**Calling variants in RNAseq**

The workflow is based on GATK Best Practices workflow for SNP and indel calling on RNAseq data( <https://software.broadinstitute.org/gatk/guide/article?id=3891>)

1. STAR 2-pass alignment

The index generation and alignment in 1st pass with STAR are described in RNAseq QC. In 2nd pass a new index is then created using splice junction information contained in the file SJ.out.tab from the first pass with sjdbOverhang set to 75 and resulting index is then used to produce the final alignments.

2nd pass reference index: /groups/sorger/cchris/ref\_genome\_2pass

2nd pass sam files: /n/scratch2/cc400/bam\_2pass

2. Add read groups, sort, mark duplicates, and create index

<https://broadinstitute.github.io/picard/>

Sam files generated in the first step are converted to bam format and then put through the usual Picard processing steps: adding read group information, sorting, marking duplicates and indexing.

2nd pass bam files: /groups/sorger/cchris/bam\_files

2nd pass bam files after marking duplicates: /groups/sorger/cchris/bam\_files\_markdup

3. Split'N'Trim and reassign mapping qualities

<https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_rnaseq_SplitNCigarReads.php>

Splits reads into exon segments (getting rid of Ns but maintaining grouping information) and hard-clip any sequences overhanging into the intronic regions with SplitNCigarReads in GATK.

Output bam files: /groups/sorger/cchris/bam\_files\_dupSplit

4. Indel Realignment

<https://software.broadinstitute.org/gatk/events/slides/1212/GATKwh0-BP-2-Realignment.pdf>

Output bam files: /groups/sorger/cchris/bam\_files\_dupSplit\_realign

Reference vcf file: <ftp://ftp.ensembl.org/pub/release-85/variation/vcf/homo_sapiens/Homo_sapiens_somatic.vcf.gz>

5. Base Recalibration

<https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_bqsr_BaseRecalibrator.php>

Output bam files: /groups/sorger/cchris/bam\_files\_dupSplit\_realign\_BQSR

6. Variant calling with HaplotypeCaller

<https://software.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php>

Output VCF files: /groups/sorger/cchris/vcf\_files

7. Variant filtering

<https://software.broadinstitute.org/gatk/guide/tooldocs/org_broadinstitute_gatk_tools_walkers_filters_VariantFiltration.php>

filtering based on Fisher Strand values (FS > 30.0) and Qual By Depth values (QD < 2.0) and filter clusters of at least 3 SNPs that are within a window of 35 bases between them.

Output VCF files: /groups/sorger/cchris/vcf\_files\_filtered

8. Variation annotation with Ensembl Variant Effect Predictor

<http://www.ensembl.org/info/docs/tools/vep/script/index.html>

Output vef files: /groups/sorger/cchris/vef\_files

All scripts used is stored at /groups/sorger/cchris/src