MASSEY UNIVERSITY
Institute of Fundamental Sciences

LGTmate

User's manual

Table of Contents

Icons used in this manual:


Instructions for the GUI version of LGTmate


Instructions for the command line version of LGTmate

1. Introduction

The rationale behind LGTmate

The *LGTmate* pipeline is designed to identify putative LGTs (Lateral Gene Transfer events) in any animal, yeast, or plant protein dataset. It is simple to install under Windows, Linux and MacOS environments, and can be executed using a GUI front-end or command line. The program has been designed to perform well on standard desktop computers and the output is easily accessible with standard spread sheet software (Excel® or Libre office). The goal of LGTmate is to provide an easy to use tools for life scientists that allow them to identify putative LGTs that can then be verified using independent molecular and evolutionary methods.

1.1 Overview of how LGTmate works

The core feature of *LGTmate* is the presentation of BLASTP results in a format that allows efficient exploration and manipulation of the data using ubiquitous spreadsheet software.

Firstly, the program performs a search of the proteome under investigation ("**query**") against three separate databases called "**source group**", "**broad group**", and "**narrow group**". The source group database provides the BLASTP information used to identify possible LGT donor species, with the user able to choose between bacteria, fungi or plants. The broad and narrow group databases contain proteomes from species more closely related to the organisms under investigation, and provide a means of identifying hits to highly conserved slowly evolving genes that would incorrectly appear to be LGTs. For example, in a search for insects LGTs we would include representative proteomes from mammals to nematodes in the broad group, whilst the narrow group would contain sequences from arthropods only. Please see Section 4 for a detailed explanation of the setup of these important databases.

Secondly, the obtained BLAST bit scores are inserted into a table in combination with other information pertinent to the discovery of LGTs.

Thirdly, the BLAST bit scores and other useful information, such as the scaffold name, gene position, and the protein sequence itself are combined in a tab-delimited file that is best viewed using Excel® or LibreOffice Calc.

A detailed explanation of all steps is provided in Section 3, but for impatient uses a quickstart guide is provided in Section 2.

1.2 Important setup information

LGTmate requires no pre-installed software* and the installation procedure is very simple: the user can download the compiled binaries' package corresponding to their preferred operative system (Windows, Linux, Mac) from http://sourceforge.net/projects/lgtmate and extract it in a chosen folder.

Warning: due to BLAST limits, it is strongly advised to extract the *LGTmate* package in a folder that does not contain spaces in its path. E.g.:

C:\Documents and settings\my folder ✗

C:\Path_without_spaces\my_folder ✓

Once the downloaded compressed folder is expanded it will have the structure shown in Figure 1.

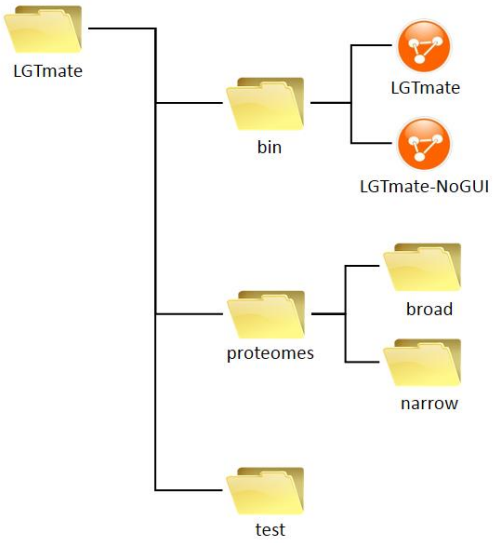

Figure 1

- ***bin*** : this folder contains all the files and folders needed for the execution of *LGTmate,* in particular the executable files *LGTmate* and *LGTmate-NoGUI*. **These files cannot be moved outside this folder, so it is recommended that the user creates a shortcut** (Right click on the file → Create a shortcut), **to be placed in a handy location**. The *blast_db/* folder contains the BLAST databases for the **sources** organisms, bacteria, plants and fungi.

  Advanced users: the *src/* folder contains the Python source code.

