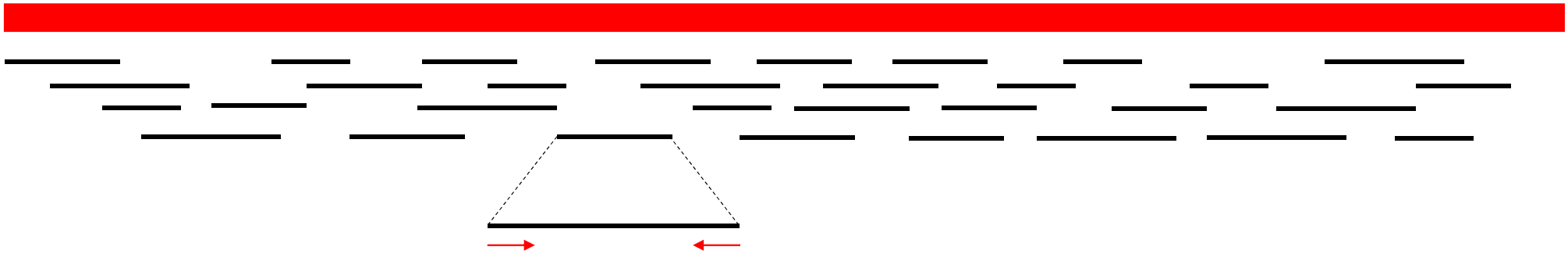
2018年
第三回バイオインフォマティクス実習

RNA-seqデータ解析

先端医科学研究センター
バイオインフォマティクス解析室　中林潤

DNA sequence



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

GCATCGATCGAGC
GCATGCCGCAT
AGGTGCATG
…AGGTGCATGCCGCATCGATCGAGC…

ファイルの取得

- reference genome
- シーケンスデータ

  GEO データベースaccession number GSE84686

  CXCL8 (+) T cell single cell RNA-seq

  SRR3939298.sra

  sra ファイルで配布されている

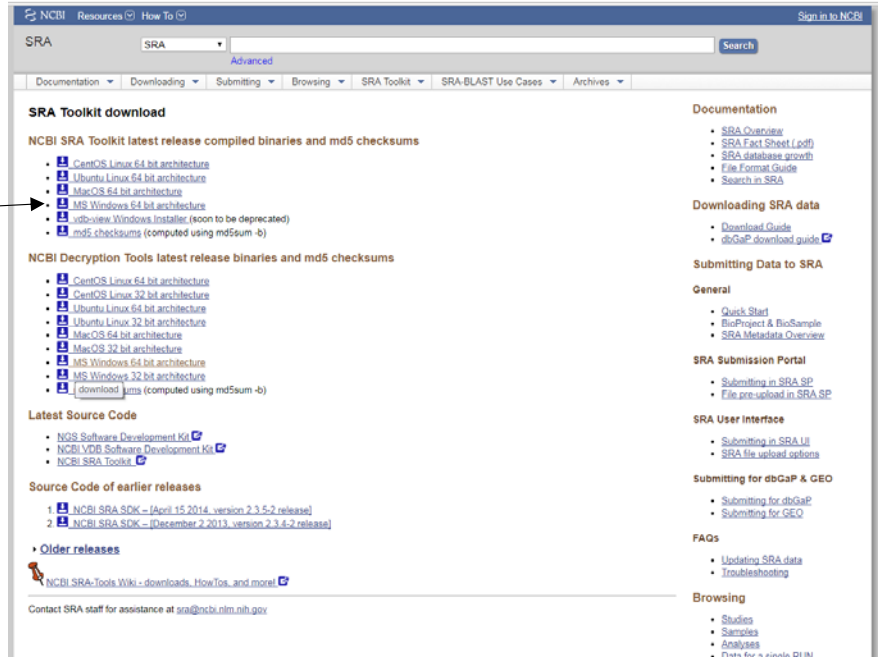
  sra→fastqへ変換して解析に使用する

SRAToolkit download

- http://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft

MS Windows 64 bit architectureを選択

ダウンロードしたファイルを展開する
だけで使用できます。



sraファイルを保存したフォルダで、shit+右クリック
“コマンドウィンドウをここで開く“を選択
C:¥fastq-dump.exeが保存されたフォルダ名¥fastq-dump.exe SRRファイル名 --split-filesと入力

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{{

Cygwin

- cygwin
  windows上で動作するUnix環境の一つ

- 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)

