xenome(1) Xenome User Manual

Bryan Beresford-Smith, Andrew Bromage,
Thomas Conway, Jeremy Wazny

September 12, 2012

NAME

xenome - a tool for classifying reads from xenograft sources.

Version 1.0.1

SYNOPSIS

xenome index -T 8 -P idx -H mouse.fa -G human.fa

xenome classify -T 8 -P idx —pairs —host-name mouse —graft-name human -i in_1.fastq -i in_2.fastq

xenome help

DESCRIPTION

Shotgun sequence read data derived from xenograft material contains a mixture of reads arising from the host and reads arising from the graft. Xenome is an application for classifying the read mixture to separate the two, allowing for more precise analysis to be performed.

Xenome uses host and graft reference sequences to characterise the set of all possible k-mers according to whether they belong to:

- only the graft (and NOT the host)
- only the host (and NOT the graft)
- both references
- neither reference
- the subset of the host (or graft) k-mers which is one base substitution away from being in the graft (or host) - we call these k-mers "marginal"

Given a read, or read pair, xenome will calculate which of the above categories its k-mers belong to, and classify it as one of: graft, host, both, neither, or ambiguous.

Xenome has two distinct stages, which are embodied in two separate commands: 'index' and 'classify'. Before reads can be classified, an index must be constructed from the graft and host reference sequences. The references must be in FASTA format, and may optionally be compressed (gzip).

```
xenome index -M 24 -T 8 -P idx -H mouse.fa -G human.fa
```

A xenome index consists of a number of related files which can be identified by a user-specified prefix, e.g. 'idx' in the above command. The prefix may contain '/' characters, allowing the index to be in a sub-directory. (Any such sub-directory must already exist - xenome will not create it.) For example, the set of files comprising an index with prefix 'idx' are:

```
idx-both.header
idx-both.kmers-d0
idx-both.kmers-d1
idx-both.kmers.header
idx-both.kmers.high-bits
idx-both.kmers.low-bits.lwr
idx-both.kmers.low-bits.upr
idx-both.lhs-bits
idx-both.rhs-bits
```

Once an index is available, reads can be classified according to whether they appear to contain graft or host material. In the simplest case, Xenome can classify each read from a single source file individually.

```
xenome classify -P idx -i in.fastq
```

This step produces a file for each read category, containing all of the reads which have been assigned that classification:

```
ambiguous.fastq
both.fastq
graft.fastq
host.fastq
neither.fastq
```

Input files are base-space reads in FASTA or FASTQ format or in a format with one read per line and in either plain text or compressed format (gzip).

The files produced are in the same format as the input file, with all of the input read data preserved. i.e. if the input reads are in FASTQ format, the reads written to each of the output files will also be in FASTQ format.

Multiple input files may be specified, but all inputs in the same format will be written to the same set of output files.

```
xenome classify -P idx -i inA.fastq -i inB.fastq -I inC.fasta
```

The above will result in the following set of files:

```
ambiguous.fasta
ambiguous.fastq
both.fasta
both.fastq
graft.fasta
graft.fastq
host.fasta
host.fastq
neither.fasta
neither.fastq
```

Each of the FASTQ files contains a mixture of reads from inA.fastq and inB.fastq. The FASTA files contain reads from inC.fasta.

If the combining of input reads from separate files is not desired, xenome should be run separately for each input. The output from different runs can be distinguished by prefixing the filenames with a distinct string.

```
xenome classify -P idx -i inA.fastq --output-filename-prefix A
xenome classify -P idx -i inB.fastq --output-filename-prefix B
```

Running these two commands yields:

```
A_ambiguous.fastq
A_both.fastq
A_graft.fastq
A_host.fastq
A_neither.fastq
B_ambiguous.fastq
B_both.fastq
B_graft.fastq
B_host.fastq
B_neither.fastq
```

Xenome can also process pairs of reads.

```
xenome classify -P idx --pairs -i in_1.fastq -i in_2.fastq
```

