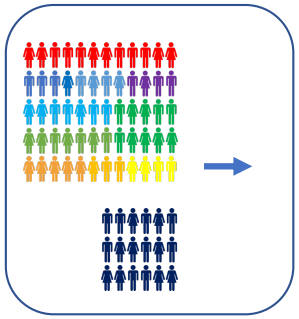
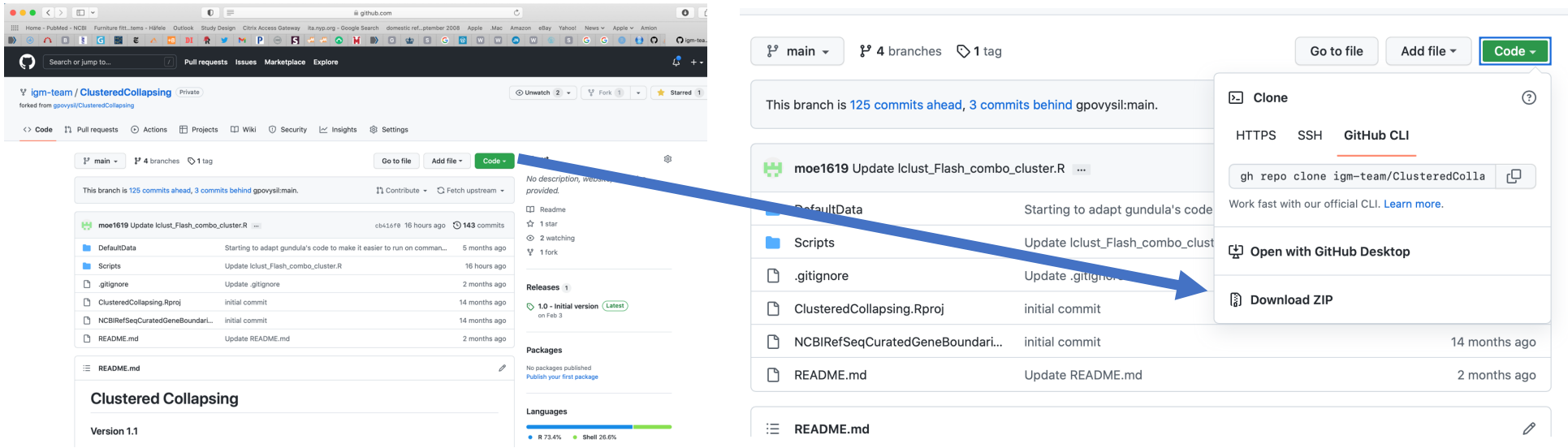
Collapsing User Guide

v20220713

Setup



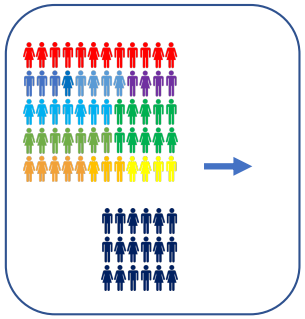
https://github.com/igm-team/ClusteredCollapsing