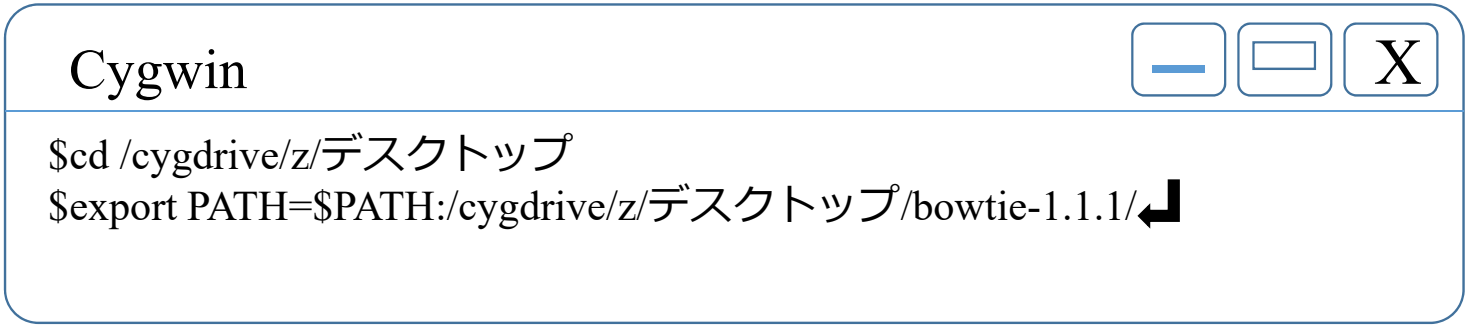
Bowtie

- 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

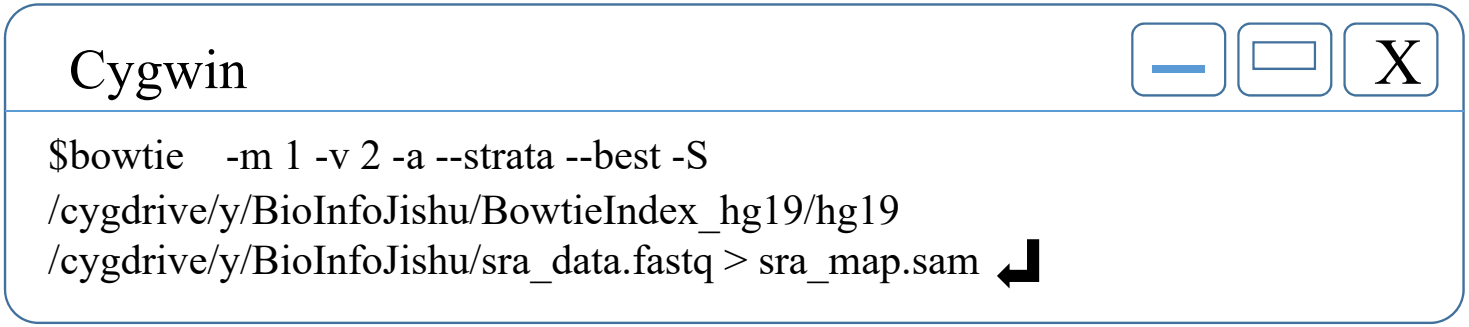
- http://bowtie-bio.sourceforge.net/index.shtml
  Latest releaseから最新版をダウンロード
  bowtie-1.1.1-mingw.x86_64.zip
  展開するだけで使用可能