This results in a pair of files for each read category. The two reads of each pair are written to the corresponding '_1' and '_2' files respectively.

```
ambiguous_1.fastq
ambiguous_2.fastq
both_1.fastq
both_2.fastq
graft_1.fastq
graft_2.fastq
host_1.fastq
host_2.fastq
neither_1.fastq
neither_2.fastq
```

If desired, more specific names can be used in place of 'host' and 'graft'.

```
xenome classify -P idx -i in.fastq --graft-name human --host-name mouse
```

This will cause xenome to produce the following files.

```
ambiguous.fastq
both.fastq
human.fastq
mouse.fastq
neither.fastq
```

In addition to generating sets of output files, the classify command produces statistics about the number and proportion of reads assigned to each category. These are printed to standard out at the end of a run and look as follows:

```
Statistics
B       G       H       M       count     percent   class
0       0       0       0       1900      0.938267  "neither"
0       0       0       1       21        0.0103703 "both"
0       0       1       0       28491     14.0696   "definitely host"
0       0       1       1       7366      3.63751   "probably host"
0       1       0       0       91895     45.38     "definitely graft"
0       1       0       1       30059     14.8439   "probably graft"
0       1       1       0       282       0.139259  "ambiguous"
0       1       1       1       330       0.162962  "ambiguous"
```

```
1       0       0       0       2878      1.42123   "both"
1       0       0       1       254       0.125431  "probably both"
1       0       1       0       610       0.301233  "definitely host"
1       0       1       1       5815      2.87159   "probably host"
1       1       0       0       3843      1.89777   "definitely graft"
1       1       0       1       27775     13.716    "probably graft"
1       1       1       0       99        0.0488886 "ambiguous"
1       1       1       1       883       0.436047  "ambiguous"

Summary
count     percent   class
153572    75.8377   "graft"
42282     20.8799   "host"
3153      1.55703   "both"
1900      0.938267  "neither"
1594      0.787157  "ambiguous"
```

Both tables contain a single heading line, followed by rows of TAB-separated elements; a format suitable for loading into R or a spreadsheet.

Each row represents the number and proportion of reads assigned to a particular class. The B, G, H, and M fields represent the presence (1) or absence (0) of k-mers belonging to the both, graft, host and marginal k-mer subsets, according to the reference index.

The Statistics table contains 16 rows; one for each possible combination of k-mer classes present within a read. The first row of the above table, indicates that for the given input, 1,900 reads (or pairs) - 0.938267% of the total reads - contained no k-mers that belonged to the B, G, H, or M k-mer subsets, and are accordingly neither host nor graft reads. Similarly, the fourteenth line states that 27,775 reads (or pairs) - 13.716% of the total - contained k-mers that belong to the B, G, M, but not H subsets, and are therefore “probably graft” reads.

In the Summary table, the B, G, H, and M columns are removed, and the classes from the Statistics table have been collapsed into the five shown; the definitely/probably graft/host classes are combined into just graft/host classes. Notice that the different read output files, described earlier, correspond exactly to these classes.

OPTIONS COMMON TO ALL COMMANDS

The following options can be used with all of the *xenome* commands and are therefore not listed separately for each command.

-h, –help Show a help message.

-l *FILE*, –log-file *FILE* Place to write progress messages. Messages are only written if the -v flag is used. If omitted, messages are written to stderr.

-T *INT*, –num-threads *INT* The maximum number of *worker* threads to use. The actual number of threads used during the algorithms depends on each implementation. *xenome* may use a small number of additional threads for performing non cpu-bound operations, such as file I/O.

–tmp-dir *DIRECTORY* A directory to use for temporary files. This flag may be repeated in order to nominate multiple temporary directories.

-v, –verbose Show progress messages.

-V, –version Show the software version.

COMMANDS AND OPTIONS

xenome index

xenome index [-k *INT*] [-M *INT*] -P *PREFIX* -G *FASTA-filename* -H *FASTA-filename*

Build the xenome reference index from the graft and host reference sequences. The input files must be in FASTA format. They may be gzip compressed, in which case the filename suffix must be *.gz*.

The k-mer size may be specified using the *-k* flag. If omitted, xenome defaults to k=25.

