InteMAP Manual

What is InteMAP?

InteMAP (Integrated Metagenomic Assembly Pipeline) is a pipeline which integrates individual assemblers for assembling metagenomic short sequencing reads. By integrating three assemblers IDBA_UD, ABySS, and CABOG, InteMAP takes advantage of their strengths on assembling species with special sequencing coverage depths, and generates a high quality assembly.InteMAPis suitable for unix-like system with gcc and python installed.

InteMAP was designed to work with paired-end reads produced by the Illumina Genome Analyzer.

Before using InteMAP, you are recommended to filter your data to remove low quality reads or contaminant reads via special tools, such as PRINSEQ (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi).

Prerequisites

To use InteMAP, you will need the following programs in your PATH:

- IDBA_UDrelease 1.1.0 or later (http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/)
- CABOG (also named Celera or wgs-assembler) release 7.0 or later (http://wgs-assembler.sourceforge.net/wiki/index.php?title=Main_Page)
- ABySS release 1.2.7 or later (http://www.bcgsc.ca/platform/bioinfo/software/abyss)
- Quake version 0.3 (JELLYFISH version 1.1 is required, not 2.0) (http://www.cbcb.umd.edu/software/quake/index.html) (http://www.cbcb.umd.edu/software/jellyfish/)
- MUMmer release 3.0 (http://mummer.sourceforge.net/)
- Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)

Note: the program ABySS requires Boost, sparsehash, and Open MPI. Please refer to the instruction of ABySS.

Make sure the binaries of these programs are in directories in your PATH environment variable.

You also need python to run InteMAP.

- ## Obtaining InteMAP

Download the InteMAP source package from http://bioinfo.ctb.pku.edu.cn/InteMAP.

- ## Installation

The InteMAP package includes the C++ programs and some scripts written in python and bash. The C++ programs include implementation of the assembly-merging algorithm and other toolsused by InteMAP. To use InteMAP, you only need to build the C++ programs. To do this, execute these commands on any unix-like platform after you download the InteMAP package:

```
$ tar –zxvf InteMAP_v1.0.tar.gz
$ cd InteMAP_v1.0
$ cd src/mergeassembly
$ make
$ cd ../../
$ make
```

Now, InteMAP is ready. Suppose your InteMAP directory is InteMAP-dir, you can use InteMAP by executing the python script:

```
$ python InteMAP-dir/runInteMAP.py [arguments]
```

Optionally, you can make the python script executable andadd your InteMAP directoryto your shell's PATH, so that whenever you run InteMAP from the command line, you will not have to specify the entire path. First, change the script file's mode toexecutable:

```
$ chmod +x runInteMAP.py
```

Then add the InteMAP directory to your shell's PATH. After that, you can use InteMAP directly:

```
$ runInteMAP.py [arguments]
```

- ## Test InteMAP

Here is an example to test InteMAP:

```
$ cd [your InteMAP directory]
$ cd example
$ python ../runInteMAP.py Ecoli-files libraryinfofiletotparafile
```

It will cost about 10 minutes to finish the task.The output file is ‘out.fa’ that keeps the final contigs in fasta format.

- ## Preparing sequence files

Right now InteMAPsupports paired-end reads in fastq format. The paired reads should be in separate files. A paired of reads must be named with the suffixes with /1 and /2 to identify the first and second read. In addition, the orientation of pared end reads should be 5’-3’ <-> 3’-5’ (the usual case).

To use InteMAP, prepare two files: **Seq-file**and **Library-info-file. Seq-file**keeps the directories of sequencing read files. **Library-info-file**keeps the information of the library for sequencing reads. Two paired read files from one library should be written in one line and separated by a space. If you have more than one library of sequences, use multiple lines and make sure each library use one line. The **Library-info-file** is prepared for running CABOG in InteMAP pipeline, and the information recorded in each line will be input as arguments for running ‘fastqToCA’ in CABOG program on the command line.To see the details of the arguments of ‘fastqToCA’, execute ‘fastqToCA’ on the command line. Following three options are mandatory:

-libraryname p	The UID of the library
-insertsizei d	Mates are on average i +- d bp apart
-type t	What type of fastq: ‘sanger’ – QV’s are PHRED, offset=33, NCBI SRA data. ‘illumina’ – QV’s are PHRED, offset=64

All the information for each library should be written in one line, and the order of the libraries must be consistent with the order in **Seq-file** if there are more than one libraries.

