# 解析の実践 （blast、mapping、assemble）

# 検索前の準備

1. gw.ddbj.nig.ac.jpにログイン
   ssh [user name]@gw..ddbj.nig.ac.jp

2. 解析ノードにログイン
   qlogin

~はホームディレクトリ
（例：/home/hidekih15）
ホームディレクトリの表示；pwd

3. Pathの設定
   emacs ~/.bashrc
   PATH="$PATH":/usr/local/bin:/usr/local/pkg/bowtie2/currentという行を追加

4. ファイル名の補完
   set autolist

5. .bashrcの反映
   source .bashrc

（プログラム本体）
＊Blast (version 2.2.26)
/usr/local/bin/blastall
＊Bowtie 2 (2.0.0-beta6)
/usr/local/pkg/bowtie2/current/bowtie2
＊SOAPdenovo (1.05)
/usr/local/bin/soapdenovo

```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

alias emacs='emacs -nw'
alias rm='rm -i'
alias cp='cp -i'
alias mv='mv -i'

set autolist

PATH="$PATH":/usr/local/bin:/usr/local/pkg/bowtie2/current
```

# BLAST検索

1. 使用するデータ

(1) 問い合わせ配列 （クエリー）
/home/hidekih15/lecture/data/BLAST/query/

＊遺伝子の塩基配列：test_nt.fa
＊遺伝子のアミノ酸配列：test_aa.fa

(2) データベース （サブジェクト）

＊NIG SuperComputerに登録されているDBを対象とした場合
/usr/local/seq/blast/uniprot/swissprot # 例：SWISSPROT

＊手元の配列を対象とした場合
/home/hidekih15/lecture/data/BLAST/db/S_aureus_N315_chr.fa # ゲノム塩基配列
/home/hidekih15/lecture/data/BLAST/db/S_aureus_N315_ORFs_aa.fa # 遺伝子のアミノ酸配列

2. データのコピー

(1) cd ~ #ホームディレクトリへの移動
(2) mkdir test #解析用ディレクトリの作成
(3) cd test
(4) mkdir BLAST
(5) cd BLAST
(6) mkdir query # 1から6は'mkdir –p ~/test/BLAST/query'でも可能。
(7) cd query
(8) cp /home/hidekih15/lecture/data/BLAST/query/test_nt.fa .
(9) cp /home/hidekih15/lecture/data/BLAST/query/test_aa.fa .
# 8と9は'cp /home/hidekih15/lecture/data/BLAST/query/*.fa .'でも可能

3. BLASTライブラリの作成

(1) cd ~/test/BLAST
(2) mkdir db
(3) cd db
(4) cp /home/hidekih15/lecture/data/BLAST/db/*.fa .
(5) formatdb -i S_aureus_N315_chr.fa -p F # 塩基配列の場合
formatdb -i S_aureus_N315_ORFs_aa.fa -p T # アミノ酸配列の場合

4. BLASTの実行

(1) cd ~/test/BLAST/query/

(2) blastall -p [program name] -a [# of CPUs] -d [Library file name] -i [query filename] -o [output filename] # 基本コマンド

例）クエリーがアミノ酸配列（test_aa.fa）、データベースがSWISSPROT（アミノ酸配列）の場合

blastall -p blastp -a 8 -F F -d /usr/local/seq/blast/uniprot/swissprot -i test_aa.fa -o test_vs_swissprot.bp # 結果の閲覧：less test_vs_swissprot.bp

例）クエリーが塩基配列（test_nt.fa）、デーベースがSWISSPROT（アミノ酸配列）の場合

blastall -p blastx -a 8 -F F -d /usr/local/seq/blast/uniprot/swissprot -i test_nt.fa -o test_vs_swissprot.bx

例）クエリーが塩基配列（test_nt.fa）、データベースが塩基配列の場合

blastall -p blastn -a 8 -F F -d ~/test/BLAST/db/S_aureus_N315_chr.fa -i test_nt.fa -o test_vs_N315_chr.bn

例）クエリーがアミノ酸配列（test_aa.fa）、データベースがアミノ酸配列の場合

blastall -p blastp -a 8 -F F -d ~/test/BLAST/db/S_aureus_N315_ORFs_aa.fa -i test_aa.fa -o test_vs_S_aureus_N315_ORF.bp

BLASTのプログラムの種類

| プログラム名 | 問い合わせ配列（クエリー） | データベース（サブジェクト） |
|---|---|---|
| BLASTN | 塩基配列 | 塩基配列 |
| BLASTP | アミノ酸配列 | アミノ酸配列 |
| TBLASTN | アミノ酸配列 | 塩基配列 |
| BLASTX | 塩基配列 | アミノ酸配列 |

# BLASTのオプション表示

blastall

-p Program Name [String]
-d Database [String]
default = nr
-i Query File [File In]
default = stdin
-e Expectation value (E) [Real]
default = 10.0
-m alignment view options:
0 = pairwise,
1 = query-anchored showing identities,
2 = query-anchored no identities,
3 = flat query-anchored, show identities,
4 = flat query-anchored, no identities,
5 = query-anchored no identities and blunt ends,
6 = flat query-anchored, no identities and blunt ends,
7 = XML Blast output,
8 = tabular,
9 tabular with comment lines
・・・

