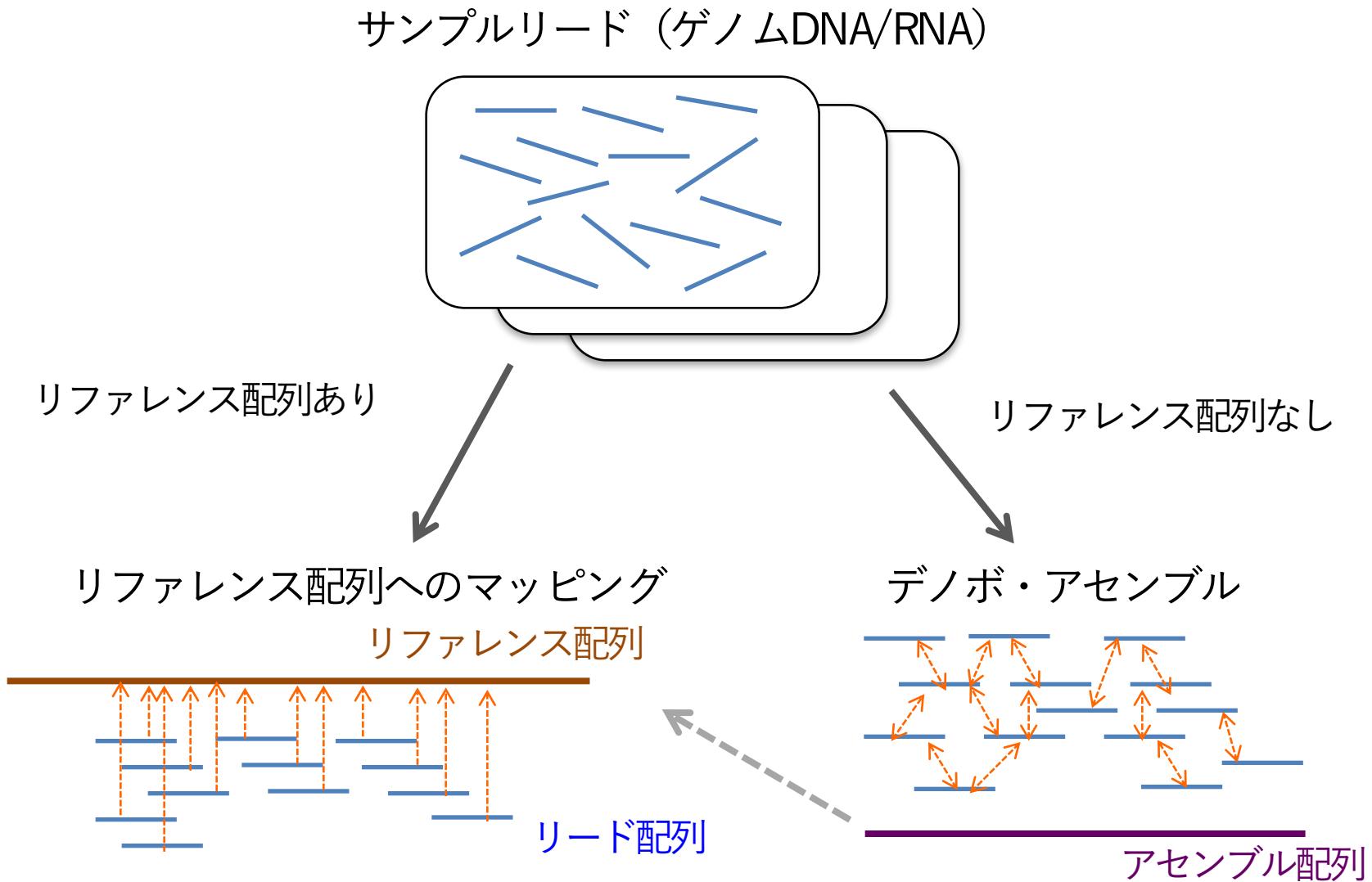


NGS基本ツール

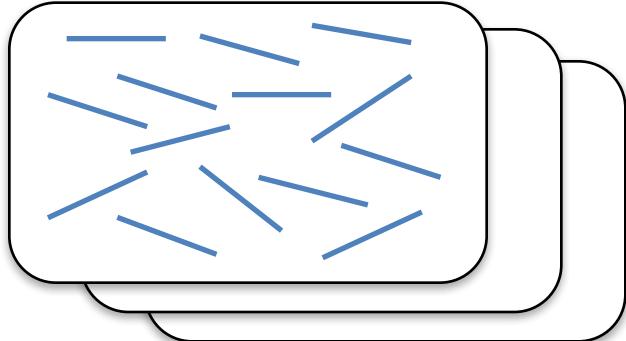
Bowtie2
SAMtools
HT-Seq

基礎生物学研究所 情報管理解析室
西出 浩世 

次世代シーケンサデータ処理の概要



サンプルリード（ゲノム DNA/RNA）



リファレンス配列あり

リファレンス配列へのマッピング

FASTQファイル (配列+クオリティ)

遺伝子アノテーション GFF(GTF)ファイル

```
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";
```

ゲノム（リファレンス）配列 FASTAファイル

```
>chr
ACGTCATTCCTGACGCCAAAGGGCAATAAGCT
CCTGCGGATTTAAAAAGTGCTGCTCATAGTAC
TCTCAACCTGCTCCCTGCTGAGTAATTBAAA
TTTATGCTTACGGCTCTAAATCTTTAACCAAA
TATGACCCTGGCTGGCTGATATAAAATTCAG
AGTACACACACCCAAACGCTTACGCCACACC
ATTACCAACCCATACCTTACCAACGGTACGG
```

CHD	VN:1.0	SO:unsorted									
ESQ	SN:chr	IN:4639675									
@FG	ID:bowtie2	RN:bowtie2	VN:2.2.4	CL:"/bio/bin/bowtie2-alig							
SRR1515276.40	0	chr	4423609	42 51M * 0 0 0	GGATTCCTTACCTCA						
