# AlignDB expansion for multiple bacteria genomes

This tutorial describes several scripts helping analysis multiple bacteria genomes. All these scripts extend AlignDB and AlignDB::Multi’s ability and can be seen as an expansion.

## Data source

NCBI lists 1650 Complete Microbial Genomes, which are our source data. <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>

You should download the following data from NCBI ftp server first: 1) the taxonomy database, <ftp://ftp.ncbi.nih.gov:21/pub/taxonomy/> ; 2) the packed bacteria genomes <ftp://ftp.ncbi.nih.gov:21/genomes/Bacteria/> .

Unzip taxdmp.zip in taxonomy to a separate directory, namely e:\data\bacteria\taxdmp\ .

Copy lproks\_0.txt, lproks\_1.txt, lproks\_2.txt, ReadMe.txt, SameSpecies.gi, summary.txt, all.fna.tar.gz and all.gbk.tar.gz to a separate directory, namely e:\data\bacteria\bacteria\ .

Unpack all.fna.tar.gz to its current path.

Unpack all.gbk.tar.gz to its current path.

## Step by step

From now on, the current working directory (cwd) will be AlignDB\bac\.

### Aligndb.ini

Add the follow contents to aligndb.ini.

[bac]

db=bac

base\_dir=d:/data/bacteria/bacteria

taxon\_dir=d:/data/bacteria/taxdmp

strain\_file=bac\_strains.csv

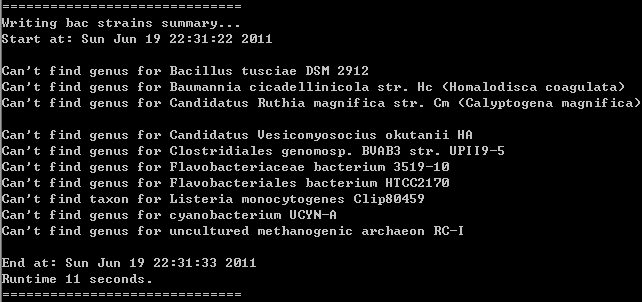
seq\_file=bac\_seqs.csv

species\_file=bac\_species.csv

### bac\_strains.pl

This script will incorporate bacteria strain and taxonomy information, and write two files bac\_strains.csv for strains and bac\_seqs.csv for sequences.

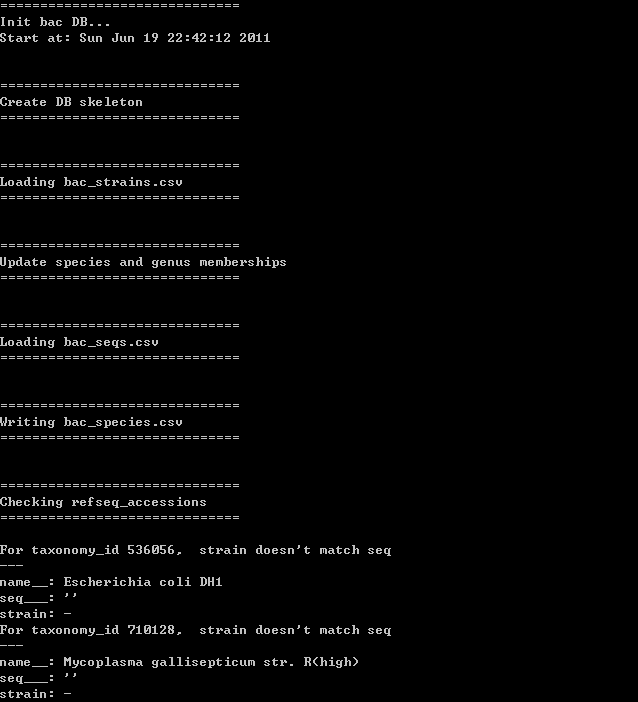
>perl bac\_strains.pl



### bac\_db.pl

This script will read contents from bac\_strains.csv and bac\_seqs.csv, create a database for easy manipulation. It will write a file, bac\_species.csv, containing genus and species information.

>perl bac\_db.pl





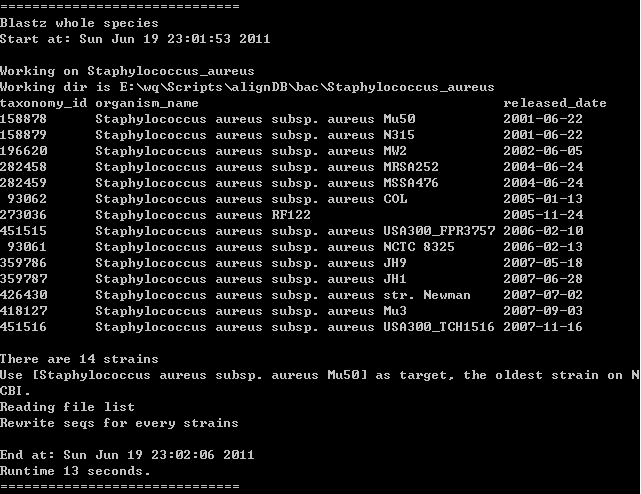
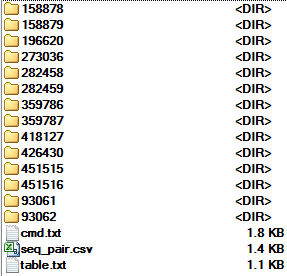
There will be a lot of error messages, you can just ignore them.

