

# 2015年度第一回 バイオインフォマティクス実習

Bowtieを用いたシーケンスデータのマッピング

# Cygwin

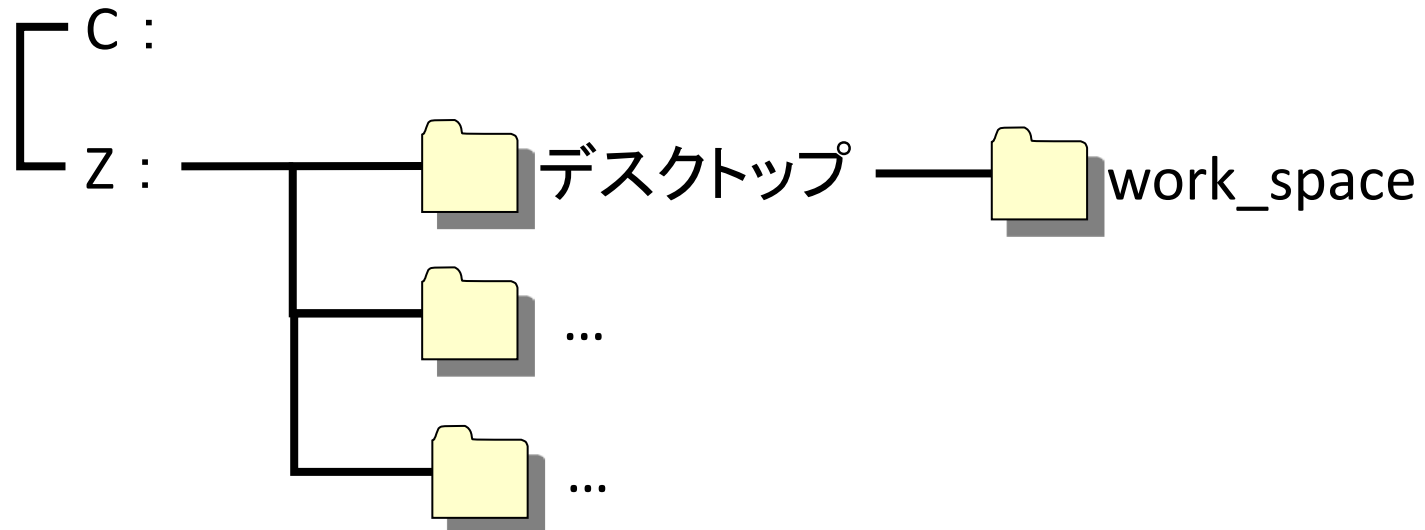
- cygwin  
windows上で動作するUnix環境の一つ
- [www.cygwin.com](http://www.cygwin.com)で配布しているsetup.exeをダウンロード
- 実行してインストール
- Cygwinターミナル(端末)を起動  
スタートメニュー → 2.ネットワークツール → 仮想UNIX端末(cygwin64)

# 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

# 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

# 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

# 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

# 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



# 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	a	b	r	a	c	a	d	a	b	r	a	\$
$X = \text{abracadabra}$	b	r	a	c	a	d	a	b	r	a	\$	a
末尾に\$を付ける	r	a	c	a	d	a	b	r	a	\$	a	b
左に一文字ずつシフトして	a	c	a	d	a	b	r	a	\$	a	b	r
ローテーション	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
$BWT(X) = \text{ard\$rcaaaabb}$	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

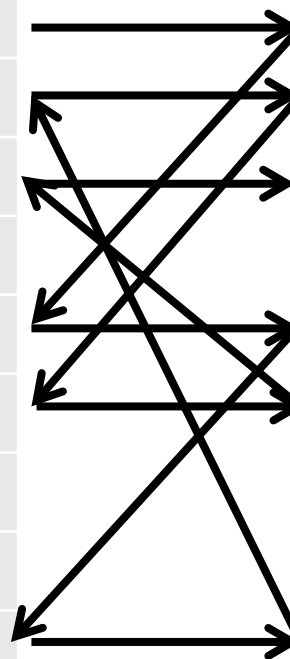
(複数個ある  
文字には番号を  
付けてある)

