**Steps in the tsinfer analysis**

1) **Merging vcf files** by samples

* Start with .vcf.gz
* Extract sampleIDs for each file for each chromosome
* Compare files and secure there are no overlaps
* Filter vcfs to remove overlapping sampleIDs   
  (I guess this could become an if statement)
* Compress and index
* Merge files
* what are the checks here? The main point here is not having duplicated samples because they will become a problem further down the line

2) **Phasing**

* If files are compressed and indexed, it is just calling SHAPEIT4
* what are the checks here? Ensure there is only one file/chromosome   
  Ensure that files are per chromosome

3) **Prepare Files for tsinfer**

* we start with compressed and indexed vcf files
* whether we start with one or multiple vcfs –if there is only vcf → we split it into chromosome vcfs
* we decompress the files
* next, we put in the ancestral information
  + the ancestral file need to be standardized
  + for the sites that don’t have an ancestral allele known, we leave a black space in the VCF (make sure this script runs fine and do what expected!)
* - we compress and index the files

4) **Infer the trees**

* prepare the samples file
* what steps do each of us take
* standardize the meta file (to read in the populations)
* infer by chromosome (this is probably true for everyone)
* what about tsdate?? Are we putting here as well or having a separate one?   
  One issue I constantly have is (if sample to dated.tree are all on the same scrips) is when it crash because of wrong time/memory set ups and I have to run the full script! But I guess that is the point of snakemake so we are probably good having it all together??

5) **Do statistics on the trees**

* GNN
* Fst
* Tajima’s D
* IBD
* …