CHSIM Quick Reference

Typical usage:

```
chsim --input sp.fa --output data.fa --chsim_iters 10 --chsim_nperiter 100
```

Input file

Input file is a set of species sequences. Each must have a label formatted like this:

```
>sp3/ab=8.0/name=Clostridiumnexile
```

Species are assigned an integer identifier (3 in the above example) which must be 0, 1 ... etc. A relative abundance is specified by /ab=xx/, where xx is a floating-point number.

Output file

The output file contains all species and all chimeras. Chimeras are labeled like this:

```
>ch5118/ab=0.01/sp5:0-85/sp85:86-249/N=2/top=sp85:94.4%
```

Annotations include N=m, where m is the number of segments, and the parent species & coordinates for each segment. The top=spxx:yy% annotation says that the closest species is spxx and the identity with that species is yy%.

Accepts

Optionally, criteria can be specified for the number of segments and divergence (identity with the top species). If given, then chimeras meeting these criteria are written to a separate output file specified by the --outacc option. The --outacc n option says to terminate simulation after n such examples have been output. For example:

```
chsim --input sp.fa --output data.fa --chsim_iters 10 --chsim_nperiter 100 --outacc acc.fa \
  --chsim_divlo 90 --chsim_divhi 95 --chsim_minm 4 --chsim_maxm 4 --chsim_acc 10
```

Options and parameters

Command-line option	Description
--intput filename	Species (FASTA).
--output filename	Output file with species and chimeras (FASTA).
--outacc filename	Output file with chimeras meeting accept criteria (FASTA).
--chsim_acc n	Maximum number of accepted chimeras. Simulation terminates when this number have been accepted.
--chsim_minm m	Minimum number of segments for accept.

Command-line option	Description
--chsim_maxm m	Maximum number of segments for accept.
--chsim_mindiv p	Minimum identity with top parent for accept, as percentage.
--chsim_maxdiv p	Maximum identity with top parent for accept, as percentage.
--chsim_iters n	Number of PCR iterations. Default 10.
--chsim_nperiter n	Number of chimeras to create per iteration. Default 256.
--chsim_expab n	If specified, then abundances specified in the input file are ignored and an exponential abundance distribution is imposed. Species are selected in a random order and are assigned abudances n, n/2, n/4... until a value <= 2.0 is reached. Remaining species are assigned abundance 2.0.
--chsim_abfactor f	Chimera abundance = f x abundance_parent_1 x abundance_parent_2. Default 10^{-8}.
--chsim_chab f	Chimera abundance is set to f, regardless of parent abundances. By default, f=0 which means determine from parent abundances and --chsim_abfactor.
--chsim_abnoisepct p	Average noise to add to/subtract to chimera abundance. Specified as an integer percentage. Default 0.
--flank n	Do not form a crossover closer than n letters to one end of a parent sequence. Default 10.
--k n	Form crossovers at identical n-mers, weighted by the abundance of the 10mer in the pool. Default 10.
--randseed s	Random number seed. Integer in range 0 .. 2^{32} - 1. By default, the seed is set based on the clock and process id so varies in each run; this option allows reproducible runs.