Conversion of molecular information to computer and human interpretable information has necessitated the development of file formats for this data. The current field standard – the fastq format, traces back to a file format developed for the purpose aligning protein and nucleic acid sequence data, the fasta format. The fasta format provided two pieces of information: sequence name, and sequence. The fastq format improves on this by providing additional information on the quality of the inferred read. An overview and history of the file format can be found here: . Because most short read data generated currently is via Illumina sequencing, most tools are primarily geared towards dealing with this format, reviewed here: This information allows us to do a few useful things: assess data quality, remove low quality data, and incorporate variation in data quality into models used for inferring genotypes and alleles. This rest of this note will focus on the first two.

During the generation of sequencing data, especially via any high-throughput sequencing method, some errors will exist in some reads. These errors exist at the level of individual bases, stretches of bases, and entire sequences. They can be introduced through DNA degradation, mysterious metabolites, sequencing error via machine or chemistry problems, or library preparation. We can use the tool FastQC to assess and visualize these problems.

There are many options to visualize here, but the Q score across reads, duplication level, and adapter content tend to be most important. As a general rule for each of these: You want Q above 30 as much as possible, Q above 20 is acceptable, and once you’re below Q20 your sequence is “bad” - your reads are heading into the “1 in 10 basecalls are wrong” zone, which is a lot of error. For duplication level: you want duplication levels consistent with your expected coverage. If you anticipate 10X coverage and see duplication levels at that rate, great! But also don’t worry too much – the genome is full of repetitive stuff, and some of it is going to be “duplicated” at really high rates. Last – adapter content. This is where your