In the previous example, ‘Ecoli-files’ is the **Seq-file**and ‘libraryinfofile’ is the **Library-info-file.**

- ## Usage

After preparing the **Seq-file**and **Library-info-file,** run InteMAP:

$ python InteMAP-dir/runInteMAP.py Seq-file Library-info-file [options]

or,if you have added InteMAP directory in your PATH:

$ runInteMAP.pySeq-file Library-info-file [options]

There are two ways to specify options.

1. Assign options via command line.

Options can be assigned via command line:

-t, --thread_num [int]
Number of threads, default=2

-q, --quality_start {33,64}
Quality value ASCII start, default=33

-o, --outfile [string]
Output contig file name, default='out.fa'

-l, --minHighCovLength [int]
Indicate the high coverage length threshold, above which InteMAP will deem that high-coverage species exist in the community, and run IDBA-UD and ABySS for high coverage sequences, default=1000000

--k_for_abyss [int]
kmer size for ABySS, default=61

--min_k_for_idba [int]
Option 'mink' for IDBA-UD, default=23

--high_cov_idba [int]
lower bound for IDBA-contigs of high coverage, default=20

--low_cov_idba [int]
higher bound for IDBA-contigs of low coverage, default=50

--high_cov_abyss [int]
lower bound for ABySS-contigs of high coverage, default=20

-c, --clearance
Make clearance of intermediate output files, default=False

In previous example, you can also run InteMAP:

$ runInteMAP.pyEcoli-files libraryinfofile -l 0

2.Assign parameters via files.

InteMAP integrates several programs, each program having specific options. Options for each step of InteMAP can be written in individual specific files, and the names (directories) of these files must be written in a master file, i.e.**Specfile**. Besides, **Specfile**is also used to keep the options for the InteMAP pipeline, such as the option 'minHighCovLength'.

After the **Specfile** is ready, run InteMAP:

```
$ runInteMAP.pySeq-file Library-info-file Specfile
```

2.1. Options for master file**Specfile**(optional)**:**

The options should be written in the pattern:
option [value]
in individual lines exclusively.

2.1.1. Options forspecifying daughter files for individual programs integrated in the InteMAP pipeline.

abyssparafile [file name/directory(String)]
The file keeps the options for running ABySS in the InteMAP pipeline. ('InteMAP-dir/Para/abyssparafile' by default)

bowtie2parafile [file name/directory(String)]
The file keeps the options for running bowtie2 in the InteMAP pipeline. ('InteMAP-dir/Para/bowtie2parafile' by default)

cabogspecfile [file name/directory(String)]
The file keeps the options for running CABOG in the InteMAP pipeline. ('InteMAP-dir/Para/cabogspecfile' by default)

idbahighparafile[file name/directory(String)]
idbalowparafile [file name/directory(String)]
The InteMAP pipeline runs IDBA-UD twice. The first time runs IDBA-UD on 'total read set' (See InteMAP paper for details) for low coverage. The second time runs IDBA-UD on corrected read set (See [InteMAP paper] for explaination) for high coverage. *idbahighparafile* and *idbalowparafile* specify the options for running IDBA-UD for high coverage and low coverage, respectively. ('InteMAP-dir/Para/idbaparafile' by default)

quakeparafile [file name/directory(String)]

The file keeps the options for running the Quake in the InteMAP pipeline. ('InteMAP-dir/Para/QuakeParaFile' by default)

2.1.2. Global options for InteMAP pipeline

clearance [Value(Int)]
If specified NOT 0, clearance of the files output by intermediate steps will be made. (0 by default)

