This repository contains the code to analyse scSNV-seq data.

The following scripts and notebooks analyse the DNA modality:

1. MB_all_barcodes_count.R: extracts gRNA, puroR and iBAR barcodes, using read groups and sequences extracted from reads mapped to the gRNA and iBAR amplicons by the MissionBio tapestri pipeline
2. MB_all_barcodes.Rmd: assignes gRNAs, puroR and iBAR barcodes to cells from counts obtained by MB_all_barcodes_count.R
3. run_gatk.sh: genotype calling for each cell barcode
4. summarise_gatk_output.R: summarise vcf output files obtained for each cell barcode from run_gatk.sh to a list in R
5. MB_genotype_gRNA_processing_per_cell.Rmd: processing of genotype data per cell
6. GT_per_barcode.Rmd: This script processes per-cell genotype calls using iBAR and puroR-barcode information. The output is summarised genotype information per barcode. Further, consequence and impact of genotypes are predicted using VEP (W. McLaren et al. *The ensembl variant effect predictor*. Genome Biol., 17 (2016)) and actual genotype information. (using input file gRNA_info_coordinates.csv)
7. add_amplicons_to_metadata.Rmd: add to the barcode meta data obtained from GT_per_barcode.Rmd the name of the Mission Bio amplicon that covers the gRNA associated with the barcode
8. core_functions_MissionBio.R: functions for analysis of genotype at the cell and barcode level used in the notebooks

The following scripts and notebooks analyse the RNA modality and combine the two modalities:

large_non_genotyped_base_editing_screen: QC, gRNA calling and and downstream analysis of non-genotyped data set; QC and gRNA calling were performed in the same way for the genotyped data set

genotyped_screen: barcode calling, analysis, integration with non-genotyped data, integration of DNA and RNA modalities

The folder gRNA_barcode_number_plots contains notebooks to plot distributions of barcodes and gRNAs for both of the screens.