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

辞書式に  
並べ替え



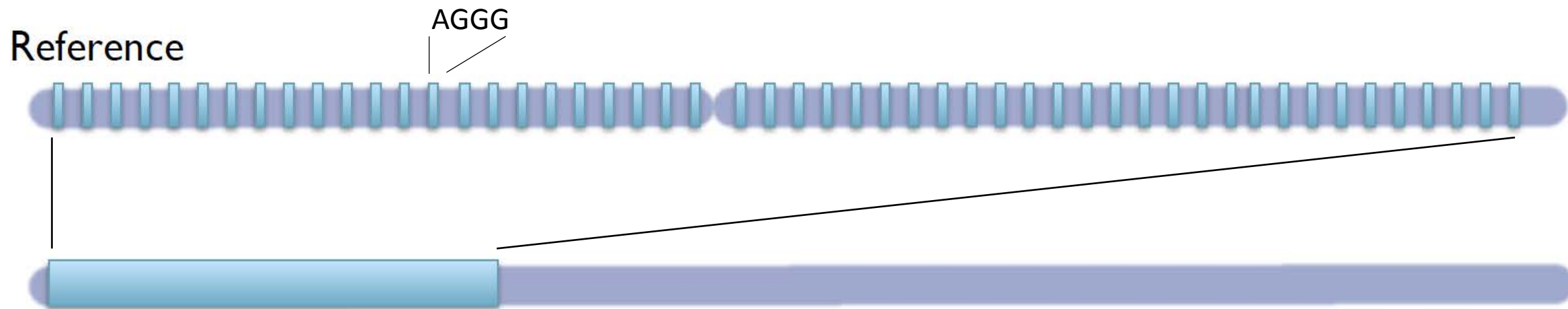
L	F
a0	\$
r0	a0
d	a1
\$	a2
r1	a3
c	a4
a1	b0
a2	b1
a3	c
a4	d
b0	r0
b1	r1



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



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
...	

# 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

# Burrows Wheeler transform

original

$X = \text{ctgaaactggt}$

0	c	t	g	a	a	a	c	t	g	g	t	\$
1	t	g	a	a	a	c	t	g	g	t	\$	c
2	G	a	a	a	c	t	g	g	t	\$	c	t
3	a	a	a	c	t	g	g	t	\$	c	t	g
4	a	a	c	t	g	g	t	\$	c	t	g	a
5	a	c	t	g	g	t	\$	c	t	g	a	a
6	c	t	g	g	t	\$	c	t	g	a	a	a
7	t	g	g	t	\$	c	t	g	a	a	a	c
8	g	g	t	\$	c	t	g	a	a	a	c	t
9	g	t	\$	c	t	g	a	a	a	c	t	g
10	t	\$	c	t	g	a	a	a	c	t	g	g
11	\$	c	t	g	a	a	a	c	t	g	g	t

# Burrows Wheeler transform

$X = \text{ctgaaactggt\$}$

$BWT(X) = \text{tgaa\$attggcc}$

<b>0</b>	<b>\$</b>	c	t	g	a	a	a	c	t	g	g	<b>t</b>	<b>11</b>
<b>1</b>	<b>a</b>	a	a	c	t	g	g	t	\$	c	t	<b>g</b>	<b>3</b>
<b>2</b>	<b>a</b>	a	c	t	g	g	t	\$	c	t	g	<b>a</b>	<b>4</b>
<b>3</b>	<b>a</b>	c	t	g	g	t	\$	c	t	g	a	<b>a</b>	<b>5</b>
<b>4</b>	<b>c</b>	t	g	a	a	a	c	t	g	g	t	<b>\$</b>	<b>0</b>
<b>5</b>	<b>c</b>	t	g	g	t	\$	c	t	g	a	a	<b>a</b>	<b>6</b>
<b>6</b>	<b>g</b>	a	a	a	c	t	g	g	t	\$	c	<b>t</b>	<b>2</b>
<b>7</b>	<b>g</b>	g	t	\$	c	t	g	a	a	a	c	<b>t</b>	<b>8</b>
<b>8</b>	<b>g</b>	t	\$	c	t	g	a	a	a	c	t	<b>g</b>	<b>9</b>
<b>9</b>	<b>t</b>	\$	c	t	g	a	a	a	c	t	g	<b>g</b>	<b>10</b>
<b>10</b>	<b>t</b>	g	a	a	a	c	t	g	g	t	\$	<b>c</b>	<b>1</b>
<b>11</b>	<b>t</b>	g	g	t	\$	c	t	g	a	a	a	<b>c</b>	<b>7</b>