良く使うオプション
-e：E-valueの閾値（例：-e 1e-10）

-m：テーブル形式の表示（例：-m 8）
-v：リストの最大表示数（例：-v 5）
-b：アライメントの最大表示数（例: -b 5）

（FASTAのコマンド）
fasta36 -Q test_aa.fa /home/hidekih15/lecture/data/BLAST/db/S_aureus_N315_ORFs_aa.fa > test_vs_S_aureus_N315_ORF.fasta36

# イルミナリードのマッピング（Bowtie 2）

1. 用いるファイル

リファレンスゲノム （黄色ブドウ球菌N315株）
/home/hidekih15/lecture/data/BLAST/db/S_aureus_N315_chr.fa
# NCBIのgenomeから入手； http://www.ncbi.nlm.nih.gov/genome/154?project_id=57837

イルミナリード （100 bp paired-end reads； DRAから入手したもの）
/home/hidekih15/lecture/data/illumina/
MRSA_SRR583008_50x_1.fastq
MRSA_SRR583008_50x_2.fastq

Fastqファイルの例 （.fastq、.fq）
@SRR583008.2:73:B00NUACXX:3:1101:1939:1927 （リード名）
GTATTTCCTAGGACATATAGAGGAACATTCAACGTAAAATTCCCG･･･ （配列）
+ （コメント行）
CBBFFFFFHHHHFHIJJJJJJJIJJJJJJJJJJJIJJIJJJJJJI･･･ （QV値）
http://en.wikipedia.org/wiki/FASTQ_format

2. 作業ディレクトリ
mkdir –p ~/Mapping/ref
cd ~/Mapping/ref
cp /home/hidekih15/lecture/data/BLAST/db/S_aureus_N315_chr.fa .

3. リファレンスのインデクシング
bowtie2-build -f S_aureus_N315_chr.fa S_aureus_N315_chr.fa > S_aureus_N315_chr.log

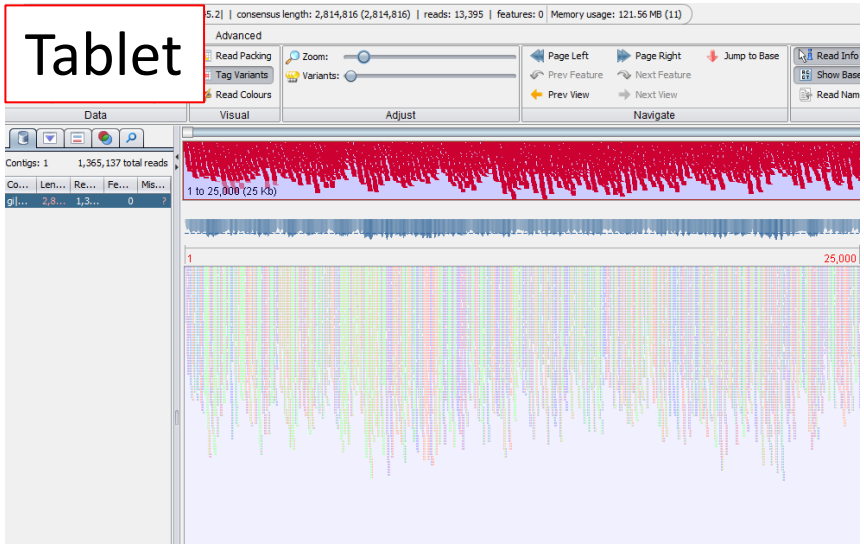
4. リードのアライメント
cd ~/Mapping
bowtie2 -I 200 -X 500 -x ./ref/S_aureus_N315_chr.fa -1 ~/lecture/data/illumina/MRSA_SRR583008_50x_1.fastq -2 ~/lecture/data/illumina/MRSA_SRR583008_50x_2.fastq -p 4 -S S_aureus_pe275.sam >& MRSA_SRR583008_50x_bowtie2.out

## オプションの説明 ##
bowtie2 -I [minimum insert size] -X [maximum insert size] -x [indexed reference name] -1 [paired-end read_1] -2 [paired-end read_2] -p [number of threads] -S [name of output sam file]

# 出力結果

718541 reads; of these:
  718541 (100.00%) were paired; of these:
    66347 (9.23%) aligned concordantly 0 times
    634069 (88.24%) aligned concordantly exactly 1 time
    18125 (2.52%) aligned concordantly >1 times
    ----
    66347 pairs aligned concordantly 0 times; of these:
      26852 (40.47%) aligned discordantly 1 time
    ----
    39495 pairs aligned 0 times concordantly or discordantly; of these:
      78990 mates make up the pairs; of these:
        71945 (91.08%) aligned 0 times
        4958 (6.28%) aligned exactly 1 time
        2087 (2.64%) aligned >1 times
94.99% overall alignment rate

マッピング結果のビューワ：　Tablet, IGV
Tablet:  http://bioinf.scri.ac.uk/tablet/
  samtools index S_aureus_pe275.bam
IGV:  http://www.broadinstitute.org/igv/



