From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: FTP information for assembly data
Date: December 13, 2016 at 10:04 AM
To: Caroline Cusick ccusick@broadinstitute.org
Cc: James Bochicchio jboch@broadinstitute.org

Awesome, thanks Caroline.

Cheers,

Matt

On Dec 13, 2016, at 9:24 AM, Caroline Cusick <ccusick@broadinstitute.org> wrote:

> Hi Matt,
>
> These should be all set and in the folder for the FTP (the file names will be XXX.tar.gz, and will match the key). If you could just let me know when you're all set, I'll make sure to remove the compressed duplicates.
>
> thanks!
> Caroline
>
> On Thu, Dec 8, 2016 at 2:14 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:
>
>> HI Matt,
>>
>> I'm queueing the tarballs up now, I'll send you another email when they're complete.
>>
>> thanks!
>> Caroline
>>
>> On Wed, Dec 7, 2016 at 3:27 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>>
>>> Hi Caroline, when you get a chance, would you compress the raw data files?
>>>
>>> Thanks!

> > > Matt
> > >
> > > > On Dec 6, 2016, at 5:20 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
> > > >
> > > > Hi Caroline, thanks! I’ll move those over onto my server tomorrow morning.
> > > >
> > > > Best,
> > > >
> > > > Matt
> > > >
> > > > > On Dec 6, 2016, at 4:42 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:
> > > > >
> > > > > Hi Matt,
> > > > >
> > > > > I just wanted to follow up with the FTP log in for the Raw Data. I've also re-attached the key that indicates the library type for each sequencing run. Let me know if you run into any issues
> > > > >
> > > > > access by: ftp://ftp.broadinstitute.org/
> > > > > username: SSF-1728
> > > > > password: f3aexw29
> > > > >
> > > > > thanks!
> > > > > Caroline
> > > > >
> > > > > On Mon, Nov 21, 2016 at 11:25 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
> > > > >
> > > > > > Great, thanks Terry!
> > > > > >
> > > > > > I was able to access the files. I’ll grab those and get them off your FTP server ASAP.
> > > > > >
> > > > > > Thanks for your hard work on this.

Happy T-Day,

Matt

> On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
>
> 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:
>
> <image.png>
>
> and then within this "lau-ant" folder 7 folders, one for each sample, for example:
>
> <image.png>
>
> 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

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_assembly_stats_20161118.xlsx>
> > > > >
> > > > > --
> > > > > Caroline Cusick
> > > > > Product Coordinator
> > > > > Broad Institute of MIT and Harvard
> > > > > 75 Ames St | Cambridge MA 02142 USA
> > > > > <SSF-1728_KeyforCollaborator.xlsx>
> >
> > --
> > Caroline Cusick
> > Product Coordinator
> > Broad Institute of MIT and Harvard
> > 75 Ames St | Cambridge MA 02142 USA
>
> --
> Caroline Cusick
> Product Coordinator
> Broad Institute of MIT and Harvard
> 75 Ames St | Cambridge MA 02142 USA

From: **Caroline Cusick** ccusick@broadinstitute.org
Subject: Re: FTP information for assembly data
Date: December 13, 2016 at 9:24 AM
To: Lau, Matthew K. matthewklau@fas.harvard.edu
Cc: James Bochicchio jboch@broadinstitute.org

CC

Hi Matt,

These should be all set and in the folder for the FTP (the file names will be XXX.tar.gz, and will match the key). If you could just let me know when you're all set, I'll make sure to remove the compressed duplicates.

thanks!
Caroline

On Thu, Dec 8, 2016 at 2:14 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:

> HI Matt,
>
> I'm queueing the tarballs up now, I'll send you another email when they're complete.
>
> thanks!
> Caroline
>
> On Wed, Dec 7, 2016 at 3:27 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>
>> Hi Caroline, when you get a chance, would you compress the raw data files?
>>
>> Thanks!
>>
>> Matt
>>
>>> On Dec 6, 2016, at 5:20 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>>>
>>> Hi Caroline, thanks! I’ll move those over onto my server tomorrow morning.

> > > Best,
> > >
> > > Matt
> > >
> > > > On Dec 6, 2016, at 4:42 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:
> > > >
> > > > Hi Matt,
> > > >
> > > > I just wanted to follow up with the FTP log in for the Raw Data. I've also re-attached the key that indicates the library type for each sequencing run. Let me know if you run into any issues
> > > >
> > > > access by: ftp://ftp.broadinstitute.org/
> > > > username: SSF-1728
> > > > password: f3aexw29
> > > >
> > > > thanks!
> > > > Caroline
> > > >
> > > > On Mon, Nov 21, 2016 at 11:25 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
> > > >
> > > > > Great, thanks Terry!
> > > > >
> > > > > I was able to access the files. I’ll grab those and get them off your FTP server ASAP.
> > > > >
> > > > > Thanks for your hard work on this.
> > > > >
> > > > > Happy T-Day,
> > > > >
> > > > > Matt
> > > > >
> > > > > > On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
> > > > > >
> > > > > > Hello again-
> > > > > > It was a pleasure to meet you in the meeting today. Below

