

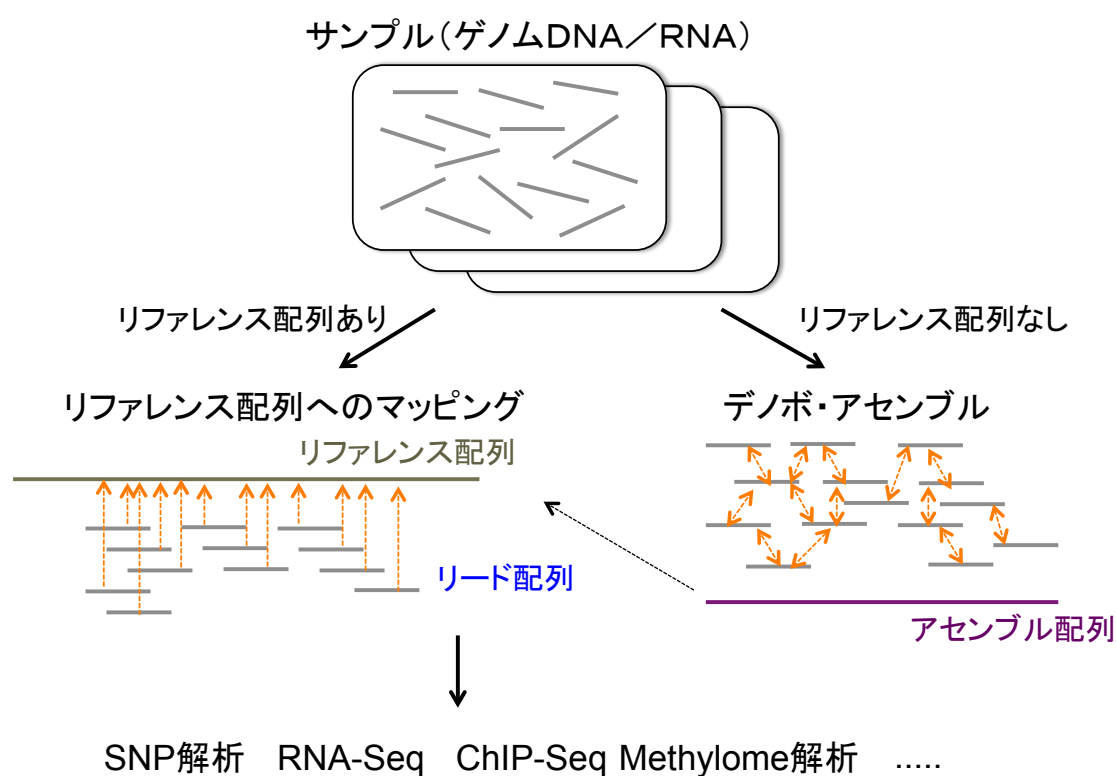
# 次世代シーケンサ用 データ解析コマンド

基礎生物學研究所

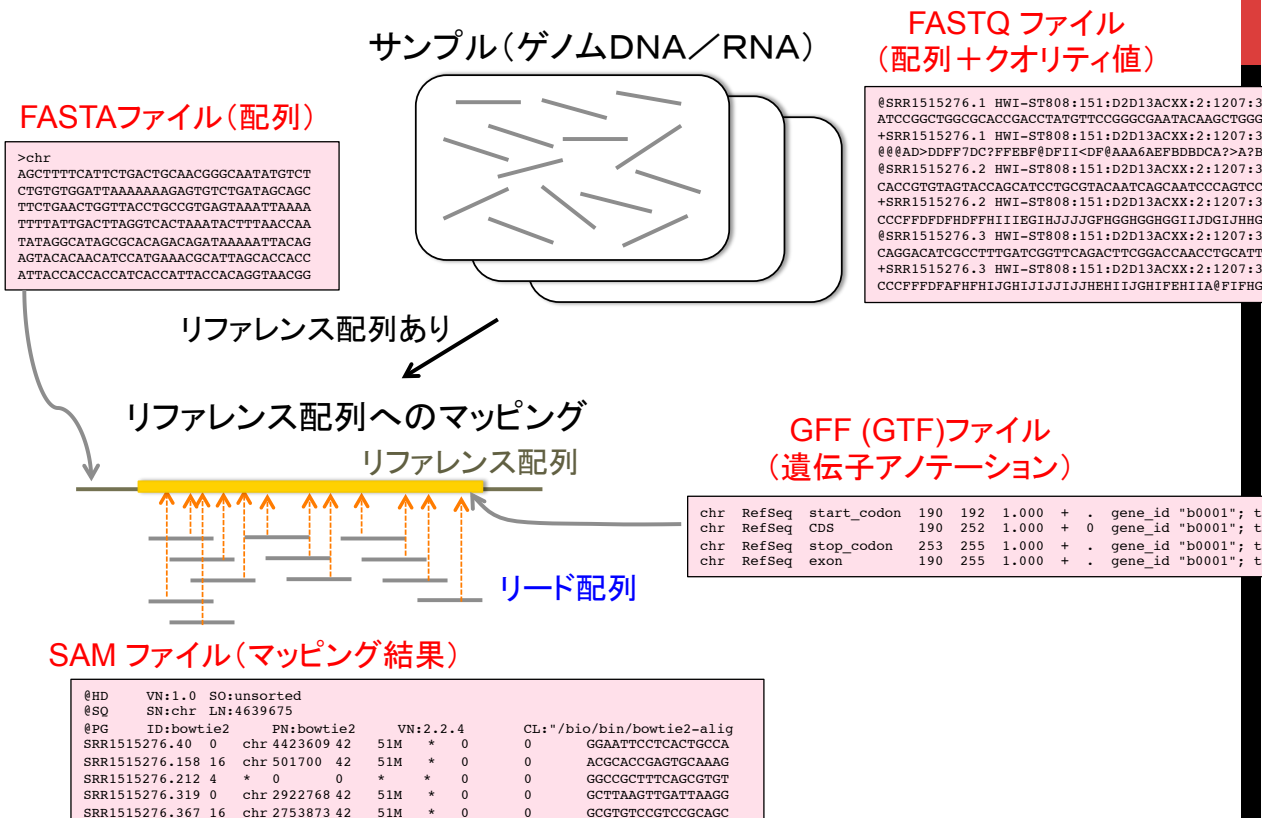
ゲノムインフォマティクストレーニングコース

内山 郁夫 ([uchiyama@nibb.ac.jp](mailto:uchiyama@nibb.ac.jp))

## 次世代シーケンサデータ処理の概要



# 次世代シーケンサデータ処理の概要



## 配列データファイル: FASTA format

```
>NM_174292.2 Bos taurus crystallin gamma S (CRYGS) mRNA
TGCACCAAACATGTCTAAAGCTGGAACCAAAATTACTTTCTTTGAAGACAAAAC
TTTCAAGGCCCGCCACTATGACAGCGATTGCGACTGTGCAGATTTCCACATGTACC
TGAGCCGCTGCAACTCCATCAGAGTGGAAGGAGGCACCTGGGCTGTGTATGAAAG
GCCCAATTTTGCTGG
>NM_174294.1 Bos taurus casein kappa (CSN3) mRNA
TTCACCTACAGTGGAAGGCCAACTGAACCTACTGCCAAGCAAGAGCTGACGGTC
ACAAGGAAAGGTGCAATGATGAAGAGTTTTTTCCTAGTTGTGACTATCCTGGCAT
TAACCCTGCCATTTTTTGGGTGCCAGGAGCAAAACCAAGAACAACCAATACGCTG
TGAGAAAGATGAAAG
```

> 配列ID (説明)  
塩基配列

- 配列ファイルの標準フォーマット
- >で始まる行がタイトル行、その後に配列が続く
- タイトル行の最初の単語が配列ID、以降は説明 (省略可)
- タイトル行の長さに制限はないが、途中で改行は入らない
- 配列は途中で改行が入ってもよい

## 配列データファイル: FASTQ format

@ERR004063.1 IL33_2678:6:1:0:902/2	← @配列ID (説明)
CTGTAAATATGTACAGGATATTTCTGACCATTCTTC	← 塩基配列
+	← + (配列ID 説明)
CCCDCC8@DD8DDCC5?,D7??&=-4&9*&&+-.'.	← クオリティ値

