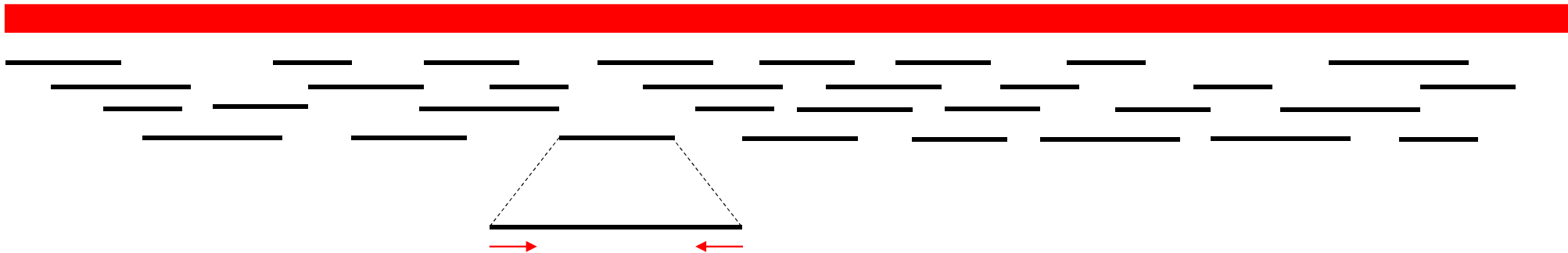




# 2016年度第一回 バイオインフォマティクス実習

次世代シーケンスデータのマッピング

# DNA sequence



length of genome :  $G$   
#reads :  $N$   
length of each read :  $L$   
coverage  $C = (N \times L) / G$

GCTGATCGT  
GATGCTAGCTGCT  
ATCGAGCGCGATG  
GCATCGATCGAGC  
GCATGCCGCAT  
AGGTGCATG  
...AGGTGCATGCCGCATCGATCGAGCGCGATGCTAGCTGCTGATCGT...

# ファイルの取得

- Index用ゲノムreference  
マウス染色体1番

<http://hgdownload.cse.ucsc.edu/goldenPath/mm10/chromosomes/chr1.fa.gz>

- シーケンスデータ  
GEO データベースaccession number GSE65976  
SRR1508230.sra ,SRR1508234.sra  
SRR1805875.sra, SRR1805876.sra

課題配布フォルダ → bioinfojisyu → chr1.fa  
part\_SRR1805875.fastq

# Cygwin

- cygwin  
windows上で動作するUnix環境の一つ
- [www.cygwin.com](http://www.cygwin.com)で配布しているsetup.exeをダウンロード
- 実行してインストール
- Cygwinターミナル(端末)を起動  
スタートメニュー → 2.ネットワークツール → 仮想UNIX端末(cygwin64)

# Bowtie

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

Latest releaseから最新版をダウンロード

bowtie-1.1.1-mingw.x86\_64.zip

展開するだけで使用可能

課題配布フォルダのbowtie-1.1.1フォルダを各自のデスクトップにコピー

# FASTQ format

- 1行目: @配列ID
- 2行目: 塩基配列
- 3行目: +配列ID 説明
- 4行目: クオリティー値 (シーケンスエラーの生じる確率)

@Seq-ID

AGGTGCATCGATGCGCGAATAAT

+

!1''\*)++)+//?"AAA{{

# Bowtie

- マッピングツールの一つ
- Burrows Wheeler transformを利用している
- 高速である
- メモリの消費は少ない
- 並列化に対応している

# Burrows Wheeler transform

- The BWT applies a reversible transformation to a block of input text. The transformation does not itself compress the data, but reorders it to make it easy to compress with simple algorithms such as move-to-front coding.

Burrows M, Wheeler DJ (1994) A block-sorting lossless data compression algorithm. Technical report 124. Palo Alto, CA: *Digital Equipment Corporation*.



# Burrows Wheeler transform

original

X = abracadabra

末尾に\$を付ける

左に一文字ずつシフトして

ローテーション

a	b	r	a	c	a	d	a	b	r	a	\$
b	r	a	c	a	d	a	b	r	a	\$	a
r	a	c	a	d	a	b	r	a	\$	a	b
a	c	a	d	a	b	r	a	\$	a	b	r
c	a	d	a	b	r	a	\$	a	b	r	a
a	d	a	b	r	a	\$	a	b	r	a	c
d	a	b	r	a	\$	a	b	r	a	c	a
a	b	r	a	\$	a	b	r	a	c	a	d
b	r	a	\$	a	b	r	a	c	a	d	a
r	a	\$	a	b	r	a	c	a	d	a	b
a	\$	a	b	r	a	c	a	d	a	b	r
\$	a	b	r	a	c	a	d	a	b	r	a

