The following analysis is done based on the graphical outputs of FASTQC package (outputs are in the attached Zip file, named FASTQC).

Overall GC content of the reads

In Bliss 52%, in Restq 39%.

This difference in the GC content of the reads is probably due to the primer design and targeted sequencing, in Restq.

From Per base sequence quality

Base calling in Restq is done with higher qualities, compared to the Bliss. However qualities in both files are acceptable.

Per sequence quality scores

Due to the higher base calling qualities in Restq, mean sequence qualities in Restq are also higher than Bliss. (Distribution of mean sequence qualities is shifted to the right and takes larger phred scores in Restq)

Per base sequence content

In Bliss, sequence content across all bases are approximately the same (those little variations at the start positions are negligible)

In Restq, we observe a high variations in the sequence contents of the first nucleotides (6 nts), compared to the next positions of the read. My gesture is that this occurs due to the library preparation protocol and your specially designed primers to target some special DNA sequences which are complementary to the primer. These regions are consequently cleaved using the restriction enzyme. Therefore, this is a technical bias and we cannot correct it by trimming the reads.

Sequence Duplication Levels

In Bliss, the sequence duplication level is quite standard, based on my previous experience in analyzing such genomic data. Indeed, some repetitive read sequences may appear due to PCR or the genome-wide duplications.

In the case of Restq, a high level of sequence duplication is observed (percent of reads remaining if de-duplicated is only 25.49% which is quite low). My inference is that as we have targeted some special genomic regions – due to the library preparation protocol - these regions are probably sequenced with a high coverage and also amplified in PCR.

However, I'm not saying that only 25% of the studied chromosome is sequenced, because it depends on the frequency of the targeted genomic regions i.e. regions which are complementary to the designed primers, along the studied chromosome. This can be discussed only after mapping reads to the reference genome.

Overrepresented sequences

In Bliss, we observed no over represented sequence. This is quit usual in a standard sequencing protocol.

In Restq, we have observed 7 overrepresented sequences, where 5 sequences are RNA PCR Primer. No source is identified for the other sequences and these could be biologically meaningful e.g. a motif, and it may worth to find their source, in a consequent analysis.

Kmer Content

In Bliss and Restq, we have observed a number of k-mers at different positions along the reads. The positions of k-mers in Bliss differ from those positions in Restq, except for a k-mer that is observed at position 17.

In the next step, I will map the short reads to the reference genome.

Since I'm not given the name of the organism from which these reads are coming, I have to identify its name i.e. either human or mice, based on the yesterday discussion.

Abnormalities and variants between these two fastq files are analyzed only after mapping short reads to the reference genome.