# Manual

**Prerequisites**

Perl 5.8 or later, with DBI module installed.

MySQL (version 5 or later).

For debian/ubuntu, Perl and DBI module should have been installed by default.

To check the installation status of those modules, run the following command in the terminal:

$ perldoc DBI

To install MySQL on debian/ubuntu:

$ apt-get install mysql-server

## Setting up the MySQL database

### Create the MySQL user and databases

1. Without root permission

In this section, MYSQL\_DB\_DIR is where you want to place the MySQL data.

mkdir -p MYSQL\_DB\_DIR/data MYSQL\_DB\_DIR/temp

Edit MYSQL\_DB\_DIR/data/my.cnf, make sure it has below contents:

[client]

user=root

socket=MYSQL\_DB\_DIR/mysql.sock

[mysqld]

datadir=MYSQL\_DB\_DIR/data

socket=MYSQL\_DB\_DIR/mysql.sock

tmpdir=MYSQL\_DB\_DIR/temp

log-error=MYSQL\_DB\_DIR/mysql.log

pid-file=MYSQL\_DB\_DIR/mysql.pid

#skip-networking

port=93306 #keep accordingly with get\_species\_taxids.pl and other .pl files

bind-address=192.168.4.98 #mgmt01

Start this MySQL daemon:

mysqld\_safe --defaults-file="MYSQL\_DB\_DIR/data/my.cnf" --no-auto-restart

Or shutdown it:

mysqladmin --defaults-file="MYSQL\_DB\_DIR/data/my.cnf" shutdown

Log into MySQL daemon:

mysql --defaults-file="MYSQL\_DB\_DIR/data/my.cnf"

1. With root permission

mysql -uroot -p

1. Once you have logged into MySQL, create a MySQL user and two databases for blast2hgt:

>insert into mysql.user(Host,User,Password) values("localhost","test",password("1234"));

>CREATE DATABASE taxondb;

>CREATE DATABASE taxnode;

>GRANT ALL ON taxondb.\* TO 'test'@'localhost' IDENTIFIED BY '1234';

>GRANT ALL ON taxonnode.\* TO 'test'@'localhost' IDENTIFIED BY '1234';

>flush privileges;

>QUIT;

### Import required data to MySQL

1. Download taxdump.tar.gz (nodes.dmp and names.dmp are used):

$ cd $path\_to\_blast2hgt/install\_database/

$ wget <ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz>

And decompress to $path\_to\_blast2hgt/install\_database/;

1. Get all taxid from NCBI nr database:

$ install\_database/extract\_acc.taxid\_from\_nr.sh

1. Now import them into MySQL databases:

$ cd install\_database && ./importdatabase.sh

This step may take hours to finish depending on the hardware performance;

However, if it lasts for a week, please check.

## Building (split) BLAST databases (nr or nt) for each taxonomy group

See species\_taxids/create\_species\_taxids.sh.

Or use 'get\_species\_taxids.sh' (officially released by NCBI), if you have a good internet connection to NIH.

## BLAST against each taxonomy group

An example script is placed in bin/blastp\_each.taxids.final.sh, please modify it according to your study design and HPC environment.

## Running blast2hgt

See 0run.sh for detail, and modify it when needed.

At this step, a table (\*.rp.tsv) containing the genes (proteins), and their horizontal transfer probabilities (measured by alien index / bitscore difference / ratio of bitscore difference), as well as donor/recipient, will be produced. Users are required to manually screen HGT candidates from this table.

## Reconstruction of phylogenetic trees to verify HGT

To confirm the HGT candidate, a phylogenetic tree is the golden standard.

### Balanced sequence sampling from each taxonomy group

This part gets sequences from each taxonomy group, using a balanced sampling strategy. The top *n* best BLAST hits (based on bitscore) are selected.

lineage2idgroup3.batch.pl: get a maximum of 10 sequences per taxonomy group; and 1 sequence per species within self-group. Recommended to new users.

lineage2idgroup3.unlimited.pl: unlimited sequences per taxonomy group.

Command:

perl phy/lineage2idgroup3.batch.pl [representatives.txt] [\*.lineage] [id\_file]

[representatives.txt]: the defined taxonomy groups, could be slightly modified based on study design.

[\*.lineage]: the merged $query.nr.lin and $query.nr.taxid produced by 0run.sh. Here it could be referred as:

`cat $query.nr.lin $query.nr.taxid`.

[id\_file]: A file containing HGT candidate IDs obtained from step 4. One ID per line.

At the end of this part, a \*.id.org file will be generated for each HGT candidate listed in [id\_file].

### Sequence extraction and phylogenetic tree reconstruction

Examples can be found in below folders:

Sequence extraction: phy/phy\_verify.sh.

Phylogenetic tree reconstruction: phy/phy.sh.

You may need to modify above two files according to your HPC environment.