Mapping Metagenomic Reads to Reference Sequences (Cyverse) Version 3

Benjamin Bolduc

Abstract

Mapping metagenomic reads from Ocean Sampling Day (OSD) 2014 against NCBI's ViralRefSeq alongside viral sequences identified from the Tara Oceans survey using VirSorter.

Citation: Benjamin Bolduc Mapping Metagenomic Reads to Reference Sequences (Cyverse). **protocols.io** dx.doi.org/10.17504/protocols.io.gv2bw8e
Published: 31 Dec 2016

Guidelines

One of the most commonly used procedures for analyzing viral metagenomic data is to map their reads (or reads from another dataset) against a set of references, often those from the read assembly. For example, if one wanted to know how well-represented viruses in NCBI's Viral Reference Sequences (ViralRefSeq) were in ocean viromes, they could map reads from lots of ocean viral metagenomes against ViralRefSeq. This is generally done using Bowtie2 or BWA, by selecting a reference set of sequences, and then providing paired or unpaired reads to Bowtie2/BWA. *Then* the results must be processed/filtered to generate coverage tables. Dealing with setting up multiple reads files (10 paired metagenomes = 10 alignment runs) and the processing those read files can be challenging (not to mention requring computational resources).

In this protocol, we'□□ll be using reads fromÂ Ocean Sampling Day (OSD) 2014 and map them against ViralRefSeq alongside viral sequences identified from the Tara Oceans survey using VirSorter.

Before start

To run this protocol, users must first register for Cyverse account. All data (both inputs and outputs) are available within Cyverse's data store at /iplant/home/shared/iVirus/ExampleData/

All source code is available at the Sullivan lab bitbucket repository, MAVERICLab - either *-bowtiebatch or *-read2refmapper. * denotes either docker or stampede-based repos.

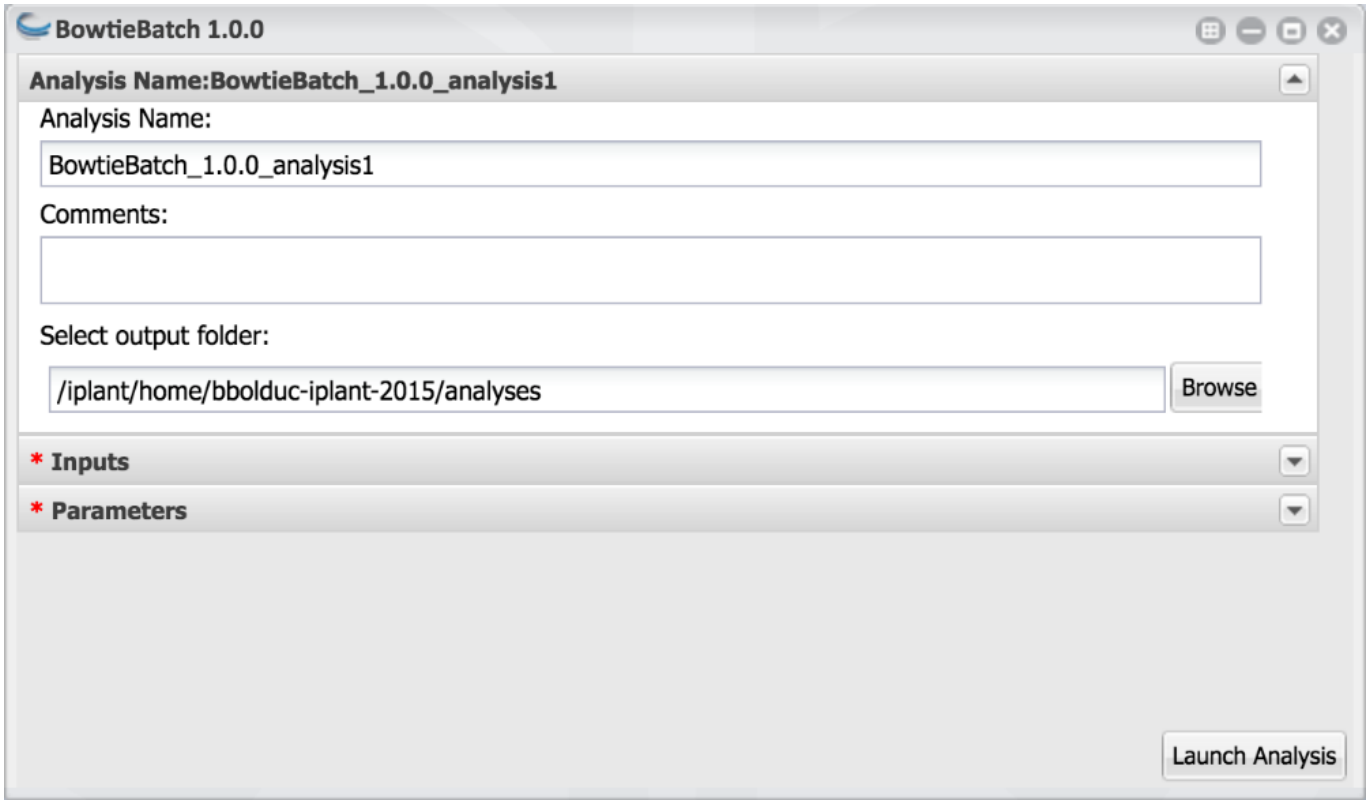
Protocol

BowtieBatch

Step 1.

Open BowtieBatch

Open BowtieBatch-1.0.0 from 'Apps'



BowtieBatch

Step 2.

Select Inputs

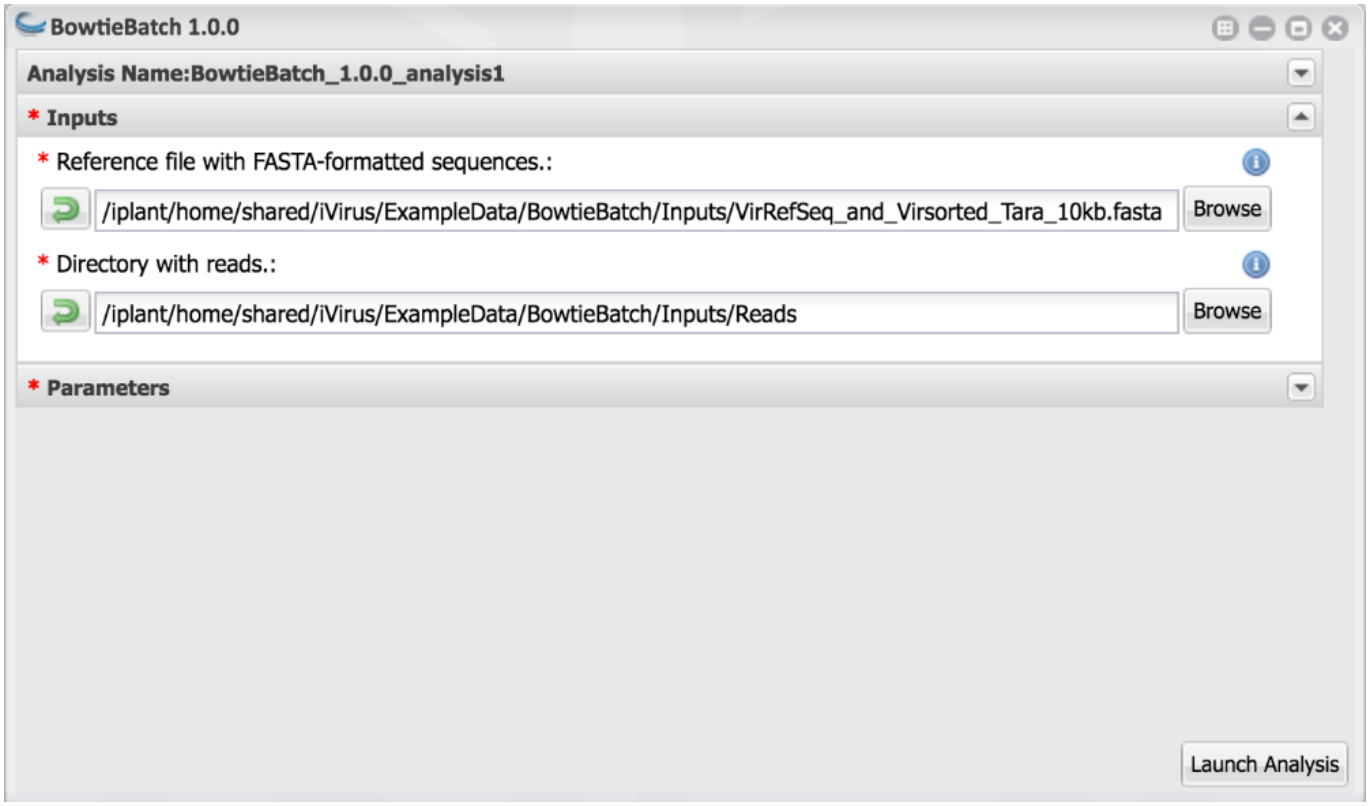
Select the 'Inputs' tab.