SRR1515276.158	16	chr	501700	42 51M * 0 0 0	ACCCACCGAGCTTAAG						
SRR1515276.212	4	*	0	0 * * 0 0 0	CCCCCTTCATCCTGCT						
SRR1515276.319	0	chr	2922768	42 51M * 0 0 0	CCTTAAAGTTAATTAAG						
SRR1515276.367	16	chr	2753873	42 51M * 0 0 0	CCGCGTGTGCCCCPACC						
SRR1515276.411	0	chr	3440721	42 51M * 0 0 0	ACCCGATTAATTCTICA						
SRR1515276.434	0	chr	4198737	42 51M * 0 0 0	CCCCGGTACCCAGTCGG						

マッピング結果 SAM ファイル

ショートリードマッピング用ツール

- ハッシュテーブルによるインデックス
 - リード配列にインデックスづけ
 - MAQ, RMAP など
 - リファレンス配列にインデックスづけ
 - SSAHA2, MOSAIK, NovoAlign など
- Burrows Wheeler 変換に基づくインデックス (FM index)
 - リファレンス配列にインデックスづけ
 - Bowtie/Bowtie2, BWA, SOAP など
- splice-aware aligner
 - spliced exon からなるリードも正確にマッピングできる
 - TopHat, HISAT, STAR, GSNAP など
 - (TopHatやHISATはBowtieを内部的に使っている)

