Notes why and how to add the ABL L298L data:

3.26.2023 Haider Inam

Summary: Our ABL SM Library was made on REFSEQ ABL background for almost everything. But for 16 residues from 290-305 of the kinase, the library has a synonymous SNP on position 298. This SNP causes our in-house variant caller to think that there are two mutants on each read for these positions (one L298L and the other one being Twist's SNV/MNV). These are automatically discarded by our variant caller. I could have fixed the variant caller to account for this discrepancy, but the easier approach is to just do two separate variant calling pipelines for the L298L and the non-L298L regions.

* No L298L means a reference was used that had the normal CTG at residue 298 instead of a CTA.
* L298L means a CTA at residue 298 instead of the normal CTG.

To combat the L298L problem:

* We align our reads on a non-L298L reference and then an L298L reference. This is why you will see an L298L folder in each directory on the AWS instance (for ABL region 1 samples).
  + The two types of datasets come from the same consensus family calls, i.e. the dunovo consensus family files.
  + The dunovo consensus calls were aligned first to a reference without L298L and then to a reference with L298L
* For the reads aligned to the L298L reference, we extract the mutants called between ABL residues 290-305 and merge them with the non-L298L-called variants
  + The add\_l298l.R function (in the code directory) takes in a non-L298L variant caller outputs and L298L variant caller outputs, and merges them.
* To automatically add L298L for all the samples in a directory, you can use the following R code (this code is also present in the ABL\_SM\_CRISPR\_Cut\_Analysis.Rmd file):

source("code/variantcaller/add\_l298l.R")

for(i in c(1:18)){

sample=paste("sample",i,sep = "")

# sample="sample1"

input\_df\_nol298l=read.csv(paste("data/Consensus\_Data/novogene\_lane18/",sample,"/nol298l/duplex/variant\_caller\_outputs/variants\_unique\_ann.csv",sep=""))

input\_df\_l298l=read.csv(paste("data/Consensus\_Data/novogene\_lane18/",sample,"/l298l/duplex/variant\_caller\_outputs/variants\_unique\_ann.csv",sep=""))

output\_df=add\_l298l(input\_df\_nol298l,input\_df\_l298l)

write.csv(output\_df,

paste("data/Consensus\_Data/novogene\_lane18/",sample,"/duplex/variant\_caller\_outputs/variants\_unique\_ann.csv",sep = ""))

}