@ERR004063.2 IL33_2678:6:1:0:1059/2	
AAATGGAGACTTATTCCTTGTCTTTGGGTAATCAATC	
+	
@@=@;>CC;<;DD>>7AA>88>DE>AC96@;&<C7>	

@ERR004063.3 IL33_2678:6:1:0:1320/2	
AGCATTTGGAGTGGCTTTTTTTTGTTTTTTTTTTAAA	
+	
07ECEEE=CDCA<AC108D@0(*4(,:0;6*0*(4((	

- 配列とクオリティ値をひとつにまとめたもの
- 各データの先頭は「@」、塩基配列とクオリティ値配列の間に「+」
- クオリティ値は「アスキーコード」に従って文字列で表示される  
→ 数字を使う場合と比べてサイズが圧縮され、高速な処理が可能
- 塩基配列、クオリティ値配列ともに1行で書くのが基本
- @や+で始まる行がタイトル行やクオリティデータ開始行であるとは限らない(@や+はクオリティ値の中にも現れるので)

## クオリティ値

- PHRED Quality: エラー率  $p_e$  に対して、以下で定義される

$$Q_{PHRED} = -10 \log_{10}(p_e)$$

例)  $Q_{PHRED}=30$  の場合、エラー率は  $10^{-3}$

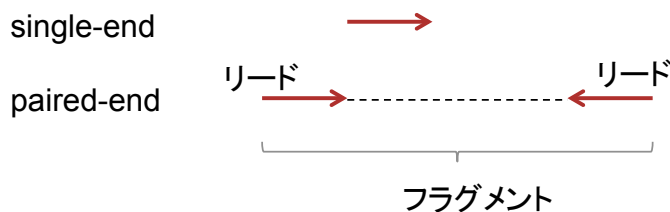
- $Q_{PHRED}+33$  の値が、以下のアスキーコード表に従って文字列として表示される

	30	40	50	60	70	80	90	100	110	120
0:		(	2	<	F	P	Z	d	n	x
1:		)	3	=	G	Q	[	e	o	y
2:	(SP)	*	4	>	H	R	\	f	p	z
3:	!	+	5	?	I	S	]	g	q	{
4:	"	,	6	@	J	T	^	h	r	
5:	#	-	7	A	K	U	_	i	s	}
6:	\$	.	8	B	L	V	'	j	t	~
7:	%	/	9	C	M	W	a	k	u	(DEL)
8:	&	0	:	D	N	X	b	l	v	
9:	'	1	;	E	O	Y	c	m	w	

## リファレンス配列へのマッピング

### Bowtie, BWA, SOAP などのコマンド

- 長大なリファレンス配列に大量の短いリード配列を若干のミスマッチを許して照合する
- リファレンス配列に対して、あらかじめ全文検索インデックスを作成することにより高速に検索を行う
- paired-end read に対応。insert sizeに制約をつけられる



## Bowtie コマンド

- BowtieとBowtie2がある。後者はギャップを考慮した検索を行うので、感度がより高い

- インデックスの作成

```
bowtie2-build 配列ファイル インデックス名
```

- マッピングの実行

(single-end read の場合)

```
bowtie2 -x インデックス名  
          -u リードファイル -s 出力ファイル
```

(paired-end read の場合)

```
bowtie2 -x インデックス名  
          -1 リード1 -2 リード2 -s 出力ファイル
```

(改行せずに1行で打つ)

## 実習

データ: 大腸菌RNA-Seqデータ(GEO:SRP044366)の一部を抜粋したもの

ディレクトリ: `~/data/2_ngs`

リードファイル: `test_fastq/ecoli.[1-12].fastq`

大腸菌ゲノム配列: `ecoli_genome.fa`

大腸菌遺伝子テーブル: `ecoli.gtf`

- 作業ディレクトリに移動

```
$ cd ~/data/2_ngs
```

## Bowtie実習

- リファレンス配列(`ecoli_genome.fa`)に対してインデックスを作成する

```
$ bowtie2-build ecoli_genome.fa ecoli_genome
```