マッピング



cd : デスクトップフォルダに移動
export PATH : bowtie-1.1.1フォルダ内のbowtieにパスを通す

マッピング



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

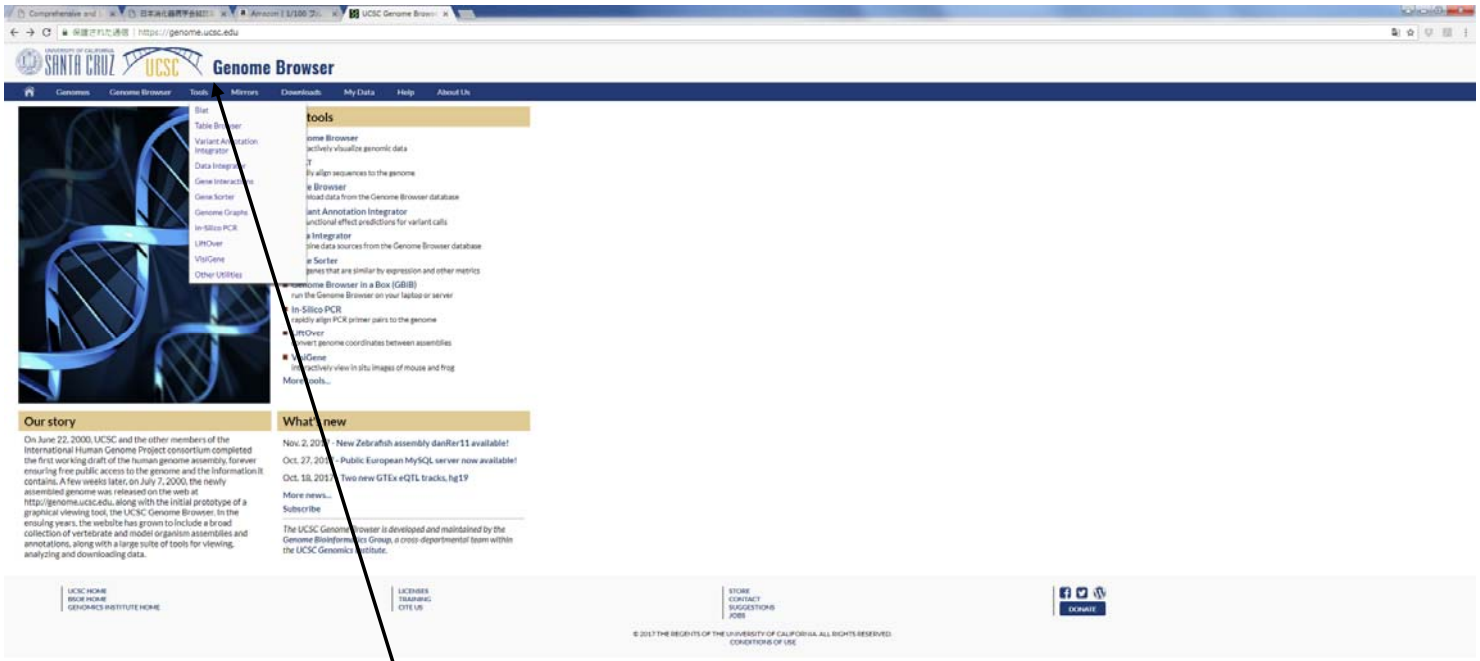
-m 1 : 1リードを1か所にマッピングする
-v 2 : ミスマッチを2個まで許容する
-a :候補の配列を全て列挙する
--best :ベストマッチの場所にマッピング
--strata :
-S : 結果をsamファイル形式で出力