* Python 2.7 is required to run the source scripts, for advanced users and debugging.

- ***proteomes*** : this folder contains two sub-folders, **narrow** and **broad**, each containing **.fasta* files of organisms, respectively more or less related to the query, that the user wants to scan for LGTs.

- ***test*** : this folder contains the files needed for a "test run", to check if the program is working normally on the system with the provided dummy databases.

2. Quick start guide

This section is guide for impatient people; however, we strongly recommended that the user reads Section 4 regarding how to correctly setup the source, broad and narrow databases as they are critically important for proper data generation!

1) Add sequences in **.fasta* format to the broad and narrow database folders located in the proteomes directory (Fig. 1).

2) Open the GUI application by clicking on the **LGTmate** executable inside the *LGTmate/bin* folder (Fig. 1)

3) Select the query proteome and optional GFF file

4) Choose one or more Source kingdoms for LGT donors

5) Select a output folder

6) Hit "Go!"

Run time will vary depending on the size of the search databases and the speed of the computer. Once the program is complete a text file containing the results will be produced for each Source kingdom selected in step 4, these results are best viewed in Excel® or LibreOffice Calc. A detailed description of the output tables is found in Section 3.4.

3. Execution

The program's workflow; explanation of main commands and functions

The program comes in both a GUI and a command-line (No-GUI) version. Each version has its own advantages: the GUI is surely more straight-forward and easier to use, while the command-line allows more control to the advanced user. In any case, the workflow and speed of execution will be the same.

3.1 The GUI

To open the graphical interface (Fig.2), double click either on the *LGTmate* executable inside the *bin/* folder, or on a previously made shortcut (see Section 1.2).

→ The **"query"** proteome, in *fasta* format.

→ *Optional:* A file containing information about genes for the **query** proteome, in *GFF* format.

→ Selection of the **"source"** database(s) for possible LGT donors.

→ *Optional:* it is possible to exclude species contained in the *proteomes/broad* and *proteome/narrow* folders from the search, in order to skip the BLASTP search for close relatives.

→ Output folder selection

→ Number of processors to be used (defaults to 1), and start/stop buttons.

→ Console emulator: useful to check the state of the program, and follow its execution.

Figure 2

Once all the necessary fields are filled, and at least one **source** kingdom has been checked, the "***GO!***" button will become clickable. Once started, the process can be stopped using the abort button next to it.

NOTE: during some steps (i.e. the BLAST analysis), the aborting procedure can take a long time.

3.2 Workflow

The *LGTmate* algorithm proceeds in a stepwise fashion through the following workflow:

1. Input reading

In order to run, the program will need:
• The **"query"** proteome under investigation for LGTs, in *fasta* file format;
• The **"source"** kingdoms' databases against which the proteome will be screened for LGTs.
• *Optional:* Information about the proteome's genes and their location, in *GFF* file format;
• *Optional:* The names of species contained in the customizable folders *proteomes/broad* and *proteomes/narrow* that the user wants to **exclude** from the search (e.g. in case they are evolutionarily too close to the query, in order to avoid trivial results, see Section 4.1). If not specified, *LGTmate* will screen against all the species contained in these folders.
• *Optional:* The number of processors to be used for calculations. The more processors working, the faster the program will be. If not specified, one processor will be used.

2. Input processing

After having collected the needed information, *LGTmate* will analyze the query and the GFF files (if provided), in order to obtain the gene sequences.

3. *Broad* and *narrow* databases construction

The program will dynamically build the *broad* and *narrow* databases using the files found in the *proteomes/broad* and *proteomes/narrow* folders, possibly excluding any selected by the GUI in Step 1.

4. **BLASTP execution**

   LGTmate will now execute the BLASTP tool to screen the query's genes against the *narrow, broad* and *source* databases, looking matches that could possibly indicate lateral gene transfer events.

5. **Output generation**

   Finally, the results will be written into a .txt file, that can be viewed and edited using the most common spreadsheet software (Microsoft Excel®, OpenOffice Calc, LibreOffice Calc). See Output description (Section 3.4) for more information.

