



Version 1.4.1 Last updated 20th March 2015

This document is released under the GNU Free documentation license version 1.3 (<https://www.gnu.org/copyleft/fdl.html>)

BamM is a c library, wrapped in python, that parses BAM files.

The code is intended to provide a faster, more stable interface to parsing BAM files than PySam, but doesn't implement all / any of PySam's features.

Do you want all the links that join two contigs in a BAM?

Do you need to get coverage?

Would you like to just work out the insert size and orientation of some mapped reads?

Then BamM is for you!

For impatient people

1. `$ bamm make -d <reference.fna> -c read1.R1.fq.gz read1.R2.fq.gz ...`
2. `$ bamm parse -c covs.tsv -l links.tsv -i inserts.tsv -b mapping.bam`
3. `$ bamm extract -g BIN_1.fna -b mapping.bam`

*BamM make produces indexed and sorted BAM files which contain only reads that mapped.

Table of Contents

Installation.....	3
Dependencies.....	3
Installing BamM.....	3
Testing your installation.....	4
BamM command line overview.....	5
Making BAM files.....	5
Passing extra information to BamM.....	6
Getting information from a collection of BAM files.....	7
Finding the insert size and relative orientation of paired reads.....	8
Creating a coverage profile from several BAM files.....	9
Finding reads that link 2 contigs.....	10
Extracting reads from BAM files.....	10
Using BamM as a python library.....	12
Calculating coverage profiles, insert types or linking reads.....	12

Installation

The BAM parsing is done using c and a few external libraries. This slightly complicates the BamM installation process but fortunately not too much.

Dependencies

If you're running 'BamM make' or you want to run nosetests after installation you'll need to have BWA and Samtools installed. Installation of these tools is really straightforward. You can find the code and instructions at:

Samtools: <http://sourceforge.net/projects/samtools>
(BamM is tested against v1.2)

BWA: <https://github.com/lh3/bwa> (BamM is tested against v0.7.12)

The two external dependencies of BamM; libcfu (v1.0.3) and htlib (v1.2.1) are bundled with the source. I strongly advise against trying to install libcfu independently but there may be some reasons to use a local install of htlib. The notes below are for installing on Ubuntu but they should be transferable to other *nix systems. **Be aware that some users have reported difficulty installing the dependencies (htlib) and BamM on OSX**, you may need some sysadmin-fu (or need to find someone with a suitable amount of sysadmin-fu) to get things going properly. I strongly doubt this code will work on a Windows system.

First, you need git, zlib, numpy, a C-compiler and some friends. On Ubuntu this looks like:

```
$ sudo apt-get -y install git automake build-essential zlib1g-dev python2.7-numpy libtool
```

Installing BamM

Get the latest version from github (BamM is not available on pip):

```
$ git clone https://github.com/Ecogenomics/BamM.git
```

If you are using the bundled htlib and libcfu or if these tools are installed system-wide then installation is very straight forward:

```
$ sudo python setup.py install
```

If you installed one or more of these libraries locally then you need to tell setup.py where they are:

```
$ python setup.py install --with-libhts-lib /path/to/htslib  
--with-libhts-inc /path/to/htslib --with-libcfu-inc /path/to/libcfu/include/  
--with-libcfu-lib path/to/libcfu/lib/
```

Relative paths are OK. You can add the --prefix flag to setup.py to install BamM locally. Once done, don't forget to add BamM to your PYTHONPATH. Also, if htlib and libcfu are in non-standard places and you plan to access the C code, you'll need to mess with your LD_LIBRARY_PATH.

Testing your installation

After install you can test your installation using nosetests.

First, make sure pip, nose, bwa and samtools are installed:

```
$ sudo aptitude install python-pip  
$ sudo pip install nose pysam
```

Now test BamM:

```
$ cd BamM  
$ nosetests
```

You *should see output like this

```
$ nosetests  
.....  
-----  
Ran 45 tests in 4.574s  
  
OK
```

If you have any errors then you should try fix them now (or they may bite you downstream).

BamM command line overview

I've wrapped the python in a script / library called BamM.

BamM has 3 modes; 'make', 'parse' and 'extract'. The first option allows you to make BAM files. The second option lets you derive coverage profiles or linking information. The final option lets you extract reads that map to a set(s) of contigs.

Making BAM files

BamM make

Description:

Map several read sets onto a single reference with one command.

