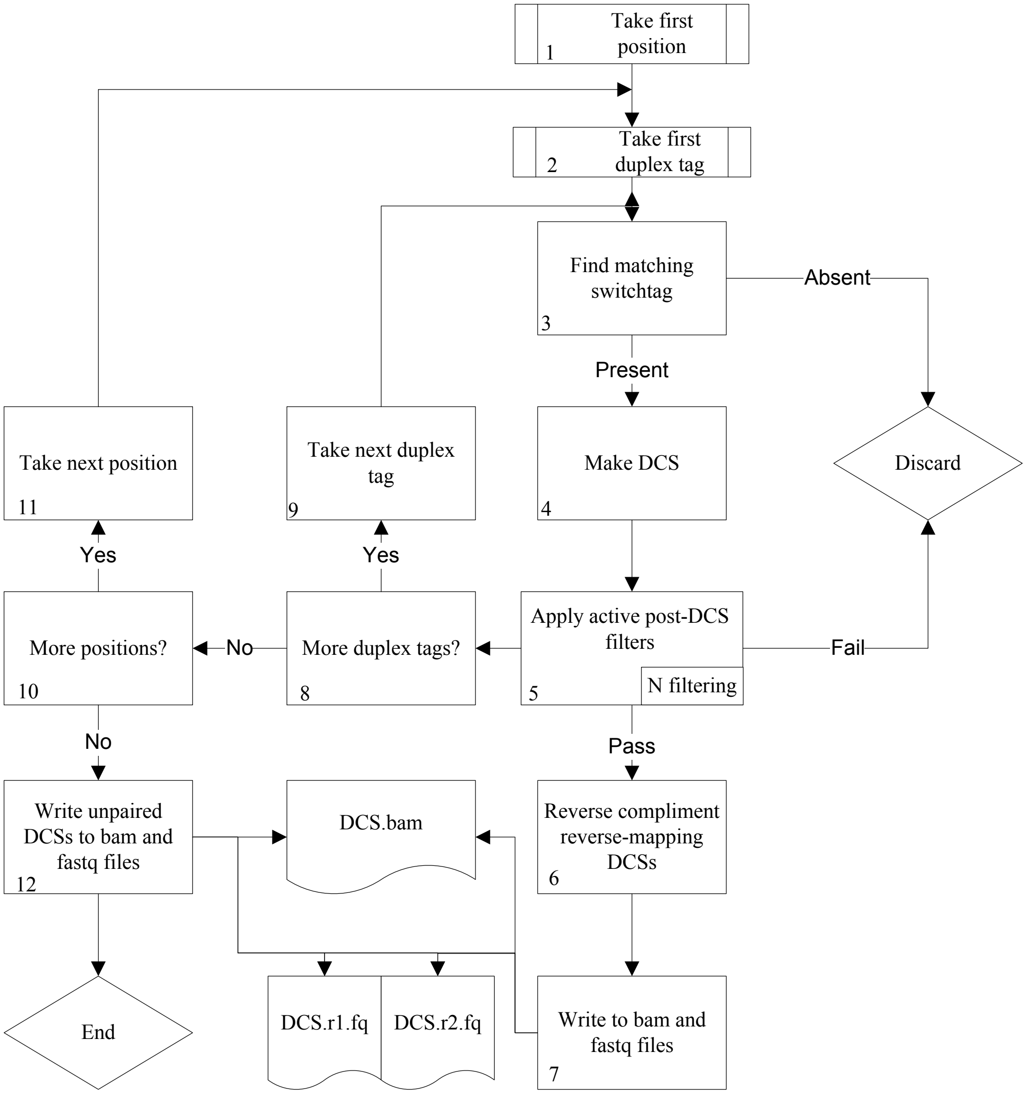

Algorithm for DuplexMaker.py

DuplexMaker.py takes the SSCS.bam file generated by ConsensusMaker.py as input. The script begins by loading all SSCS reads that map to the same genomic position (i.e. reads sharing the same POS field) into memory. All SSCS's needed for the creation of the final DCS read will map to the same position in the reference genome. The 24-nt duplex tag is parsed from the read header (i.e. QNAME field). The script then attempts to find the matching read by transposing the two 12-nt halves of tag sequence and comparing this new tag to every remaining tag associated with reads mapping to the same position. For example, if the original tag has a sequence AGTC, the transposed form of the tag would be TCAG. The TCAG tag will be compared to all the remaining tags mapping at the same genomic coordinates. If a SSCS for the transposed tag (TCAG in this case) is present at the given genomic position, the original read (tag AGTC) and the transposed tag associated read (tag TCAG) are compared to one another at each nucleotide position. If the two positions match, the given nucleotide base is used in the resultant DCS. If the positions do not match, an N is placed in the resultant DCS. After a

DCS is created, the two tags are removed from further consideration and are no longer used to form DCS reads. Following duplex making, the DCS reads may be optionally filtered based on the proportion of positions containing N's. The sequence of the final DCS read with FLAG scores of with 65, 69, 73, 97, 99, 129, 133, 137, 161, and 163 is changed to its reverse complement. Once a read's mate pair (i.e. shares the same tag sequence) has been processed, the two reads are written, in FASTQ format, to two files (one read is written to one file and the mate pair is written to the other FASTQ file), a well as to a .bam file. Waiting to write DCS reads to their respective files is done in order to keep the mate pairs in register in the FASTQ files, thus maintaining proper read pairing. If one read of a read-pair fails to form a DCS, then a dummy pair consisting of N's is written in its place. This step is done to ensure that reads remain properly paired. Once all tags at a given position have been examined, the program moves on to the next reference position and repeats the process.