3.3 The command line

To use the command line version of *LGTmate*, open a terminal window (Ctrl+Alt+T in Linux, Start→Accessories→CommandPrompt in Windows), navigate to the **bin/** program folder and execute **LGTmate-NoGUI**.

3.3.1 The Configuration File

The user must supply the required files (see Section 3.2, Step 1) by creating a **configuration file** (see Fig. 3 and example file in ***test/configure.txt***) containing the following lines:

1. **LIBRARY line:** custom library name, in square brackets
2. **FASTA line:** path to the **query** *fasta* file;
3. **GFF line:** path to the *GFF* file relative to the query;
4. **KINGDOM line: source** databases against which screen the query for LGTs; insert ***p*** for *plants*, ***f*** for *fungi* and ***b*** for *bacteria*. Letters must be **comma separated.**
5. **EXCLUDE line:** optional, the name of *fasta* files in the *proteomes/broad* and/or *proteomes/narrow* folder to be excluded from the search. File names must be **comma separated.**

```
[Lib1]                          → LIBRARY line
FASTA=../test/dp_subset.fa      → FASTA line
GFF=../test/dp_subset.gff       → GFF line
KINGDOM=p,f                     → KINGDOM line
EXCLUDE=danio_rerio.fasta       → EXCLUDE line
```

Figure 3

A single configuration file can contain multiple libraries for multiple experiments, one after another.

3.3.2 Launching LGTmate

The program can be launched using the following command.

Linux:
```
./LGTmate -c [path to config file] -p [number of processors] -v
```

Windows:
```
LGTmate_NoGUI.exe -c [path to config file] -p [number of processors] -v
```

Arguments:
-c **:** indicates the path to the configuration file (see previous section)

-p : optional; indicates the number of processors to be used for calculations. Defaults to 1.

-v : optional; if present, *LGTmate* will run in "verbose" mode. Mainly for debug purposes.

The option **–h** will provide a short help.

All the **results** will be stored in a folder called "**final_output**" that can be found in the same location as the configuration file.

3.4 Output description

LGTmate stores results into as many text files as the number of **source** databases chosen, with this name format:

[name of query fasta file]_[source]_pep_tab.txt

i.e. if the query is the file "*drosophila.fasta*" and it is screened against fungi, the resulting file will be named*: drosophila_fungi_pep_tab.txt*

The author recommends to open the results file with a spreadsheet editor like Microsoft Excel® or LibreOffice Calc.

Once opened, the user will find a table similar to the one shown in Fig.4.

gene#	chm	pep_name	st_pos	end_pos	user_flag	broad	narrow	fungi	broad_bitscore	narrow_bitscore	fungi_bitscore
1	DPSCF300225	DPOGS200016-PA	115452	123493		1	1	1	404	1544	141
2	DPSCF300225	DPOGS200017-PA	194542	199629		0	1	0	0	1050	0
3	DPSCF300225	DPOGS200015-PA	216606	217651	CHECK	0	1	1	0	281	400
4	DPSCF300225	DPOGS200018-PA	219553	219920		0	0	0	0	0	0
5	DPSCF300225	DPOGS200014-PA	233892	239358		1	1	0	184	377	0

Figure 4

Each row of the table shows information for a given sequence in the **query** proteome. The "gene#" column is used to sort protein by order within a scaffold. The column headings "chm", "pep_name", "st_pos" and "end_pos" provide details of the source scaffold, protein name, and its position coordinates. The "user_flag" column allows user mark-up of the data, and is also used by *LGTmate* to label proteins with "CHECK" if the source BLAST bit score is greater than that obtained for both the broad and narrow databases (which indicates a strong LGT candidate) and is greater than or equal to 100. The next six columns provide information on the BLAST hit score for this protein obtained from BLASTp searches against the broad, narrow and source databases represented in binary format or as bit-scores; finally, the last four columns show the IDs of the best hits for each database and the peptide sequence, respectively.