Required arguments:

-d --database <file.fna> name of fna file to map reads onto

And at least one of:

-i --interleaved <interleaved.fa> [<interleaved.fa> ...] shuffled reads
-c --coupled <coupled.fa> [<coupled.fa> ...] paired files
-s --single <single.fa> [<single.fa> ...] singleton read files

Optional arguments:

-p --prefix	"	prefix to apply to BAM files
-o --out_folder	'.'	write to this folder
--index_algorithm	auto	algorithm bwa uses for indexing
--alignment_algorithm	mem	algorithm bwa uses for alignment
--extras	"	pass extra arguments to BWA
-k --keep	False	keep all the database index files etc after
-K --kept	False	assume the indices already exist, don't re-make
-f --force	False	force overwrite of index files if they are present
--output_tam	False	output TAM file instead of BAM file
-t --threads	1	max number of threads to use
-m --memory	2GB/thread	maximum memory to use per bwa process Suffix K/M/G recognized
--show_commands	False	show all commands made by the wrapper
--quiet	False	suppress output from the mapper
--silent	False	suppress all output

Example usage:

```
$ bamm make -d my_assembly.fa -i ileaved_1.fastq.gz ileaved_2.fastq.gz -c  
paired_R1.fastq.gz paired_R2.fastq.gz -s unpaired.fastq.gz [-t 20] [-v]
```

This command will make 4 BAM files by mapping the two interleaved read sets, the one paired read set and the singleton read set onto the reference my_assembly.fa

The code calls BWA and Samtools to produce a set of sorted and indexed BAM files. If you specify -t <threads> then BamM will pass this onto BWA and Samtools.

Use -v to get more verbose output.

NOTE: To save space, the final BAM files contain only mapped reads.

NOTE: Output files are automatically named based on the names of the read files and the database, however you can specify the output directory and a prefix to append to the beginning of all output files.

Passing extra information to BamM

```
$ bamm make -d my_assembly.fa -i ileaved_1.fastq.gz ileaved_2.fastq.gz -  
extras "mem:-k 25"
```

```
$ bamm make -d my_assembly.fa -i ileaved_1.fastq.gz ileaved_2.fastq.gz -  
extras "aln:-l 25,sampe:-s -a 100"
```

The first command will pass the extra argument "-k 25" to bwa mem. The second command will pass the extra arguments "-l 25" to bwa aln and "-s -a 100" to bwa sampe.

WARNING

Values passed using --extras are not checked by BamM. This represents a significant security risk if BamM is being run with elevated privileges. Thus you should NEVER run 'bamm make' as root or some other powerful user, ESPECIALLY if you are providing access to multiple / unknown users.

Getting information from a collection of BAM files

BamM parse

Description:

BamM parse lets you find the insert size(s) and relative read orientation(s) associated with a collection BAM files, it lets you find paired reads that link contigs together and it also lets you create coverage profiles of individual contigs across multiple BAM files.

Required arguments:

`-b --bamfiles <bamfile> [<bamfile> ...]` space separated list of BAM files to parse

Optional arguments:

<code>-l --links</code>	<code>"</code>	filename to write pairing links to
<code>-i --inserts</code>	<code>"</code>	filename to write insert distributions to
<code>-c --coverages</code>	<code>"</code>	filename to write coverage profiles to
<code>-n --num_types</code>	<code>1</code>	number of insert/orientation types per BAM
<code>-m --coverage_mode</code>	<code>pmean</code>	how to calculate coverage* (req --coverages)
<code>-r --cutoff_range</code>	<code>?</code>	used to calculate upper / lower rejection cut offs when calculating coverage
<code>--length</code>	<code>ANY</code>	minimum Q length
<code>--base_quality</code>	<code>ANY</code>	base quality threshold (Qscore)
<code>--mapping_quality</code>	<code>ANY</code>	mapping quality threshold
<code>--max_distance</code>	<code>ANY</code>	maximum allowable edit distance from query to reference
<code>--use_secondary</code>	<code>False</code>	use reads marked with the secondary flag
<code>--use_supplementary</code>	<code>False</code>	use reads marked with the supplementary flag
<code>-v, --verbose</code>	<code>False</code>	be verbose
<code>-t --threads</code>	<code>1</code>	maximum number of threads to use

The 'cutoff_range' variable is used for trimmed mean and outlier mean. This argument takes at most two values. The first is the lower cut off and the second is the upper. If only value is supplied then lower == upper.

