Starter cookbook to processing genome w/ annotation pacakges for NCBI

Katherine Amrine

02 February 2015

Alrighty, here is my step-by-step and explanation for what I did to do a WGS annotated genome submission for NCBI

Step 1 is to notify NCBI you will be submitting data

create a bioproject submission per genome.

Specify that you will be submitting WGS, Annotated, SRA, etc.

Next create a biosample submission

include details on your sample

Next start the WGS submission process

here, I think they will ask you for a prefix, and in parenthesis, it will say "leave blank if you want NCBI to assign one for you". I am paraphrasing because I didn't realize this was important until I had already missed this step. It is important because that prefix will be what you have to start all your gene names with. When you get down to the `newIDs.new.ncbi.txt` file I ask for in Step3, this prefix is what needs to go in this file.

also, they will ask you for the type of evidence that you have for the strings of N's within sequences. First off, N's do NOT occur as singletons, so they are not ambiguous bases. Ns refer to gaps of the size of the string of Ns if they ask you.

WHAT YOU WILL NEED

-fasta file with scaffolds with concise names this means that you really need to slim down names to as simple of identifiers as possible. -gff version 2 file (with the scripts I provide here) you don't need version 2, but I used it and wrote my scripts to use this type of file.

WHAT I GIVE YOU

a crappy script to do the first conversion of the gff file to the tbl

file required for the tbl2asn script that will ultimately make your asn submission file

another crappy script to substitute in the names necessary

Step 2 is to pre-process the peptide files

isolate the peptide sequences with internal stops (check number of seqs)

```
fasgrep -s "[^A-Z].+" <peptide_file.fas> > <file_with_internal_stop_sequences.fas>
```

remove peptide sequences with internal stops (confirm you are missing seqs)

```
fasgrep -vs "[^A-Z].+" <peptide_file.fas> > <peptiled_file_minus_internal_stops.fas>
```

remove annotated genes with internal stops from gff file (confirm correct unique ids are removed)

```
egrep -f <(egrep ">" <file_with_internal_stop_sequences.fas> | cut -c 2- |cut -f1,2 -d'
```

make sure there are no sequences in your gff file. It should only be coordinates for the following scripts.

Step 3 is to fix the scaffolds files

remove trailing ends (no one cares where these are, just truncate them)

```
fassub -s 'N{1,}$' '' <scaffolds_file_original.fas> > <scaffolds_no_trailing.fas>
```

find sequences with leading n's and their numbers, put into file

```
fasgrep -s "^N" <scaffolds_no_trailing.fas> | perl -ne 'BEGIN{$start=0;}if(/^\>/){$fini
```

change coordinates for leading N sequences in gff file

for this script *changecoords.pl*, you need the file `leading_Ns.txt` but it takes the gff file and fixes all the coordinates

```
./change_coords.pl <new_gff_minus_internal_stop_peptides.gff> > <new_gff_with_no_errors
```

and then remove the leading Ns from the sequences

```
fassub -s '^N{1,}' '' <scaffolds_no_trailing.fas> <scaffolds_no_trailing_leading.fas>
```

convert files to NCBI-ready with my messy table scripts

the first script takes the gff as is, and puts the information into table formt

```
~/bin/katie_gff2tbl.pl <new_gff_with_no_errors.gff> > <genome_annot_from_gff.tbl-sub>
```

the second script takes a file called `gene_name_conversions.txt` and substitutes the correct annotations into the tbl file from the previous command. check *c-strain-annotation.sub.txt* for an example of the format of this file, which is tab-delimited and expects a header line. the subsequent command just changes gff identifiers to NCBI identifiers

```
/bin/katie_gff2tble2.pl <genome_annot_from_gff.tbl-sub> > <genome_annot_from_1gff.tbl>
perl -pe 's/gene_id/protein_id/' <genome_annot_from_1gff.tbl> > <genome_annot_from_2gfi
```

change the gene names to ones that are permissible by NCBI.

file to annotate the specific identifier assigned by NCBI. `newIDs.new.ncbi.txt` look at *newIDs.new.txt* for an example on how to do it. remember, the prefix that you entered in the WGS submission will go at the beginning of your new gene IDs. In the example file, I have the number followed by a dash. USE AN UNDERSCORE, NOT A DASH.

```
~/bin/change_IDs.pl <genome_annot_from_2gff.tbl> > <final.tbl>
```

miscellaneous help

these were to clean up things listed as problematic in the discrepancy report
this example shows deleting weird names that muck up the translator like "AF443189_2" and "homolog" and "af376000_1" in the .tbl annotations

```
perl -pe 's/\"//g;s/AF443189_2//g;s/, putative//g;s/, //g;s/homolog$/-like protein/g;s/
```

Step 4 is to submit

run most recent version of tbl2asn.

Need to go download mac.tbl2asn or whatever it is called Enec.sbt came from
the NCBI website creating an sbt file for me at http://www.ncbi.nlm.nih.gov/WebSub/template.cgi

change things that are different, like paired-ends, organism name, isolate name, check your options, etc.

```
./mac.tbl2asn -i <> -t Enec.sbt -M n -a r10k -j"[organism=Erysiphe necator][isolate=c]"
```

check errorsummary.val for fatal errors

submit .sqn and .fas file

Step 5 is to fix the errors they find

adapter contamination

You will likely have some adapter contamination they find with your submission. To fix this, Copy the lines from the email NCBI sends out into a file named *adapters_to_remove.txt* (check this file for an example of the input). The next script is still in very raw form. You want to edit the lines that specify your input files and your output files within the script. The standard output of this script is the scaffolds in fasta format. A new gff file will be created within the script. Bioperl is required.

```
./remove_NCBI_adaptors.pl > <new_trimmed_scaffolds.fas>
```