3.5 Typical workflow for analyzing LGTmate output

When using standard spread sheet software (we highly recommend LibreOffice Calc) one typical workflow is as follows. First, use a descending sort of BLAST bit scores in the source group column, followed by ascending sorts of the corresponding broad and narrow animal databases. Immediately, strong candidate LGTs can be identified based on the protein having much larger BLAST hit score in the source column versus those obtained in the broad and narrow databases. For these strong candidate proteins, the user can manually add a searchable character (for example "@") so they can be easily identified when the table is resorted. Next, if the optional GFF file was provided, an ascending sort on the "gene#" order column is performed to re-capitulate the order of proteins along the scaffolds. Next a text search can be used to identify proteins with the LGTmate flag "CHECK" and/or the user specified search character. The information included in the table can be used to investigate the position of the candidate LGT within a scaffold, the nature of genes surrounding it. Finally, one can use the attached peptide sequence to perform an independent BLASTp searches against the complete NR database at NCBI or EBI. This latter step is important because for practical reasons the LGTmate databases used to create Blast hit scores are much less representative than those hosted at NCBI.

Strong LGT candidates (I) have a significantly larger Source database blast hit scores than those obtained from the broad and narrow sets(II) are surrounded by genes clearly derived from vertical descent (III) are in large well assembled scaffolds (lacking long or frequent stretches of missing data). Conversely, contaminating scaffolds are readily identified by all of peptides in a scaffold either having a larger source BLAST hit score or a 0-0-1 binary pattern for broad, narrow, and source groups, respectively. Other artifacts like chimeric proteins and BLAST scores based on low complexity proteins sequences can also be visually identified based on information returned by the manual NCBI BLAST search.

4. Databases Setup

As described in the previous sections, *LGTmate* uses three databases for the BLAST search at the core of the LGT detection algorithm: the ***narrow, broad*** and **source** DBs.

The first two, located in the folder "*proteomes/narrow"* and *"proteomes/broad",* are represented by a simple collection of *fasta* files of species more or less related to the query, respectively. Therefore, they are easily and quickly customizable, as shown in the following section.

The **source** databases, on the other hand, are a large collection of proteomes for the three kingdoms of **bacteria, plants** and **fungi**, and should be more "stable" and not frequently modified; they are saved in the ordinary BLAST-db format, and not directly editable. For download size reasons, the databases are not included in the LGTmate package; they should be downloaded either using the *dbDownloader* tool (recommended) or manually (for advanced users) – see Section 4.2.

4.1 Customizing *narrow* and *broad* DBs

The ***narrow*** and ***broad*** databases are automatically generated from protein sequences found in the two folders "*proteomes/narrow"* and *"proteomes/broad",* and can be customized by the user by **simply adding or removing *fasta* files of species closely related to the organism under investigation**.

For example, in a search for insect LGTs, the user would include representative proteomes from mammals and nematodes in the *broad* group, whilst the *narrow* group would contain sequences from only the arthropods. It is important that species used in these databases are carefully chosen, especially for the *narrow* group as inclusion of species closely related to the organisms under investigation will "hide" LGTs events that occurred in their common ancestor. For this reason the *LGTmate* GUI displays the sequences currently found in the *broad* and *narrow* database folders and allows the user to easily exclude sequences as required (under the option "*I want to exclude some organisms from BLAST search"*).

HINT: Use the species names as the file names (but use "_" in place of spaces) for sequences in the broad and narrow directories as you may want to exclude these sequences in subsequent runs.

Well-known sources of proteome files in *fasta* format are:

ftp://ftp.ncbi.nih.gov/genomes/ (for each species, look for the "protein" folder, and the *.fa* or *.faa* compressed sequence file)

http://www.uniprot.org/taxonomy/complete-proteomes (for instructions, please refer to this page).

4.2 Downloading the *source* DBs

Before the first LGTmate run on the user's machine, the **source** databases must be populated by downloading their up-to-date versions from the web; once obtained, they will be used for all the future LGTmate runs.