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:

<image.png>

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

<image.png>

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 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_assembly_stats_20161118.xlsx>

>>>> --
>>>> Caroline Cusick
>>>> Product Coordinator
>>>> Broad Institute of MIT and Harvard
>>>> 75 Ames St | Cambridge MA 02142 USA
>>>> <SSF-1728_KeyforCollaborator.xlsx>

> --
> Caroline Cusick
> Product Coordinator
> Broad Institute of MIT and Harvard
> 75 Ames St | Cambridge MA 02142 USA

--
Caroline Cusick
Product Coordinator
Broad Institute of MIT and Harvard
75 Ames St | Cambridge MA 02142 USA

From: **Caroline Cusick** ccusick@broadinstitute.org
Subject: Re: FTP information for assembly data
Date: December 8, 2016 at 2:15 PM
To: Lau, Matthew K. matthewklau@fas.harvard.edu
Cc: Terrance Shea tshea@broadinstitute.org, Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org

HI Matt,

I'm queueing the tarballs up now, I'll send you another email when they're complete.

thanks!
Caroline

On Wed, Dec 7, 2016 at 3:27 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:

> Hi Caroline, when you get a chance, would you compress the raw data files?
>
> Thanks!
>
> Matt
>
> > On Dec 6, 2016, at 5:20 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
> >
> > Hi Caroline, thanks! I'll move those over onto my server tomorrow morning.
> >
> > Best,
> >
> > Matt
> >
> > > On Dec 6, 2016, at 4:42 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:
> > >
> > > Hi Matt,
> > >
> > > I just wanted to follow up with the FTP log in for the Raw Data. I've also re-attached the key that indicates the library type for

> > > each sequencing run. Let me know if you run into any issues
> > >
> > > access by: ftp://ftp.broadinstitute.org/
> > > username: SSF-1728
> > > password: f3aexw29
> > >
> > > thanks!
> > > Caroline
> > >
> > > On Mon, Nov 21, 2016 at 11:25 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
> > >
> > > > Great, thanks Terry!
> > > >
> > > > I was able to access the files. I'll grab those and get them off your FTP server ASAP.
> > > >
> > > > Thanks for your hard work on this.
> > > >
> > > > Happy T-Day,
> > > >
> > > > Matt
> > > >
> > > > > On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
> > > > >
> > > > > 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:

<image.png>

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

<image.png>

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

> > > > > 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_assembly_stats_20161118.xlsx>
> > >
> > > --
> > > Caroline Cusick
> > > Product Coordinator
> > > Broad Institute of MIT and Harvard
> > > 75 Ames St | Cambridge MA 02142 USA
> > > <SSF-1728_KeyforCollaborator.xlsx>

--
Caroline Cusick
Product Coordinator
Broad Institute of MIT and Harvard
75 Ames St | Cambridge MA 02142 USA

From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: FTP information for assembly data
Date: December 7, 2016 at 3:27 PM
To: Caroline Cusick ccusick@broadinstitute.org
Cc: Terrance Shea tshea@broadinstitute.org, Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org

Hi Caroline, when you get a chance, would you compress the raw data files?

Thanks!

Matt

> On Dec 6, 2016, at 5:20 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>
> Hi Caroline, thanks! I'll move those over onto my server tomorrow morning.
>
> Best,
>
> Matt
>
> > On Dec 6, 2016, at 4:42 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:
> >
> > Hi Matt,
> >
> > I just wanted to follow up with the FTP log in for the Raw Data. I've also re-attached the key that indicates the library type for each sequencing run. Let me know if you run into any issues
> >
> > access by: ftp://ftp.broadinstitute.org/
> > username: SSF-1728
> > password: f3aexw29
> >
> > thanks!
> > Caroline
> >
> > On Mon, Nov 21, 2016 at 11:25 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
> >
> > > Great, thanks Terry!

> > > Great, thanks Terry!
> > >
> > > I was able to access the files. I'll grab those and get them off your FTP server ASAP.
> > >
> > > Thanks for your hard work on this.
> > >
> > > Happy T-Day,
> > >
> > > Matt
> > >
> > > > On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
> > > >
> > > > 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:
> > > >
> > > > <image.png>
> > > >
> > > > and then within this "lau-ant" folder 7 folders, one for each sample, for example:
> > > >
> > > > <image.png>
> > > >
> > > > 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

> > > > 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_assembly_stats_20161118.xlsx>
>
> > --
> > Caroline Cusick
> > Product Coordinator
> > Broad Institute of MIT and Harvard
> > 75 Ames St | Cambridge MA 02142 USA
> > <SSF-1728_KeyforCollaborator.xlsx>

From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: FTP information for assembly data
Date: December 6, 2016 at 5:20 PM
To: Caroline Cusick ccusick@broadinstitute.org
Cc: Terrance Shea tshea@broadinstitute.org, Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org

Hi Caroline, thanks! I’ll move those over onto my server tomorrow morning.