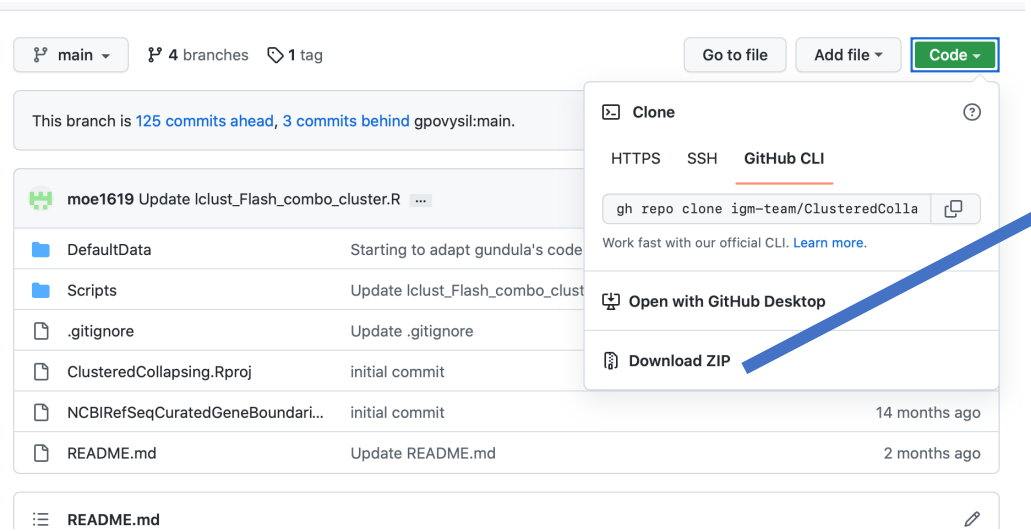


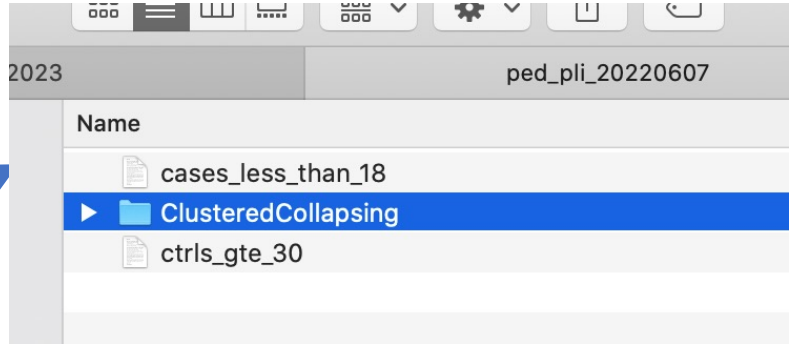
- Povysil et al. Nat Rev Genet. 2019

Background: Rare-Variant Collapsing Workflow

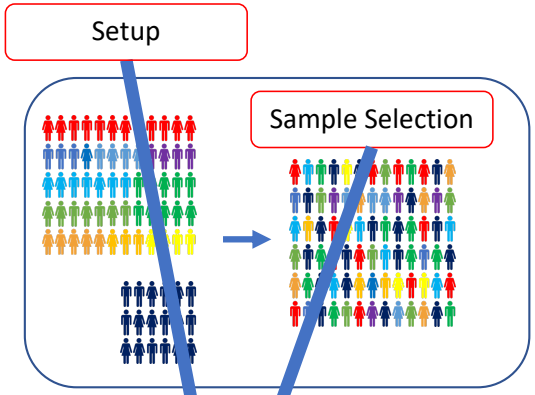
Setup

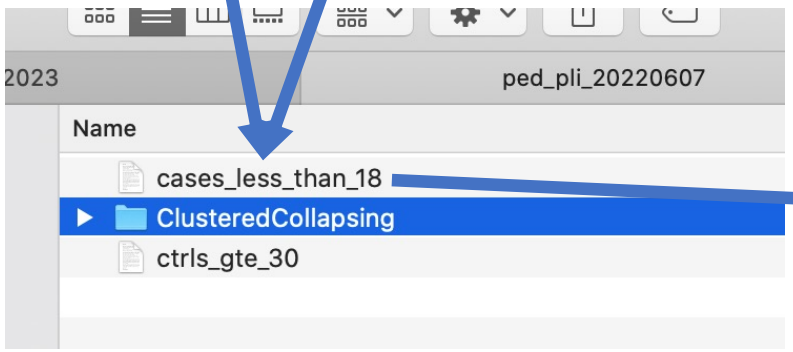






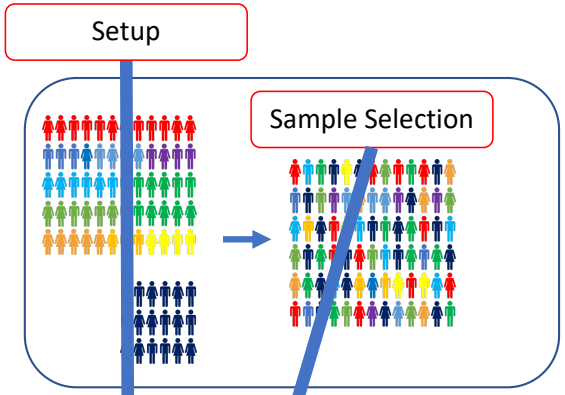
Background: Rare-Variant Collapsing Workflow

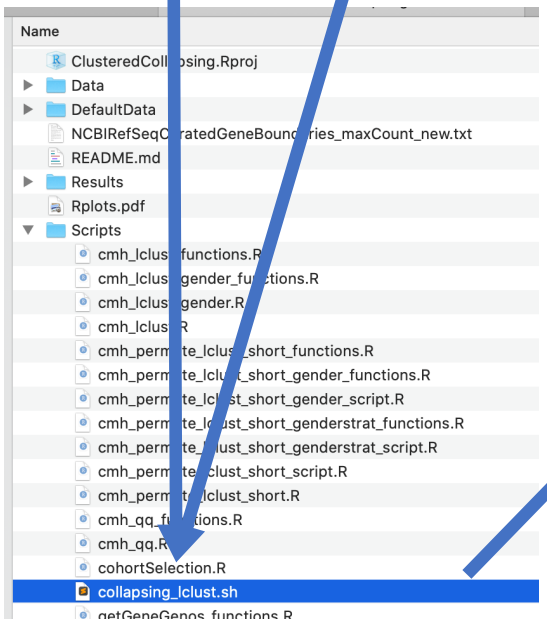




- Povysil et al. Nat Rev Genet. 2019

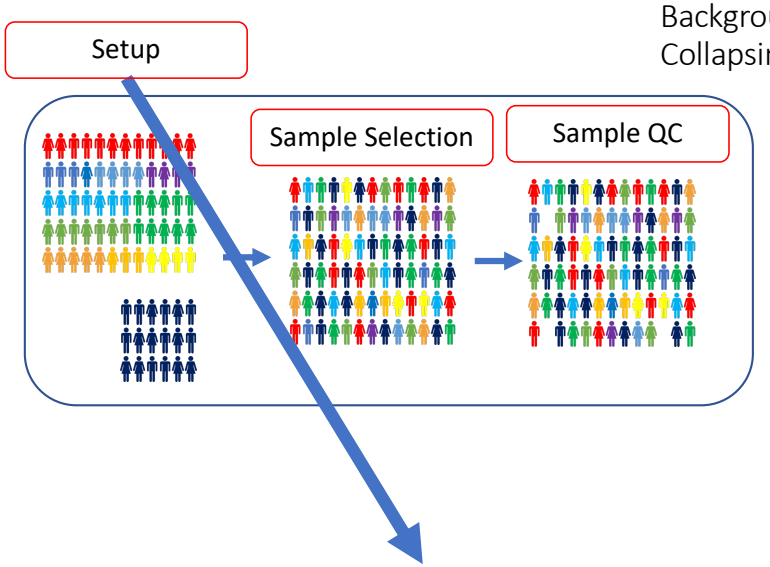
Background: Rare-Variant Collapsing Workflow