# Burrows Wheeler transform

	F											L
アルファベット順にソート	\$	a	b	r	a	c	a	d	a	b	r	a
\$はaよりも先	a	\$	a	b	r	a	c	a	d	a	b	r
$BWT(X) = ard\$rcaaaabb$	a	b	r	a	\$	a	b	r	a	c	a	d
	a	b	r	a	c	a	d	a	b	r	a	\$
	a	c	a	d	a	b	r	a	\$	a	b	r
	a	d	a	b	r	a	\$	a	b	r	a	c
	b	r	a	\$	a	b	r	a	c	a	d	a
	b	r	a	c	a	d	a	b	r	a	\$	a
	c	a	d	a	b	r	a	\$	a	b	r	a
	d	a	b	r	a	\$	a	b	r	a	c	a
	r	a	\$	a	b	r	a	c	a	d	a	b
	r	a	c	a	d	a	b	r	a	\$	a	b

# Burrows Wheeler 逆変換

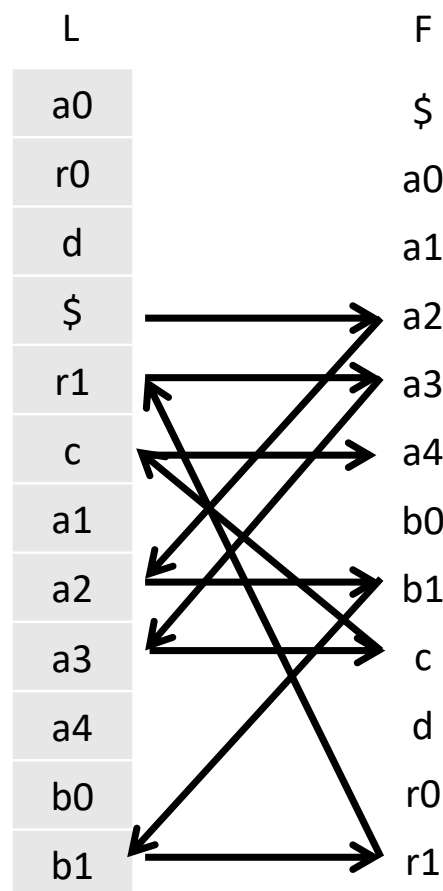
(複数個ある  
文字には番号を  
付けてある)

L
a0
r0
d
\$
r1
c
a1
a2
a3
a4
b0
b1

辞書式に  
並べ替え



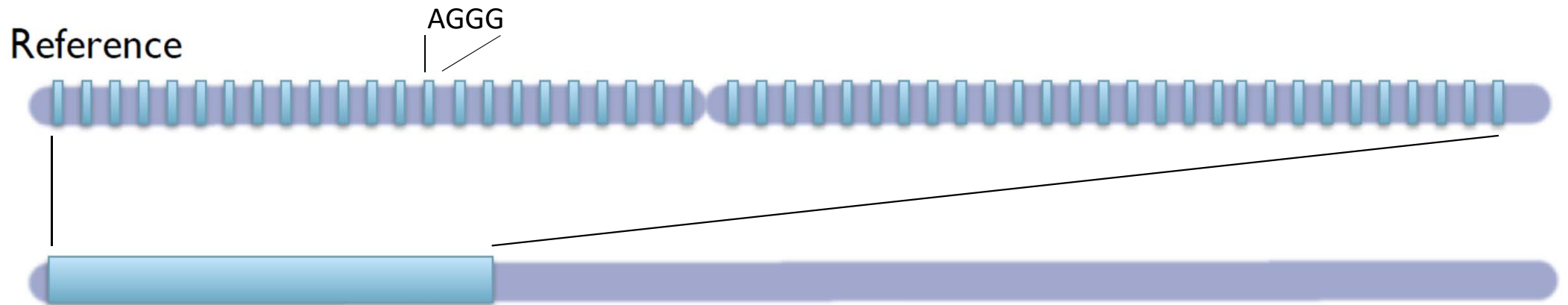
F
\$
a0
a1
a2
a3
a4
b0
b1
c
d
r0
r1



LはFの一文字前  
\$から出発して  
L→F→L→L...と  
たどっていくと  
オリジナルの文字列を  
復元できる

a2b1r1a3ca4...