Best,

Matt

> On Dec 6, 2016, at 4:42 PM, Caroline Cusick <ccusick@broadinstitute.org> wrote:
>
> Hi Matt,
>
> I just wanted to follow up with the FTP log in for the Raw Data. I've also re-attached the key that indicates the library type for each sequencing run. Let me know if you run into any issues
>
> access by: ftp://ftp.broadinstitute.org/
> username: SSF-1728
> password: f3aexw29
>
> thanks!
> Caroline
>
> On Mon, Nov 21, 2016 at 11:25 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>
>> Great, thanks Terry!
>>
>> I was able to access the files. I’ll grab those and get them off your FTP server ASAP.
>>
>> Thanks for your hard work on this.
>>
>> Happy T-Day,
>>
>> Matt

> > > On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
> > >
> > > 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:
> > >
> > > <image.png>
> > >
> > > and then within this "lau-ant" folder 7 folders, one for each sample, for example:
> > >
> > > <image.png>
> > >
> > > 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_assembly_stats_20161118.xlsx>

> --
> Caroline Cusick
> Product Coordinator
> Broad Institute of MIT and Harvard
> 75 Ames St | Cambridge MA 02142 USA
> <SSF-1728_KeyforCollaborator.xlsx>

From: **Caroline Cusick** ccusick@broadinstitute.org
Subject: Re: FTP information for assembly data
Date: December 6, 2016 at 4:42 PM
To: Lau, Matthew K. matthewklau@fas.harvard.edu
Cc: Terrance Shea tshea@broadinstitute.org, Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org

Hi Matt,

I just wanted to follow up with the FTP log in for the Raw Data. I've also re-attached the key that indicates the library type for each sequencing run. Let me know if you run into any issues

access by: ftp://ftp.broadinstitute.org/
username: SSF-1728
password: f3aexw29

thanks!
Caroline

On Mon, Nov 21, 2016 at 11:25 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:

> Great, thanks Terry!
>
> I was able to access the files. I'll grab those and get them off your FTP server ASAP.
>
> Thanks for your hard work on this.
>
> Happy T-Day,
>
> Matt
>
>> On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
>>
>> 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:

the following username and password:

- username aadata
- password CILanalysis154

You should see the following folder:

<image.png>

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

<image.png>

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_assembly_stats_20161118.xlsx>

--
Caroline Cusick
Product Coordinator
Broad Institute of MIT and Harvard
75 Ames St | Cambridge MA 02142 USA

SSF-1728_Keyfo
rCollab...tor.xlsx

From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: FTP information for assembly data
Date: November 21, 2016 at 11:25 AM
To: Terrance Shea tshea@broadinstitute.org
Cc: Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org, Caroline Cusick ccusick@broadinstitute.org

Great, thanks Terry!

I was able to access the files. I'll grab those and get them off your FTP server ASAP.

Thanks for your hard work on this.

Happy T-Day,

Matt

> On Nov 18, 2016, at 4:14 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
>
> 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:
>
> <image.png>
>
> and then within this "lau-ant" folder 7 folders, one for each sample, for example:
>
> <image.png>

> 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_assembly_stats_20161118.xlsx>

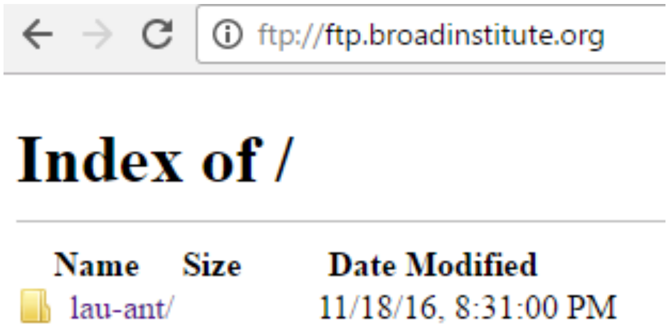
From: **Terrance Shea** tshea@broadinstitute.org
Subject: FTP information for assembly data
Date: November 18, 2016 at 4:15 PM
To: matthewklau@fas.harvard.edu
Cc: Sarah Towey stowey@broadinstitute.org, James Bochicchio jboch@broadinstitute.org, Caroline Cusick ccusick@broadinstitute.org

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

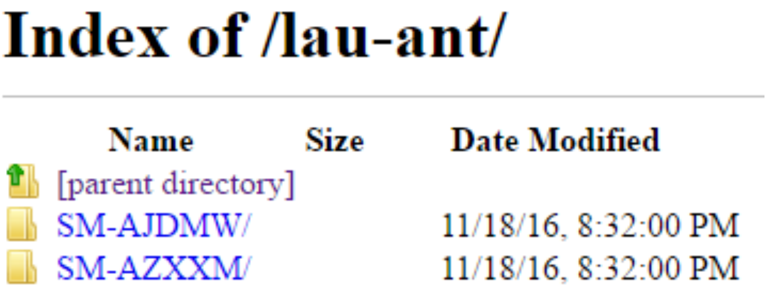
You should see the following folder:



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 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