

From: **Terrance Shea** tshea@broadinstitute.org 

Subject: FTP information for assembly data

Date: November 18, 2016 at 4:15 PM

To: matthewklau@fas.harvard.edu

Cc: Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org, Caroline Cusick ccusick@broadinstitute.org



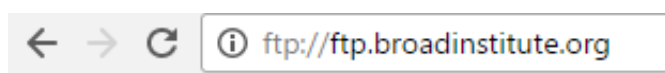
Hello again-

It was a pleasure to meet you in the meeting today. Below described the contents of the assembly handoff and how to access it.


The assembly data is available via <ftp://aadata@ftp.broadinstitute.org> with the following username and password:

- username **aadata**
- password **CILanalysis154**

You should see the following folder:






Index of /

Name	Size	Date Modified
 lau-ant/		11/18/16, 8:31:00 PM

and then within this "lau-ant" folder 7 folders, one for each sample, for example:

Index of /lau-ant/

Name	Size	Date Modified
 [parent directory]		
 SM-AJDMW/		11/18/16, 8:32:00 PM
 SM-AZXXM/		11/18/16, 8:32:00 PM

I have attached the assembly metrics sheet which was reviewed in the meeting. The "pre_post_pilon" tab shows the assembly metrics before and after running Pilon (assembly improvement tool that is run after ALLPATHS-LG assembly). The "filtered_assembly_stats" shows the metrics after likely contaminant contigs/scaffolds were removed.

There are two versions of the assembly in each sample directory, pre- and post- contamination filtering.

Each sample directory contains a contigs file (contigs.fasta), scaffolds file (scaffolds.fasta), and agp file (assembly.agp). These are pre-contamination removal. The agp file describes how contigs are linked within scaffolds.

Each sample directory contains a "gaemr" analysis directory and within this is a "chart", "table", and "html" subdirectory. The "index.html" file (within gaemr/html folder) may be opened in a browser to view. The GAEMR analysis corresponds to the previously described "contigs.fasta" and "scaffolds.fasta" files. One additional gaemr file is found in the "chart" directory and that is a blast bubble plot done at the superkingdom level (default level is genus). This plot gives an approximation for the number of bacterial and viral contigs found.

Likely contaminants (generally either Wolbochia and/or Mycoplasma) were found during the analysis of the GAEMR output. Contaminants scaffolds were removed and these are listed in the "remove.txt" file. The assembly with contaminants removed is found in the "filtered.contigs.fasta", "filtered.scaffolds.fasta", and "filtered.agp" set of files. The contigs/scaffolds removed (not in the filtered.contigs.fasta file) may be found in the "filtered.removed.fasta" file. GAEMR was not re-run on the contamination-filtered assembly.

Please let us know if you encounter any problems in accessing the data. Also, if all goes smoothly, please let us know once you have all of this as we can then remove it from the FTP area.

I gather that Caroline or Jim will be in touch once back from their break regarding raw data handoff.

Terry



SSF-1728_basic
_assem...18.xlsx

