# NAME

```
new_mergeCounts.pl
```

# SYNOPSIS

%perl new_mergeCounts.pl x y z

x = <LMER>

y = < all | genic | intergenic >

z = < chr = chromosomes | plasmids >

ie new_mergeCounts.pl 4 all chr

Script 2 motifCounts.pl must have already been run

The packages DBI and DBD_mysql must be installed

# DESCRIPTION

## SUMMARY

This script takes the various motif counts created by the motifCounts.pl script and combines them into an single .xls file for use in satistical anaylsis and also adds some additional metadate

## INPUT

### motifCount .xls files

This script takes input in the form of .xls files created by the script motif.pl and stored up one level in the directory ../2_countMotifs/cached_counts. The xls file is in the format of tab seperated values containing the precentage of bases that are G or C and the number of times each motif is found in the chunk the .xls file repersents

### Command line Arguments

The program takes 3 command line arguments. The first argument tells what size LMER's to use. The second specifies what part of the genome to look for data from which may be entire genome, the genic portions, or the intergenic portions. Finally the third argument says whether to read chromosomes or plasmids.

### 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

The program outputs a single .xls file containing the motif counts for every single chunk of the genomes of every organism, as well as the cgroup,genus, kingdom, salinity, species, tempRange, xGCpercent , and xProportionGENIC for every single chunk.

## AUTHORS

```
 Mark D. LeBlanc
 Donald W. Bass
```

## MODIFICATION HISTORY

### 6/17/2010 (nkf)

Removed Wheaton database log-in info from the mysql_dbh subroutine. Users must replace the generic data with their own database access info.

### 6/14/10 (dwb)

Updated overlap detection code to handle genes overlapping the OR

### 6/03/10 (dwb)

Made code theoretically platform independent

### 6/02/10 (dwb)

Wrote pod documentation

### 08/05/2009 (mdl)

added two new metadata columns: temp_range and salinity

### 05/11/2009 (mdl) --

need to rotate rows->cols; R does not like columns with two types of data; this will now produce columns with either numeric (motif counts) or text (metadata); hopefully, this will speed-up R's read.table()

also changed output to print COMMA-DATA so last column will *not* have a final comma (messes with mKahn in R)

### 04/23/2009 (mdl)

added read of %GC from header line of each wordCount (input) file; this metadata is now added to other metadata for each column (chunk)

### 03/31/2009 (mdl)

need to handle the situation where genes within a chunk overlap each other; NOTE: currently, i'm only peeking back at the previous gene for potential overlap; if in a series of genes (1,2,3), gene3 overlaps gene1, i currently will *not* catch that :(

### 02/12/2009 (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

### 01/12/2009 (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

### 01/09/2009 (mdl)

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

**01/06/2009 (mdl)**

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

**11/19/2008 (mdl)**

hack together a script to merge individual vectors of Lmers

## COPYRIGHT INFORMATION