Variants Can be Fun: A Code Sample Request

For the below code sample, feel free to email any questions and to use any informational resources (books, internet, etc.). Feel free to use libraries for obvious tasks, like file parsing. Do not use libraries specifically designed for dealing with files in the VCF format.

A VCF (variant call format) file contains a description of the genetic variants that were found for a specific individual. Following the lengthy header, each record corresponds to a variant at a specific genomic position. This highly variable format is meant to be adaptable to many different kinds of information that a user might wish to store while still adhering to the prescribed format. Here are some of the kinds of information that can be included in a VCF file:

1. Genomic Position - The location on a given chromosome, in this case, where a variant occurs.

2. SNV - A SNV is a Single Nucleotide Variation. In other words, the variant occurs at only one genomic position.

3. Indel - An Indel is either an insertion or a deletion, meaning that starting at the given genomic position several nucleotides have either been inserted or deleted.

4. Read Depth - One of the key features of the next generation sequencing technology is that each genomic location is sequenced more than once. Each "pass" over a group of positions is known as a read. Having each position contained in several reads allows us to reach a consensus among the reads as to what nucleotide was found at the given position. The read depth is simply the number of reads that contain the given genomic position.

Three VCF files are provided:
NA12878.QC_RAW_OnBait.vcf
NA12891.QC_RAW_OnBait.vcf
NA12891.QC_RAW_OnBait_truncated.vcf - a smaller VCF file for testing code quickly.

Please write a small project that answer the below questions using the provided VCF files:

1. How many SNVs and how many Indels appear in the provided VCF files?

2. What is the mean raw read depth for all SNVs in the provided VCF files?

3. How many SNVs appear on each chromosome in the provided VCF files?

4. SNV Concordance is defined as the percentage of identical SNVs between two VCF files. For example, if both file one and file two have an ‘A’ on chromosome 2 at position 1134, then they

would have a concordant SNV at that position. What is the SNV concordance between the two provided VCF Files?