Filter-idba-high[Value(Int)]
Filter-idba-low[Value(Int)]
Filter-abyss [Value(Int)]
The InteMAP pipeline filters the assembled contigs within a certain range of sequencing coverage and merges them together.
Filter-idba-low specifies the upper bound of the coverage for IDBA contigs of low coverage (50 by default); *Filter-idba-high* specifies the lower bound for IDBA-contigs of high coverage (20 by default); and *Filter-abyss* specify the lower bound for ABySS-contigs of high coverage (20 by default).

minHighCovLength [Value(Int)]:
This option specifies the minimal value of the total length of high-coverage contigs determined by InteMAP after running IDBA-UD on the 'total read set', above which InteMAP will deem that high-coverage species exist in the community, and run IDBA-UD and ABySS on the high coverage sequences. (1000000 by default)

output [file name/directory(String)]:
This specify the output contig file name. (‘out.fa’ by default)

2.2. Options for daughter files

2.2.1. Options for**abyssparafile**

All the options must be written in one line, as this line will be input on the command line when running ABySS.
The way to specify the options for **abyssparafile**is '[option1]=[value1] [option2]=[value2] ...'

InteMAP uses the program 'abyss-pe' only when running ABySS, which means ABySS in InteMAP only assemble paired-end reads. Except the input read files

that are specified by InteMAP, all other options for 'abyss-pe' can be specified in **abyssparafile**. See instructions of ABySS for the options of 'abyss-pe'.

Some important options are:
k=[value(Int)] indicates the kmer size for running ABySS in InteMAP; (k=61 by de fault)
n=[value(Int)] indicates the minimal number of mate pairs links to merge two contigs when scaffolding when running ABySS in InteMAP; (n=5 by default)
np=[value(Int)] indicates the number of threads; (np=2 by default)

2.2.2. Options for bowtie2parafile

Options for running bowtie2 can be checked from the help files of the command 'bowtie2'. The only options supported in the bowtie2parafile are: “--phred64” or “--phred33”. These two options are mutually exclusive.
“--phred64” means qualities of fastq are Phred+64,while “--phred33” means qualities are Phred+33 (by default in InteMAP).

2.2.3. Options for cabogspecfile

cabogspecfile keeps the options for running CABOG in InteMAP. The options and the way to specify them can be found from the [CABOG webpage]:
http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=SpecFiles
The default cabogspecfile is 'InteMAP-dir/Para/cabogspecfile'

2.2.4. Options for idbahighparafile and idbalowparafile

All the options must be written in one line.
Options for running IDBA-UD for high coverage and low coverage.Execute 'idba_ud' on the command line to see options supported by IDBA-UD. Except the input read files and the output directory, all other options can by specified in the idbahighparafile/idbalowparafile.
By default, InteMAP uses 'InteMAP-dir/Para/idbaparafile' for both option inputs for running 'idba_ud'.
Important options: '--num_threadsarg(=0)', number of threads; '--mink arg(=20)', minimum k value (In InteMAP, mink=23 by default)

2.2.5. Options of quakeparafile

InteMAP run Quake in several steps: 'jellyfish count', 'jellyfshdump', 'correct'. The options in quakeparafile include the options for all these steps. Items of options are written in individual lines. Each line has only one item of option specified in the pattern as below:

[option] [value]
Options include:

quality-start [int]	Starting ASCII for quality values (33 by default in InteMAP),
kmer [int]	Length of mer (17 by default in InteMAP), used as the option '-m' in 'jellyfish count', the option '-m' in 'jellyfish merge', and the option '-k' in 'correct'.
hash_size [uint64]	Hash size (2G by default in InteMAP), used as the option '-s' in both 'jellyfish count' and 'jellyfish qmerge'.
Threads[int]	number of threads (2 by default), used as the option '-t' in 'jellyfish count' and the option '-p' in 'count'.

- **Output**

InteMAP output contigs in fasta format. The output file name is ‘out.fa’ by default or specified by users via option ‘-o/--outfile’.