NAME

getMetaData.pl

SYNOPSIS

% perl getMetaData.pl <SCRIPT DATA TO USE>

```
 Example:  % perl getMetaData.pl  1
```

Means use the results from Script 1 (cutter.pl) as input. The data directory is set to 1_cutter/split_texts/data_all_chr.

```
 Another example:  % perl getMetaData.pl 0
```

Sets the data directory to 0_extract_from_dB/data_all_chr. Input comes from the results of Script 0 (extract_ALL_chrs).

DESCRIPTION

Summary

GetMetaData.pl is the fifth of a suite of scripts designed to assist in the analysis of DNA. This script assumes that all other scripts in our software pipeline have already been run.

This particular script quieries a database to gather metadata about the bugs in the data directory. Data gathered includes the organism's reference sequence, super kingdom, group, genus, species, strain, oxygen requirements, habitat, temperature range, and pathogenic data.

Please note that this script requires the DBI module and the DBD-mysql module.

INPUT

Genomic .fna files

This script takes input in the form of .fna files produced by the scripts extract_ALL_chrs.pl, cutter.pl, and/or motifCounts.pl.
The fna files are text files primarily containing the letters A, T, C, and G.

Command line Arguments

The script takes a single argument dictating which script's data directory to use as input.

```
 Use this table as a reference for valid inputs:
 Value            Data Directory
   0   ----------- ../0_extract_from_dB/data_all_chr/
   1   ----------- ../1_cutter/split_texts/data_all_chr/
   2   ----------- ../2_countMotifs/cache_counts/
```

Connecting to Your Database

Once you have a database ready you will need to make a few minor edits to the script so
it can connect to your database. Search this file for the mysql_dbh subroutine:

```
 sub mysql_dbh {
```

Modify these four lines, replacing this generic data with your database's access info:

```
 my $db        = 'test';
 my $host      = 'localhost';
 my $user      = 'GenomicsUser';
 my $pass      = '';
```

For $db enter the name of the (MySQL) database you are using.
   Ex: 'GenomeDatabase'

For $host enter the name or address of the server your database is on.
   Ex: 'WheatonGenomics'

For $user enter your MySQL username on the database.
   Ex: 'wsmith'

For $pass enter the user's corresponding password (or leave it blank).
   Ex: 'lollipop'

OUTPUT

This script produces a tab-delimitted Excel (.xls) file in the folder 4_extractGroupPhylum/results. The file will be named <SCRIPT DATA TO USE>_metadata.xls.

Ex: Using the results from Script 2 would produce 2_metadata.xls.

```
Partial example:
ref_seq     superKingdom   Group           Genus            ...   Habitat
NC_009925   Bacteria       Cyanobacteria   Acaryochloris    ...   Aquatic
NC_008009   Bacteria       Acidobacteria   Acidobacteria    ...   UNKNOWN
```

AUTHORS

Mark LeBlanc
Nick Faulconer

MODIFICATION HISTORY

June 17, 2010 (nkf) --

Removed Wheaton database info from the mysql_dbh subroutine.
Replaced it with generic data to be filled in by future users.

June 4, 2010 (nkf) --

Removed most of the next/exit hacks and the associated code.
Fixed the run-time stopwatch.
Got rid of the old argument list (they didn't actually do anything).
Added a new argument that lets the user pick which script's results to use.
Improved this documentation.

June 3, 2010 (nkf) --

```
Made this script platform-independent.
Cleaned up the internal documentation a bit.
Wrote this POD (readme file).
```

April 16, 2009 (mdl) --

```
Split from mergeCounts.pl
HACK to just dump bugs with their GROUP(phylum)
see NEXT and EXIT (hacks)
```

February 12, 2009 (mdl) --

```
adding dB hit to fetch protein table data to compute %genic per chunk;
plan is to add this to the metadata at the end of each chunk's vector
```

January 12, 2009 (mdl) --

```
 don't read the total number of motifs per file in first line; start with
AAAA <count>, etc;
 this will make it consistent with the way we count in the NORMALIZED
version
```

January 09, 2009 (mdl) --

```
adding the dB hits to fetch metadata;
assumes we are running LOCALLY on lexomics (see mysql_dbh() at bottom)
```

January 06, 2009 (mdl) --

```
adding metadata at the bottom of columns;
for now, only:  "genus", "species", "strain" names
```

November 19, 2008 (mdl) --

```
hack together a script to merge individual vectors of Lmers
```

COPYRIGHT INFORMATION