For **Reference file with FASTA-formatted sequences:** this is a FASTA-formatted file with reference sequences to map reads against

- Navigate to *Community Data --> iVirus --> ExampleData --> BowtieBatch --> Inputs*. Select *VirRefSeq_and_Virsorted_Tara_10kb.fasta* Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/BowtieBatch/Inputs into the navigation bar and select the fasta file.

For **Directory with reads**: this is a directory containing the reads files that are to be mapped

- Navigate to *Community Data --> iVirus --> ExampleData --> BowtieBatch --> Inputs*. Select the *Reads* folder. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/BowtieBatch/Inputs into the navigation bar and select the reads folder.



NOTES

Benjamin Bolduc 31 Dec 2016

TIP: Reads directory can contain any number of reads files. They can be paired, unpaired, gzipped or uncompressed. They *do* have to have some semblence of ordering, either computer or human-natural sorting.

BowtieBatch

Step 3.

Select Parameters

The default options will be sufficient for this example, however additional options are explain below.

File extension: Extension of the reads. GZIP files are automatically recognized.

Interleaved files: Are reads interleaved? i.e. Are forward and reverse reads in the same file? These read files contain both pairs, usually Forward1, Reverse1, Forward2, etc... By using this option users can automatically handle interleaved files w/out worrying about having another app handling this conversion.
Read type: What type are the reads? *Paired*, *unpaired*, a *mix* of both?

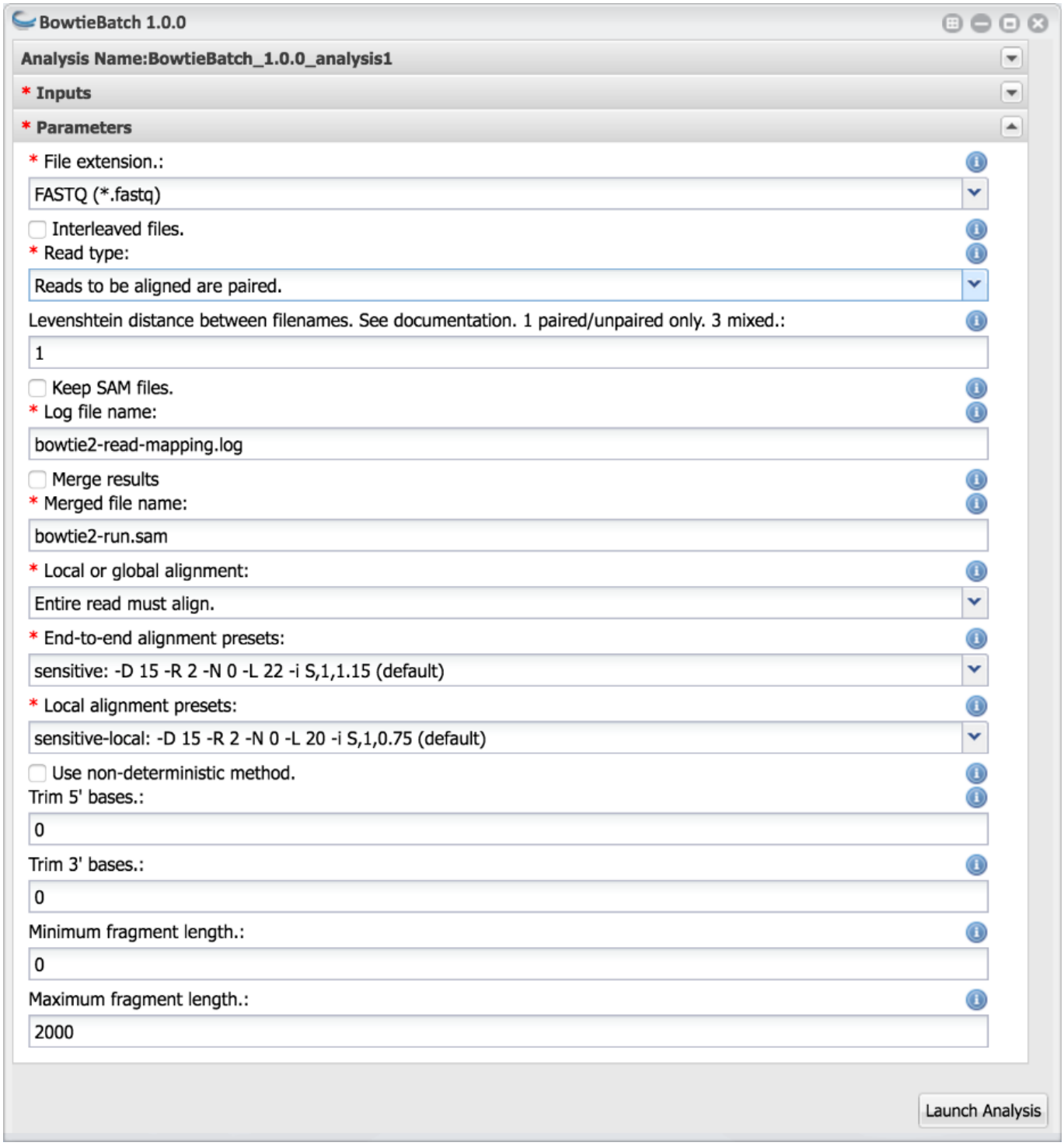
Levenshtein distance: "Distance" between the read names. i.e. "CAT" and "BAT" have a distance of 1 because the C/B difference. Generally, read pairs/unpairs differ by 1 and mixed differ by 3. If reads do not group correctly, try changing this value.

Keep SAM files: Bowtie2 generates SAM files as a result of mapping reads. However, these take up a lot of space and aren't kept if they're not needed. Users can select if they'd like to keep these files around in case they have other analyses that use SAM files. For this example, they are unnecessary.

Log file name: Information about the run's processing will be stored with this file name. This file *can become very large* as it contains trimming information.
Merge results: Bowtie2 can create a single SAM file that combines all the individual paired/unpaired alignments. Because the next tool in this pipeline (Read2RefMapper) requires individually separated files, do not select this.

Merged file name: Name to use if results are combined.

The remaining options are bowtie specific, please consult the Bowtie2 documentation.