インデックス？



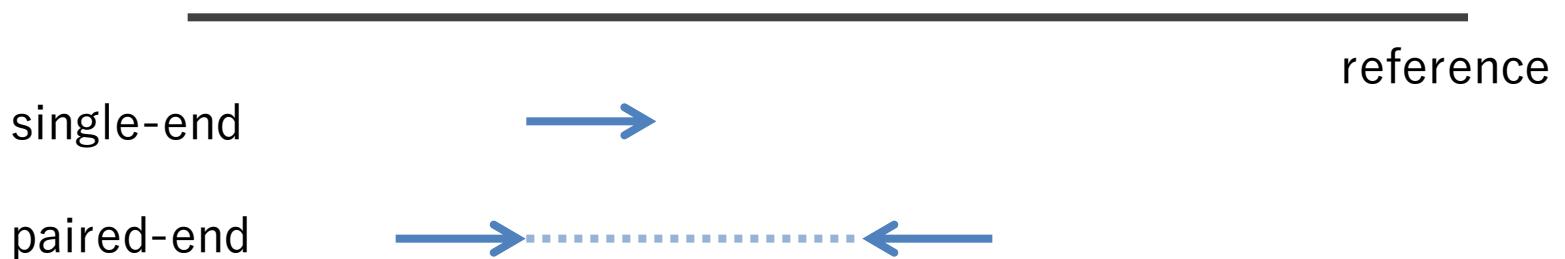
辞書における
インデックスタブ

- 索引、目次、見出し
- ファイルのどの辺りに何が書いてあるかの指標
- インデックスを作成すると別ファイルができるのは、分厚い本の「別冊目次」ができるイメージ
- 欲しい情報を探すのにファイル（本）を先頭から総ナメして探さなくてもよい

リファレンス配列へのマッピング

Bowtie, BWA, SOAP など

- 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

- Burrows-Wheeler 変換に基づくインデックスを利用したショートリードのマッピングプログラム
- BowtieとBowtie2がある。後者はギャップを考慮した検索を行い、感度がより高い。また、検索の方針が単純化されて分かりやすくなるなど、多くの点で改良されている。
- シーケンスのリード長が長い（50bp以上）時はBowtie2の方が一般に検索効率がよく、精度も高い。リード長が短い（50bp未満）時はBowtieの方が検索効率または精度がいい場合もある。



JOHNS HOPKINS
UNIVERSITY

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).



Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.



リファレンス配列のインデックスを作成

bowtie2-build

```
bowtie2-build リファレンス配列ファイル インデックス名
```

- 実行すると、インデックスとして、
 - ✓ インデックス名.n.bt2 (n=1-4)
 - ✓ インデックス名.rev.m.bt2 (m=1-2)の、計 6 つのファイルが作成される
- 配列ファイルはカンマで区切って複数を指定可能

Bowtie2実習 1 : **bowtie2-build**

実習用ディレクトリ `~/data/4_ngs` に移動して `ls` で中を見る

```
$ cd ~/data/4_ngs  
$ ls
```

- リファレンス用ゲノムデータ (FASTA形式) `ecoli_genome.fa`
- bowtie2用インデックスの作成 (インデックス名: `eco`)

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

- インデックスから元の配列データを再構築

```
$ bowtie2-inspect eco | less
```

マッピングの実行 bowtie2

- 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 リードファイル1
-2 リードファイル2 -S 出力ファイル
(実際は改行せずに 1 行で打つ 😈)
- リードファイルはカンマ区切りで複数を指定可能

Bowtie実習2 : **bowtie2**

- リード配列 (FASTQ 形式, single-end read)

ecoli.fastq

リファレンス配列のインデックス名 (実習1で作ったもの)

eco

- bowtie2の実行 (改行せず実行すること 

```
$ bowtie2 -x eco -U ecoli.fastq  
-S eco_bowtie2.sam
```

マッピング結果：SAMフォーマットファイル

```
$ less -S eco_bowtie2.sam
```

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 chr1 7 30 8M2I4M1D3M = 37 39 TTAGATAAAAGGATACTG *
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

Bowtie2: その他のオプション

- **-h** ヘルプを表示する
- **-a** 全てのアライメントを表示する
- **-p 整数** 指定した数のCPUコアを使って実行する
- **-f** リードがFASTA形式のファイルである
- 他、Bowtie2マニュアル詳細

<http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml>

Samtools

Samtools

Home

Download ▾

Workflows ▾

Documentation ▾

Support ▾

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

SAM/BAMフォーマットのマッピング結果を扱うコマンド群

1. フォーマット変換

SAM（テキスト） ⇄ BAM（バイナリ）の変換

2. データのソート, 索引付け

3. データ抽出

特定のリードの選出

統計情報収集（発現量解析）

Samtools の起動

\$ samtools

```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.3 (using htslib 1.3)

Usage:   samtools <command> [options]

Commands:
-- Indexing
dict          create a sequence dictionary file
faidx         index/extract FASTA
index         index alignment

-- Editing
calmd         recalculate MD/NM tags and '=' bases
fixmate       fix mate information
reheader      replace BAM header
rmdup         remove PCR duplicates
targetcut     cut fosmid regions (for fosmid pool only)
addreplacerg  adds or replaces RG tags
```