# SNPs/indels の検出（Samtools）

1. samアライメントファイルの bamフォーマットへの変換とリファンレンス上の位置に従ったソート
   samtools view -Sb S_aureus_pe275.sam | samtools sort - S_aureus_pe275
   （S_aureus_pe275.bamファイルが出力される）

sam (Sequence Alignment / Map)
bam (Binary version of a sam file)

2. SNPs/indelsの抽出
   samtools mpileup -uBf ./ref/S_aureus_N315_chr.fa S_aureus_pe275.bam | bcftools view -vc -i 0.1 - > S_aureus_pe275.vcf

3. vcf ファイルのフィルタリング（variant qualityに基づく）
   awk '$6>=100' S_aureus_pe300.vcf > S_aureus_pe300.filt-Q100.vcf

VCFファイルの例

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | | |
|---|---|---|---|---|---|---|---|---|---|
| scaffold7 | 1158 | . | A | C | 31 | . | DP=36; .. | GT:PL:GQ | 0/1:61,0,54:56 |

・・・

http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

DP=36：クオリティ
DP4=15,0,20,0：厚み
REFと同一リード数：15（＋）、0（－）
ALTと同一リード数：20（＋）、0（－）

SNPアノテーション

SnpEff, SnpSift: http://snpeff.sourceforge.net
Annovar: http://www.openbioinformatics.org/annovar/

# イルミナリードのアセンブル

1. 用いるファイル
イルミナリード
＊101 bp paired-end reads（インサートサイズ：180 bp）
/home/hidekih15/lecture/data/illumina/
frag_1.fastq
frag_2.fastq
＊ 37 bp mate-pair reads（インサートサイズ：3,500 bp）
shortjump_1.fastq
shortjump_2.fastq

2. SOAPdenovo用の設定ファイル（configure.txt；次ページ）の作成
アセンブル前に設定ファイルを準備する必要がある。
リードファイルの場所、リード長、アセンブル手順等の設定をconfigure.txtファイルに記入。
configureファイルは、http://soap.genomics.org.cn/soapdenovo.html を参考にして独自に作成。

```
cd ~/lecture/Assembly
cp ~/lecture/data/SOAPdenovo/configure.txt .  # PEのみの場合

SOAPdenovo-31mer all -s configure.txt -K 31 -d -D -L 500 -o S_aureus_pe -p 4 > S_aureus_pe.log

cp ~/lecture/data/SOAPdenovo/configure_pe_mp.txt .  # PE と MPを混ぜた場合
SOAPdenovo-31mer all -s configure_pe_mp.txt -K 31 -d -D -L 500 -o S_aureus_pe_mp -p 4 > S_aureus_pe_mp.log

## オプションの説明 ##
 オプションの説明はhttp://soap.genomics.org.cn/soapdenovo.html を参照。
 ## 特に、-K と -mの値を変えて試す。リード長とリードカバレッジによって最適値が異なる。
SOAPdenovo-127merのみの実行でマニュアル表示。
```

# configure.txt

```
[LIB]
#maximal read length
max_rd_len=101
#average insert size
avg_ins=275
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=3
#use only first 100 bps of each read
#rd_len_cutoff=100
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32
q1=/home/hidekih15/lecture/data/illumina/MRSA_SRR583008_50x_1.fastq
q2=/home/hidekih15/lecture/data/illumina/MRSA_SRR583008_50x_2.fastq
```

# configure_pe_mp.txt

```
[LIB]
#maximal read length
max_rd_len=100
#average insert size
avg_ins=300
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=3
#use only first 100 bps of each read
#rd_len_cutoff=100
#in which order the reads are used while scaffolding
rank=1
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32
q1=/home/hidekih15/lecture/data/illumina/MRSA_SRR583008_50x_1.fastq
q2=/home/hidekih15/lecture/data/illumina/MRSA_SRR583008_50x_2.fastq

[LIB]
#maximal read length
max_rd_len=37
#average insert size
avg_ins=3500
#if sequence needs to be reversed
reverse_seq=1
#in which part(s) the reads are used
asm_flags=2
#use only first 100 bps of each read
#rd_len_cutoff=37
#in which order the reads are used while scaffolding
rank=2
# cutoff of pair number for a reliable connection (at least 3 for short insert size)
pair_num_cutoff=2
#minimum aligned length to contigs for a reliable read location (at least 32 for short insert size)
map_len=32
q1=/home/hidekih15/lecture/data/illumina/shortjump_1.fastq
q2=/home/hidekih15/lecture/data/illumina/shortjump_2.fastq
```

# 出力結果

S_aureus_pe.log： ログファイル
S_aureus_pe.contig： コンティグ配列
S_aureus_pe.scafSeq： スキャフォールド配列
S_aureus_pe.scaf： スキャフォールドにおけるコンティグの位置情報

配列情報の解析 （EMBOSS） http://emboss.sourceforge.net/

infoseq -sequence S_aureus_pe.scafSeq -outfile S_aureus_pe.scafSeq.infoseq
infoseq -sequence S_aureus_pe.contig -outfile S_aureus_pe.contig.infoseq