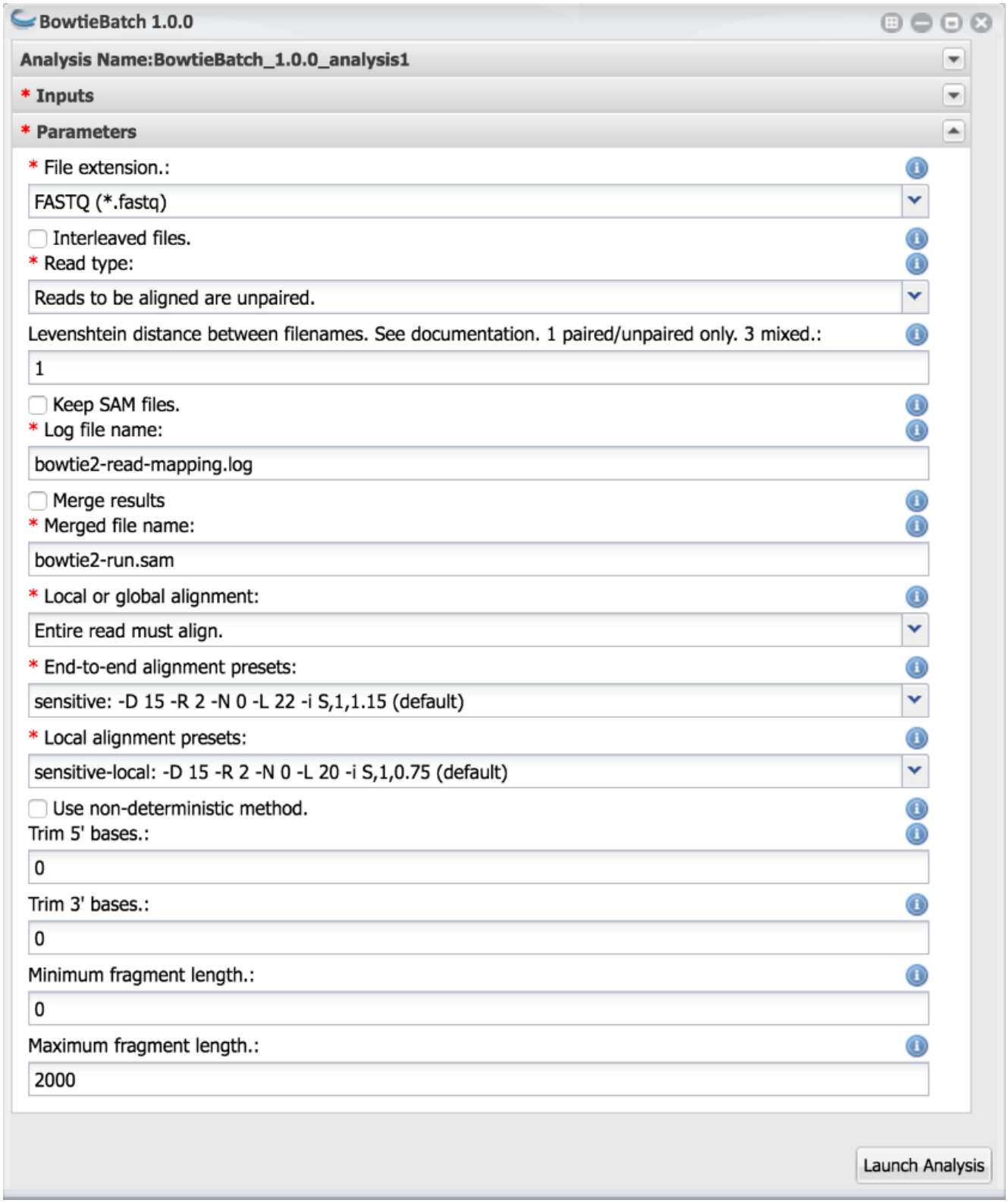
NOTES

Benjamin Bolduc 31 Dec 2016

While the default options will suffice for this example, play around with a few of the options (excluding file extension).

Benjamin Bolduc 31 Dec 2016

Reads can be paired, unpaired or a mix of both. In this example there are 2 sets of paired reads. Alternatively, one could set all the reads to be unpaired.



BowtieBatch

Step 4.

Launch Analysis

Run the job!

Depending on the number of files to align and/or the number of sequences in the reference database, the job could take many hours.

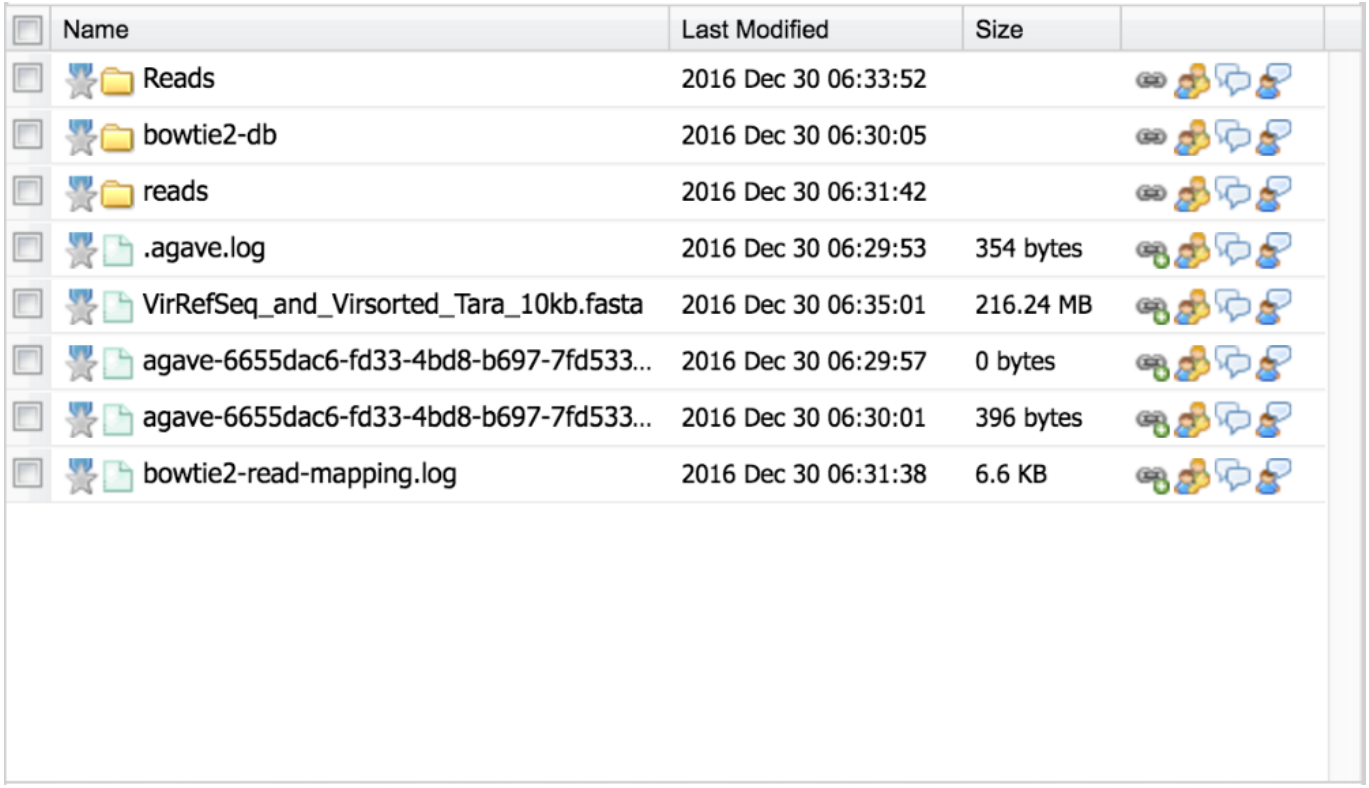
BowtieBatch

Step 5.

Results

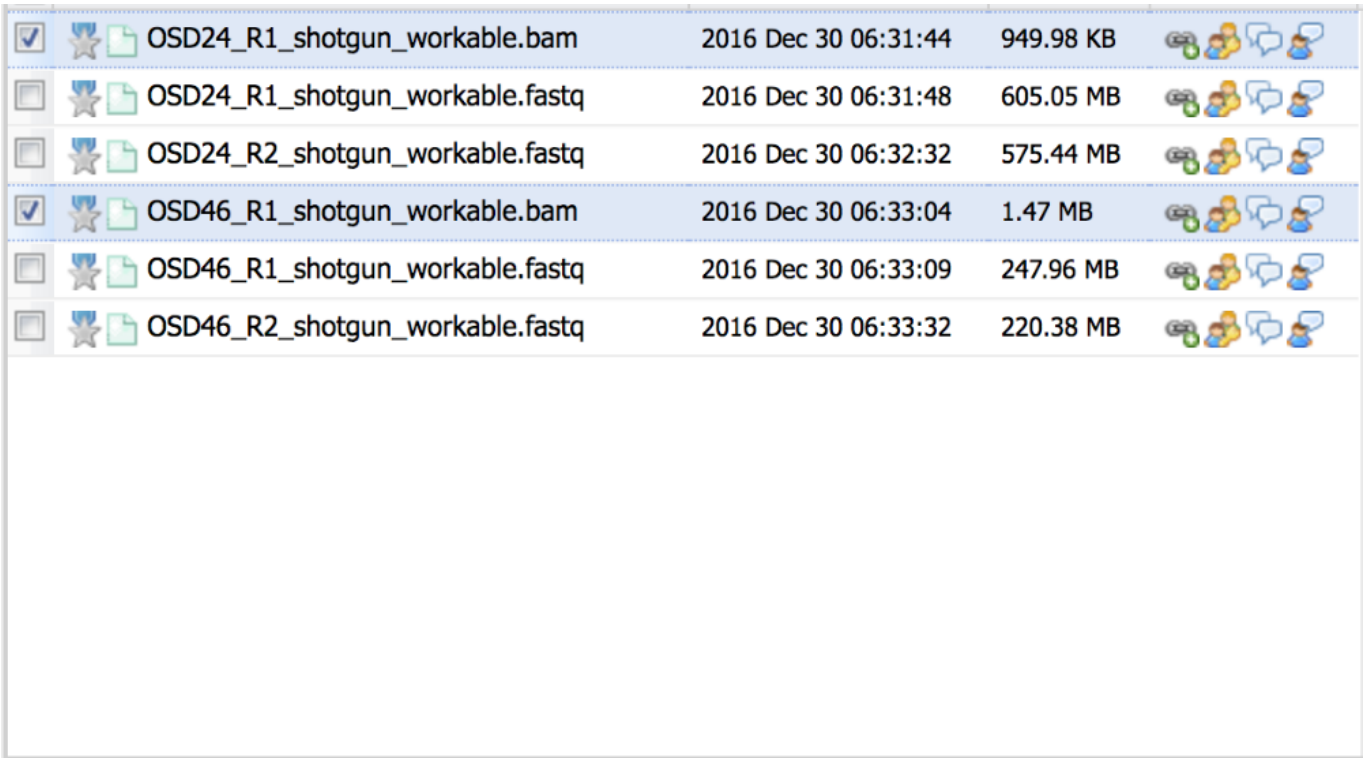
Expected results can be found from the 'output' directory of BowtieBatch. They'll consist of 3 directories:

