

NGS 基本フォーマットとツール 補足と復習

基礎生物学研究所
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 (uchiyama@nibb.ac.jp)

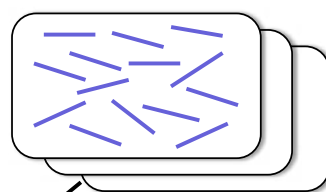
ショートリードのマッピング

ゲノム配列
(リファレンス reference 配列)

形式 (配列)

```
>chr
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAAGAGTGTGATAGCAGC
TTCTGAACCTGGTTACCTGCCGTGAGTAAATTTAA
TTTATTTGACTTACCTGACTTAATACTTTAACCAA
TATAGGCATAGCCGACAGACAGATAAAATACAG
ACTACACAACATCCATGAAACGCAATTAGCACCAC
ATTACCACCACCATTACCATTACCACAGGTAACGG
```

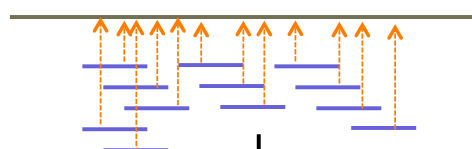
サンプル (ゲノムDNA/RNA)
(リード read 配列)



形式
(配列 + クオリティ値)

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631
ATCCGGCTGGCGCACCACCTATGTTCCGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513
CACCGTGTAGTACCAGCATCCTGCGTACAAATCAGCAATCCAGTCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513
CCCFDFDFHDFHIIIEGIHJJJGPHGGHGGHGIJJDGIJHHGGHHIH
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530
CAGGACATCGCCTTTGATCGGTTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530
CCCFDFDFAFHFHJGHJIJJJJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

リファレンス配列へのマッピング



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 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 GCGTGTCCGTCGCCAGC
SRR1515276.411 0 chr 3440721 42 51M * 0 0 ACGGCATAATTTCTTGA
```

復習: **bowtie2** 用インデックスの作成

実習用ディレクトリ `~/data/IU` に移動

- 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`

- リード配列 (FASTQ 形式; paired-end)

`etec_1.fq`

`etec_2.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の実行 (出力: `etec_bowtie2.sam`)

\$ `bowtie2`

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		リファレンス配列に関する情報	
@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:2.2.7 Cl:"/bio/bin/bowtie2-align-s --wrapper basic-0 -x etec -1 etec_1.fastq -2 etec_2.fastq"	
SRR345261.25	89 ETEC_chr 3758170 1 49M = 3758170 0 ACACGGCGCATGGCTG... ###7ED>EBDBDDE,E...	AS:i:-1 XS:i:-1 XN:i:0 XM	
SRR345261.25	133 ETEC_chr 3758170 0 * = 3758170 0 NNNNNNNNNNNNNNN... #####	YT:Z:UP YF:Z:NS	
SRR345261.50	73 ETEC_chr 4361458 1 49M = 4361458 0 CAACGCTTAATCGGAA... :HEGDFHHHH#BGG=B...	AS:i:0 XS:i:0 XN:i:0 XM	
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 CGGTGGATGCCCTGCG... DDDDBD6DB>BB>	AS:i:-2 XS:i:-2 XN:i:0 XM	