4.2.1 Automatic download using *dbDownloader (recommended)*

GUI version
Once the GUI is open, the user can click on "Menu" in the top-left corner, and click again on the *dbDownloader* button. A small window will open (Fig. 5).

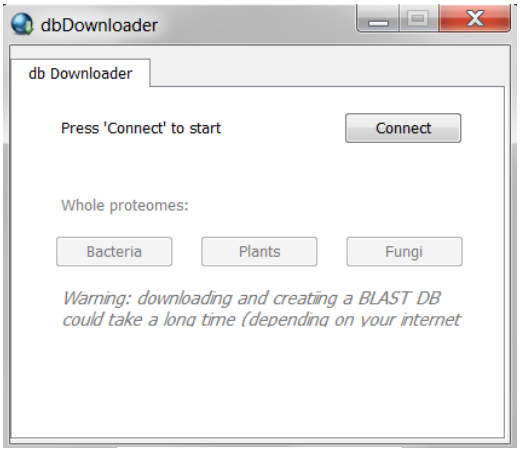

Figure 5

From here, with a single click on the "***Connect***" button it is possible to connect to the ftp sites of NCBI and ENSEMBL; then, the user can choose which database to download by clicking on the correspondent button (in case a window pops up asking the permission to overwrite the present databases, answer "Yes"). *dbDownloader* will take care of downloading the files, unzip and join them, and convert them into BLAST-db format. After this, the window can be closed and LGTmate will be ready.
Warning: depending on your internet connection speed, the download can take a long time, up to hours.

Command line version

To execute *dbDownloader* from the command line, the user must issue the following command:

Linux:

```
./LGTmate  --db-download [kingdom]
```

Windows:

```
LGTmate-NoGUI  --db-download [kingdom]
```

where [kingdom] must be replaced either by the word **bacteria, plants**, or **fungi**, e.g. (Linux):

```
./LGTmate  --db-download bacteria
```

4.2.2 Manual download and customization (advanced users)

In order to manually set up the **source** databases, the user must have a collection of proteomes in *fasta* format for each kingdom. These can be obtained, for example, from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/) or the ENSEMBL ftp site (ftp://ftp.ensemblgenomes.org/pub/). Once downloaded and ucompressed, in case of multiple *fasta* files it is recommended to join them in a single long file for each kingdom (e.g. using the **cat** command in Linux), and then, using the **command line**, feed this file to the *Makeblastdb* utility supplied in the "*bin/blast_bin/"* folder.

E.g. for bacteria (for other **kingdoms** just replace the word "*bacteria"* with "*fungi"* or *"plants"*):

Linux (32-bit):

```
makeblastdb32 -in [kingdom_fasta_file] -max_file_sz 5000000000 -dbtype
prot -title 'bacteria_db' -out ./bacteria_db
```

Linux (64-bit):

```
makeblastdb -in [kingdom_fasta_file] -max_file_sz 5000000000 -dbtype prot
-title 'bacteria_db' -out ./bacteria_db
```

Windows (32-bit):

```
makeblastdb.exe -in [kingdom_fasta_file] -max_file_sz 5000000000 -dbtype
prot -title 'bacteria_db' -out ./bacteria_db
```

Windows (64-bit):

```
makeblastdb64.exe -in [kingdom_fasta_file] -max_file_sz 5000000000 -dbtype
prot -title 'bacteria_db' -out ./bacteria_db
```

This command will produce **3 files**, which in the case of *bacteria* will be: *bacteria_db.phn, bacteria_db.pin, bacteria_db.psq*. **These resulting files must be manually moved to the *"bin/blastdb" folder for LGTmate to work correctly.***

5. Troubleshooting

1. **As soon as the program starts, I get errors like:**

```
'<some path>' is not recognized as an internal or external command, operable
program or batch file'
Traceback (most recent call last):
    File "<string>", line 506, in run
    etc..
```

Solution: In order to solve this, please make sure that the *LGTmate* folder path does not contain any white spaces. In Windows, it is very likely that folders like "Documents and settings" are the cause of the problem. See program setup.