- Reads: containing the original files
- bowtie2-db: containing the bowtie2-formatted sequences
- reads: any uncompressed version of the reads (those files actually used by bowtie2) and BAM files. If **Keep Sam files** was selecting, they'll be SAM files there as well.



Name	Last Modified	Size
Reads	2016 Dec 30 06:33:52	
bowtie2-db	2016 Dec 30 06:30:05	
reads	2016 Dec 30 06:31:42	
.agave.log	2016 Dec 30 06:29:53	354 bytes
VirRefSeq_and_Virsorted_Tara_10kb.fasta	2016 Dec 30 06:35:01	216.24 MB
agave-6655dac6-fd33-4bd8-b697-7fd533...	2016 Dec 30 06:29:57	0 bytes
agave-6655dac6-fd33-4bd8-b697-7fd533...	2016 Dec 30 06:30:01	396 bytes
bowtie2-read-mapping.log	2016 Dec 30 06:31:38	6.6 KB

Content of reads directory:



Additionally, the read mapping log file contains very useful information, such as the overall read mapping statistics:

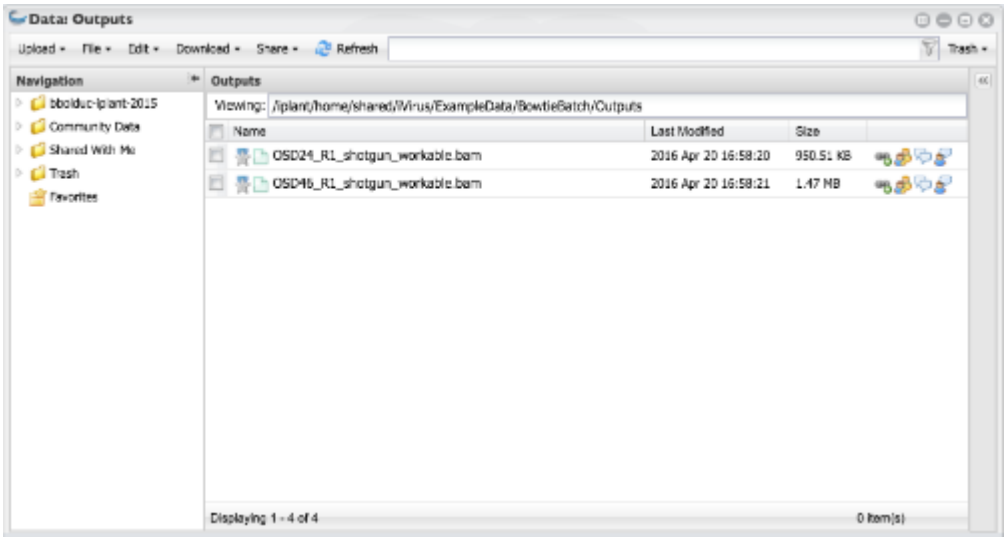
```
Created reads directory: /work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-66
Created bowtie2-db directory: /work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-0
Directory content for reads:
[]
Directory contents for bowtie2-db directory:
[]
Workable Files
['Reads/OSD46_R2_shotgun_workable.fastq.gz',
 'Reads/OSD46_R1_shotgun_workable.fastq.gz',
 'Reads/OSD24_R2_shotgun_workable.fastq.gz',
 'Reads/OSD24_R1_shotgun_workable.fastq.gz']
Workable Reads
['Reads/OSD46_R2_shotgun_workable.fastq.gz',
 'Reads/OSD46_R1_shotgun_workable.fastq.gz',
 'Reads/OSD24_R2_shotgun_workable.fastq.gz',
 'Reads/OSD24_R1_shotgun_workable.fastq.gz']
Executing cp VirRefSeq_and_Virsorted_Tara_10kb.fasta /work/03548/bb_t4k0/bbolduc-iplant-2015/job-390312567509

Executing bowtie2-build -f /work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-
Building a SMALL index

Bowtie2 base db: /work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-6655dac6-f
Sorted files:
{0: {'paired': ['Reads/OSD24_R1_shotgun_workable.fastq.gz',
                'Reads/OSD24_R2_shotgun_workable.fastq.gz']},
 1: {'paired': ['Reads/OSD46_R1_shotgun_workable.fastq.gz',
                'Reads/OSD46_R2_shotgun_workable.fastq.gz']}}
Executing zcat Reads/OSD24_R1_shotgun_workable.fastq.gz > /work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125

Executing zcat Reads/OSD46_R1_shotgun_workable.fastq.gz > /work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125

Reads in their appropriate locations?
{0: {'paired': ['/work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-6655dac6-f
                '/work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-6655dac6-f
 1: {'paired': ['/work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-6655dac6-f
                '/work/03548/bb_t4k0/bbolduc-iplant-2015/job-3903125675091029530-242ac114-0001-007-6655dac6-f
Executing bowtie2 -q --phred33 --end-to-end --sensitive -p 16 -I 0 -X 2000 --no-unal -x /work/03548/bb_t4k0/b
perl: warning: Setting locale failed.
perl: warning: Please check that your locale settings:
    LANGUAGE = (unset),
    LC_ALL = "en_US.UTF-8",
    LANG = "en_US.UTF-8"
    are supported and installed on your system.
perl: warning: Falling back to the standard locale ("C").
1214226 reads; of these:
  1214226 (100.00%) were paired; of these:
    1211266 (99.76%) aligned concordantly 0 times
    2174 (0.18%) aligned concordantly exactly 1 time
    786 (0.06%) aligned concordantly >1 times
    ----
    1211266 pairs aligned concordantly 0 times; of these:
      42 (0.00%) aligned discordantly 1 time
    ----
    1211224 pairs aligned 0 times concordantly or discordantly; of these:
      2422448 mates make up the pairs; of these:
        2421921 (99.98%) aligned 0 times
        365 (0.02%) aligned exactly 1 time
        162 (0.01%) aligned >1 times
0.27% overall alignment rate

Executing bowtie2 -q --phred33 --end-to-end --sensitive -p 16 -I 0 -X 2000 --no-unal -x /work/03548/bb_t4k0/b
perl: warning: Setting locale failed.
perl: warning: Please check that your locale settings:
    LANGUAGE = (unset),
    LC_ALL = "en_US.UTF-8",
    LANG = "en_US.UTF-8"
    are supported and installed on your system.
perl: warning: Falling back to the standard locale ("C").
406075 reads; of these:
  406075 (100.00%) were paired; of these:
    403125 (99.27%) aligned concordantly 0 times
    2513 (0.62%) aligned concordantly exactly 1 time
    437 (0.11%) aligned concordantly >1 times
    ----
    403125 pairs aligned concordantly 0 times; of these:
      103 (0.03%) aligned discordantly 1 time
    ----
    403022 pairs aligned 0 times concordantly or discordantly; of these:
      806044 mates make up the pairs; of these:
        805075 (99.88%) aligned 0 times
        785 (0.10%) aligned exactly 1 time
        184 (0.02%) aligned >1 times
0.87% overall alignment rate

Program Complete
```

Notice the grouping of the paired files, under "Sorted files." It shows 2 groups, 0 and 1, that have OSD24 and OSD46 as read groups.

