1 Methylome construction

1.1 Pre-mapping Processing

Inverted duplicated reads: The data from paired-end sequencing experiment includes two types of reads - T-rich reads and A-rich reads. They are usually kept in two separate files, one labeled with "_1" and the other one labeled with "_2" respectively. T-rich reads are sometimes referred to as $5'$ reads or mate 1 and A-rich reads are sometimes referred to $3'$ reads or mate 2. If you have a fragment of double-strands DNA to be sequenced, the base pairs starting from $5'$ of the positive strand will be sequenced and go to the $5'$ read, and parallelly, the base pairs starting from $5'$ of the complementary strand will be sequenced and go to the $3'$ read, illustrated as following (a1: adaptor1, a2: adaptor2). And after the following mapping, the $5'$ read and the complementary counterpart of the $3'$ read are mapped to the reference genome along different regions (sometimes are overlapped) respectively. But, if for some reasons (we are not clear about the reasons yet), during the libraries preparation process, the $5'$ end of the second strand sticks to the $3'$ end of the first strand. As the results, the generated **read1** and **read2** will be identical, and **read2** in fact is a "fake" read, because the complementary counterpart of **read2** can't be mapped to the reference genome except for some coincidences with small chances. We call such pair of reads "inverted duplicats".

Normal case

5′-a1-TTAATCGCAT----------TTCTCTAAAT-a2-3′
3′-a2-AATTAGCGTA----------AAGAGATTTA-a1-5′

⇓ Sequencing..

read1
5′-a1-TTAATCGCAT-a2-3′ ----------
----------3′-a2-AAGAGATTTA-a1-5′
read2

Inverted duplicates generated

5′-a1-TTAATCGCAT----------TTCTCTAAAT-a2-a1-ATTTAGAGAA----------ATGCGATTAA-a2-3′

⇓ Sequencing..

read1
5′-a1-TTAATCGCAT-a2-3′ ----------
----------3′-a2-TACGCTAATT-a1-5′
read2

How badly inverted duplicates will affect the mapping results is depending on the mapping methods. With respect to our methods provided by `rmapbs`, the fake **read2** will cause losing the information of **read1** too even **read1** is useful. In order to not waste normal reads, `inverted-dups` works to detect the inverted duplicated reads and mask **read2** with "Ns" so that **read2** could be mapped to anywhere in the genome and will not encumber the mapping of **read1**. The criteria is pretty simple: the mating reads with the frequency of the identical base pairs exceeding the cut-off will be suspected as inverted duplicates. You may use `-c` to specify the cut-off (default is 0.9) - higher cut-off will give more conservative results. And you may generate 3 types of output files: use `-o` to generate the report that shows the overlapping percentage in each read; use `-s` to generate a quality summary of the reads; use `-m` to generate new reads file with inverted duplicates masked (if you want to do so, please ensure enough disk space for new reads file, which can be very large.) Below is an example of using `inverted-dups` to detect the inverted duplicates:

```
$ ./inverted-dups -c 0.95 -o Human_ESC.invd_report -s Human_ESC.invd_stat \
-m Human_ESC_new_2.fq Human_ESC_1.fq Human_ESC_2.fq
```