SRR345261.75	133 ETEC_chr 4362922 0 * = 4362922 0 NNNNNNTTTNNNTCGG... #####	YT:Z:UP YF:Z:NS	
SRR345261.100	73 ETEC_chr 679991 42 49M = 679991 0 GTGGTTAATGAGTCC... GGGGGGGGB=ED=EEG...	AS:i:0 XN:i:0 XM:i:0 XO	
SRR345261.100	133 ETEC_chr 679991 0 * = 679991 0 NNNNNNCACGNTAGT... #####	YT:Z:UP YF:Z:NS	
SRR345261.125	73 ETEC_chr 4376280 42 49M = 4376280 0 CTCAGGATGAGGTCAC... EEEE=B<<@BDEEDE:...	AS:i:0 XN:i:0 XM:i:0 XO	
SRR345261.125	133 ETEC_chr 4376280 0 * = 4376280 0 NNNNNTTTCNTTAG... #####	YT:Z:UP YF:Z:NS	
SRR345261.150	89 ETEC_chr 779844 42 49M = 779844 0 TTCAGGAAACCTGAA... B@D>ECC?B@ECC>...	AS:i:-5 XN:i:0 XM:i:1 XO	
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 CCGCTTGGCGGGGCCA... EDE<8??;?@DGGDDE...	AS:i:0 XN:i:0 XM:i:0 XO	
SRR345261.175	163 ETEC_chr 3605113 42 49M = 3605306 242 DGGDFDGGGGGGEGD... AS:i:-3 XN:i:0 XM:i:13 XO		
SRR345261.200	77 * 0 0 * * 0 0 AAAAAAAAAAAAAA... #####	YT:Z:UP	
SRR345261.200	141 * 0 0 * * 0 0 AAAAAAAAAAAAAA... 8@#####	YT:Z:UP	
SRR345261.225	83 ETEC_chr 2879707 1 49M = 2879600 -156 CACACACAGCGTGAC... 8?D8BEBGD@GG8GCE...	AS:i:0 XS:i:0 XN:i:0 XM	
SRR345261.225	163 ETEC_chr 2879600 1 49M = 2879707 156 CCCACCTTCCTCAGT... GGGBGDEGG@GG<G8...	AS:i:-1 XS:i:-1 XN:i:0 XM	
SRR345261.250	99 ETEC_chr 4361346 1 49M = 4361525 228 GTACTTTCAGCGGGGA... ECE=>EC?FDG<EGDA...	AS:i:0 XS:i:0 XN:i:0 XM	
SRR345261.250	147 ETEC_chr 4361525 1 49M = 4361346 -228 CCGGCTCAACCTGGG... #####BC...	AS:i:0 XS:i:0 XN:i:0 XM	

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

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 ~/data/IU

- SAMファイル
etec_bowtie2.sam

- SAMからBAMへの変換 (出力ファイル名: etec_bowtie2.bam)

\$ samtools

- 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

\$ samtools

復習: **BAM**ファイルのインデックスづけ

実習用ディレクトリ ~/data/IU

- **BAMファイル**
etec_bowtie2.bam

- リファレンス配列上の位置の順にソート
(出力ファイル: etec_bowtie2_sorted.bam)

\$ samtools

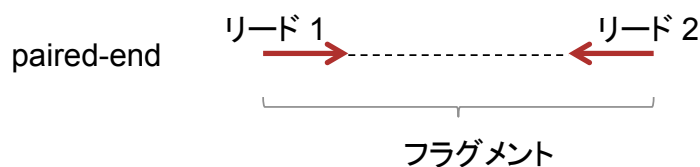
- ソートされたBAMファイルに対してインデックスの作成

\$ samtools

- インデックスを使って、リファレンスの染色体配列 (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

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

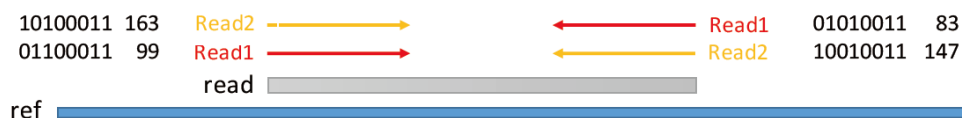
```
% samtools flags 83
```

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

各フラグの説明を表示

