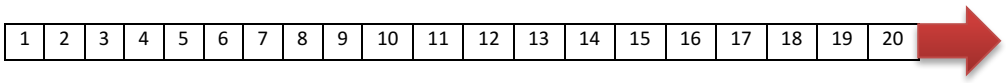
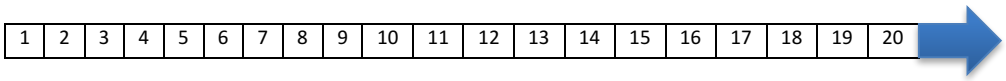
Methodology for identifying uniquely mapped reads.

1. For example, let assume there are two pair end reads: read1 and read2. Box represents bases and number represents their indexes.

   Read1:



   Read2:



2. For identifying uniquely mapped reads, it is important to first identify uniquely mapped bases.

   In this example, for simplicity read1 maps on + strand and read2 maps on – strand.

   Let's say read1 has CIGAR = 15M5S and read2 has CIGAR = 8S12M.

   For read1, mapped base indexes are = [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15]

   For read2, which maps on – strand, CIGAR is first reversed and then mapped bases are identified. In this case, reverse CIGAR would be 12M8S, and indexes of mapped bases would be = [1,2,3,4,5,6,7,8,9,10,11,12]

   This procedure is performed with every read.

3. Mapping quality (MAPQ):

   MAPQ equals to -10log10Pr[mapping position is wrong].

   Here are some known facts about value it reports.

   1. MAPQ = 0: If reads mapped to multiple places with same score, BWA puts it at a random place with MAPQ=0. **These reads are not uniquely mapped and should be filtered out**. This may also explain why we observe relatively high coverage at Chr1.

   2. MAPQ >= 30: (1 of 1000 alignments could be wrong) is considered very good score and considered as a uniquely mapped read by most bam processing software.

3. **This step is performed only for clipped reads.** For 0 < MAPQ < 30: Only those reads with this MAPQ is considered which has XA: flag (alternate alignment) or SA: flag (chimeric alignment) information. Both XA: and SA: flag has strand and CIGAR for alternate alignments.
    - Mapped bases from CIGAR information were identified as described in step2.
    - Common bases mapping to more than one position is then identified.
    - If numbers of common bases exceed 5 (not necessarily in consecutive order), read is considered **not uniquely mapped.**