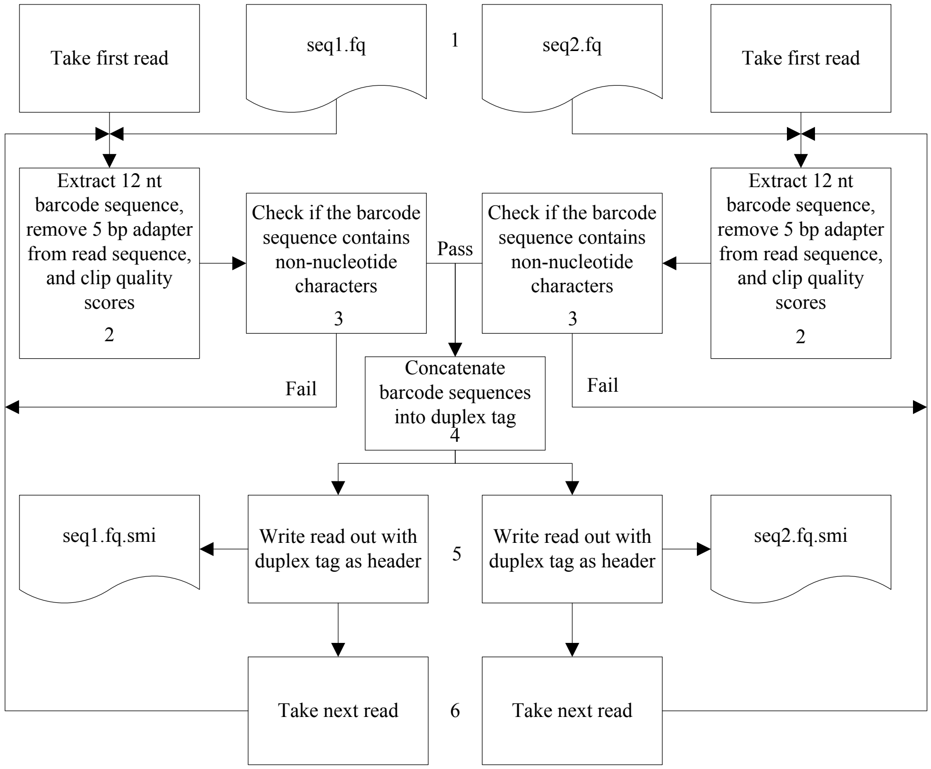
Take first read
seq1.fq
1
seq2.fq
Take first read
Extract 12 nt barcode sequence, remove 5 bp adapter from read sequence, and clip quality scores
2
Check if the barcode sequence contains non-nucleotide characters
3
Pass
Check if the barcode sequence contains non-nucleotide characters
3
Extract 12 nt barcode sequence, remove 5 bp adapter from read sequence, and clip quality scores
2
Fail
Concatenate barcode sequences into duplex tag
4
Fail
seq1.fq.smi
Write read out with duplex tag as header
5
Write read out with duplex tag as header
seq2.fq.smi
Take next read
6
Take next read

### Algorithm for tag_to_header.py

The tag_to_header.py program proceeds through the two raw FASTQ files in parallel with in-register reads for the read 1 and read 2 files being processed simultaneously. The first 17-nt of each read consists of a 12-nt random tag, followed by an invariant five base spacer sequence (Figure 1A). Prior to extracting the tag sequence, reads can be optionally filtered on the presence of a valid adapter sequence (i.e. having the sequence TGACT) (Step 2). If one or both reads have a spacer that fails this this optional filter, the entire read pair is ignored. It should be noted that using this option on data from a low quality sequencing run can lead to a significant loss of data. We rarely invoke this option in our data analysis pipeline. Next, the two 12-nt tag sequences and the invariant 5-nt spacer sequence are parsed from each of the paired reads (Step 3). The tags are checked to ensure that they are only composed of valid nucleotides (A, G, C, or T) (Step 4). If a tag fails this filter, both read pairs sharing the tag are discarded. If both tags pass the basic quality filter, the final duplex tag is formed by concatenated the 12-nt tag from read 2 (i.e. seq2.fq) to the 12-nt tag from read 1 (seq1.fq) (Step 5). For example, if the seq1.fq barcode sequence is ‘AGCT’ (the remaining bases of the tag are removed for clarity) the seq2.fq barcode sequence is CCAT, the resulting duplex tag is AGCTCCAT. The read is then written out to a new FASTQ file (seq1.fq.smi) with the duplex tag appended to the header region (Step 6).

Nature Protocols: doi:10.1038/nprot.2014.170