The plan of RNA variant calling benchmarking project,

1. Download the RNAseq data with different depth and/or platforms for at least three samples from the GIAB. (-after looking I can only find the sample that were used in 2 other articles NIST ID: HG001 and it has high confidence VCF file from genomic data available
2. If there is no VCF files available for the same samples, we will download the genomic data for comparison.
3. Map the samples to the reference genome using STAR.
4. 2-pass STAR, a new index is then created using splice junction information contained in the file SJ.out.tab from the first pass:
5. using the SAM output we will mark the duplicates using Picard tolls.
6. Splitting the reads into exon segments using SplitNCigarReads
7. Indel realignment
8. Base recalibration
9. Variant calling using GATK to get the BAM output.
10. Using BEDTools to compare the BAMout to the DNA vcf\_output.
11. Calculating the false positive and true positives and false and true negatives to determine the specificity and the sensitivity of using the RNAseq data to call for variants