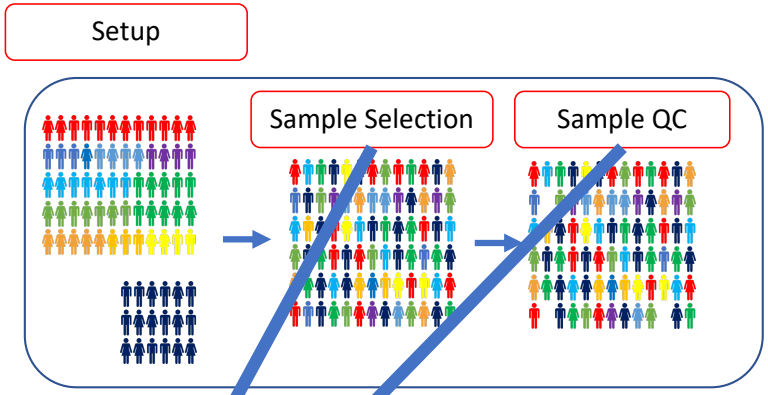


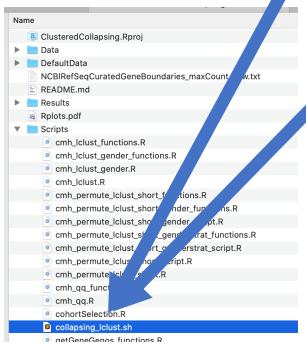

```
#\!/bin/bash
# 2021-05-20
# Run Collapsing Analysis


################ VARIABLES USER MUST MODIFY ###################
PROJECT="/nfs/projects/ped_pli/ped_pli_20220607" # path to your project
dragen_csv_path="/PATH/20220211_indragendb.csv" #path to specific dragen output but don't need to use. can
case_group="age_analysis" #for naming purposes.
case_path="$PROJECT/cases_less_than_18.txt" #path to cases. necessary
################  ###################

################ DO NOT MODIFY ###################
atav="/nfs/goldstein/software/sh/atav.sh --email" # remove --email if you don't want to get them
source /usr/local/igm/non-atav-tools/R-4.1.0_with_gcc_10-x86_64/R-4.1.0_with_gcc_10-x86_64-ENV.sh #under co
Rscript_var="/usr/local/igm/non-atav-tools/R-4.1.0_with_gcc_10-x86_64/bin/Rscript" #under construction. Thi
cohortSelection_var=$PROJECT/ClusteredCollapsing/Scripts/cohortSelection.R
lclust_flash_var=$PROJECT/ClusteredCollapsing/Scripts/lclust_Flash.R
lclust_flash_combo_cluster_var=$PROJECT/ClusteredCollapsing/Scripts/lclust_Flash_combo_cluster.R
cmh_lclust_var=$PROJECT/ClusteredCollapsing/Scripts/cmh_lclust.R
cmh_qq_var=$PROJECT/ClusteredCollapsing/Scripts/cmh_qq.R
sample_counts_table=$PROJECT/ClusteredCollapsing/Scripts/table_sample_counts.R
umap_clusters_fig=$PROJECT/ClusteredCollapsing/Scripts/plot_umap_table.R
```

Background: Rare-Variant Collapsing Workflow




```
jm4279@dev2:~$ cd /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/
jm4279@dev2:/nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing$
```



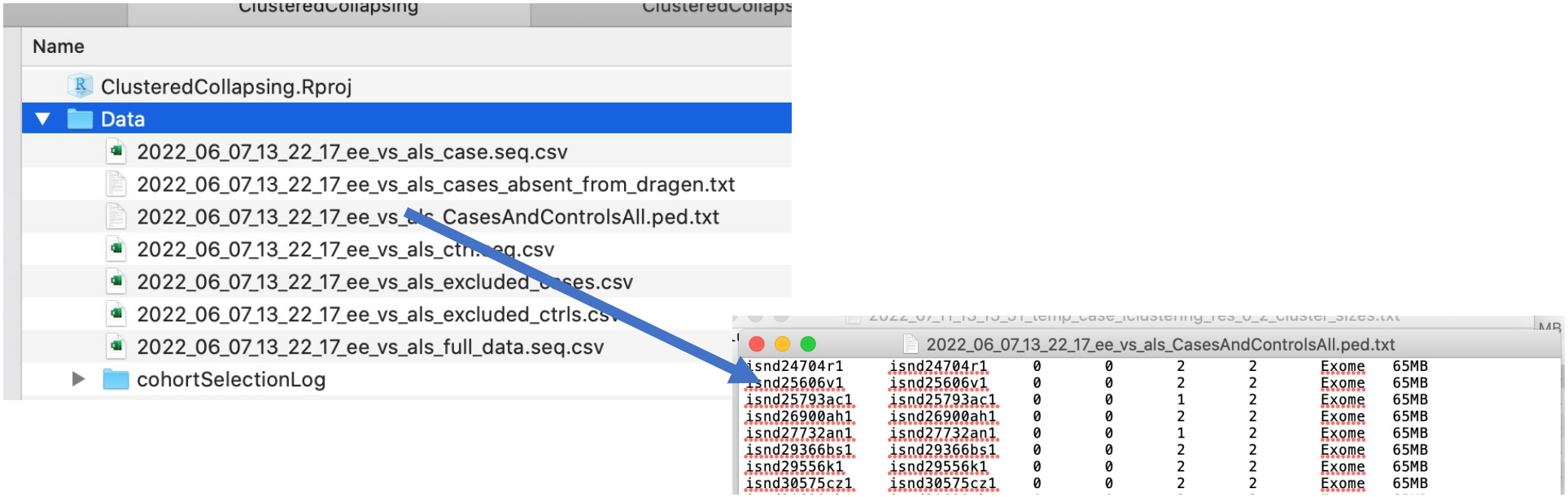


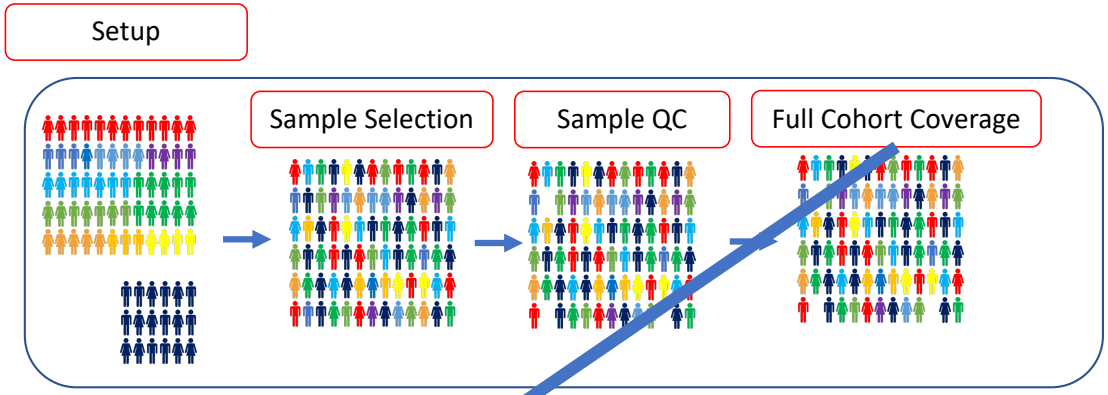

```
[jm4279@dev2:~$ cd /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/
[jm4279@dev2:/nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing$ ./Scripts/collapsing_lclust.sh cohortSelection
```

