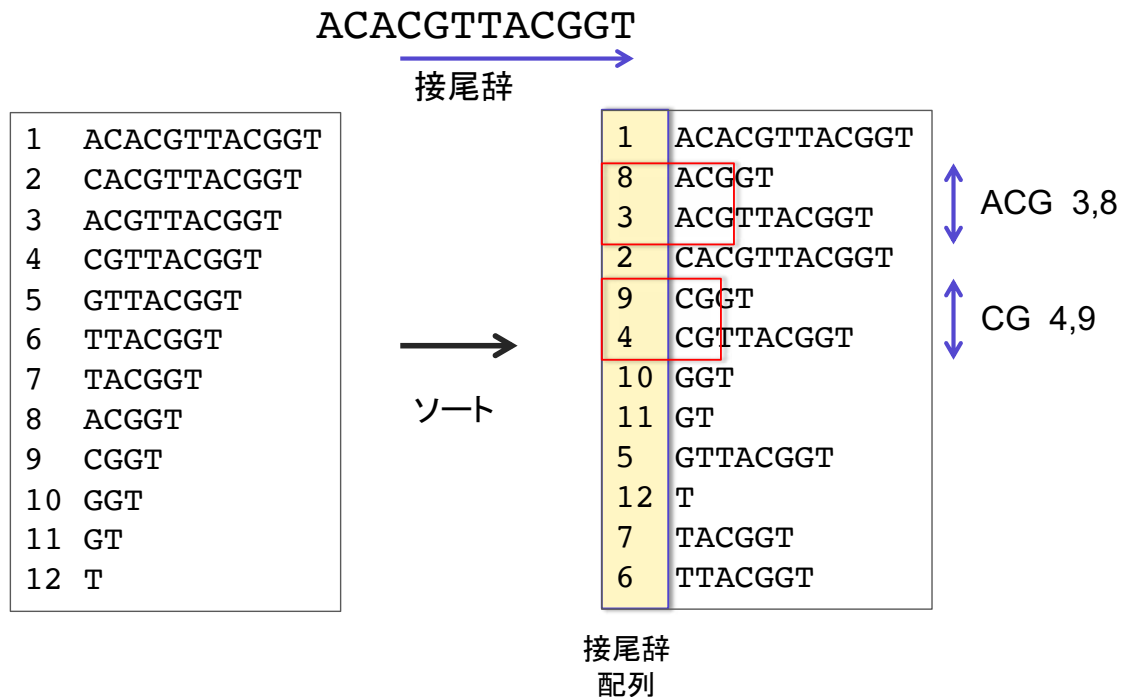


③ 見つかったseedを 延長してアライメント

ACACGTTACGGT.....
CGTTGCA

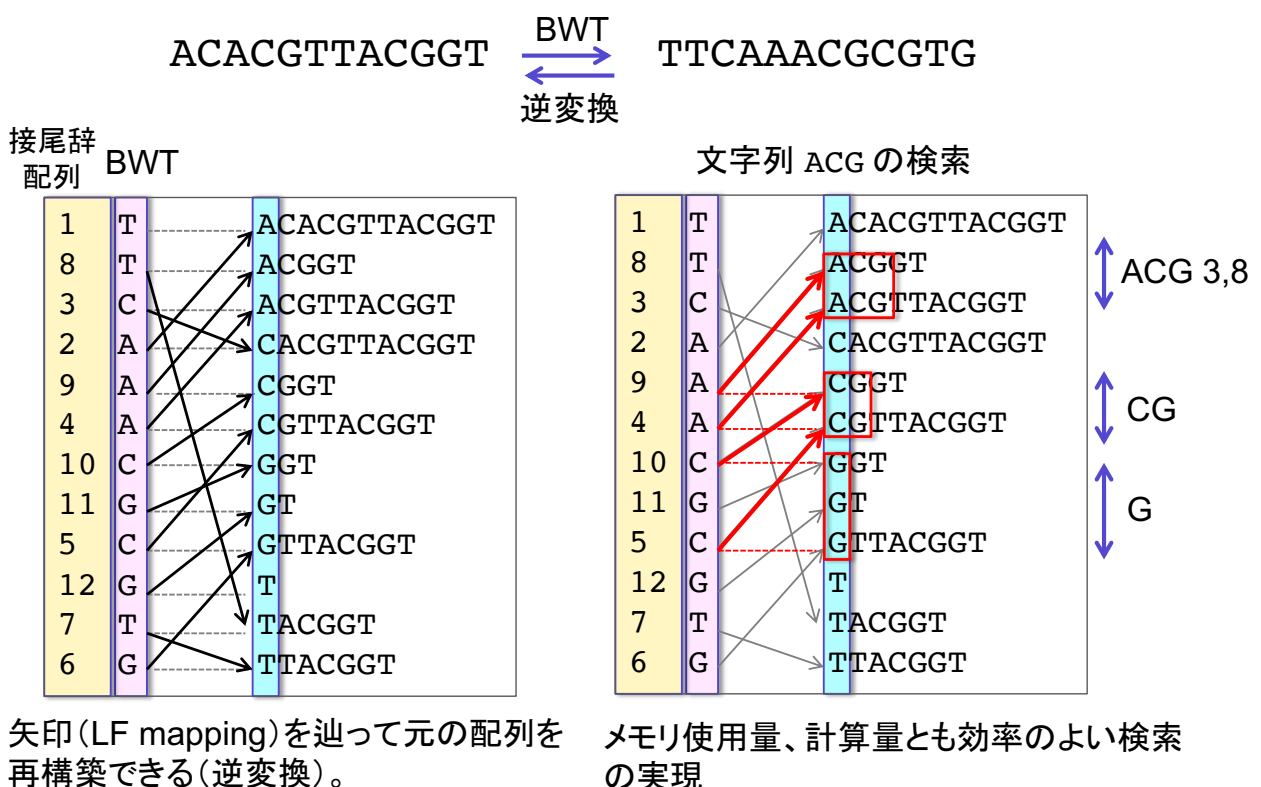
インデックスを使った高速検索

接尾辞配列 (suffix array)

