Genomics data analysis exercise

Provided are two files, (1) WGS_mutations.bed, is a list of somatic mutations from a cancer sample, where each line is one mutation in the sample, and (2) Replication_timing.bed, is a file containing the replication timing for all regions in the human genome. Replication timing refers to whether a region in the genome is replicated early or late when a cell undergoes replication. Each line in the file is a region in the genome and the 4th column is the replication time where the higher the value, the earlier the region is replicated (i.e. early replicating regions have the highest values and late replicating regions have the lowest values). Both files are in BED format. Using the two files, please answer the following questions and provide a brief explanation of how you reached the answer:

(1) How many mutations are in the cancer genome?
(2) What is the median size (in base pairs) of regions in the replication timing file and the total size of all regions together?
(3) List the earliest replicating region in the genome (i.e. region with the highest replication value).
(4) How many mutations are there in the 10,000 latest replicating regions (i.e. lowest replication value)?
(5) Is the mutation rate higher in early or late replicating regions?
(6) What type of DNA repair defect is this cancer likely to have based on the difference in mutation rate of the early and late replicating regions? (You will probably need to do literature search)
(7) *Bonus*: What cancer type is this sample from?

Hints: The BED format is used to store information about regions in genomes. Column 1 is the chromosome, column 2 is the start position and column 3 is the end position of a region. Additional columns provide information about each region. For questions 4 and 5, you will need to use bedtools. The bedtools tutorial (http://quinlanlab.org/tutorials/bedtools/bedtools.html) provides all information necessary to use bedtools to get the answer for 4 and 5.