2. **I get a warning telling me "WARNING: LGTmate could not find some [any] correspondences for the ID or NAME fields of the GFF file in the fasta file"**

LGTmate uses the GFF file to obtain important coordinate information for each protein in the query file. To do this there needs to be an identifier shared between the last column of the GFF file and the fasta header of each sequence (protein name/ID or NAME). Unfortunately, due to the flexibility associated with information placed in GFF files there is no guarantee the sequence identifier used will exactly match the name in the fasta file. LGTmate includes three different matching routines which could solve trivial mismatches, but in some cases manual correction could be necessary, otherwise LGTmate will ignore the GFF input.

The example below (Fig. 6) shows a common mis-encoding between the information in the final column of the GFF and the header of the corresponding fasta file. Such situations are easily fixed using "find and replace" features to remove of modify the sequence IDs so they are found in an identical format. In some cases though a script might be required to correctly pair up GFF and fasta file sequence ID's, however, this is should be a trivial task for a colleague with rudimentary scripting knowledge.

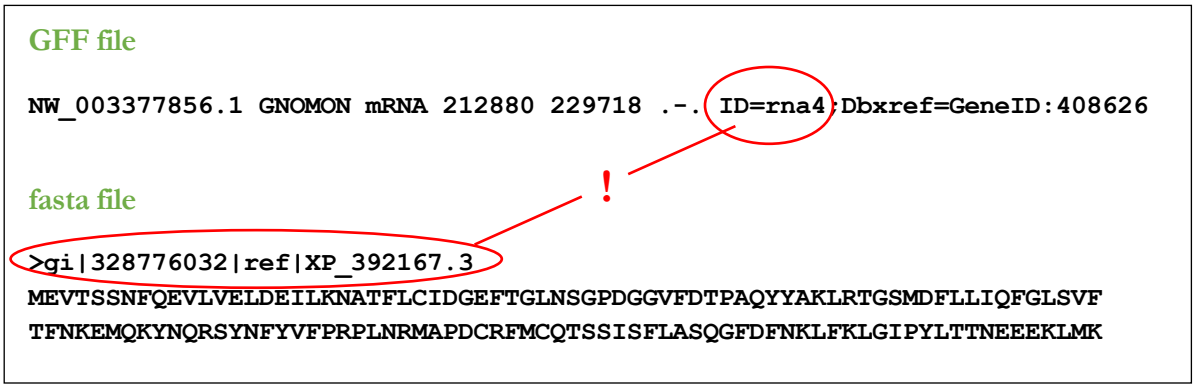

Figure 6

6. Appendix

6.1 Setting up and executing a test run

The downloaded package comes with a set of "dummy" files and databases useful for testing the program and for debug purposes. Specifically, the user can find these files in the ***test/*** folder:

- *configure.txt :* configuration file, needed for **command-line** execution. This file can be useful as a template for future runs.
- *dp_subset.fa :* proteome of a sample insect, in *fasta* format.
- *dp_subset.gff* : optional GFF file relative to the insect's proteome.
- *test_blast_db/*: folder containing reduced versions of the **source** db

To run this test using the **GUI** version of *LGTmate*, the user must:

1. Copy all the files contained in *test_blast_db/* inside *bin/blast_db*
2. Open LGTmate
3. Fill the "*Fasta file*" field and optionally the "*GFF file"* field with the path to *dp_subset.fa* and *dp_subset.gff* respectively
4. Check one or more **source** kingdoms
5. Indicate an **output** folder
6. Click "*GO!".*

To test the **command line** version, copy all the files contained in *test_blast_db/* inside *bin/blast_db,* open a terminal window (see Section 3.3) and execute:

Linux:
```
./LGTmate -c [path to configure.txt] -p [number of processors]
```

Windows
```
LGTmate_NoGUI -c [path to configure.txt] -p [number of processors]
```

The results will be saved into the folder "*test/final_output".*

NOTE: once the tests are completed, it is highly recommended to use the *dbDownloader* tool to automatically download **bacterial, plant and fungal databases** in order to obtain reliable results (see Section 4.2).