NOTES

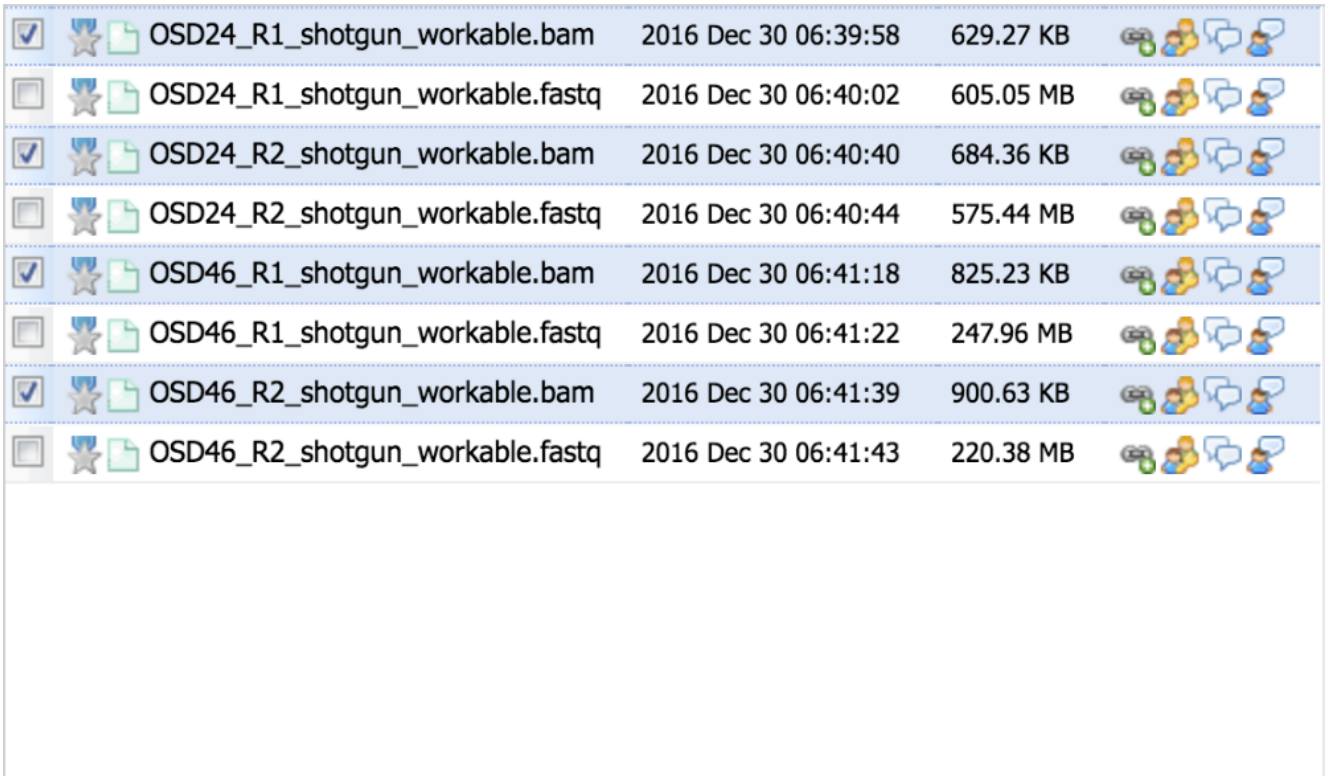
Benjamin Bolduc 31 Dec 2016

A sample output directory is located at: /iplant/home/shared/iVirus/ExampleData/BowtieBatch/Outputs



Benjamin Bolduc 31 Dec 2016

If the reads were selected to be unpaired, a BAM file would be associated with each input file.



And each file would be grouped individually:

