Arguments common to all programs:

-h Show a help message and exit. Useful for entering commands because

it shows all the arguments. Should only be entered by itself.

Required arguments:

--infile (or --infile1, --infile2)

Input files. For everything except tag\_to\_header, should be a sorted

.bam file; tag\_to\_header takes two .fq files.

--outfile (or --outfile1, --outfile2)

Output files. For everything except tag\_to\_header, should be a .bam

file; tag\_to\_header outputs two .fq files.

Optional arguments:

--read\_out How often you want to be told what the program is doing. Defaults to 1000000. In most cases, this is higher than the number of SSCS reads.

tag\_to\_header.py

usage:

tag\_to\_header.py --infile1 INFILE1.fq --infile2 INFILE2.fq --outfile1 OUTFILE1.fq.smi

--outfile2 OUTFILE2.fq.smi --barcode\_length BLENGTH

--spacer\_length SLENGTH [--read\_out ROUT]

[--adapter ADAPTERSEQ]

Required arguments:

--barcode\_length

Length of the duplex tag sequence. [12]

--spacer\_length

Length of the spacer sequences used. [5]

Optional arguments:

--adapter Optional: Spacer sequence for filtering on the

presence of the spacer. This could be thrown off by

low quality scores.

ConsensusMaker.py

usage:

ConsensusMaker.py --infile INFILE --tagfile TAGFILE --outfile OUTFILE

--minmem MINMEM --maxmem MAXMEM --cutoff CUTOFF

--Ncutoff NCUTOFF --readlength READ\_LENGTH

--read\_type READ\_TYPE --filt FILT [--isize ISIZE]

[--read\_out ROUT] [--rep\_filt REP\_FILT]

Required arguments:

--tagfile output tagcounts file

--minmem Minimum number of reads allowed to comprise a consensus. [3]

--maxmem Maximum number of reads allowed to comprise a consensus. [1000]

--cutoff Percentage of nucleotides at a given position in a read that must be identical in order for a consensus be called at that position. [0.7]

--Ncutoff With --filt 'n', maximum fraction of Ns allowed in a consensus [1.0]

--readlength

Length of the input read that is being used. [80]

--read\_type

A string specifying which types of read to consider. Read types:

n: Neither read 1 or read 2 mapped.

m: Either read 1 or read 2 mapped, but not both.

p: Both read 1 and read 2 mapped, not a propper pair.

d: Both read 1 and read 2 mapped, propper pair.

s: Single ended reads

['dpm']

--filt A string indicating which filters should be implemented. Filters:

s: Softclipping filter.

o: Overlap filter.

n: N filter.

['os']

Optional arguments

--isize maximum distance between read pairs

--rep\_filt Remove tags with homomeric runs of nucleotides of length x. [9]

DuplexMaker.py

usage:

DuplexMaker.py --infile INFILE --outfile OUTFILE --Ncutoff NCUTOFF

--readlength READ\_LENGTH --barcode\_length BLENGTH

[--read\_out ROUT]

Required arguments:

--Ncutoff Maximum percentage of Ns allowed in a consensus [1.0]

--readlength

Length of the input read that is being used. [80]

--barcode\_length

Length of the duplex tag sequence. Should match the value in

tag\_to\_header. [12]

HammingFilt.py (for this program, --outfile is optional)

usage:

HammingFilt.py --infile INFILE [--outfile OUTFILE] [--read\_out ROUT]

Details of arguments:

--minmem and --maxmem set the range of family sizes (constrained by cigar score)

that can be used to make a consensus sequence. Examples use --minmem of 3

and --maxmem of 1000

Example 1:

Ten reads (readlength = 80) have a particular barcode. Of these ten, nine

of them have a cigar string of 80M, while one has a cigar string of

39M1I40M. Only the nine with a cigar string of 80M are sent on to be

made into a SSCS.

Example 2:

Three reads (readlength 80) have a particular barcode. Of these, two have

a cigar string of 80M, and one has a cigar string of 20M1D60M. No

SSCS results.

Example 3:

A family with over 1000 members exists. A random sample of 1000 reads

from that family is used to make a SSCS.

--cutoff sets the strictness of the consensus making.

Example (--cutoff = 0.7):

Four reads (readlength = 10) are as follows:

Read 1: ACTGATACTT

Read 2: ACTGAAACCT

Read 3: ACTGATACCT

Read 4: ACTGATACTT

The resulting SSCS is:

ACTGATACNT

--Ncutoff, with --filt n enabled, sets the maximum percentage of Ns allowed in a

SSCS.

Example (--Ncutoff = .1, --readlength = 20):

Two SSCSs are generated as follows:

SSCS 1: ACGTGANCTAGTNCTNTACC

SSCS 2: GATCTAGTNCATGACCGATA

SSCS 2 passes the n filter (10%) with 1/20 = 5% Ns, while SSCS 1 does

not with 3/20 = 15% Ns.

--readlength sets the length of the reads imputed. If this value is set incorrectly, the

program will often crash with an error message about sequence length not

matching quality score length, or will output an empty SSCS bam file.

--read\_type sets which reads are considered to have 'good' flags. Options are:

d: Paired-end reads where both reads in the pair map, and where the two are

properly paired (read 2 maps in the opposite direction and on the opposite strand

from read 1). Flags are 99, 83, 163, and 147.

p: Paired-end reads where both reads in the pair map, but the two are not properly

paired. Flags are 97, 81, 161, 145, 129, 65, 177, and 113.

m: Paired-end reads where only one read in the pair maps. Flags are 181, 117,

137, 133, 73, 89, 69, and 153.

n: Paired-end reads where neither read in the pair maps, and single end unmapped

reads. Flags are 141, 77, and 4.

s: Single end mapped reads. Flags are 0 and 16.

--filt sets which filters are used. Options are:

o: Overlap filter. Filters out any read pairs which overlap. Only works on reads

of type d (see above).

s: Softclipping filter. Filters out any reads which have been soft-clipped in

alignment. This avoids later problems with hard-clipping.

n: N filter. Filters out consensus sequences with a higher percentage of Ns than

the threshold imposed by --Ncutoff. Without this option, --Ncutoff doesn't do

anything.

--isize

If not -1, sets the maximum distance between read 1 and read 2 for the two to not

be considered unpaired. Only works if --read\_type is 'd'