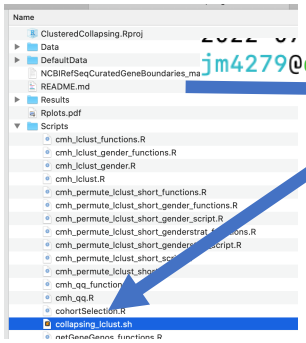
```
################ run cohortSelection.R to get input file ##################
if [[ $1 = "cohortSelection" ]]
then
  echo "Creating cohort ..."

  #path to Rscript
  #path to cohortSelection R file
  #path to dragenDB downloaded from sequence
  DetailedPhenotype_excl_regex_var="epilep|seizure"
  ctrl_path="$PROJECT/ctrls_gte_30.txt"

  # broad phenotypes to include in controls

  # Learn about the input options for cohortSelection.R
  $Rscript_var $cohortSelection_var --help

  # run cohort selection
  $Rscript_var $cohortSelection_var --case_list_path $case_path --ctrl_list_path $ctrl_path --case_group $case_group #--DetailedPhenotype_excl_regex
  $DetailedPhenotype_excl_regex_var # --sample_path_w_filename $dragen_csv_path

fi
```

Sample QC

```
[jm4279@dev2:/nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing$ ./Scripts/cohortSelection.R --help
./Scripts/cohortSelection.R: line 22: Usage:
  cohortSelection.R --case_list_path=<case_list_path> [--ctrl_list_path=<ctrl_list_path>] [--sample_path_w_filename=<sample_path_w_filena
me>] [--case_group=<case_group>] [--exclude_list=<exclude_list>] [--BroadPhenotype_include=<BroadPhenotype_include>] [--acceptable_exome_
kits=<acceptable_exome_kits>] [--DetailedPhenotype_excl_regex=<DetailedPhenotype_excl_regex>] [CCDSBasesCov10X_var=<CCDSBasesCov10X_var>]
 [DBSNPOverlapSNVs_var=<DBSNPOverlapSNVs_var>] [DBSNPOverlapIndels_var=<DBSNPOverlapIndels_var>]

  Options:
  -h --help
  --case_list_path=<case_list_path> path to text file without header with list of cases to be included
  --ctrl_list_path=<ctrl_list_path> path to text file without header with list of ctrls to be included [default: NA]
  --sample_path_w_filename=<sample_path_w_filename> path to csv file downloaded from https://sequence.igm.cumc.columbia.edu/ [default: /n
fs/goldstein/software/atav_home/data/sample/igm_lims_sample_master_list.tsv]
  --case_group=<case_group> label for analysis. example might "picu" or "IPF" [default: temp_case]
  --exclude_list=<exclude_list> specific samples to exclude [default: NA]
  --BroadPhenotype_include=<BroadPhenotype_include> broad phenotypes to include in controls. Must uses _ instead of space. [default: heal
thy_family_member,control,control_mild_neuropsychiatric_disease]
  --acceptable_exome_kits=<acceptable_exome_kits> acceptable exome capture kits [default: NexteraRapidCapture,65MB,Roche,RocheV2,IDTERPv1
,IDTERPv1Plus,IDTxGEN,IDTERPv1mtDNA,AgilentCRE,AgilentV4,Agilentv5,AgilentV5,AgilentV5UTR,AgilentV6,MedExome,N/A,Genome_v1]
  --DetailedPhenotype_excl_regex=<DetailedPhenotype_excl_regex> regexp string to exclude things in detailed phenotype [default: abcdefg]
  --CCDSBasesCov10X_var=<CCDSBasesCov10X_var> [default: 90] samples must have at least this % of the CCDS region covered by 10x
  --DBSNPOverlapSNVs_var=<DBSNPOverlapSNVs_var> [default: 0.85] must have at least this % overlap of SNVs with DBSNP
  --DBSNPOverlapIndels_var=<DBSNPOverlapIndels_var> [default: 0.80] must have at least this % overlap of indels with DBSNP
```


ClusteredCollapsing
Name
ClusteredCollapsing.Rproj
Data
2022_06_07_13_22_17_ee_vs_als_case.seq.csv
2022_06_07_13_22_17_ee_vs_als_cases_absent_from_dragen.txt
2022_06_07_13_22_17_ee_vs_als_CasesAndControlsAll.ped.txt
2022_06_07_13_22_17_ee_vs_als_full_data.seq.csv
cohortSelectionLog
2022_06_07_13_22_17_ee_vs_als_CasesAndControlsAll.ped.txt
isnd24704r1 isnd24704r1 0 0 2 2 Exome 65MB
isnd25606v1 isnd25606v1 0 0 2 2 Exome 65MB
isnd25793ac1 isnd25793ac1 0 0 1 2 Exome 65MB
isnd26900ah1 isnd26900ah1 0 0 2 2 Exome 65MB
isnd27732an1 isnd27732an1 0 0 1 2 Exome 65MB
isnd29366bs1 isnd29366bs1 0 0 2 2 Exome 65MB
isnd29556k1 isnd29556k1 0 0 2 2 Exome 65MB
isnd30575cz1 isnd30575cz1 0 0 2 2 Exome 65MB