# Backward Search

ctgaaactgggt\$から  
配列“ctgg”を検索

$\underline{R}(W)$ : 文字Wが出現する  
最初の列

$\overline{R}(W)$ : 文字Wが出現する  
最後の列

$\underline{R}(g) = 6$

$\overline{R}(g) = 8$

0	\$	c	t	g	a	a	a	c	t	g	g	t	11
1	a	a	a	c	t	g	g	t	\$	c	t	g	3
2	a	a	c	t	g	g	t	\$	c	t	g	a	4
3	a	c	t	g	g	t	\$	c	t	g	a	a	5
4	c	t	g	a	a	a	c	t	g	g	t	\$	0
5	c	t	g	g	t	\$	c	t	g	a	a	a	6
➡ 6	g	a	a	a	c	t	g	g	t	\$	c	t	2
7	g	g	t	\$	c	t	g	a	a	a	c	t	8
➡ 8	g	t	\$	c	t	g	a	a	a	c	t	g	9
9	t	\$	c	t	g	a	a	a	c	t	g	g	10
10	t	g	a	a	a	c	t	g	g	t	\$	c	1
11	t	g	g	t	\$	c	t	g	a	a	a	c	7

# Backward Search

	<b>0</b>	\$	c	t	g	a	a	a	c	t	g	g	t	<b>11</b>
$\underline{R}(aW) = C(a) + O(a, \underline{R}(W) - 1) + 1$	<b>1</b>	a	a	a	c	t	g	g	t	\$	c	t	g	<b>3</b>
$\overline{R}(aW) = C(a) + O(a, \overline{R}(W))$	<b>2</b>	a	a	c	t	g	g	t	\$	c	t	g	a	<b>4</b>
$C(a)$ : $a$ よりもアルファベットの 小さい文字数	<b>3</b>	a	c	t	g	g	t	\$	c	t	g	a	a	<b>5</b>
$O(a, i)$ : BWT列内で <i>i</i> 列目までの $a$ の数	<b>4</b>	c	t	g	a	a	a	c	t	g	g	t	\$	<b>0</b>
	<b>5</b>	c	t	g	g	t	\$	c	t	g	a	a	a	<b>6</b>
$\underline{R}(gg) = C(g) + O(g, \underline{R}(g) - 1) + 1$ $= 5 + O(g, 6 - 1) + 1$ $= 5 + 1 + 1$ $= 7$	<b>6</b>	g	a	a	a	c	t	g	g	t	\$	c	t	<b>2</b>
	<b>7</b>	g	g	t	\$	c	t	g	a	a	a	c	t	<b>8</b>
	<b>8</b>	g	t	\$	c	t	g	a	a	a	c	t	g	<b>9</b>
$\overline{R}(gg) = C(g) + O(g, \overline{R}(g))$ $= 5 + O(g, 8)$ $= 5 + 2$ $= 7$	<b>9</b>	t	\$	c	t	g	a	a	a	c	t	g	g	<b>10</b>
	<b>10</b>	t	g	a	a	a	c	t	g	g	t	\$	c	<b>1</b>
	<b>11</b>	t	g	g	t	\$	c	t	g	a	a	a	c	<b>7</b>

# backward search

$$\underline{R}(tgg) = C(t) + O(t, \underline{R}(gg) - 1) + 1$$

$$= 8 + O(t, 7 - 1) + 1$$

$$= 8 + 2 + 1$$

$$= 11$$

$$\overline{R}(tgg) = C(t) + O(t, \overline{R}(gg))$$

$$= 8 + O(t, 8)$$

$$= 8 + 3$$

$$= 11$$

<b>0</b>	<b>\$</b>	c	t	g	a	a	a	c	t	g	g	<b>t</b>	<b>11</b>
<b>1</b>	<b>a</b>	a	a	c	t	g	g	t	\$	c	t	<b>g</b>	<b>3</b>
<b>2</b>	<b>a</b>	a	c	t	g	g	t	\$	c	t	g	<b>a</b>	<b>4</b>
<b>3</b>	<b>a</b>	c	t	g	g	t	\$	c	t	g	a	<b>a</b>	<b>5</b>
<b>4</b>	<b>c</b>	t	g	a	a	a	c	t	g	g	t	<b>\$</b>	<b>0</b>
<b>5</b>	<b>c</b>	t	g	g	t	\$	c	t	g	a	a	<b>a</b>	<b>6</b>
<b>6</b>	<b>g</b>	a	a	a	c	t	g	g	t	\$	c	<b>t</b>	<b>2</b>
<b>7</b>	<b>g</b>	g	t	\$	c	t	g	a	a	a	c	<b>t</b>	<b>8</b>
<b>8</b>	<b>g</b>	t	\$	c	t	g	a	a	a	c	t	<b>g</b>	<b>9</b>
<b>9</b>	<b>t</b>	\$	c	t	g	a	a	a	c	t	g	<b>g</b>	<b>10</b>
<b>10</b>	<b>t</b>	g	a	a	a	c	t	g	g	t	\$	<b>c</b>	<b>1</b>
<b>11</b>	<b>t</b>	g	g	t	\$	c	t	g	a	a	a	<b>c</b>	<b>7</b>