オプション/引数無しで起動すると Samtools の基本的な使い方が表示される

基本的な使い方： \$ samtools *command options*

Samtools の起動: コマンド簡易マニュアル

```
$ samtools view
```

```
Usage: samtools view [options] <in.bam>|<in.sam>|<in.cram> [region ...]
```

Options:

```
-b      output BAM
-C      output CRAM (requires -T)
-l      use fast BAM compression (implies -b)
-u      uncompressed BAM output (implies -b)
-h      include header in SAM output
-H      print SAM header only (no alignments)
-c      print only the count of matching records
-o FILE output file name [stdout]
-U FILE output reads not selected by filters to FILE [null]
-t FILE FILE listing reference names and lengths (see long help) [null]
-L FILE only include reads overlapping this BED FILE [null]
-r STR  only include reads in read group STR [null]
```

- コマンドを付けてオプション無しで実行するとそのコマンドのマニュアルが表示される
- 詳細は <http://www.htslib.org/doc/samtools.html> を参照のこと

SAM/BAM 変換

samtools view options...

- SAMファイルからBAMファイルの作成

```
$ samtools view -bS eco_bowtie2.sam -o eco_bowtie2.bam
```

- BAMをSAMに変換して less コマンドで表示

```
$ samtools view eco_bowtie2.bam | less
```

- BAMファイルを less で読もうとすると...?

```
$ less eco_bowtie2.bam
```

- SAMファイルに比べてBAMファイルのサイズは?

```
$ ls -l eco_bowtie2.*
```

BAM ファイルのソート

`samtools sort options...`

- ソート

- マッピングデータをリファレンス配列上の位置順に並び替える
- 位置順のキー：
 - 染色体
 - アライメントの先頭の塩基の位置

```
$ samtools sort eco_bowtie2.bam -o  
eco_bowtie2_sorted.bam
```

- ソートされたBAMファイルをSAMに変換してlessで表示

```
$ samtools view eco_bowtie2_sorted.bam | less
```

- 元のSAMファイルの表示と比較

```
$ less eco_bowtie2.sam
```

ソートされたBAM (SAM) ファイルに インデックス (索引) を付ける

`samtools index options...`

- 先にソートされている必要がある
- インデックスは .bai という拡張子付きの別ファイルで生成される。
- 「bamファイル名.bai」が作成されたのを ls コマンドで確認

```
$ samtools index eco_bowtie2_sorted.bam
```

```
$ ls eco_bowtie2_sorted*
```

ここから先はソート & インデックス付与したbamファイルを使う

ソート & インデックス付与したbamファイルを使って

指定した領域内のマッピング結果を表示

```
$ samtools view eco_bowtie2_sorted.bam chr:200-500
```



染色体名：開始位置 - 終了位置

マッピング統計情報収集 1

samtools idxstats *options...*

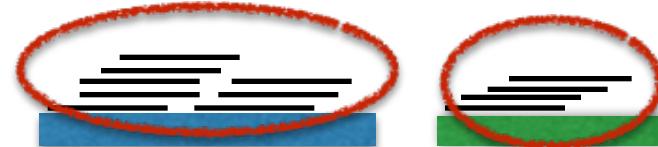
- 染色体毎にマップされたリード数を得る

```
$ samtools idxstats eco_bowtie2_sorted.bam
```

染色体名	染色体配列長	マップされた リード数	片側のみマップさ れたリード数
chr	4639675	326754	0
*	0	0	3364



マップされなかったリード数
染色体名が "*" として表示される



マッピング統計情報収集 2

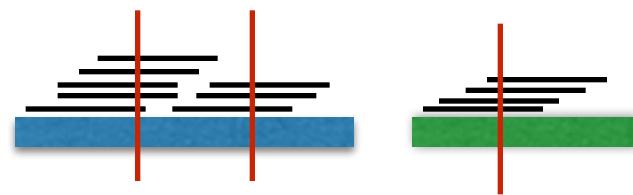
samtools depth *options...*

- 深度（マップされた回数）の統計情報を得る

```
$ samtools depth eco_bowtie2_sorted.bam
```

染色体名	位置	深度（マップされた回数）
------	----	--------------

chr	2753929	1533
chr	2753930	1470
chr	2753931	1446
chr	2753932	1101
chr	2753933	922
chr	2753934	918