Example usage:

Finding the insert size and relative orientation of paired reads

```
$ bamm parse -b file.bam
```

Produces output like this:

#file	insert	stdev	orientation	supporting
file.bam	899.7514	14.7167	IN	10000

The *IN* orientation indicates that this is an Illumina-style paired-end (PE) library with an insert of ~900 bp and a standard deviation of ~15 bp. Illumina-style mate pair (MP) libraries will typically have orientation *OUT*.

Many MP libraries also have a shadow library which looks like someone added some PE reads to the mix. You can tell BamM to look for more than one insert type by specifying the *-n* option:

```
$ bamm parse -n 2 -b mate_pair_file.bam
```

Produces output like this:

#file	insert	stdev	orientation	supporting
mate_pair_file.bam	2524.4540	729.1291	OUT	10000
mate_pair_file.bam	251.6253	44.7241	IN	10000

Multiple BAM files are separated using spaces. The *-n* argument is space separated too. By default BamM prints this info to stdout. Use the *-i* argument to specify a file to write the results to.

```
$ bamm parse -b pe_file.bam mp_file.bam -n 1 2 -i inserts.tsv
```

This command will analyse the reads in *pe_file.bam* and try to find one insert type and the reads in *mp_file.bam* and try to find two insert types. The resulting table will be written to the file *inserts.tsv*.

Creating a coverage profile from several BAM files

When passed the '-c <filename>' argument, BamM will produce a table of coverage values for each BAM file. This is referred to as a coverage profile.

```
$ bamm parse -c coverage.tsv -m <COV_MODE> -b f1.bam f2.bam f3.bam
```

Produces this output in the file 'coverage.tsv'

#contig	Length	f1.bam	f2.bam	f3.bam
contig_1	946	103.0000	327.0000	369.0000
contig_3	1147	130.0000	492.0000	778.0000
contig_5	1465	228.0000	643.0000	970.0000
contig_7	168	34.0000	82.0000	102.0000
contig_9	4045	899.0000	1756.0000	2649.0000

The -t option indicates the maximum number of threads BamM will use. This option speeds up the process but you should only use as many threads as you have BAM files. If you have 6 BAM files then you'll see no improvement when using -t 6, -t 7 or -t 700.

Coverage calculation modes

BamM implements several coverage calculation methods. The user can choose the method using the -m argument.

opmean: Outlier pileup coverage: average of reads overlapping each base, after bases with coverage outside mean ± 1 standard deviation have been excluded. The number of standard deviation used for the cutoff can be changed with --coverage_range.

pmean: Pileup coverage: average of number of reads overlapping each base

tpmean: Trimmed pileup coverage: average of reads overlapping each base, after bases with in the top and bottom 10% have been excluded. The 10% range can be changed using --coverage_range.

counts: Absolute number of reads mapping

cmean: Like 'counts' except divided by the length of the contig

pmedian: Median pileup coverage: median of number of reads overlapping each base

Finding reads that link 2 contigs

When passed the '-l <filename>' argument, BamM will find paired reads that link contigs.

```
$ bamm parse -l links.tsv -b f1.bam f2.bam f3.bam
```

Produces this output in the file links.tsv:

#cid_1	cid_2	len_1	pos_1	rev_1	len_2	pos_2	rev_2	file
contig_2203	contig_3479	1664	334	0	3873	2866	0	f2.bam
contig_2203	contig_3479	1664	384	0	3873	2818	0	f2.bam
contig_2203	contig_3479	1664	383	0	3873	2831	0	f2.bam
contig_2203	contig_3479	1664	349	0	3873	2864	0	f2.bam
contig_2203	contig_3479	1664	338	0	3873	2862	0	f2.bam

The first (non-header) line is interpreted like this:

contig_2203 is linked to contig_3479.

The first read is towards the start of contig_2203 (len_1 == 1664, pos_1 == 334) and is in the same orientation as the contig (rev_1 == 0)

The second (paired) read is towards the end of contig_3479 (len_2 == 3873, pos_2 == 2866) and is also in the same orientation as the contig (rev_2 == 0)

The linking information was extracted from file: f2.bam.

The lines following this one describe other links between the two contigs.

NOTE: the -i, -c and -l options are not mutually exclusive and can be run at the same time.

Extracting reads from BAM files

BamM extract

Description:

Extract reads that map to collections of contigs from a collection of BAM files.