### bac\_pre\_aligndb.pl

Because AlignDB.pm need taxonomy and sequence information stored in taxon.csv and chr\_length.csv, separately.  
> perl bac\_pre\_aligndb.pl

### bac\_bz.pl

This script is the most complicated one in the suite. For example, you want to analysis strains of Staphylococcus aureus. You should following these steps:

1. Visit ncbi taxonomy browser, <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>;
2. In the “search for” entry, type “Staphylococcus aureus”, as “complete name”, and press “Go”;
3. NCBI will list a page contains all strains of Staphylococcus aureus. After click the “Staphylococcus aureus” link, you got the full taxonomy information of Staphylococcus aureus. You should find that the taxonomy ID of Staphylococcus aureus is 1280.
4. Run bac\_bz.pl for the first time.  
   >perl bac\_bz.pl -p 1280  
   
5. Some interesting information displayed. There are 14 strains with complete genomes in Staphylococcus aureus. Each strains’ taxonomy id, full name and genome released date were listed.
6. A directory named “Staphylococcus\_aureus” will be generated. It contains these directories and files:  
     
   The directories named after the taxonomy id of each strains, in which there are genome sequences copied from e:\data\bacteria\bacteria\all.fna\. The table.txt file contains the same info as step 4. We will talk about cmd.txt and seq\_pair.csv later.
7. Because of our previous knowledge of Staphylococcus aureus, we pick taxid 93062, Staphylococcus aureus subsp. aureus COL, as “target” genome, which means this genome is the most reliable one. Then we pick taxid 273036, Staphylococcus aureus RF122, as “reference” genome, which means this one is the most divergent one, because it’s another subspecies strains. At last, we exclude taxid 282458, Staphylococcus aureus subsp. aureus MRSA252, because it’s in the same subspecies with other strains, but is more divergent than 273036. To make 273036 more reliable as the outgroup, we can only drop 282458.
8. After many scientific decisions (^\_^), we delete the “Staphylococcus\_aureus” directory and re-run bac\_bz.pl  
   >perl bac\_bz.pl –p 1280 -t 93062 -r 273036 -e 282458
9. In the regenerated “Staphylococcus\_aureus” directory, we will find 13 numerical directories contain genomes. Let’s take a look at cmd.txt.  
     
   It contains four parts. The first is the bac\_bz.pl command line. The second is batch blastz processing. The third will join all pairwise blastz results into a multiple alignment. The forth is cleanup temperate databases. We will execute them one by one.

### seq\_pair\_batch.pl

As we described in bac\_bz.pl, this step generate pairwise blastz alignments between target genome and others. If your computer is powerful enough, you can change the “-p” parallel parameter to 4 even 8.

> perl E:/wq/Scripts/alignDB/bac/../extra/seq\_pair\_batch.pl -d 1 -p 2 -f E:\wq\Scripts\alignDB\bac\Staphylococcus\_aureus\seq\_pair.csv

This step will cost about 30 minutes in my Intel Core i5 PC.

The following pairwise alignments will be generated: 93062vs273036, 93062vs158878, 93062vs158879, 93062vs196620, 93062vs282459, 93062vs451515, 93062vs93061, 93062vs359786, 93062vs359787, 93062vs426430, 93062vs418127, 93062vs451516.

To get more accurate alignments, you should do repeatmasker and trf (tandom repeat filter) on all the genome sequences. We just skip these.

### join\_dbs.pl

We join all the pairwise alignments here.

>cd Staphylococcus\_aureus

> perl E:/wq/Scripts/alignDB/bac/../extra/join\_dbs.pl --no\_insert 1 --trimmed\_fasta 1 --length 1000 --reduce\_end 10 --goal\_db Staphylococcus\_aureus --outgroup 0query --target 0target --queries 1query,2query,3query,4query,5query,6query,7query,8query,9query,10query,11query --dbs 93062vs273036,93062vs158878,93062vs158879,93062vs196620,93062vs282459,93062vs451515,93062vs93061,93062vs359786,93062vs359787,93062vs426430,93062vs418127,93062vs451516

Finally, we got a directory, e:\wq\Scripts\alignDB\bac\Staphylococcus\_aureus\Staphylococcus\_aureus\, containing alignments files in multi fasta format.