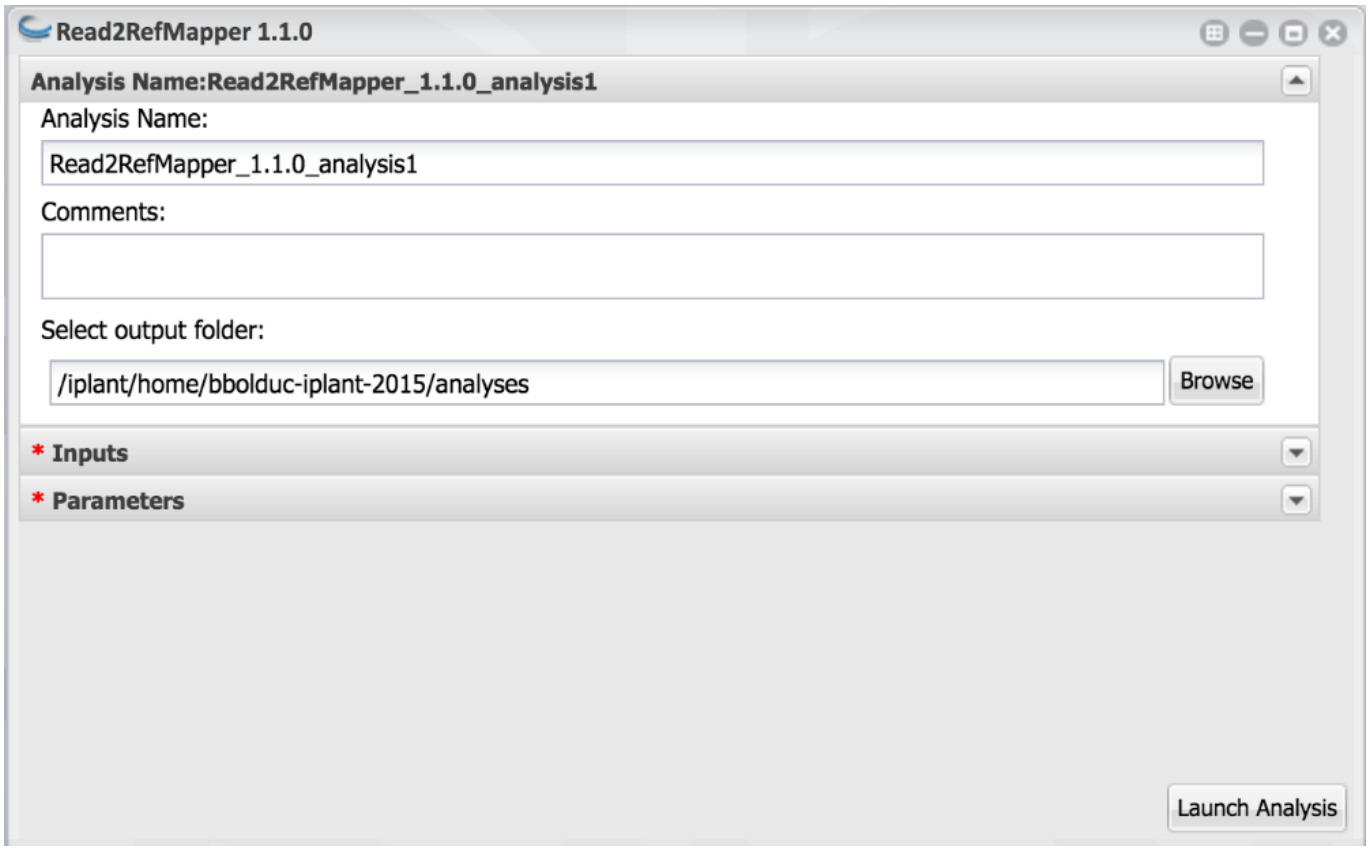
```
Created reads directory: /work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-007-
Created bowtie2-db directory: /work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-000
Directory content for reads:
[]
Directory contents for bowtie2-db directory:
[]
Workable Files
['Reads/OSD46_R2_shotgun_workable.fastq.gz',
 'Reads/OSD46_R1_shotgun_workable.fastq.gz',
 'Reads/OSD24_R2_shotgun_workable.fastq.gz',
 'Reads/OSD24_R1_shotgun_workable.fastq.gz']
Workable Reads
['Reads/OSD46_R2_shotgun_workable.fastq.gz',
 'Reads/OSD46_R1_shotgun_workable.fastq.gz',
 'Reads/OSD24_R2_shotgun_workable.fastq.gz',
 'Reads/OSD24_R1_shotgun_workable.fastq.gz']
Executing cp VirRefSeq_and_Virsorted_Tara_10kb.fasta /work/03548/bb_t4k0/bbolduc-iplant-2015/job-150022032

Executing bowtie2-build -f /work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-00
Building a SMALL index

Bowtie2 base db: /work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-007-3257762
Sorted files:
{0: {'unpaired': ['Reads/OSD24_R1_shotgun_workable.fastq.gz']},
 1: {'unpaired': ['Reads/OSD24_R2_shotgun_workable.fastq.gz']},
 2: {'unpaired': ['Reads/OSD46_R1_shotgun_workable.fastq.gz']},
 3: {'unpaired': ['Reads/OSD46_R2_shotgun_workable.fastq.gz']}}
Executing zcat < Reads/OSD24_R1_shotgun_workable.fastq.gz > /work/03548/bb_t4k0/bbolduc-iplant-2015/job-150

Executing zcat < Reads/OSD24_R2_shotgun_workable.fastq.gz > /work/03548/bb_t4k0/bbolduc-iplant-2015/job-150

Executing zcat < Reads/OSD46_R1_shotgun_workable.fastq.gz > /work/03548/bb_t4k0/bbolduc-iplant-2015/job-150

Executing zcat < Reads/OSD46_R2_shotgun_workable.fastq.gz > /work/03548/bb_t4k0/bbolduc-iplant-2015/job-150

Reads in their appropriate locations?
{0: {'unpaired': ['/work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-007-325776
 1: {'unpaired': ['/work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-007-325776
 2: {'unpaired': ['/work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-007-325776
 3: {'unpaired': ['/work/03548/bb_t4k0/bbolduc-iplant-2015/job-1500220321997909530-242ac114-0001-007-325776
Executing bowtie2 -q --phred33 --end-to-end --sensitive -p 16 -I 0 -X 2000 --no-unal -x /work/03548/bb_t4k0
perl: warning: Setting locale failed.
perl: warning: Please check that your locale settings:
    LANGUAGE = (unset),
    LC_ALL = "en_US.UTF-8",
    LANG = "en_US.UTF-8"
    are supported and installed on your system.
perl: warning: Falling back to the standard locale ("C").
1214226 reads; of these:
  1214226 (100.00%) were unpaired; of these:
    1210931 (99.73%) aligned 0 times
    2432 (0.20%) aligned exactly 1 time
    863 (0.07%) aligned >1 times
0.27% overall alignment rate

Executing bowtie2 -q --phred33 --end-to-end --sensitive -p 16 -I 0 -X 2000 --no-unal -x /work/03548/bb_t4k0
perl: warning: Setting locale failed.
perl: warning: Please check that your locale settings:
    LANGUAGE = (unset),
    LC_ALL = "en_US.UTF-8",
    LANG = "en_US.UTF-8"
    are supported and installed on your system.
perl: warning: Falling back to the standard locale ("C").
1214226 reads; of these:
  1214226 (100.00%) were unpaired; of these:
    1210806 (99.72%) aligned 0 times
    2511 (0.21%) aligned exactly 1 time
    909 (0.07%) aligned >1 times
0.28% overall alignment rate

Executing bowtie2 -q --phred33 --end-to-end --sensitive -p 16 -I 0 -X 2000 --no-unal -x /work/03548/bb_t4k0
perl: warning: Setting locale failed.
perl: warning: Please check that your locale settings:
    LANGUAGE = (unset),
    LC_ALL = "en_US.UTF-8",
    LANG = "en_US.UTF-8"
    are supported and installed on your system.
perl: warning: Falling back to the standard locale ("C").
406075 reads; of these:
  406075 (100.00%) were unpaired; of these:
    402472 (99.11%) aligned 0 times
    3007 (0.74%) aligned exactly 1 time
    596 (0.15%) aligned >1 times
0.89% overall alignment rate

Executing bowtie2 -q --phred33 --end-to-end --sensitive -p 16 -I 0 -X 2000 --no-unal -x /work/03548/bb_t4k0
perl: warning: Setting locale failed.
perl: warning: Please check that your locale settings:
```

Read2RefMapper

Step 6.

Open Read2RefMapper

Open Read2RefMapper-1.1.0 from 'Apps'



Read2RefMapper

Step 7.

Select Inputs

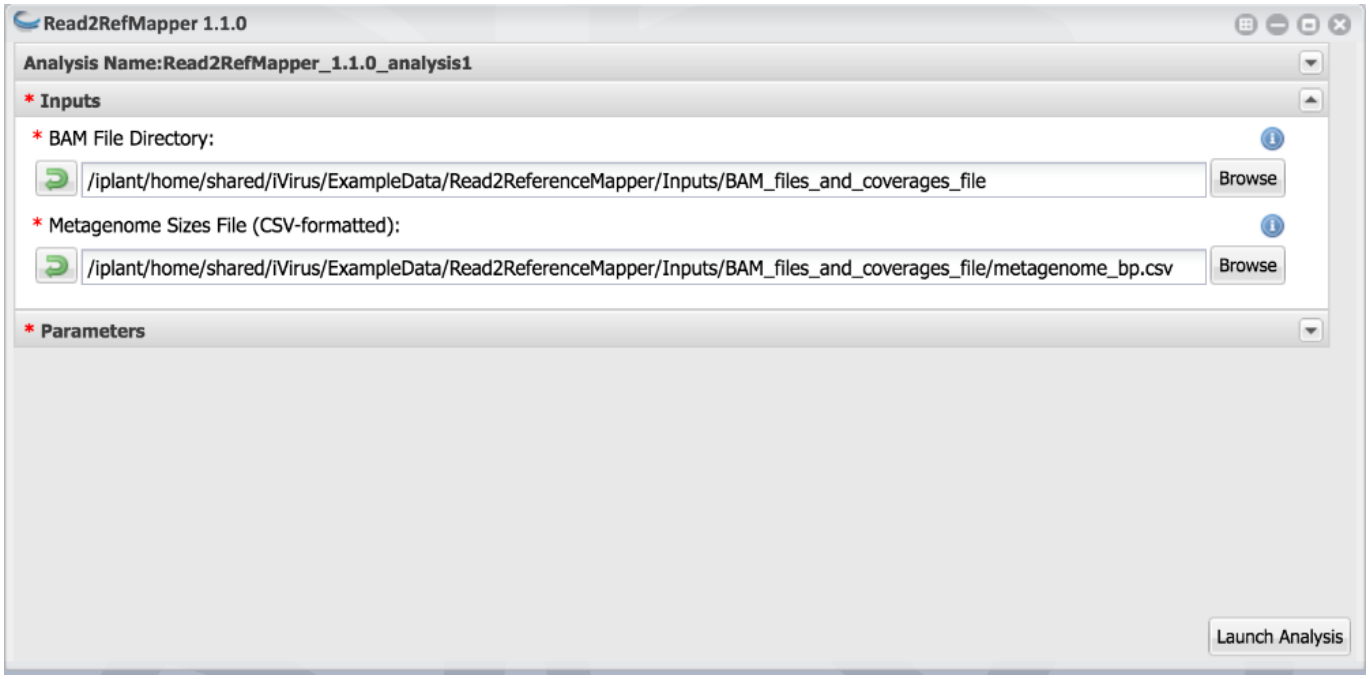
Select the 'Inputs' tab.

For **BAM File Directory (optionally including metagenome sizes file)**:

- Navigate to *Community Data* --> *iVirus* --> *ExampleData* --> Read2ReferenceMapper --> *Inputs*. Select the *BAM_files_and_coverages_file* folder. This will contain BAM files and a coverage file. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/Read2ReferenceMapper/Inputs into the navigation bar and select the folder.

