R Assignment

Your R assignment will come in two parts. The first part will be due 10/04 and the second part 10/11.

Part1

The first part of the assignment deals with subsetting dataframes and is based on the Dataset_S1.txt from the Chapter 8 directory on GitHub that we also used in class. Please answer the following questions:

1. We are interested in the Bladder cancer-associated protein encoded by the BLCAP gene located between positions 37492472 and 37527931 on human chromosome 20. How many SNPs are within this gene? Are any of them located in the same window as an exon? (Note, there are several SNP columns in the dataset. You may need to check the description of Dataset_S1.txt in the original paper (Spencer et al.2006) to choose the most appropriate.
2. We want to know if there is a dependency between the recombination rate and the nucleotide composition of the sequence. Use the `plot(x ~ y)` function we used in class to investigate this relationship. What other measurements in the dataset can depend on the GC content? Make additional plots to check your intuition.
3. Does the centromer have higher nucleotide diversity than other regions in these data? Explore by looking at the summary statistics of Pi by windows that fall in the centromere and those that do not.
4. Is the average divergence between Human and Chimp the same for the two arms of the chromosome? Not sure why I'm asking this question, but who knows! How many windows have divergence > 0.3. In what arm of the chromosome are they located?

Good luck!