- リードファイル `test_fastq/ecoli.1.fastq` をリファレンス配列上にマッピングする

```
$ bowtie2 -x ecoli_genome  
          -U test_fastq/ecoli.1.fastq  
          -S ecoli.sam
```

(改行せずに1行で打つ)

# マッピング結果ファイル: SAM format

- リファレンス配列に多数の配列をマップした結果を表す形式
- 最初の@付きの行はヘッダ行、以降はタブ区切りで記述されたマッピング結果のデータ

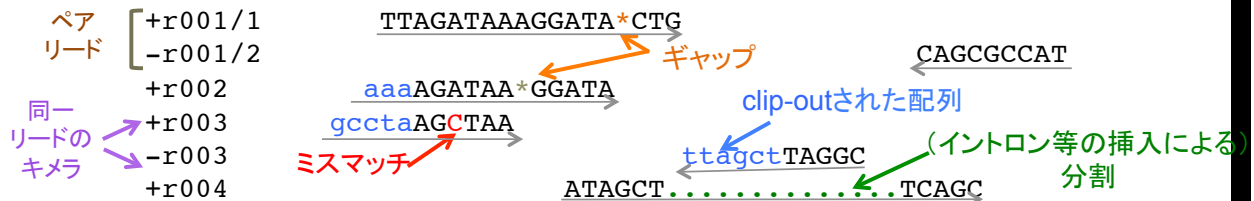
```
@HD VN:1.3 SO:coordinate
```

```
@SQ SN:ref LN:45
```

```
r001 163 chr1 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 chr1 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 chr1 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 chr1 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 chr1 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 chr1 37 30 M = 7 -39 CAGCGCCAT *
```

テンプレート名    フラグ    マップ結果    アライメント (CIGAR)    対となるリードの位置情報    リードの配列    オプション

リファレンス配列    chr1    12345678901234    5678901234567890123456789012345  
AGCATGTTAGATAA\*\*GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT



# マッピング結果ファイル: SAM format

ヘッダ行

@HD VN: バージョン番号 SO:並び順

@SQ SN: リファレンス配列名 LN:リファレンス配列長

@RG リードグループの情報、@PG プログラムの情報

```
@HD VN:1.3 SO:coordinate
```

```
@SQ SN:ref LN:45
```

```
r001 163 chr1 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 chr1 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 chr1 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 chr1 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 chr1 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 chr1 37 30 M = 7 -39 CAGCGCCAT *
```

フラグ    フラグ    マップ結果    アライメント (CIGAR)    対となるリードの位置情報    リードの配列    オプション

SAM形式各カラムの情報

1. QNAME: クエリー配列名
2. FLAG: 各種情報を保持したフラグ
3. RNAME: リファレンス配列名
4. POS: リードがマッピングされた左端位置
5. MAPQ: マッピング クオリティ  $-10\log_{10} p_e$
6. CIGAR: CIGAR 文字列(アライメント)

7. RNEXT: ペアの相方が載ったリファレンス配列名 (=は同一配列、\*は情報なし)
8. PNEXT: ペアの相方のマップ位置
9. TLEN: フラグメントの長さ
10. SEQ: リードの配列
11. QUAL: リードのクオリティ(アスキー)
12. OPT: tag:type:value で表す任意の情報

## BAM format

- SAM形式のテキストファイルをバイナリファイルに変換し、かつデータ圧縮をかけてコンパクトかつ効率的に処理できるようにしたもの。
- samtoolsを使って変換する
  - SAM→BAM変換

```
samtools view -Sb file.sam > file.bam
```

    - -s 入力がSAMファイル; -b 出力がBAMファイル
  - BAM→SAM変換

```
samtools view -h file.bam > file.sam
```

    - -h 出力にヘッダを含める

## samtools

- SAM/BAM形式のマッピング結果ファイルを扱うためのコマンド群

```
samtools <command> [options]
```

2番目の引数の<command>によって機能を切り替える

- view SAM↔BAM 変換して表示
- sort 並べかえ(デフォルトはマップされた位置で)
- index ソートされたBAMファイルをインデックスづけ
- idxstats 各リファレンス配列にマップされたリード数を表示
- depth リファレンスの各位置にマップされたリード数を表示
- mpileup リファレンスの各位置にマップされた塩基を表示