```
% samtools flags
```

Paired end readでのFLAG値



	Read2の配列である	Read1の配列である	逆鎖にマップされた	ペア相手は逆鎖にマップされた	自分がマップされていない	ペア相手がマップされていない	両方適切にマップされている	2進数表記	10進数表記
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	1	1	1	1	1	1	1	11111111	255
片方しかアラインしていない場合	0	1	0	0	1	0	0	01010011	83
	0	1	1	0	0	0	1	01100011	99
	1	0	0	1	0	0	1	10010011	147
	1	0	1	0	0	0	1	10100011	163
どっちもアラインしていない場合	0	1	0	0	1	1	0	01001001	73
	0	1	0	1	1	0	0	01011001	89
	0	1	0	0	0	1	0	01000101	69
	0	1	1	0	0	1	0	01100101	101
	1	0	0	0	1	0	0	10001001	137
	1	0	0	1	1	0	0	10011001	153
	1	0	0	0	0	1	0	10000101	133
	1	0	1	0	0	1	0	10100101	165
	0	1	0	0	1	1	0	01001101	77
	1	0	0	0	1	1	0	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、SAM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する)。

- **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、SAM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

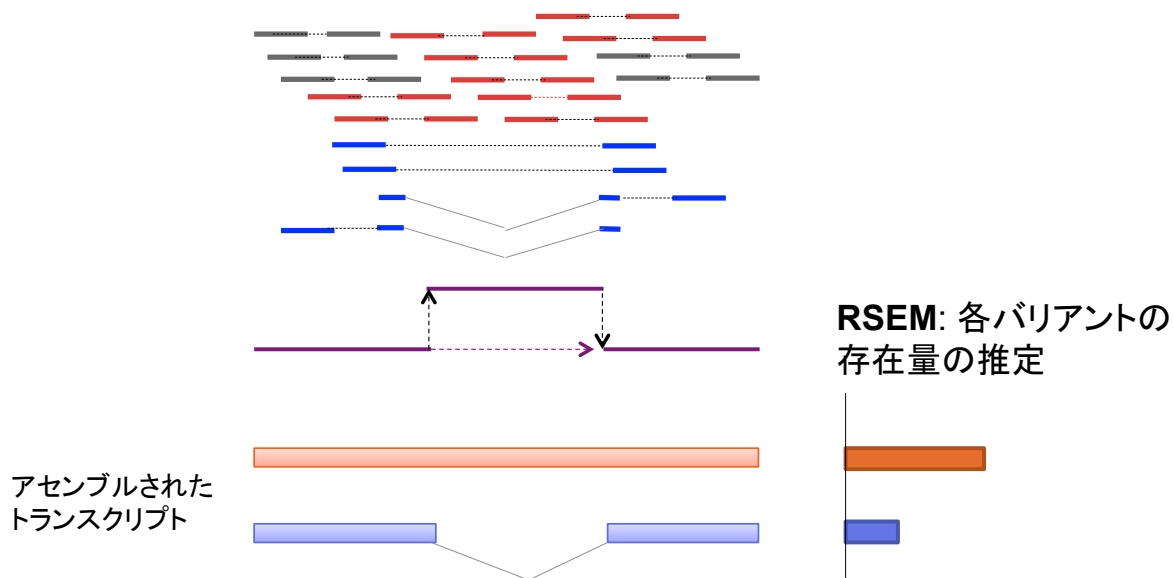
- 一般に、1つのリードは複数の箇所にマップされる。



- **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
- **-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のとすると、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



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: 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

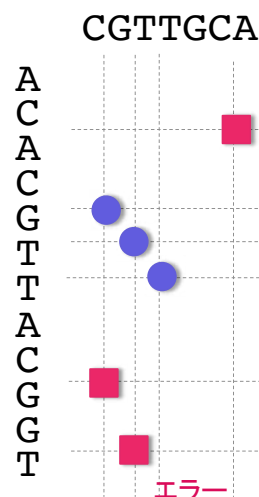
① インデックス作成

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

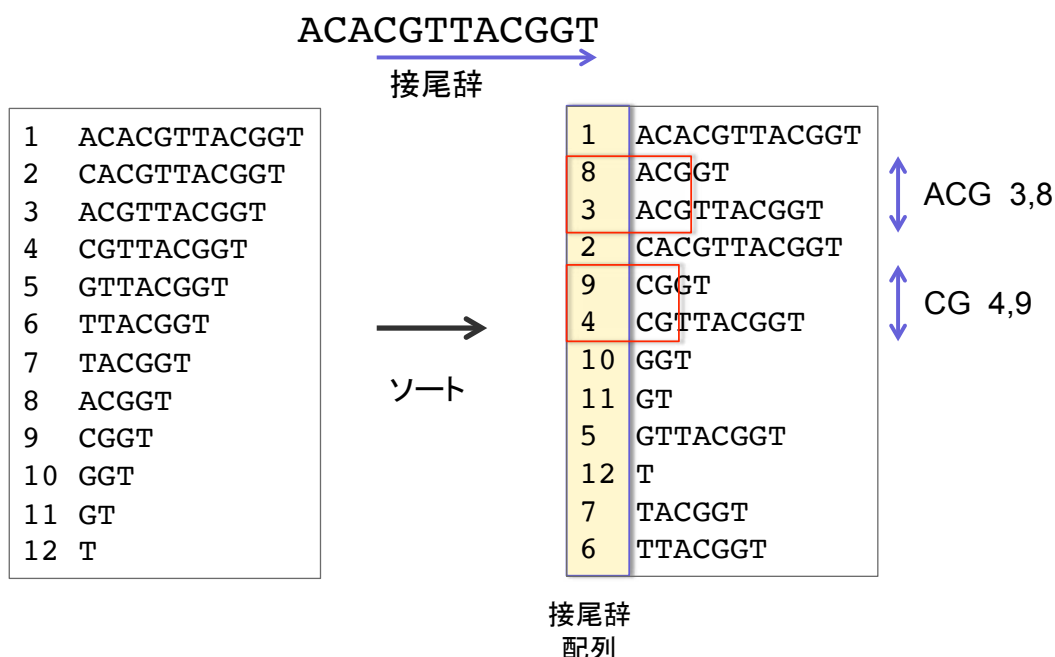


③ 見つかったseedを 延長してアライメント

ACACGTTACGGT.....
CGTTGCA

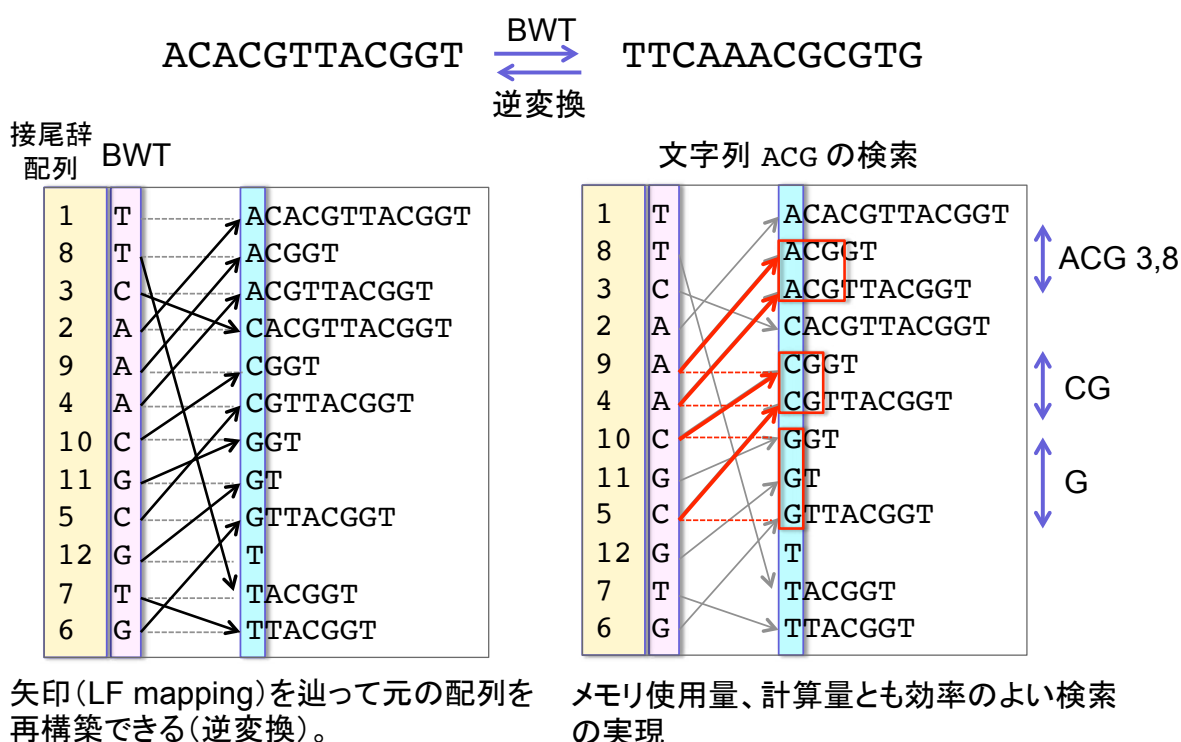
インデックスを使った高速検索

接尾辞配列 (suffix array)

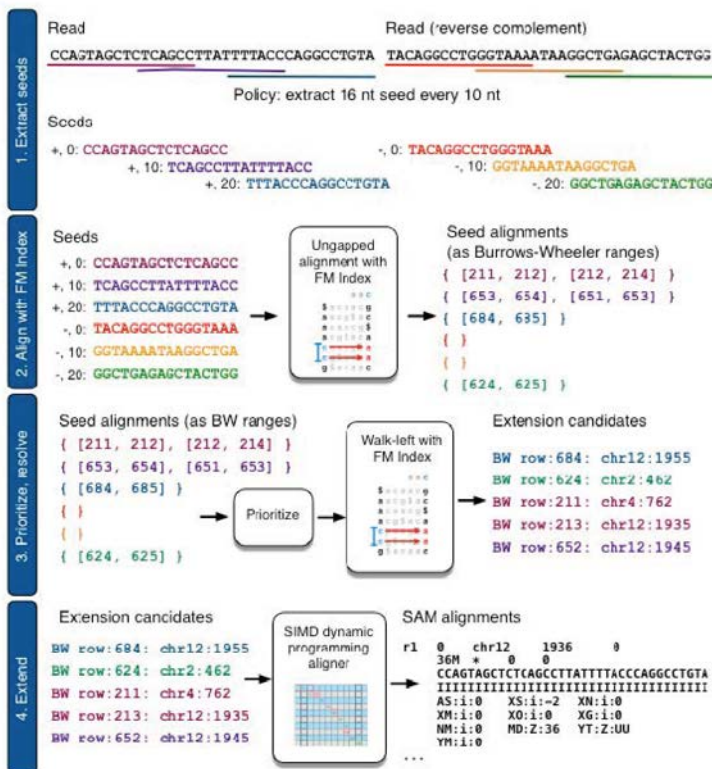


Burrows-Wheeler 変換 (BWT) に基づく

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