# backward search

$\underline{R}(ctgg) = C(c) + O(c, \underline{R}(tgg) - 1) + 1$ $= 3 + O(c, 11 - 1) + 1$ $= 3 + 1 + 1$ $= 5$	<b>0</b>	<b>\$</b>	c	t	g	a	a	a	c	t	g	g	<b>t</b>	<b>11</b>
	<b>1</b>	<b>a</b>	a	a	c	t	g	g	t	\$	c	t	<b>g</b>	<b>3</b>
	<b>2</b>	<b>a</b>	a	c	t	g	g	t	\$	c	t	g	<b>a</b>	<b>4</b>
$\overline{R}(ctgg) = C(c) + O(c, \overline{R}(tgg))$ $= 3 + O(c, 11)$ $= 3 + 2$ $= 5$	<b>3</b>	<b>a</b>	c	t	g	g	t	\$	c	t	g	a	<b>a</b>	<b>5</b>
	<b>4</b>	<b>c</b>	t	g	a	a	a	c	t	g	g	t	<b>\$</b>	<b>0</b>
	<b>5</b>	<b>c</b>	t	g	g	t	\$	c	t	g	a	a	<b>a</b>	<b>6</b>
	<b>6</b>	<b>g</b>	a	a	a	c	t	g	g	t	\$	c	<b>t</b>	<b>2</b>
	<b>7</b>	<b>g</b>	g	t	\$	c	t	g	a	a	a	c	<b>t</b>	<b>8</b>
	<b>8</b>	<b>g</b>	t	\$	c	t	g	a	a	a	c	t	<b>g</b>	<b>9</b>
	<b>9</b>	<b>t</b>	\$	c	t	g	a	a	a	c	t	g	<b>g</b>	<b>10</b>
	<b>10</b>	<b>t</b>	g	a	a	a	c	t	g	g	t	\$	<b>c</b>	<b>1</b>
	<b>11</b>	<b>t</b>	g	g	t	\$	c	t	g	a	a	a	<b>c</b>	<b>7</b>

# backward search

$$\begin{aligned}\underline{R}(ttgg) &= C(t) + O(t, \underline{R}(tgg) - 1) + 1 \\ &= 8 + O(t, 11 - 1) + 1 \\ &= 8 + 3 + 1 \\ &= 12\end{aligned}$$

$$\begin{aligned}\overline{R}(ttgg) &= C(t) + O(t, \overline{R}(tgg)) \\ &= 8 + O(t, 11) \\ &= 8 + 2 \\ &= 11\end{aligned}$$

$\underline{R}(aW) \leq \overline{R}(aW)$ : 配列  $aW$  が存在する条件

<b>0</b>	<b>\$</b>	c	t	g	a	a	a	c	t	g	g	<b>t</b>	<b>11</b>
<b>1</b>	<b>a</b>	a	a	c	t	g	g	t	\$	c	t	<b>g</b>	<b>3</b>
<b>2</b>	<b>a</b>	a	c	t	g	g	t	\$	c	t	g	<b>a</b>	<b>4</b>
<b>3</b>	<b>a</b>	c	t	g	g	t	\$	c	t	g	a	<b>a</b>	<b>5</b>
<b>4</b>	<b>c</b>	t	g	a	a	a	c	t	g	g	t	<b>\$</b>	<b>0</b>
<b>5</b>	<b>c</b>	t	g	g	t	\$	c	t	g	a	a	<b>a</b>	<b>6</b>
<b>6</b>	<b>g</b>	a	a	a	c	t	g	g	t	\$	c	<b>t</b>	<b>2</b>
<b>7</b>	<b>g</b>	g	t	\$	c	t	g	a	a	a	c	<b>t</b>	<b>8</b>
<b>8</b>	<b>g</b>	t	\$	c	t	g	a	a	a	c	t	<b>g</b>	<b>9</b>
<b>9</b>	<b>t</b>	\$	c	t	g	a	a	a	c	t	g	<b>g</b>	<b>10</b>
<b>10</b>	<b>t</b>	g	a	a	a	c	t	g	g	t	\$	<b>c</b>	<b>1</b>
<b>11</b>	<b>t</b>	g	g	t	\$	c	t	g	a	a	a	<b>c</b>	<b>7</b>