## samtools 実習1

- BAMファイルの作成

```
$ samtools view -bS ecoli.sam > ecoli.bam
```

- BAMをSAMに変換して表示

```
$ samtools view ecoli.bam |less
```

## samtools 実習2

BAMファイルは、ソートしてインデックスづけを行うことによって、より便利に使えるようになる。

- BAMファイルをソート

```
$ samtools sort ecoli.bam ecoli_sorted
```

ソート後のファイル名は、最後の引数に.bamが付加されたものになる

- ソートされたBAMファイルをSAMに変換してlessで表示

```
$ samtools view ecoli_sorted.bam |less
```

- ソートされたBAMファイルに対してインデックスを作成

```
$ samtools index ecoli_sorted.bam
```



## samtools 実習3

以下は、インデックスづけされたBAMファイルを使う

- 指定した領域内にマッピングされたリードのみを表示

```
$ samtools view ecoli_sorted.bam chr:200-500
```

染色体名:開始位置—終了位置

- 各染色体にマッピングされたリード数を表示

```
$ samtools idxstats ecoli_sorted.bam
```

マッピングされなかったリードは、染色体名が '\*' として表示される

- mpileupコマンドで、リファレンスの位置ごとにマッピングされた塩基を表示

```
$ samtools mpileup -f ecoli_genome.fa ecoli_sorted.bam | less
```

## RNA-Seq 解析

- ゲノム上にマッピングされたリードを遺伝子領域ごとに集めて数をカウント



通常、カウントした数を遺伝子の長さ、およびマップされたリード全体の数で割って標準化する

RPKM (Read Per Kilobase per Million mapped reads) または

FPKM (Fragment Per Kilobase per Million mapped reads)

## 遺伝子アノテーションファイル: **GFF format**

- ゲノム上の特徴配列(Feature segment)をタブ区切りテキスト形式で表現する標準的なフォーマット
- 最後のカラムに任意の情報を格納できるため、様々な特徴配列の情報を記述できるが、とくに遺伝子構造を記述するために特化した形式を **GTF format** という



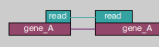


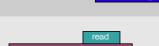
1	2	3	4	5	6	7	8	9
chr	RefSeq	start_codon	190	192	1.000	+	.	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	CDS	190	252	1.000	+	0	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	stop_codon	253	255	1.000	+	.	gene_id "b0001"; transcript_id "b0001";
chr	RefSeq	exon	190	255	1.000	+	.	gene_id "b0001"; transcript_id "b0001";

- |                             |                              |
|-----------------------------|------------------------------|
| 1. Seqname 配列名              | 6. Score スコア、省略可 (.)         |
| 2. Source 予測プログラム、データベース名など | 7. Strand スtrand(+/-)、省略可(.) |
| 3. Feature 特徴セグメントの種類       | 8. Frame 読み枠(0/1/2)、省略可(.)   |
| 4. Start開始位置                | 9. Attribute (Optional)      |
| 5. End 終了位置                 | セミコロンで区切られたタグ-値の対            |

## マッピング結果を領域ごとに 集計するコマンド: **htseq-count**

- **htseq-count** マッピングファイル (SAM) 遺伝子ファイル (GFF)

3つの照合モード: union, intersection\_strict, intersection\_nonempty

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

<http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>

## htseq-countを用いた実習

- アノテーションテーブルecoli.gtfを使って、各遺伝子にマッピングされたリード数をカウント

```
$ htseq-count ecol_i.sam ecol_i.gtf > ecol_i.htseq
```

- 遺伝子アノテーション(GFF)ファイルの中で、どのFeatureの行を使うかはオプション `-t` で、遺伝子IDとして何を使うかについては `-i` で指定する。
- 標準的なGTF形式のファイルであれば、デフォルトのまま(Featureが `exon` の行を使い、遺伝子IDは `gene_id` を使う)で動作するようになっている。

## 出力結果

b0001	11
b0002	117
b0003	33
b0004	44
b0005	3
b0006	14
b0007	4
b0008	181

## 今回使ったツールとファイルのまとめ