For **Metagenome Sizes File (CSV-formatted)**:

- Navigate to *Community Data* --> *iVirus* --> *ExampleData* --> Read2ReferenceMapper --> *Inputs*. Select the *BAM_files_and_coverages_file* folder. Select metagenome_bp.csv. Alternatively, copy-and-paste the location: /iplant/home/shared/iVirus/ExampleData/Read2ReferenceMapper/Inputs into the navigation bar and select the metagenome_bp.csv file.



NOTES

Benjamin Bolduc 31 Dec 2016

NEW: Prior versions of Read2RefMapper required users to type in the filename for the metagenome sizes file and place it within the same directory. Updates to the core code have eliminated this requirement. This example still uses the old directory structure.

Read2RefMapper

Step 8.

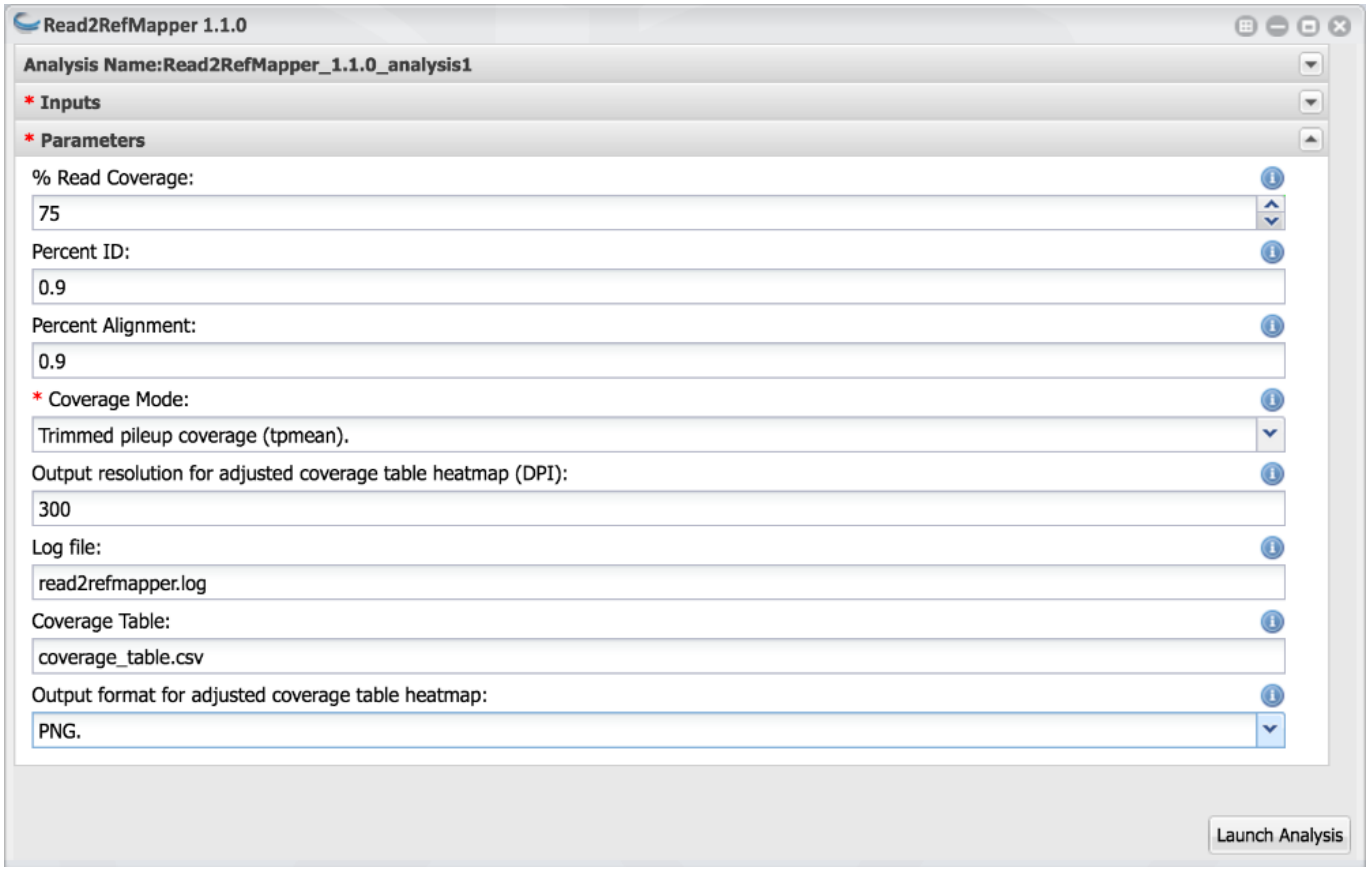
Select Parameters and Outputs

For this example, the defaults are sufficient.

% Read Coverage: Percent of the reference sequence that must be covered by reads. For example, if 75% of a reference must be covered by reads to be considered 'present' in that sample.
Percent ID: Minimum identity a read must be against the reference sequence. For example, 0.90 is 90% identity between the read and reference.
Percent Alignment: What percent of the read length must be aligned to be considered matching. For example, 0.90 is 90% of the *read length* must align to the reference. This allows users to excludes reads where only a small local alignment is present.
Coverage Mode: How coverage should be calculated.

Output resolution for adjusted coverage table heatmap (DPI):If metagenome sizes file provided, will dictate the resolution of the heatmap for the adjusted coverage.
Log file: Filename to store logging information.
Coverage table: Name of the coverage table.

Output format for adjusted coverage table heatmap: If metagenome sizes file provided, will generate a heatmap for the adjusted coverage.



Read2RefMapper

Step 9.

Launch Analysis

Run the job!

In terms of run time, Read2RefMapper is considerably faster than BatchBowtie (above).

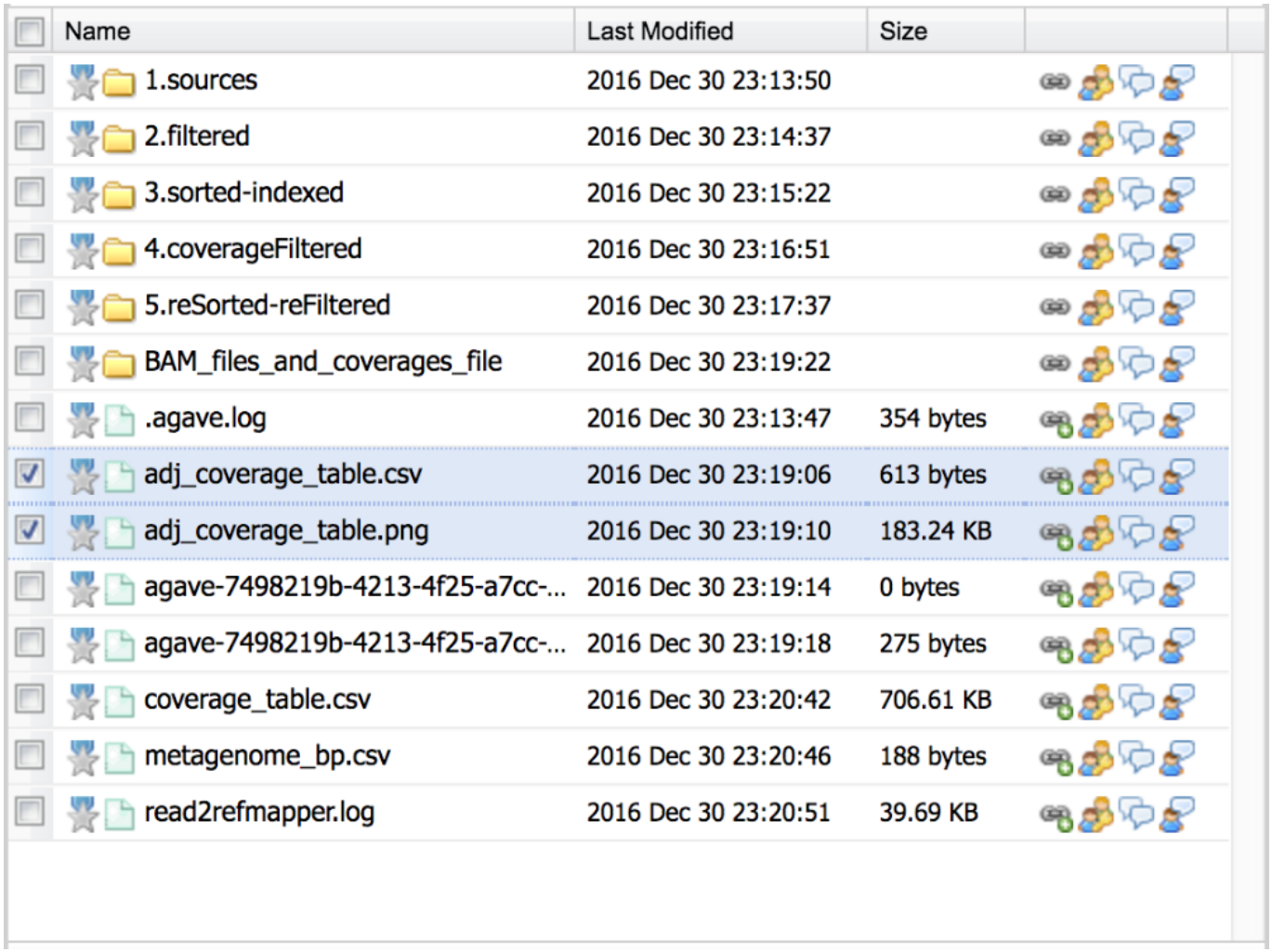
Read2RefMapper

Step 10.