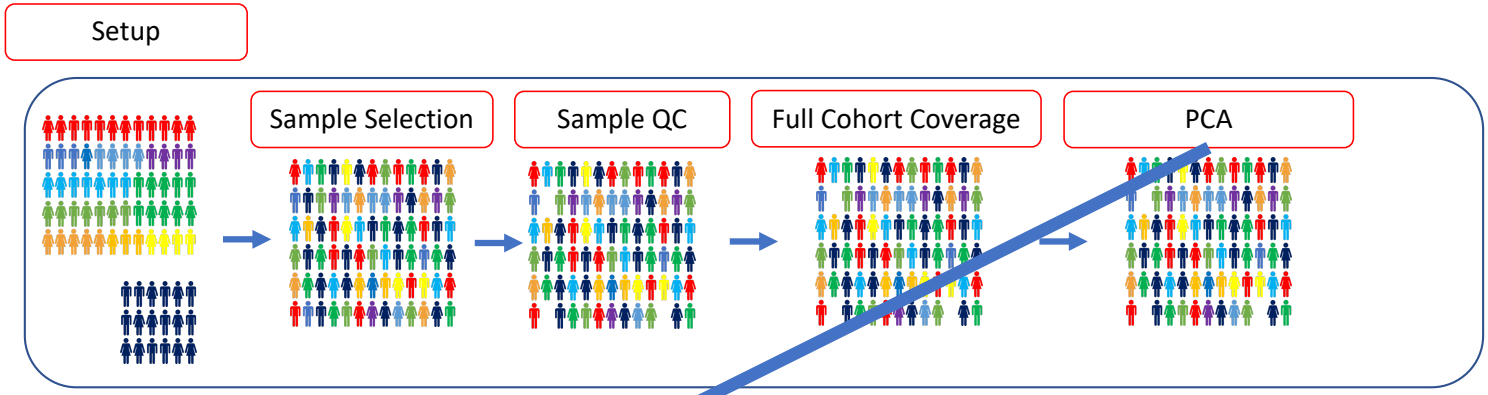


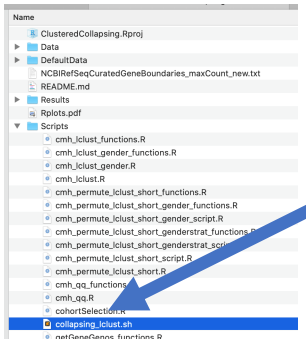



```
jm4279@qs1:~$ /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/Scripts/collapsing_lclust.sh initialCoverage
```

```
##########################################################################################
##########################################################################################
#####                              INITIAL COVERAGE: run on qs1                      #####
##########################################################################################
##########################################################################################
SAMPLES="$PROJECT/ClusteredCollapsing/Data/*.ped.txt"
if [[ $1 = "initialCoverage" ]]
then
  echo "Creating Initial Coverage..."


  # Coverage summary

  $atav --coverage-comparison --gene-boundary /nfs/goldstein/software/atav_home/data/ccds/addjusted.CCDS.genes.index.r20.hg19.txt --
    min-coverage 10 --sample $SAMPLES --out $PROJECT/ClusteredCollapsing/Results/Coverage

  # if it fails, you can run the following
  # $atav --coverage-comparison --gene-boundary /nfs/goldstein/software/atav_home/data/ccds/addjusted.CCDS.genes.index.r20.hg19.txt
  --min-coverage 10 --sample $PROJECT/Results/Coverage/*_Coverage_existing.sample.txt --out $PROJECT/ClusteredCollapsing/Results/
  Coverage
fi
```

Name	Date Modified	Size	Kind
cases_less_than_18	Jun 7, 2022 at 1:16 PM	84 KB	Plain Text
ClusteredCollapsing	Today at 10:57 AM	--	Folder
ClusteredCollapsing.Rproj	Today at 9:12 AM	190 bytes	R Project
Data	Today at 12:21 PM	--	Folder
DefaultData	Jun 7, 2022 at 1:07 PM	--	Folder
doc	Today at 10:57 AM	Zero bytes	Unix executable
NCBIRefSeqCuratedGeneBoundaries_maxCount_new.txt	Jun 7, 2022 at 1:07 PM	4.9 MB	Plain Text
README.md	Jun 7, 2022 at 1:07 PM	1 KB	Markdo...ument
Results	Yesterday at 1:04 PM	--	Folder
Coverage	Today at 10:21 AM	--	Folder
2022-06-07_13-22-36_stderr.log	Jun 7, 2022 at 1:22 PM	Zero bytes	Log File
2022-06-07_13-22-36_stdout.log	Jun 8, 2022 at 5:56 AM	621 KB	Log File
2022-06-07_13-22-43_Coverage_atav.log	Jun 8, 2022 at 5:56 AM	2 KB	Log File
2022-06-07_13-22-43_Co...rage.summary.by.exon.csv	Jun 8, 2022 at 5:56 AM	12.6 MB	comma...values
2022-06-07_13-22-43_Co...verage.summary.clean.csv	Jun 8, 2022 at 5:56 AM	960 KB	comma...values
2022-06-07_13-22-43_Co...ge_coverage.summary.csv	Jun 8, 2022 at 5:56 AM	875 KB	comma...values
2022-06-07_13-22-43_Coverage_existing.sample.txt	Jun 7, 2022 at 1:22 PM	659 KB	Plain Text
2022-06-07_13-22-43_Coverage_exon.clean.txt	Jun 8, 2022 at 5:56 AM	3.8 MB	Plain Text
2022-06-07_13-22-43_Coverage_sample.summary.csv	Jun 8, 2022 at 5:56 AM	880 KB	comma...values