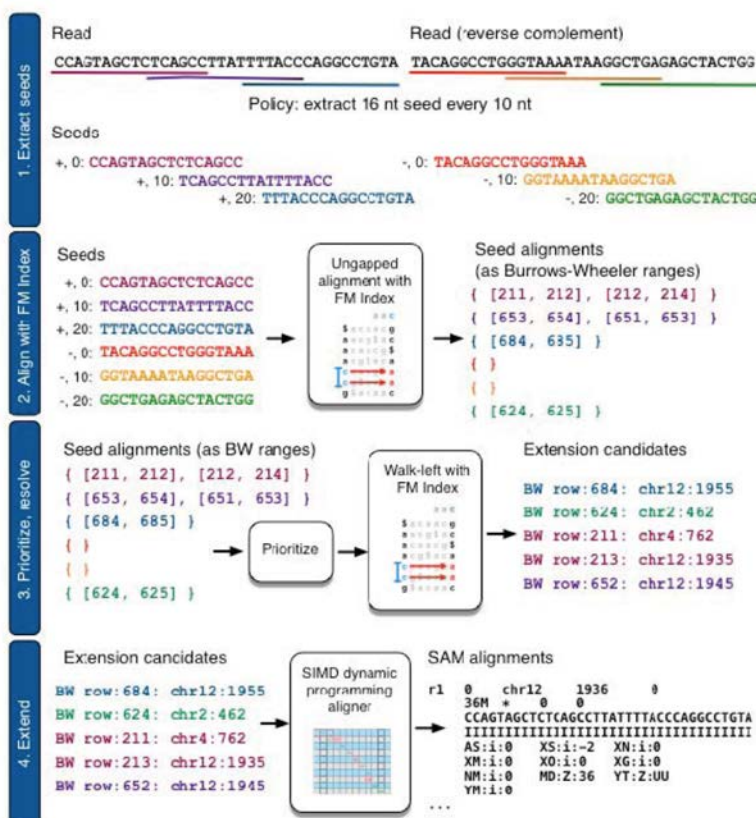


Burrows-Wheeler 変換 (BWT)に基づく

インデックス (FM-Index)



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ま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。

