This document mainly talks about Plexseq reads quality checking, processing VCF file from Mayo, and analysis of genetic data from both Mayo and Plexseq.

Alert: All the file path in the scripts (shell, python and others) are subject to my local path and should be adapted to the local path of whoever's using these scripts.

Plexseq Results Quality Checking
samtools-1.3 is used for all following process

Use *diff.py* to figure out sample ids in plexseq_diagnosis.xlsx (information saved in sample.txt) that corresponding to diff-SNPs in *quality_checking_12_13_2017 (information saved in difflist.txt),* saved in ***diffid.txt***

Use *filter.py* to filter out 102 samples out of 9024 samples we have that contain diff_SNPs, saved in copy directory.

Use *bwa.sh* to align 102 samples to hg38.fa, save all the sam files in *sam* folder

Use *sam2bam.sh* to convert sam files to bam files

Use *sort.sh* to sort, use index.sh to index all bam files, save in *sorted* folder

Use *brc.sh* to count the reads at interests (region.txt includes the snps information obtained from ncbi snps database: https://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?searchType=adhoc_search&type=rs&rs=rs4666451) All the resulting txt files are saved in s*orted* folder

Use *interpret.py* to interpret the results, saved in *213snps.xlsx*

Processing (Mayo)

1. Use */Volumes/CORSAIR/Mayo/cleaning_script/vcf2allele.py* to parse VCF into reads file

2. Use */Volumes/CORSAIR/Mayo/comparison/compare.*py to compare the reads of overlap SNPs from mayo and Plexseq

3. Use */Volumes/CORSAIR/Mayo/hwe/HWE_recessive.py* or */Volumes/CORSAIR/Mayo/hwe/HWE.py* to calculate Hardy-Weinberg equilibrium (can choose to compute from dominant or recessive)

4. Use *`/Volumes/CORSAIR/Mayo/mayo_data_snp_flip/vcf2allele_fliped.py`* to flip SNPs that are on reverse strands from Plexseq, since Mayo always report SNPs in the forward strands and we want to be consistent about it.

Analysis:

5. /Volumes/CORSAIR/TFA (Complex folder):
   a. 33 loci folder: info from 33 loci paper

   b. GFL folder: group fused lasso

   c. Old binarize method folder: old binarization method input and results 0.5 for heterozygous

   d. TWAS folder: TWAS paper information

   e. Status prediction folder: inputs and results from snp_algorithim.py and survey_data_algorithm.py

   f. Binarize_mayo.py including functions:

      1 zscore calculation
      2 binarize data
      3 make ped file (for plink)
      4 make map file(for plink)
      5 ped file transformation (get rid of "/")

   g. Overlap.py including functions:
      1 mayo/plexseq overlap ID check
      2 overlap of our snps with 33 and cis-eQTL
      3 …multiple overlap checkings upon requested

   h. Vital_match.py including functions:
      1 mayo-plexseq overlap checking
      2 vital status match

   i. Plink_mac folder: utilize Plink to calculate LD.
      Plink –bed bed.txt –ped ped.txt –out bc (see plink.log for command and processing details)
      Output *bc.fam, bc.bed and bc.bim*

      Output LD calculated by Plink: *bc.ld*
      See Bc.log for command detail

      Calculated Z-score for each SNPs
      Download gene annotation information from PAINTOR and SNPnexus.

j. Use of PAINTOR to calculate posterior probability

```
 --sam1=finalmat.csv --sam2=prob
```

PAINTOR’s annotation library has no useful information.

Encode_2321 and regbuild_2321 from SNPnexus are useful.

```
Annotaiont command: python AnnotateLocus.py --
input=Annotation_mammary --locus=locus1.txt --
out=locus1.annotation --chr=CHR --pos=POS
```

```
./PAINTOR -input input -in run -out run -Zhead Zscore -
enumerate 10 -annotations
Promoter,Enhancer,CTCF_Binding_Site,Promoter_Flanking_Region,O
pen_chromatin,TF_binding_site,H3K4me2,H3K4me3,H3K9ac,H3K27ac,H
3K36me3,DNase1,H3K27me3 -LDname ld
```