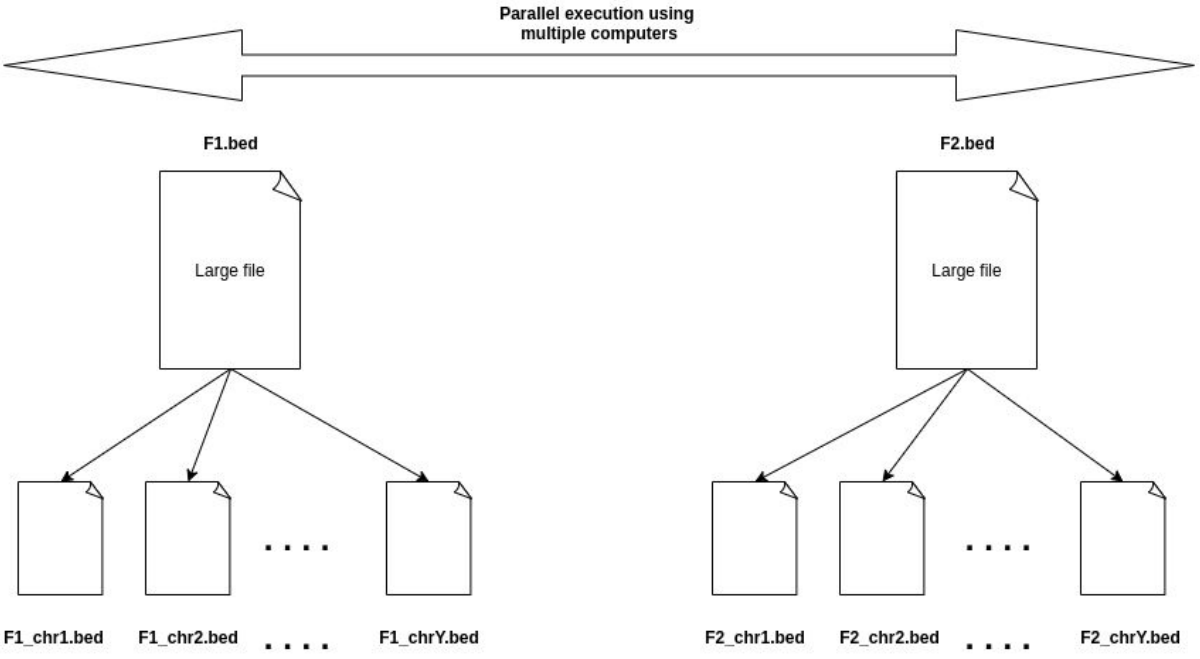
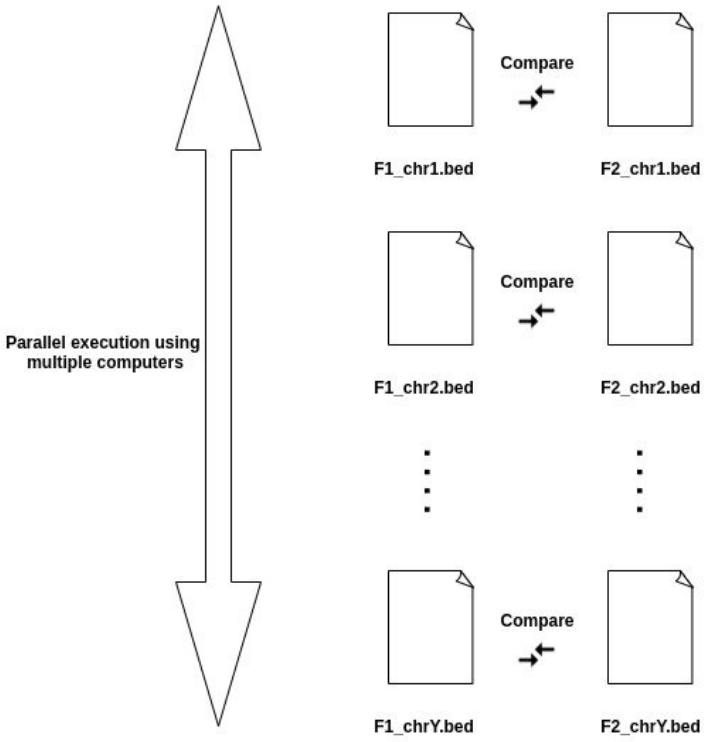
Problem 2

b) When the BED file is really large it might not be fitting into the memory of the machine in one. But if we have smaller files that have only reads of the same chromosome (pre-sorted by chromosome), those can be processed using separate, multiple computers and get smaller portions of the overall result in a parallel manner. Then pairs of files of each chromosome will be looked into (compared) rather than the pair of large files.





As shown in the above diagram, even the splitting of the large files into files sorted by chromosome can be parallelized.

Once the pairs of smaller files (pre-sorted by chromosome) are compared/processed we can merge the results and get a single output file spending only a fraction of the time taken in processing two large files. Because of this parallelization of processing operations, we can achieve a significant performance enhancement in terms of efficiency.