Required arguments:

-g --groups <group> [<group> ...]	files containing reference names (1 per line) or contigs file in fasta format
-b --bamfiles <file> [<file> ...]	BAM files to parse

Optional arguments:

-p --prefix	"	prefix to apply to output files
-o --out_folder	'.'	write to this folder
--mix_bams	False	use the same file for multiple BAM files
--mix_groups	False	use the same files for multiple group groups
--mix_reads	False	use the same files for paired/unpaired reads

<code>--interleave</code>	False	interleave paired reads in output files
<code>--mapping_quality</code>	0	mapping quality threshold
<code>--use_secondary</code>	False	use reads marked with the secondary flag
<code>--use_supplementary</code>	False	use reads marked with the supplementary flag
<code>--max_distance</code>	1000	maximum edit distance from query to reference
<code>--no_gzip</code>	False	do not gzip output files
<code>--headers_only</code>	False	extract only (unique) headers
<code>-v, --verbose</code>	False	be verbose
<code>-t --threads</code>	1	maximum number of threads to use

Example usage:

Extract all reads mapping to a particular set of contigs

```
$ BamM extract -g group1.file group2.file -b f1.bam f2.bam f3.bam
```

Will extract all reads from each of the three BAM files that map to the contigs in group1 or group2. The 'group' files can be multiple (gzipped) FASTA (like the fna files you can extract from GroopM) or lists of contig headers (one sequence per line).

Unless specified otherwise, BamM differentiates between paired and unpaired reads (from a mapping and group perspective), reads from different BAM files and reads mapping to contigs in different groups.

Paired reads may not be paired when mapped (only one read maps). Also paired reads may map to different groups / bins, so when they're extracted they are *unpaired* in a *group* sense. BamM preserves this (and other) information in the read header.

The read header has the following format:

```
>g_<group>;p_<pairing_info>;b_<bamfile>;c_<contig_id>;r_<read_id>
```

Where:

<code><group></code>	the name of the group / bin file with the contig this read maps to.
<code><pairing_info></code>	describes the pairing information about this read (see below).
<code><bamfile></code>	the name of the BAM file containing this read.
<code><contig_id></code>	the id of the contig the read maps to.
<code><read_id></code>	the id of the read as given in the BAM file.

The pairing information has the following format:

p_<code>R_<code>M_<code>G;

Where <code> is one of:

P	Paired	N	Not applicable
U	Unpaired	E	Error

For example:

p_PR_PM_PG; indicates a paired read that is mapped as a pair to contig(s) within the same group.

p_PR_PM_UG; indicated a paired read that is mapped as a pair to contig(s) that are in different groups.

p_PR_UM_NG; indicates a paired read where only this read was mapped.

p_UR_NM_NG; indicates an unpaired read.

NOTE: this command can produce A LOT of output files.

Using BamM as a python library

BamM is intended to be used as a python library in any code that needs to produce coverage profiles or paired read linking information. The dev docs will be more useful than this quick guide (see http://ecogenomics.github.io/BamM/dev_docs/). But it's here to give a taste of what you can do with BamM.

Calculating coverage profiles, insert types or linking reads

The following snippet shows how to calculate a coverage profile within your code.

```
# first import it
from BamM.bamParser import BamParser
from bamm.bamFile import BM_coverageType

# choose the type of coverage to calculate and make a parser
cov_type = BM_coverageType(CT.P_MEAN_OUTLIER, 1, 1)
BP = BamParser(cov_type)

# get a list of BAM files and parse them
bam_files = ['file1.bam', 'file2.bam']
BP.parseBams(bam_files,
              doLinks=False,          # set to False for no links
              doCovs=True,            # set to False for no coverages
              threads=2)               # 2 files so use 2 threads
```

The BamParser has an instance variable called a BamFileInfo (BFI). This object contains all the information that has been derived from the BAM files.

BP.BFI.numBams	the number of BAMs that were parsed
BP.BFI.bamFiles	path to the BAM files
BP.BFI.numContigs	the number of contigs in the BAM files
BP.BFI.contigLengths	array of contig lengths
BP.BFI.contigNames	array of contig names
BP.BFI.coverages	matrix of coverage values (numContigs x numBams)
BP.BFI.links	hash of links (see dev docs for more information)

The BamParser comes with several pre-written print functions:

BP.printBamTypes()	print insert information as detailed above
BP.printCoverages()	print coverage information
BP.printLinks()	print linking read information

All of these functions print to stdout. If you supply a file name then BamM will print the information there.