NMPDR Releasing GFF3 Files
Standard Operating Procedure NMPDR|SOP006

I. INTRODUCTION

We regularly release our data to BRC-central via GFF3 files. This page describes the steps to release the data.

II. SCOPE

This SOP applies to the procedures to upload NMPDR data to the BRC Central site.

III. APPLICABLE REGULATIONS AND GUIDELINES

NMPDR Contract	Delivery of NMPDR SOP's
BRC Metrics	Production of metrics
GO	List of GO terms
Transaction Logging	NMPDR Logging requirements

IV. RESPONSIBILITY

This SOP applies to those members of the NMPDR research team involved in uploading data. This includes the following:

Principal Investigator
Bioinformaticians

V. DEFINITIONS

The definitions found here: http://www.theseed.org/wiki/Glossary, apply to this SOP, as well as the following:

Standard Operating Procedures (SOPs): Detailed, written instructions to achieve uniformity of the performance of a specific function.
Subsystem: A collection of functional roles that together implement a specific biological process or structural complex.
FigFam: Protein families. Each family is intended to contain a set of globally similar proteins that implement the same function.
Annotation: A tuple consisting of a date, annotator name, and textual message.
Structured Annotation: An annotation where the text is structured.

There are two kinds, a) placing a gene within a subsystem and b) assigning a function to a gene. (See attachment "A" for a description of the structure).

VI. PROCESS OVERVIEW

1. Create the files
2. Upload the files

3. Check the upload

Vii. Create the files

1. Choose a machine that is up to date, and create an empty directory. For this example, we'll use the directory NMPDR

2. Run the command nmpdr2gff NMPDR. This looks through all genomes for the NMPDR flag, and if it is found then a GFF3 file is created using the seed2gff command. If you suspect that files are not created for some genomes that should have them, then the NMPDR flag has not been set.

3. If you would like to create a gff3 file of a single organism, you can use seed2gff with just that organism.

The creation takes about 30-40 seconds per genome, so you can expect it to run for some time.

Once complete you will have a directory structure that looks something like this (only the first two genomes are shown for each organism):

- NMPDR
  - Campylobacter
    - Campylobacter.coli.RM2228.gff3
    - Campylobacter.jejuni.subsp.jejuni.84-25.gff3
    - ...
  - Listeria
    - Listeria.innocua.Clip11262.gff3
    - Listeria.monocytogenes.EGD-e.gff3
    - ...
  - Staphylococcus
    - Staphylococcus.aureus.RF122.gff3
    - Staphylococcus.aureus.subsp.aureus.MRSA252.gff3
    - ...
  - Streptococcus
    - Streptococcus.pneumoniae.R6.gff3
    - Streptococcus.pyogenes.MGAS10270.gff3
    - ...
  - Vibrio
    - Vibrio.cholerae.MO10.gff3
    - Vibrio.cholerae.O395.gff3
    - ...

VIII. Uploading the files

One the creation of the GFF3 files is complete, use the brc-central validator to validate and upload the data to the site. This requires the GO::Parser PERL module. This should be part of the standard install everywhere in the SEED.

Use this command to validate and upload our data:

```
gff3_validator.pl -b NMPDR -d /path/to/directory/NMPDR -p CDS
```

One this has completed you should ftp to ftp://ftp.brc-central.org and check that the files are correct. If there are problems with the validator or upload you should email Todd Creasy at TIGR for help.