Results

Expect results can be found in the associated app's 'output' directory.

There will be several directories, corresponding to each step along the pipeline. Notably are the *coverage_table.csv*, which will have raw coverages for each contig across all metagenomes, and the *adj_coverage_table.csv*. In this example, we specified a metagenome bp file, meaning Read2RefMapper will adjust/normalize the coverage table according to the metagenome sizes. If no metagenome bp file is found, then the adjusted/normalized coverage table will not be generated!



Name	Last Modified	Size
1.sources	2016 Dec 30 23:13:50	
2.filtered	2016 Dec 30 23:14:37	
3.sorted-indexed	2016 Dec 30 23:15:22	
4.coverageFiltered	2016 Dec 30 23:16:51	
5.reSorted-reFiltered	2016 Dec 30 23:17:37	
BAM_files_and_coverages_file	2016 Dec 30 23:19:22	
.agave.log	2016 Dec 30 23:13:47	354 bytes
adj_coverage_table.csv	2016 Dec 30 23:19:06	613 bytes
adj_coverage_table.png	2016 Dec 30 23:19:10	183.24 KB
agave-7498219b-4213-4f25-a7cc-...	2016 Dec 30 23:19:14	0 bytes
agave-7498219b-4213-4f25-a7cc-...	2016 Dec 30 23:19:18	275 bytes
coverage_table.csv	2016 Dec 30 23:20:42	706.61 KB
metagenome_bp.csv	2016 Dec 30 23:20:46	188 bytes
read2refmapper.log	2016 Dec 30 23:20:51	39.69 KB

EXPECTED RESULTS

The adjusted coverage table.

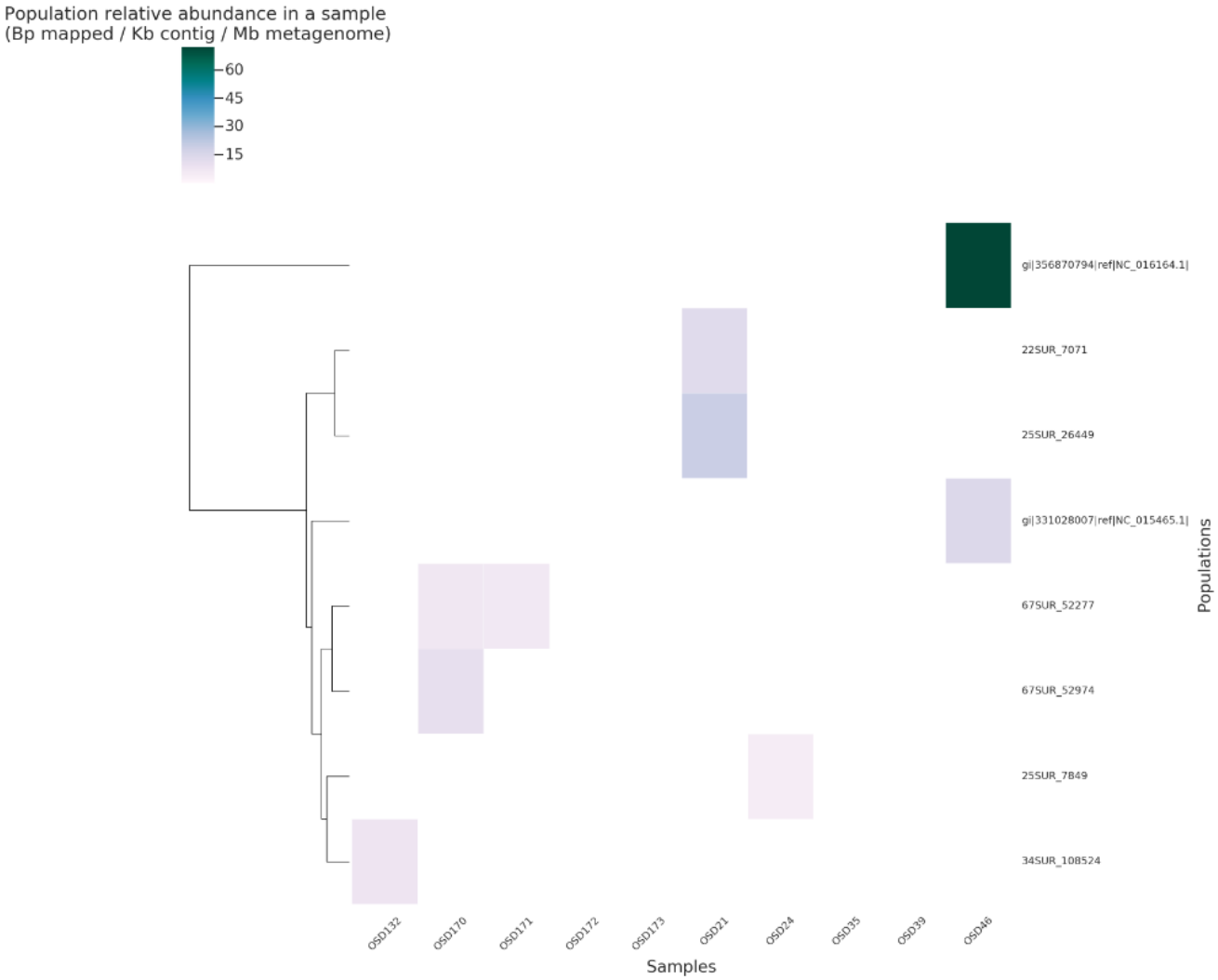
adj_coverage_table.csv

Tabular View: adj_coverage_table.csv | adj_coverage_table.csv

Save | Refresh | Line Numbers | Skip First N Line(s) 0 | First Row Header | Editable

0	1	2	3	4	5	6	7
#contig	OSD132	OSD170	OSD171	OSD172	OSD173	OSD21	OSD24
22SUR_7071	0.0	0.0	0.0	0.0	0.0	12.9200211496	0.0
25SUR_7849	0.0	0.0	0.0	0.0	0.0	0.0	5.08353390511
25SUR_26449	0.0	0.0	0.0	0.0	0.0	19.4146784438	0.0
34SUR_108524	8.4838156008	0.0	0.0	0.0	0.0	0.0	0.0
67SUR_52277	0.0	7.16997102321	6.73235754524	0.0	0.0	0.0	0.0
67SUR_52974	0.0	10.6403082552	0.0	0.0	0.0	0.0	0.0
gi|331028007|re...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
gi|356870794|re...	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The output PNG file:



NOTES

Benjamin Bolduc 31 Dec 2016

The non-adjusted coverage table results:

coverage_table.csv

Tabular View: coverage_table.csv | coverage_table.csv

0	1	2	3	4	5	6	7
#contig	Length	/work/03548/bb...	/work/03548/bb...	/work/03548/bb...	/work/03548/bb...	/work/03548/bb...	/work/03548/
109DCM_97	14114	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_614	16022	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_751	10589	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_1006	15802	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_1161	24128	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_1650	21707	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_1731	12288	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_1847	13361	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_2193	11566	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_2433	17771	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_2878	10902	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_3111	11841	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_3849	15865	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_4160	11481	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_4380	13426	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_4727	18243	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
109DCM_4943	15454	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000