マッピング統計情報収集 3

samtools mpileup *options...*

- 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  
eco_bowtie2_sorted.bam
```

chr	24242	G	14	GFIDIIGI@IEHG
chr	24243	A	15	.\$.,\$,,.....^K,	DEI=IIDID@DEIGC
chr	24244	G	14	,\$,.....^K.	EDIIIH@GEFDAE
chr	24245	C	14	,,.....^K.	DIIGIDDGBHDCHB
chr	24246	C	15	,,.....^K,^K,	EIIIEAE<GGH:BE

- bcftools と組み合わせてSNPコールに使える
- リファレンスを指定しないと表示は変わる

mpileup の読み方

記号	意味	
. (ピリオド)	純鎖 リファレンスと一致	
, (カンマ)	逆鎖 リファレンスと一致	
ACGTN	純鎖 リファレンスと不一致	
acgtn	逆鎖 リファレンスと不一致	
+数字 (数字と同数のACGTNまたは acgtn)	挿入部位	+1aは逆鎖に"a"が1塩基挿入。 +2ATは順鎖に"AT"が2塩基挿入。 部位はこの塩基と、次の塩基の間。
-数字 (数字と同数のACGTNまたは acgtn)	欠損部位	+1Aは順鎖"a"が1塩基欠損。 -atは逆鎖に"at"が2塩基欠損。
^とその後の一文字	リードの開始位置と マップ品質値	^~は"~"文字がASCIIコードの126番目なので、 126-33=93がマップ品質値。
\$	リードの終了	
* (アスタリスク)	短い欠損の最中	近傍に2以上の欠損があり、 近傍の欠損に含まれる場合。 例えば-2ATの次の位置には"**"が含まれる。
<または>	長い欠損の最中	"**"とほぼ同じ意味だが、長い欠損の場合"<または>" となる。

リファレンスファイルのインデックス作成

`samtools faidx options...`

```
$ samtools faidx ecoli_genome.fa
```

- リファレンスのインデックスは、mpileup を実行する際に使われる
- 明示的に実行しなくてもインデックスがなかったら自動で実行される
- リファレンス配列ファイル名.fai というインデックスファイルが作られる

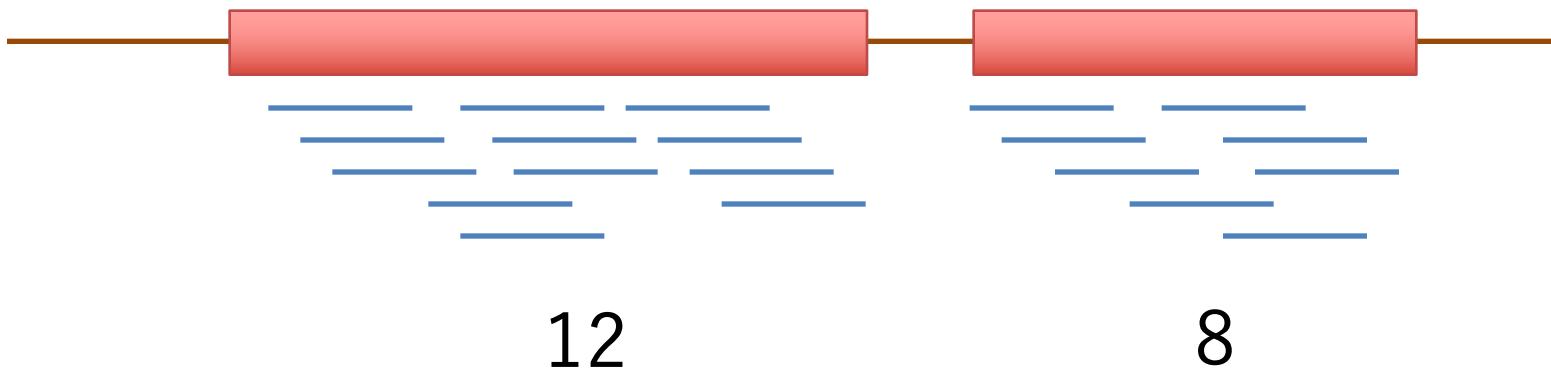
Samtools まとめ

samtools

view	リードを抽出, SAM/BAM変換
sort	ソート
index	.bamのインデックス作成
idxstats	染色体毎のマッピング状況
depth	位置毎のマッピング深度
mpileup	位置毎にマッピングされた塩基を表示
faidx	リファレンスのインデックス作成

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)

FPKM (Fragment Per Kilobase per Million mapped reads)

HTSeq: SAM, BAM を処理するコマンド群

- マッピング結果を指定した領域ごとに集計するコマンド：

htseq-count

htseq-count -f マッピングファイルのフォーマット
マッピングファイル(SAM or BAM) 遺伝子ファイル(GFF or GTF)

- **-f フォーマット** sam または bam (default: sam)

HTSeq実習：htseq-count

- 遺伝子情報ファイル 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

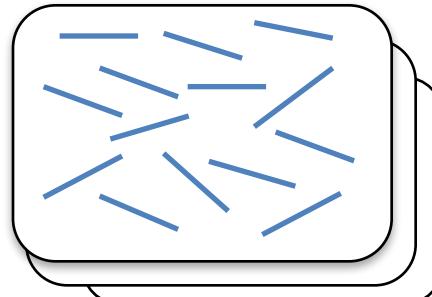
```
$ htseq-count eco_bowtie2.sam ecoli.gtf > ecoli.htseq
```

- その他 htseq-count マニュアル
<http://www-huber.embl.de/HTSeq/doc/count.html>

今回使ったツール・ファイルのまとめ

ゲノム（リファレンス）配列
FASTAファイル