# Backward Search

$$\begin{aligned}\underline{R}(ct) &= C(c) + O(c, \underline{R}(t) - 1) + 1 \\ &= 3 + O(c, 9 - 1) + 1 \\ &= 3 + 0 + 1 \\ &= 4\end{aligned}$$

$$\begin{aligned}\overline{R}(ct) &= C(c) + O(c, \overline{R}(t)) \\ &= 3 + O(c, 11) \\ &= 3 + 2 \\ &= 5\end{aligned}$$

<b>0</b>	<b>\$</b>	c	t	g	a	a	a	c	t	g	g	<b>t</b>	<b>11</b>
<b>1</b>	<b>a</b>	a	a	c	t	g	g	t	\$	c	t	<b>g</b>	<b>3</b>
<b>2</b>	<b>a</b>	a	c	t	g	g	t	\$	c	t	g	<b>a</b>	<b>4</b>
<b>3</b>	<b>a</b>	c	t	g	g	t	\$	c	t	g	a	<b>a</b>	<b>5</b>
<b>4</b>	<b>c</b>	t	g	a	a	a	c	t	g	g	t	<b>\$</b>	<b>0</b>
<b>5</b>	<b>c</b>	t	g	g	t	\$	c	t	g	a	a	<b>a</b>	<b>6</b>
<b>6</b>	<b>g</b>	a	a	a	c	t	g	g	t	\$	c	<b>t</b>	<b>2</b>
<b>7</b>	<b>g</b>	g	t	\$	c	t	g	a	a	a	c	<b>t</b>	<b>8</b>
<b>8</b>	<b>g</b>	t	\$	c	t	g	a	a	a	c	t	<b>g</b>	<b>9</b>
<b>9</b>	<b>t</b>	\$	c	t	g	a	a	a	c	t	g	<b>g</b>	<b>10</b>
<b>10</b>	<b>t</b>	g	a	a	a	c	t	g	g	t	\$	<b>c</b>	<b>1</b>
<b>11</b>	<b>t</b>	g	g	t	\$	c	t	g	a	a	a	<b>c</b>	<b>7</b>

# Backward Search

$$\underline{R}(act) = C(a) + O(a, \underline{R}(ct) - 1) + 1$$

$$= 0 + O(a, 4 - 1) + 1$$

$$= 0 + 2 + 1$$

$$= 3$$

$$\overline{R}(act) = C(a) + O(a, \overline{R}(ct))$$

$$= 0 + O(a, 5)$$

$$= 0 + 3$$

$$= 3$$

<b>0</b>	<b>\$</b>	c	t	g	a	a	a	c	t	g	g	<b>t</b>	<b>11</b>
<b>1</b>	<b>a</b>	a	a	c	t	g	g	t	\$	c	t	<b>g</b>	<b>3</b>
<b>2</b>	<b>a</b>	a	c	t	g	g	t	\$	c	t	g	<b>a</b>	<b>4</b>
<b>3</b>	<b>a</b>	c	t	g	g	t	\$	c	t	g	a	<b>a</b>	<b>5</b>
<b>4</b>	<b>c</b>	t	g	a	a	a	c	t	g	g	t	<b>\$</b>	<b>0</b>
<b>5</b>	<b>c</b>	t	g	g	t	\$	c	t	g	a	a	<b>a</b>	<b>6</b>
<b>6</b>	<b>g</b>	a	a	a	c	t	g	g	t	\$	c	<b>t</b>	<b>2</b>
<b>7</b>	<b>g</b>	g	t	\$	c	t	g	a	a	a	c	<b>t</b>	<b>8</b>
<b>8</b>	<b>g</b>	t	\$	c	t	g	a	a	a	c	t	<b>g</b>	<b>9</b>
<b>9</b>	<b>t</b>	\$	c	t	g	a	a	a	c	t	g	<b>g</b>	<b>10</b>
<b>10</b>	<b>t</b>	g	a	a	a	c	t	g	g	t	\$	<b>c</b>	<b>1</b>
<b>11</b>	<b>t</b>	g	g	t	\$	c	t	g	a	a	a	<b>c</b>	<b>7</b>

# 課題

- マウス一番染色体以外のゲノムインデックスを作成し、配列データのマッピングを行ってください
- マッピングのパラメータを変更して実行してください



# 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