# BWTの利点



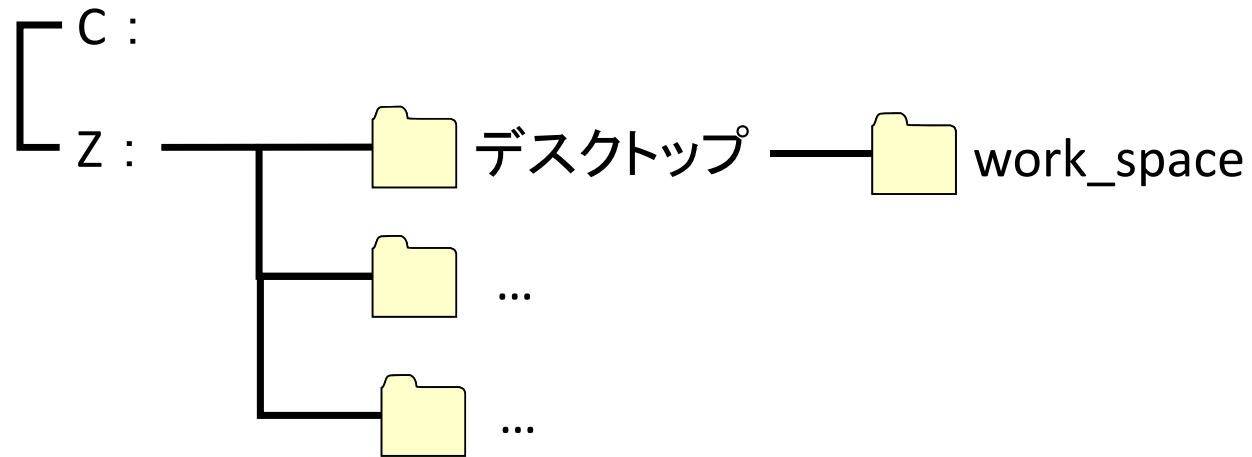
BWT( Reference )

同じ文字が固まる傾向がある  
検索しやすい  
簡単に復元できる

F      L  
GGGA ~ A  
GGGC ~ A  
GGGT ~ A  
...

# 作業ディレクトリ作成

```
Cygwin  
$ cd /cygdrive/z/デスクトップ ↵  
$ mkdir work_space ↵  
$ cd work_space ↵
```

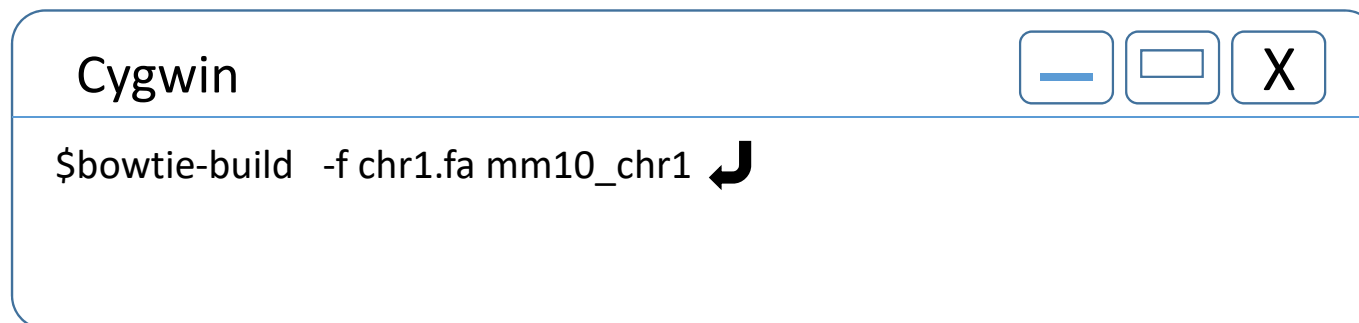


# フォルダ（ディレクトリ）の権限

- ドライブC ローカルPCのハードディスク  
読み込みのみ
- ドライブZ 各アカウントに割り当てられたハードドライブ  
ネットワーク上のハードディスク 八景キャンパスのサーバ  
読み書き実行
- ドライブY 課題配布フォルダ  
ネットワーク上のハードディスク 八景キャンパスのサーバ  
読み込みのみ

# Bowtieのコマンド1

## Index作成



```
Cygwin
$bowtie-build -f chr1.fa mm10_chr1 ↵
```

bowtie-build -f リファレンスファイル名 インデックス名  
referenceのゲノム配列をBurrows-Wheeler変換を使ってインデックス化する  
mm10\_chr1.1.ebwt  
mm10\_chr1.2.ebwt  
mm10\_chr1.3.ebwt  
mm10\_chr1.4.ebwt  
mm10\_chr1.rev.1.ebwt  
mm10\_chr1.rev.2.ebwt  
6個のファイルが作成される

# マッピング

Cygwin



```
$bowtie -m 1 -v 2 -a --strata --best -S mm10_chr1 part_SRR1805875.fastq  
mm10_chr1 SRR1805875.sam ↵
```

bowtie (option) 参照インデックス名 fastqファイル名 出力ファイル名

- m 1 : 1リードを1か所にマッピングする
- v 2 : ミスマッチを2個まで許容する
- a : 候補の配列を全て列挙する
- best : ベストマッチの場所にマッピング
- strata :
- S : 結果をsamファイル形式で出力



# SAM format

- 11列
- 3列目 : 染色体番号
- 4列目 : 位置
- 10列目 : 配列

## 第2回 予定

# Integrative Genomics Viewer(IGV)による可視化

- samtoolsでファイル変換
  - 1) sam→bam変換
  - 2) bamファイルを染色体順に並べ替え
  - 3) indexファイルの作成
- IGVへアップロード、表示