```
>chr
ACCTTCATCCTCACCCAAACGGCAATAAGCT
CICIGGATTAATTTAACAGTGCTCATACACC
TTCAGACTGGTACCCCGTGTTAAATTTAAA
TTTTAGCTAGGCTAAATACTTACCAA
TATAGGCTATACCCGCAAGCAATAAAATPCAG
AGTACACACACACACAGAACCCATTGCCACACC
```



サンプルリード（ゲノム DNA/RNA）
FASTQファイル（配列+クオリティ）

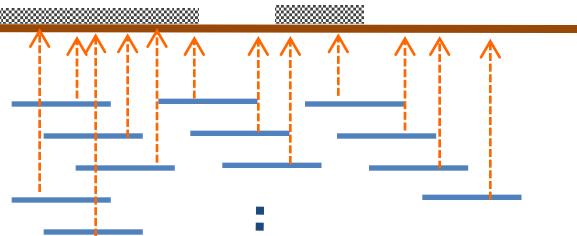
```
@SRR1515276.1 HSI-SI808:151:D2D13ACXX:2:1207:3625:88631 length=51
AICCCCCCCTCCCCCAACGGACCTAATGTCGGGGCGAATACAGCGGGGAG
+SRR1515276.1 HSI-SI808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@D>DFF7DC?FEEF@DFI<DF@AAA@FBD@CA>?P@B>B:::
@SRR1515276.2 HSI-SI808:151:D2D13ACXX:2:1207:3871:88513 length=51
CAACGCGIPGIAPOACAGTCCTGGTCAAAACACAAACGACGCGCCCGCC
+SRR1515276.2 HSI-SI808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCGFDFFDHFDEHIIIEGHJJJGFGHGHGHIJJDGJHJHGHHIH
```

インデックス作成

bowtie2-build

リファレンス配列へのマッピング

bowtie2



マッピング結果 SAM ファイル

HD	VN:1.0	SD:unsorted
@SQ	SN:chr	IN:4639675
@RG	ID:bowtie2	PN:bowtie2 VN:2.2.4 CL:"/bio/bin/bowtie2-align
SRR1515276.40	0	chr 4423609 42 51M * 0 0
SRR1515276.158	16	chr 501700 42 51M * 0 0
SRR1515276.212	4	* 0 0 * * 0 0
SRR1515276.319	0	chr 2922768 42 51M * 0 0
SRR1515276.367	16	chr 2753873 42 51M * 0 0
SRR1515276.411	0	chr 3440721 42 51M * 0 0
SRR1515276.434	0	chr 4198737 42 51M * 0 0

遺伝子アノテーション GFF(GTF)ファイル

```
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";
```

遺伝子ごとの集計

htseq-count

b0001	11
b0002	117
b0003	36

BAM ファイル

samtools

並べ替え
検索
ゲノムブラウザへ