Workflow:

1. **s2.sh Gunzip** to decompress fq.gz files.

**fastp** to trim the adapters and the low-quality reads from the sample genome files.

1. **samtools faidx** to create. fa index of the reference genome WS292 (command line)
2. **picard createsequencedictionary** for creating dictionary for the reference genome (.dict file) (command line)
3. **s3.sh bowtie2** for aligning the reads with the reference genome.

**samtools view** to convert sam(Sequence Alignment Map) files into bam(Binary Alignment Map).

**samtools sort** to sort bam files in the order of the chromosome and positions.

**bamtools filter -isProperPair true** to make sure all the read pairs are aligned and in a correct order.

1. **s5.sh picard AddOrReplaceReadGroups**  add group information to the files.
2. **s6.sh MarkDuplicates** command in gatk to remove duplicate reads.
3. **sq.sh fastqc and multiqc** to check the read quality.
4. **s7.sh HaplotypeCaller** in gatk to call the variants.
5. **s8.sh GenomicsDBImport** in gatk to consolidate all the vcf files in a database.
6. **s9.sh** GenotypeGVCFs in gatk to call variants jointly and generate vcf files.
7. **s10.sh SelectVariants** in gatk to extract SNPs and INDELs.
8. **s11.sh VariantFiltration** in gatk to apply 3X and 10X coverage filtering.

**SelectVariants** to fileter no call variants.

1. **s12.sh bcftools view max allele** to convert into bi allelic vcf files
2. **mutation.sh bcftools view count** to filter out the rows other than the mutation rows.
3. **pair.sh** List all the subline pairs in a text file named “pair.txt”.
4. **combine2.sh** and **cov.sh**  Calculate the number of base pairs covered in both sublines using the text file “pair.txt”.. Record the coverage data in two text files named “common\_3x\_summary.txt” and “common\_10x\_summary.txt”
5. **vcf\_to\_csv.R** to convert all the files into csv files. (package used: **dplyr, vcfR**)
6. **tandem repeat finder (trf) 2 5 5 75 10 0 2 -d -m -h** to identify repeat regions of the reference genome WS292. (command line)
7. **mono75.R,** to extract the mono and dinucleotide repeat regions in the reference region and convert to .csv files. (Total mono repeats, 6786116) (package used: **dplyr**) (use dat file from previous step)
8. **mono.R** split all the files into two groups: mono and non mono (package used: **dplyr, vcfR**).
9. **Order.R,** to split all the files into two groups again, o1 and o2) (package used: **dplyr**)
10. **indel.R and snp.R,** to distribute the mutations into indel and snp spectra. Calculated total for each sample (package used: **dplyr**).
11. **3Xcover.sh, 10Xcover.sh, G3xcover.sh, G10xcover.sh** using these four-custom made bash script to calculate the percentage of the total length covered by the MA genomes.
12. **cat.R** to add coverage information and delete ancestors.
13. **merge\_new.R** Combines the snp and indel files.
14. **merge\_EG.R** Combines the E and G sublines in a single file.
15. **Indel\_spectra.R** Plots insertion and deletion spectra seperately.
16. **Irate\_spectra.R** Plots SNV and Indel spectra seperately.
17. **Dummy\_reference.R** . Randomly inserted SNV and indels to generate a pseudo reference genome. (package used: **GenomicRanges,IRanges,Biostrings**)
18. **GATK**: **FastaAlternateReferenceMaker.** Pseudo-Reference genome created. (command line).
19. **fn.**R To calculate False negatives and failure to recall.
20. **sim\_plain.**R To calculate the point estimates of the rates per generation/days with and without FP correction.
21. **sim\_from\_o1.**R Testing the hypothesis of uniform mean.
22. **sd\_seperate\_o1o2.**R Testing the hypothesis of uniform variance in o1 and o2.
23. **mismatch.sh** . Diagnose the mismatched o1 lines.
24. **snpeff** . Annotate the mutations, and their putative impacts.
25. **Annotation.R and add\_line.R** . Process the annotated vcf files and convert to csv file.