In pileup format, each line represents one nucleotide of the reference sequence, so for the mouse genome, a whole-genome pileup file will be over 2.7 billion lines. This is too much data to deal with for a problem set, so we’ve given you a subset of chromosome 7 that has less than 5 thousand lines. The format of the pileup file is tab-delimited, with the following fields:

1 - Reference chromosome

2 - Reference position (1-indexed)

3 - Reference base (“forward” strand)

4 - Sequence coverage

5 - Sequence pile

6 - Sequence base qualities

The 5th field, for which this format is named, has a single character for every read that mapped to the reference genome at this position. Additionally, the pile shows whether a base is the first or last base in a read, as well as whether insertions/deletions (indels) span the reference base. Here is an example line from a pileup file, with tab-delimiters explicitly shown and bolded:

chr1**<tab>**50**<tab>**C**<tab>**34**<tab>**.$.$.$.$T,..A...,,,....+4AGGC.........,a,^F,^].^F,**<tab>**[[`p!S\_\_^\\\a`ZU`X\B\_Va^\_\_\_`^[\_a\W

Below is a table showing the valid characters in a pile, as well regular expressions one could use to extract the given characters from the pile. For indels, the below lengths are just examples. Indels can be of any length.

|  |  |  |  |
| --- | --- | --- | --- |
| **Character(s)** | **Meaning** | **Suggested Regular Expression** |  |
| . | Match to reference base on forward strand | /\./ |  |
| , | Match to reference base on reverse strand | /,/ |  |
| ACGT | Mismatch on forward strand | /[ACGT]/ |  |
| acgt | Mismatch on reverse strand | /[acgt]/ |  |
| ^x | Start of a read, where x is the mapping quality as ASCII character - 33(see below); precedes the character that indicates the nucleotide at that position from the read | /\^./ |  |
| $ | End of read; follows the character that indicates the nucleotide at that position from the read | /\$/ |  |
| +2AA | Insertion of two bases relative to the reference. ‘+’ indicates an insertion, ‘2’ means it is 2 bases in length, ‘AA’ means that two ‘A’ bases are inserted. | /\+\d+[ACGT]+/ |  |
| -3ACG | Deletion of three bases relative to the reference. Same conventions as insertions, except a ‘-‘ signifying a deletion. | /\-\d+[ACGT]+/ |  |

For the example line above, the reads map to chromosome 1, position 50. The reference base is a C and the sequencing coverage at this position is 34x. There are 4 reads that end at this position (represented by '$'), and there are 3 reads that start at this position (represented by '^x', where 'x' is an ASCII character - 33, which indicates that read's mapping quality - in this case ‘^F’, ‘^]’, and ‘^F’). There is also one 4-base insertion relative to the reference (+4AGGC). There are 2 reads that have an 'A' at this position, and one read that has a 'T'. Finally, there are 31 reads that have a 'C' at this position, the same as the reference base, so they are represented by either a dot or comma, depending on strand. Notice that only dots, commas, and 'ACGTacgt' characters not involved in mapping quality or indels contribute to overall coverage. This is because the read start (^x) and read end ($) indicators only serve as landmarks and do not contain sequence information. Indels also do not contribute to coverage since they are not present in the reference genome.

The 6th field contains the base qualities. There is one base quality character for each character contributing to coverage in the pile, so in the example above, there are 34 base quality characters. The conversion is the same as with mapping qualities, where the integer mapping quality is calculated by taking the ASCII value of the character minus 33. In Perl and Python, this can be done using the ord() function:

my $quality = ord("P") - 33; # 47 in decimal

**Challenge:**

Write a script in Perl or Python that parses the pileup file line-by-line, and will call a consensus base using two criteria: sequencing coverage and the proportion of consensus mismatches relative to the reference.  **Call a consensus base if**

(1) the sequencing coverage is greater than or equal to 5X, but less than 100X, and

(2) the percentage of consensus *mismatches* relative to the reference is greater than or equal to 80% of the total number of bases in the pile.

However, **ignore bases whose sequencing quality score is below 40**. That is, do not use them to compute either the proportion of consensus mismatches, or the total coverage.

**Methods:**

Consensus base calling from piles is partly an exercise in regular expressions. Since the piles are riddled with mapping qualities that can contain 'ACGTacgt' following a carat (^), as well as indels, you must parse the pile carefully, taking care to **only count the characters indicating a mismatch, and not indicating mapping qualities or indels**. That is, a mapping quality such as ‘^A’ that occurs in the pile should not be taken as an ‘A’ from sequencing, but rather seen as a mapping quality and ignored. Similarly, indel characters such as ‘-5ATGCC’ or ‘+2AA’ should not contribute to coverage.

Output:

The output file should have the following format:

1 - Reference chromosome

2 - Reference position (1-indexed)

3 - Calculated Consensus base