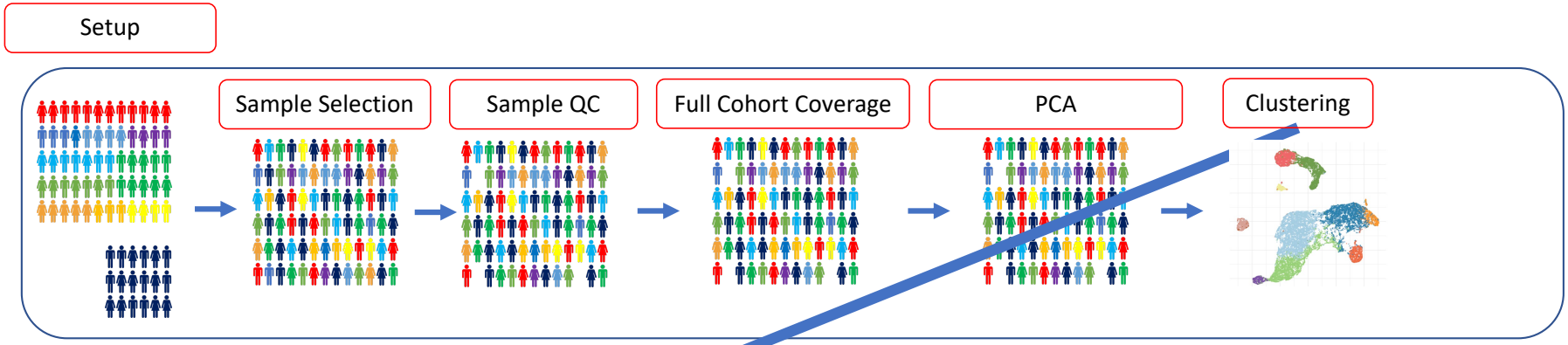


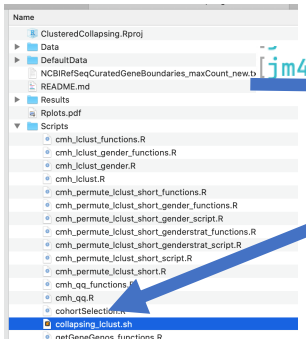



```
jm4279@qs1:~$ /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/Scripts/collapsing_lclust.sh pruning
```

```
##############################################################################
#####              KINSHIP + FlashPCA: run on qs1                        #####
##############################################################################
if [[ $1 = "pruning" ]]
then
  echo "Creating Pruning..."

  # dir="All"
  # dir2=$dir"_"

  dir=""
  dir2=""

  #Note: by default --flashpca will not perform PLINK nearest-neighbor pruning anymore use --flashpca-plink-pruning to trigger running
  plink pruning step if you want and add --flashpca-num-nearest-neighbor 200
  $atav --ped-map --variant /nfs/goldstein/software/atav_home/data/variant/informative_snps.ld_pruned.37MB.txt --kinship --flashpca --
    min-covered-case-percentage 95 --min-covered-ctrl-percentage 95 --min-coverage 10 --sample-coverage-summary $PROJECT/
    ClusteredCollapsing/Results/Coverage/*_sample.summary.csv --sample $SAMPLES --exclude-igm-gnomad-sample --out $PROJECT/
    ClusteredCollapsing/Results/KinshipFlashPCA$dir
fi
```

▼ ClusteredCollapsing	Today at 10:57 AM	--	Folder
ClusteredCollapsing.Rproj	Today at 9:12 AM	190 bytes	R Project
▶ Data	Today at 12:21 PM	--	Folder
▶ DefaultData	Jun 7, 2022 at 1:07 PM	--	Folder
doc	Today at 10:57 AM	Zero bytes	Unix executable
NCBIRefSeqCuratedGeneBoundaries_maxCount_new.txt	Jun 7, 2022 at 1:07 PM	4.9 MB	Plain Text
README.md	Jun 7, 2022 at 1:07 PM	1 KB	Markdo...ument
▼ Results	Yesterday at 1:04 PM	--	Folder
▶ Coverage	Today at 10:21 AM	--	Folder
▼ KinshipFlashPCA	Yesterday at 4:21 PM	--	Folder
2022-06-08_12-09-55_stderr.log	Jun 8, 2022 at 12:10 PM	Zero bytes	Log File
2022-06-08_12-09-55_stdout.log	Jun 8, 2022 at 12:10 PM	359 bytes	Log File
2022-06-08_12-10-09_KinshipFlashPCA_atav.log	Jun 8, 2022 at 12:10 PM	878 bytes	Log File
2022-06-08_12-12-21_stderr.log	Jun 8, 2022 at 12:12 PM	Zero bytes	Log File
2022-06-08_12-12-21_stdout.log	Jun 8, 2022 at 3:19 PM	698 KB	Log File
2022-06-08_12-12-25_KinshipFlashPCA_atav.log	Jun 8, 2022 at 3:19 PM	6 KB	Log File
2022-06-08_12-12-25_Ki...ashPCA_existing.sample.txt	Jun 8, 2022 at 12:12 PM	659 KB	Plain Text
2022-06-08_12-12-25_Ki...PCA_flashpca_eigenvalues	Jun 8, 2022 at 3:19 PM	90 bytes	Unix executable
2022-06-08_12-12-25_Ki...PCA_flashpca_eigenvectors	Jun 8, 2022 at 3:19 PM	1.6 MB	Unix executable
2022-06-08_12-12-25_KinshipFlashPCA_flashpca_pcs	Jun 8, 2022 at 3:19 PM	1.5 MB	Unix executable
2022-06-08_12-12-25_Ki...hpca_plot_eigenvalues.pdf	Jun 8, 2022 at 3:19 PM	3 KB	PDF Document
2022-06-08_12-12-25_Ki...PCA_flashpca_plot_pve.pdf	Jun 8, 2022 at 3:19 PM	3 KB	PDF Document
2022-06-08_12-12-25_Ki...pca_pruned_sample_file.txt	Jun 8, 2022 at 3:19 PM	524 KB	Plain Text
2022-06-08_12-12-25_KinshipFlashPCA_flashpca_pve	Jun 8, 2022 at 3:19 PM	123 bytes	Unix executable
2022-06-08_12-12-25_KinshipFlashPCA_flashpca.log	Jun 8, 2022 at 3:19 PM	5 KB	Log File
2022-06-08_12-12-25_KinshipFlashPCA_king.kin	Jun 8, 2022 at 3:16 PM	51 bytes	Document
2022-06-08_12-12-25_KinshipFlashPCA_king.kin0	Jun 8, 2022 at 3:18 PM	394 KB	Document






```
jm4279@dev2:/nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing$ ./Scripts/collapsing_lclust.sh clustering
```

```
##############################################################################
#####                        Louvain Clustering: Run on Dev1-4            ######
##############################################################################
if [[ $1 = "clustering" ]]
then
  echo "Creating clusters..."

  # Learn about the input options for lclust_Flash.R
  $Rscript_var $lclust_flash_var --help

  # run cluster creation
  $Rscript_var $lclust_flash_var --resolution_var 0.2 --case_group $case_group
fi

resolution="0_2"
```

