Homework 3

Instructions. As usual, you may work together on homework, but you must write up the solutions on your own. If you turn in an electronic copy, please provide it in pdf format. This homework is due by the beginning of class on Thursday, February 16th.

APOBEC3 is a family of proteins that mutagenize RNA in anthropoids, a clade of primates that includes humans. They are part of our innate immune system, protecting us from some viruses by mutagenizing them to death. A few viruses, like HIV, have evolved defensive proteins that abrogate and sometimes even capitalize on APEBEC3 function to their own benefit. In HIV, the *vif* protein product targets APOBEC3 member APOBEC3G for ubiquitination and subsequent protein degradation. One can experimentally abrogate *vif* function thus allowing APOBEC3G to "hypermutate" the virus genome, and sometimes hypermutation happens naturally, when Vif fails to function properly. APOBEC3G deaminates C to U during HIV-mediated reverse transcription of minus-strand cDNA. Subsequent synthesis of the positive strand results in G→A hypermutation.
For this homework, you are provided two datasets, one is a collection of G → A hypermutated HIV genomes and the other is a representative sample of normally replicated HIV genomes. The two samples have been balanced across the "subtypes" of HIV-1, which are diverse clades of HIV-1 that have diverged during the spread of the HIV-1 epidemic in the last several decades.

1. In this question, you will explore models to assess whether APOBEC3 function (presumably APOBEC3G because this member's known association with *vif*) is targeted to certain motifs.
   (a) One hypothesized target of APOBEC3G is GG. Using an iid model for HIV-1 genomes, can you formulate likelihood ratio test(s) to determine if and how GG is targeted in the hypermutated genomes?
   (b) The nucleotides in HIV (and really all functional genomes) are not iid. How might the unsupported assumption of iid nucleotides impact your conclusions about whether GG is an APOBEC3G motif? (Hint: If there are preferred dinucleotides in HIV-1, how would they impact your test statistic?)
   (c) Reformulate the tests of Part a using a Markov chain of order one. Do your conclusions change?

2. If Vif is not completely effective against APOBEC3G, then the virus may experience selective pressure to purge APOBEC3 targets from its genome. You will check for such evidence in this problem.
   (a) Another target of APOBEC3G (you could check it with your data) is TGGG. Estimate a whole-genome Markov chain of order one for the normally replicated genomes. Is there evidence that TGGG is under-represented in normal HIV-1 genomes?

(b) Is there evidence that the different subtypes share different local (order one) dependence structures? If so, is there more evidence for under-representation of TGGG in some subtypes than others?

3. You are told that the two sets of HIV-1 genomes are balanced across the subtypes, which means they have equal proportions of each subtype. The subtypes "B/CRF01_AE" and "C/CRF01_AE" indicate recombinants, which are mixtures of two subtypes. You may handle these any way you'd like: just be logical.

   (a) Why is it important that the two samples are balanced with respect to the subtypes?

   (b) Are the samples actually balanced? Support any claim with statistical evidence.

   (c) Accounting for differences between subtypes (regardless of whether you can reasonably reject the null hypothesis of no difference in Question 2, Part b), is there still evidence for a APOBEC3G GG motif?