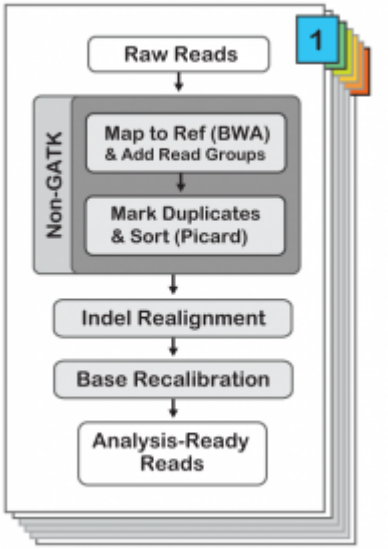
The first part of the bioinformatic process is pre-processing data of making a clean up of this, as detailed in GATK's Best Practices for Variant Discovery section BP1.0.



When you receive sequence data from your sequencing provider (whether it is an in-house service or a commercial company), the data is typically in a raw state (one or several FASTQ files) that is not immediately usable for analysis with the GATK. Even if you receive a BAM file (i.e. a file in which the reads have been aligned to a reference genome) you still need to apply some processing steps to your data to make it suitable for variant calling analysis. This section describes the pre-processing steps that are necessary in order to prepare your data for analysis, starting with FASTQ files and ending in an analysis-ready BAM file.

The steps involved are:

- Mapping and Marking Duplicates
- Local Realignment Around Indels
- Base Quality Score Recalibration (BQSR)

These steps should be performed in the order shown above. Please note that although Indel Realignment and Base Recalibration represent significant costs in terms of computational resources and runtime, we assure you that the investment will pay off with significant increases in the quality of your results.