SAM format

- 11列
- 3列目：染色体番号
- 4列目：位置
- 10列目：配列

reference gtf file

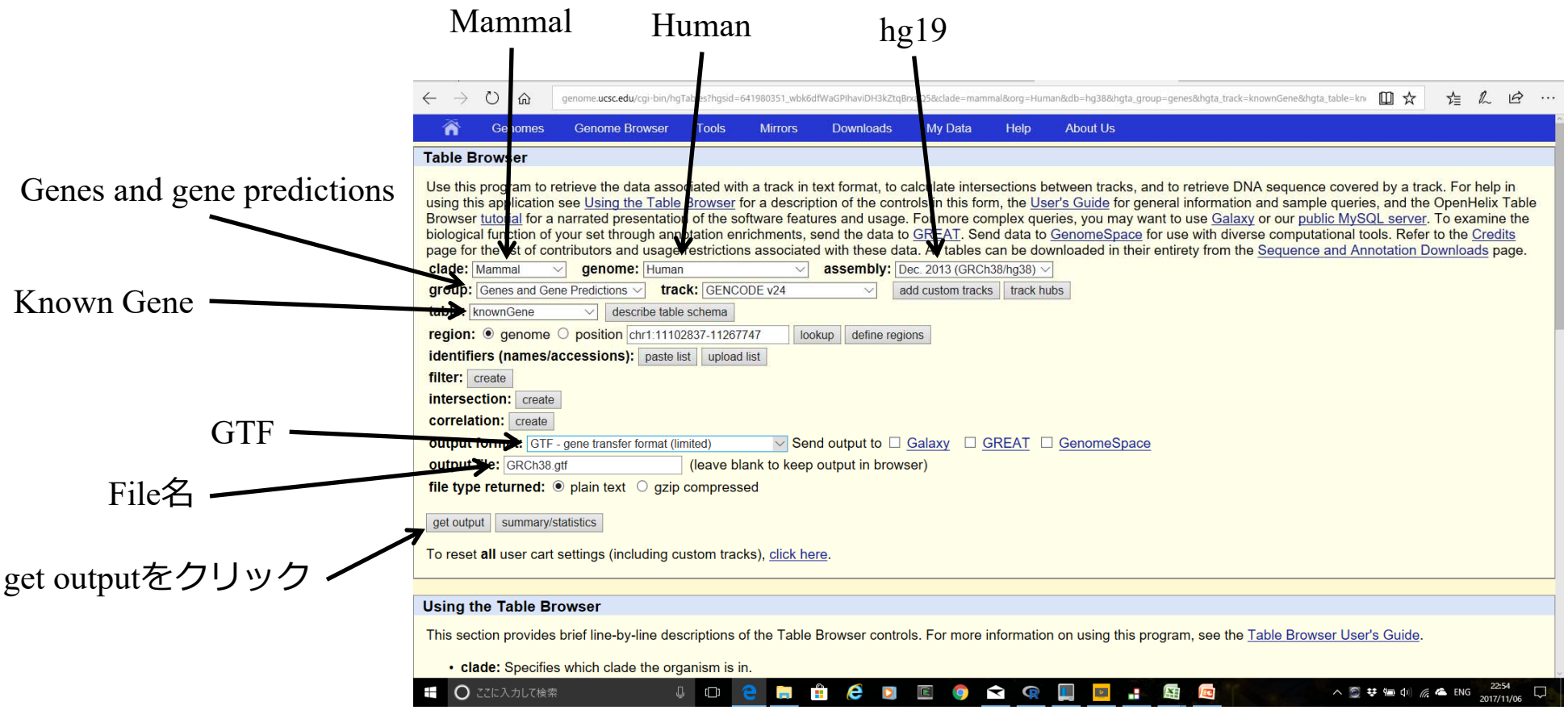
UCSC genome browser



http://genome.ucsc.edu

tools
Table Browser

Table Browser

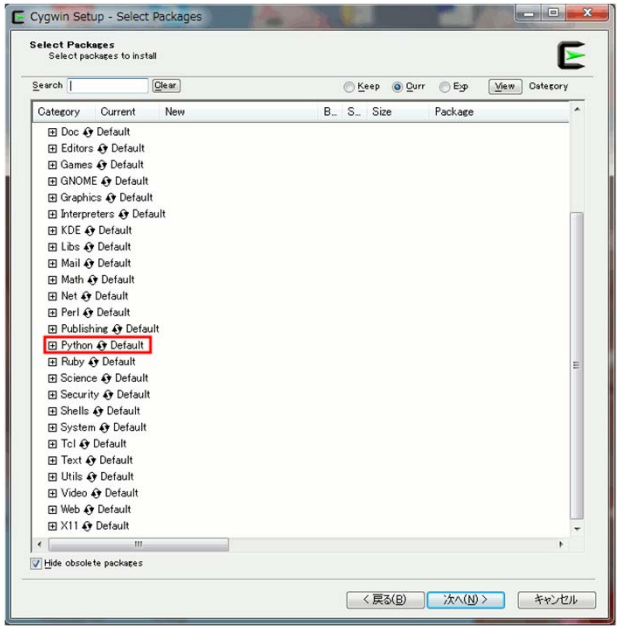


Python

- グイド・ヴァンロッサムが開発した汎用プログラム言語の一つ
- 標準ライブラリや様々な用途に使える専用の解析用ライブラリが充実している

ウィンドウズPCでの実行

- cygwin をインストール時にpythonのパッケージをチェック



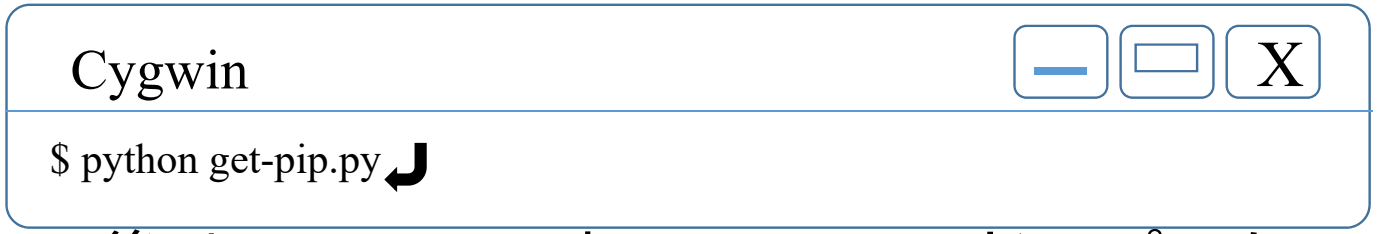
- 実習室用PCにはインストール済み
  cygwinコンソール上で実行できます

