S1 Text - Advantage of multi-threaded and pipelined SAM parsing

The sam2bam has high-performance SAM parsing. The performance of SAM parsing can limit the total throughput without the alignment database in sam2bam since sam2bam consists of a single pipeline. With the alignment database, the performance of SAM parsing is critical for the runtime required for data input to the alignment database (e.g., 43% of the total runtime of marking duplicate alignments for WGS data). If SAM parsing is very slow, the total runtime of sam2bam is dominated by the runtime of data input to the alignment database.

The sam2bam was designed to simultaneously parse multiple text lines to achieve high-performance SAM parsing. The sam2bam uses samtool's library code to parse a line text. If a single thread parsed a line text at a time by using the library code, it only provided 8–9% of the SAM parsing rate of sam2bam. Multi-threaded and pipelined SAM parsing is critical to the performance of sam2bam.