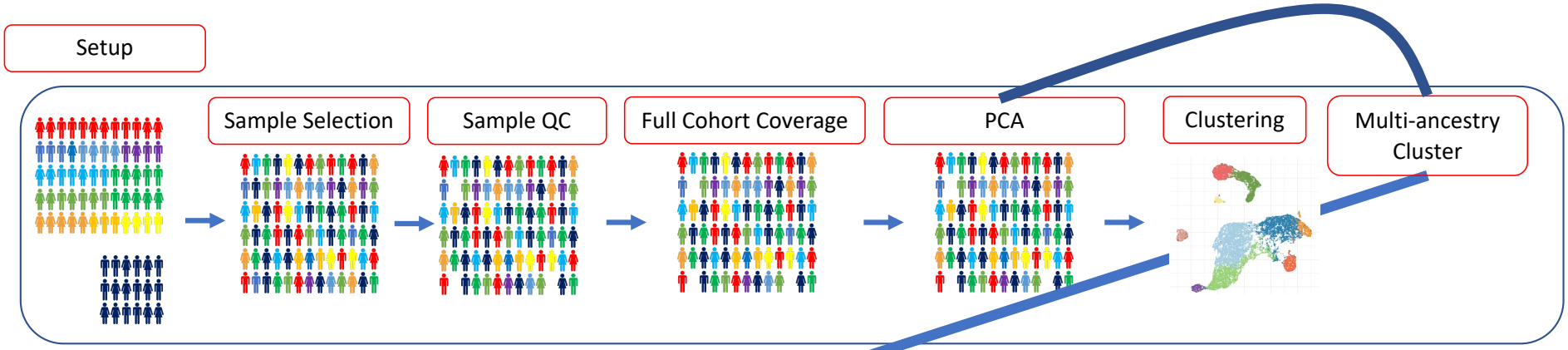
```
Usage:
  lclust_Flash.R [--case_group=<case_group>] [--resolution_var=<resolution_var>]

  Options:
  -h --help
  --case_group=<case_group> label for analysis. example might "picu" or "IPF" [default: temp_case]
  --resolution_var=<resolution_var> resolution for clustering. Typically 0.1-0.4 [default: 0.3]
jm4279@dev2:/nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing$
```

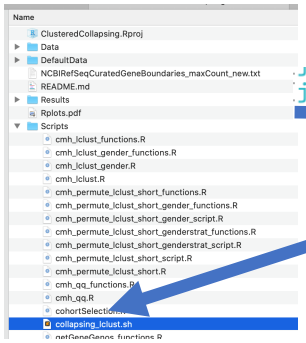
Name

- Coverage
- KinshipFlashPCA
- Plots
  - 2022_07_11_13_13_51_temp_case_Ancestry_UMAP.png
  - 2022_07_11_13_13_51_temp_case_lclust_res_0_2.RDS
  - 2022_07_11_13_13_51_temp_case_lclustering_res_0_2_cluster_sizes_gender_ratios.txt
  - 2022_07_11_13_13_51_temp_case_lclustering_res_0_2_cluster_sizes_gender.txt
  - 2022_07_11_13_13_51_temp_case_lclustering_res_0_2_cluster_sizes_ratios.txt
  - 2022_07_11_13_13_51_temp_case_lclustering_res_0_2_cluster_sizes.txt
  - 2022_07_11_13_13_51_temp_case_lclustering_res_0_2_clusters_ancestry_sizes_ratios.txt
  - 2022_07_11_13_13_51_temp_case_lclustering_res_0_2_clusters_ancestry_sizes.txt
  - 2022_07_11_13_13_51_temp_case_UMAP_0_2_Ancestry_comb.png
  - 2022_07_11_13_13_51_temp_case_UMAP_0_2.png
  - 2022_07_11_15_25_15_flashPCA_lclustering_res_0_2_cluster_9_min_sample_100_max_ratio_1_sample.txt
  - umap.RDS