Pythonのパッケージ管理

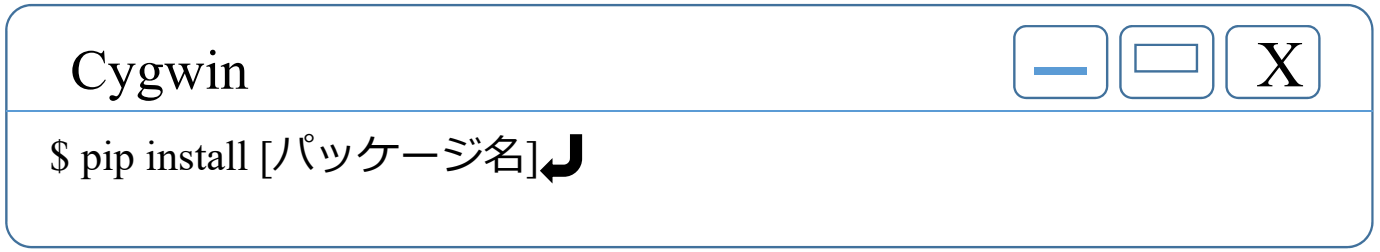
- Pipによるパッケージの管理

get-pip.pyの最新バージョンをダウンロード

get-pip.pyを保存したディレクトリで

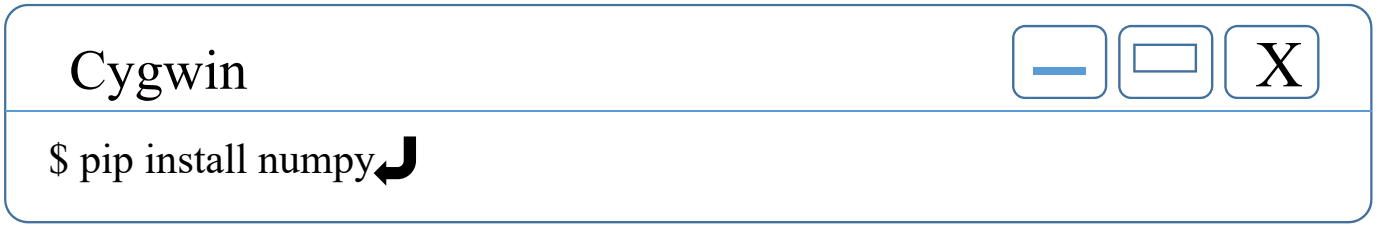


その後はコンソール上でpipコマンドでパッケージをインストールできる

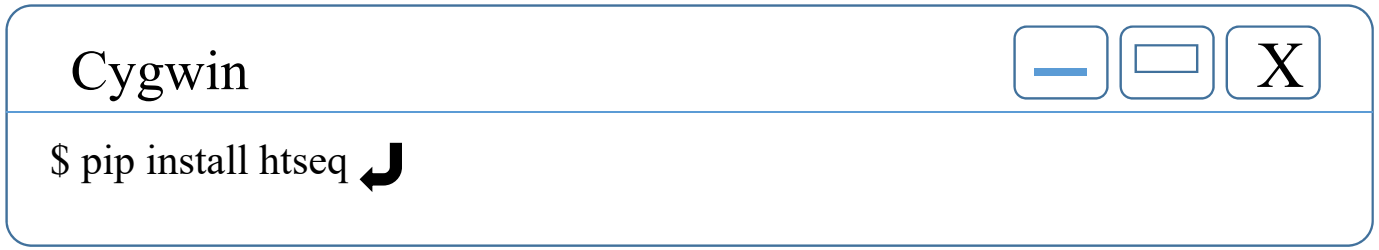


HTSeqのインストール

- numpy ライブラリのインストール



- HTSeqのインストール



HTSeqによるタグ数のカウント

- htseq-countコマンド

htseq-count マッピング後のsamファイル名　アノテーション用gtfファイル名　＞　出力ファイル名

結果

遺伝子名	カウント数
AAK1	0
AAMDC	69
AAMP	0
AANAT	0
AAR2	10
AARD	0