**Pileup Format**

Pileup format is first used by Tony Cox and Zemin Ning at the Sanger Institute. It desribes the base-pair information at each chromosomal position. This format facilitates SNP/indel calling and brief alignment viewing by eyes.

The pileup format has several variants. The default output by SAMtools looks like this:

seq1 272 T 24 ,.$.....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&

seq1 273 T 23 ,.....,,.,.,...,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+

seq1 274 T 23 ,.$....,,.,.,...,,,.,... 7<7;<;<<<<<<<<<=<;<;<<6

seq1 275 A 23 ,$....,,.,.,...,,,.,...^l. <+;9\*<<<<<<<<<=<<:;<<<<

seq1 276 G 22 ...T,,.,.,...,,,.,.... 33;+<<7=7<<7<&<<1;<<6<

seq1 277 T 22 ....,,.,.,.C.,,,.,..G. +7<;<<<<<<<&<=<<:;<<&<

seq1 278 G 23 ....,,.,.,...,,,.,....^k. %38\*<<;<7<<7<=<<<;<<<<<

seq1 279 C 23 A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<

where each line consists of chromosome, 1-based coordinate, reference base, the number of reads covering the site, read bases and base qualities. At the read base column, a dot stands for a match to the reference base on the forward strand, a comma for a match on the reverse strand, `ACGTN' for a mismatch on the forward strand and `acgtn' for a mismatch on the reverse strand. A pattern `\+[0-9]+[ACGTNacgtn]+' indicates there is an insertion between this reference position and the next reference position. The length of the insertion is given by the integer in the pattern, followed by the inserted sequence. Here is an example of 2bp insertions on three reads:

seq2 156 A 11 .$......+2AG.+2AG.+2AGGG <975;:<<<<<

Similarly, a pattern `-[0-9]+[ACGTNacgtn]+' represents a deletion from the reference. Here is an exmaple of a 4bp deletions from the reference, supported by two reads:

seq3 200 A 20 ,,,,,..,.-4CACC.-4CACC....,.,,.^~. ==<<<<<<<<<<<::<;2<<

Also at the read base column, a symbol `^' marks the start of a read segment which is a contiguous subsequence on the read separated by `N/S/H' CIGAR operations. The ASCII of the character following `^' minus 33 gives the mapping quality. A symbol `$' marks the end of a read segment. Start and end markers of a read are largely inspired by Phil Green's [CALF format](http://www.phrap.org/phredphrap/calf.pdf). These markers make it possible to reconstruct the read sequences from pileup.

SAMtools can optionally append mapping qualities to each line of the output. This makes the output much larger, but is necessary when a subset of sites are selected.