During index construction, xenome maintains a hash table of the k-mers seen so far. When this table fills, its contents are written to disk, and the table is reinitialised. The more memory xenome can use, the less often it will need to write to disk, and the faster index construction will run. By default, xenome will limit itself to 2 GB during index construction. The -M, —max-memory flag can be used to explicitly control the amount of memory available to xenome (in GB). To improve performance, this should generally be set close to the amount memory available in the system - having accounted for operating system and other overhead.

OPTIONS

-k *INT*, –kmer-size *INT* The k-mer size to use for building the graph: in version 1.0.0 this *must be an integer strictly less than 63.* If not supplied, the default value of 25 is used.

-M *INT*, –max-memory *INT* The maximum amount of memory (in GB) of memory to use. Making more memory available will reduce the number of times xenome writes intermediate index data to disk. The default is 2 GB.

-P *PREFIX*, –prefix *PREFIX* The path prefix for all generated reference index files. The prefix may contain directory separators (e.g. '/') in order to have the index files written to another directory.

-G *FILE*, –graft *FILE* The name of the FASTA file containing the graft reference sequence. If the filename ends in *.gz* it will be read as a gzip file.

-H *FILE*, –host *FILE* The name of the FASTA file containing the host reference sequence. If the filename ends in *.gz* it will be read as a gzip file.

xenome classify

xenome classify -P *PREFIX* {-I *FASTA-filename* | -i *FASTQ-filename* | —line-in *filename*}+ [—pairs] [-M *INT*] [—graft-name *STRING*] [—host-name *STRING*] [—output-filename-prefix *STRING*] [—dont-write-reads] [—preserve-read-order]

Classifies input reads according to a pre-computed k-mer index. The reads are written into separate files, according to their classification, and a breakdown of the number and proportion of reads in each class is printed.

If the total size of the index files is greater than available RAM, xenome will perform poorly. To overcome this, the -M, —max-memory flag may be used to specify the maximum amount of memory (in GB) that xenome may use at any time. If this amount is less than the size of the index structures, xenome will (effectively) partition the index into multiple subsets, each no larger than the specified maximum memory size, and classify the reads in multiple passes - with each pass using a different index subset. The results from each passes are combined, and the result is produced as usual. If run with the -v, —verbose flag, xenome will report the number of passes it will perform. Note that runtime will increase with the number of passes performed; the biggest increase will occur with the step from one pass to two.

OPTIONS

-P *PREFIX*, –prefix *PREFIX* The path prefix for all reference index files. The prefix may contain directory separators (e.g. '/') in order to have the index files written to another directory.

-I *FILE*, –fasta-in *FILE* Input file in FASTA format.

-i *FILE*, –fastq-in *FILE* Input file in FASTQ format.

—line-in *FILE* Input file with one read per line and no other annotation.

—pairs Treat reads from consecutive input files of the same type as pairs.

-M *INT*, –max-memory *INT* The maximum amount of memory (in GB) to use while classifying reads. If not specified, xenome will use as much memory as required to classify all reads in a single pass. When the maximum amount of memory is less than the size of the reference index files, xenome will need to perform multiple passes over the input data - increasing runtime.

—graft-name *STRING* The name of the graft reference to appear in filenames and statistics. If no explicit name is provided, the string "graft" is used.

—host-name *STRING* The name of the host reference to appear in filenames and statistics. If no explicit name is provided, the string "host" is used.

—output-filename-prefix *STRING* An optional prefix to apply to all output read filenames. The prefix is separated from the rest of the filename by an underscore ('_').

—dont-write-reads The reads will not be written to any files after classification, and none of the usual per-category output files will be created. The classification statistics will still be printed to standard out.

—preserve-read-order The relative ordering of reads within each output file will be the same as that in the input files. i.e. if read *r1* precedes *r2* in a single output file, then *r1* also precedes *r2* in the input. Note: If this flag is specified, the -T/—num-threads flag is ignored, and xenome will only operate with a single worker thread.

xenome help

xenome help

Prints a summary of all of the xenome commands.

—

FUTURE RELEASES

Bzip support will be introduced.