Name

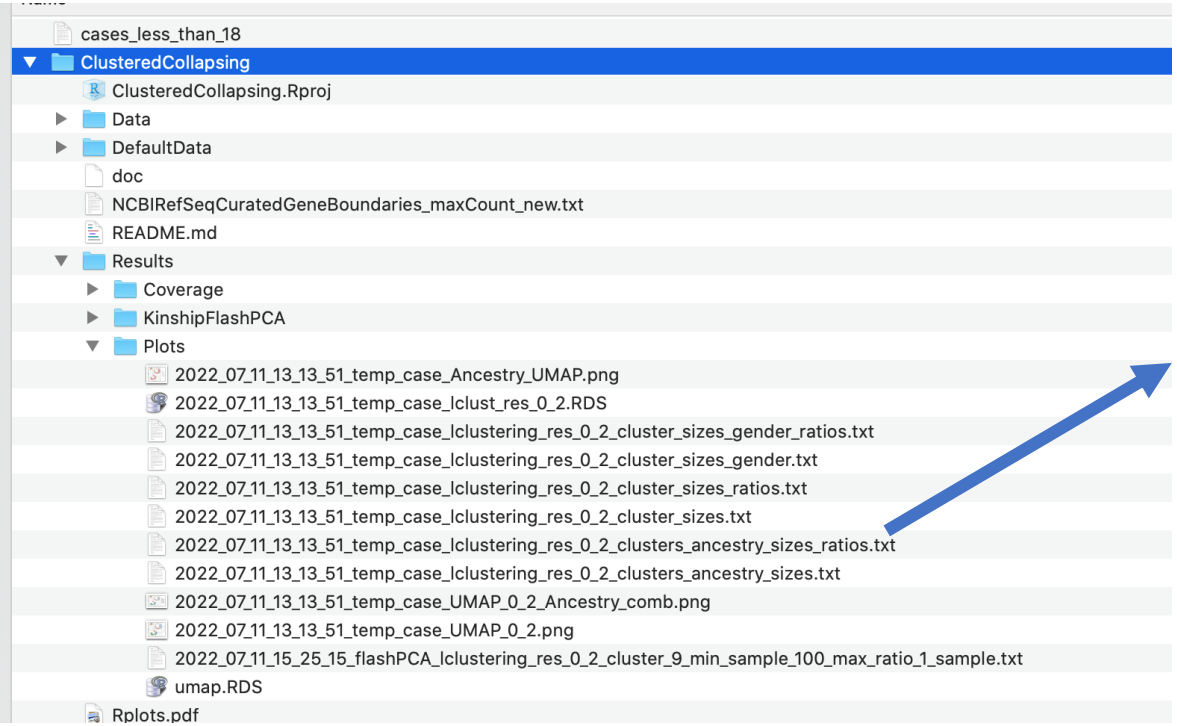
- Coverage
- Plots
- KinshipFlashPCA
  - flashPCA_lclustering_res_0_2_cluster_9_min_sample_100_max_ratio_1_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_0_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_0_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_1_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_1_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_2_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_2_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_3_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_3_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_4_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_4_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_5_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_5_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_6_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_6_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_7_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_7_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_8_sample_female.txt
  - flashPCA_lclustering_res_0_2_cluster_8_sample_male.txt
  - flashPCA_lclustering_res_0_2_cluster_0_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_1_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_2_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_3_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_4_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_5_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_6_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_7_sample.txt
  - flashPCA_lclustering_res_0_2_cluster_8_sample.txt
  - 2022-06-08 12-12-21 stdout.log

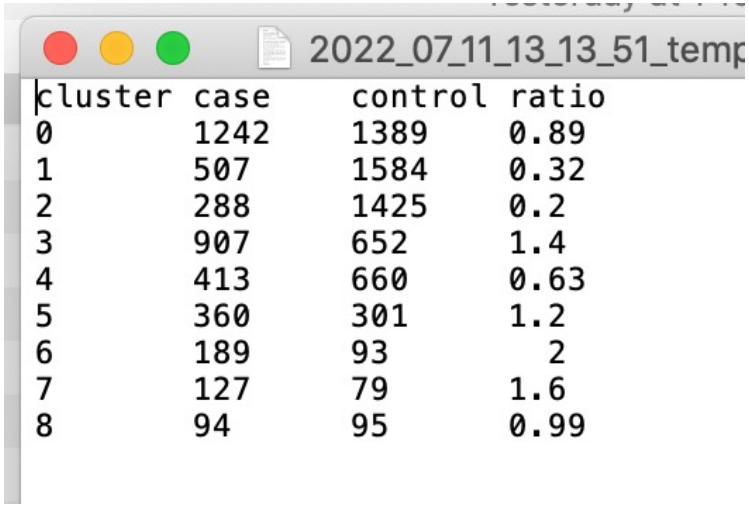




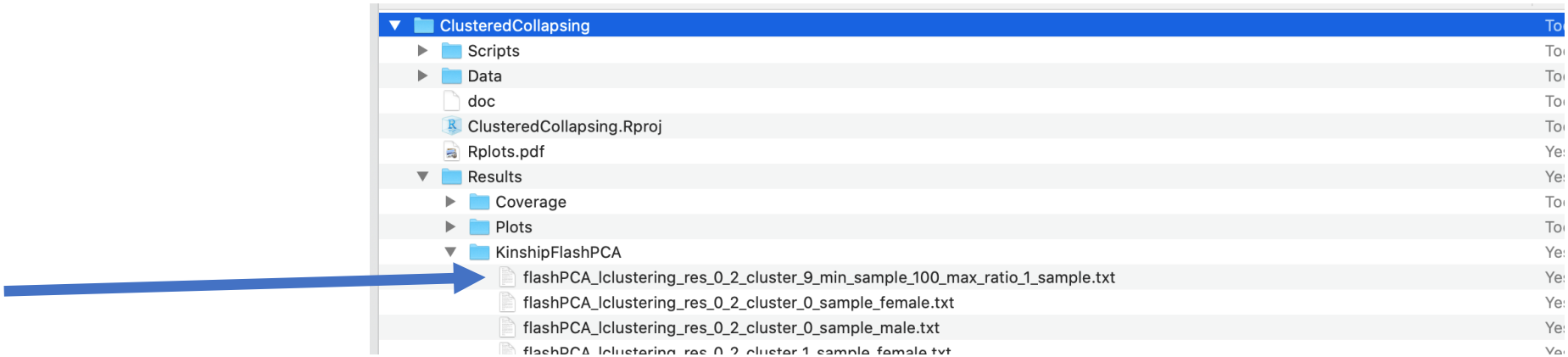

```
jm4279@dev2:/nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing$ ./Scripts/collapsing_lclust.sh clusteringMixedAncestry
```

```
##############################################################################
#####             Create a multi ancestry cluster: Run on Dev1-4            ######
##############################################################################
min_sample=100
max_ratio=1
if [[ $1 = "clusteringMixedAncestry" ]]
then
  echo "Create mixed ancestry cluster..."

  $Rscript_var $lclust_flash_combo_cluster_var --help

  # run cluster creation
  $Rscript_var $lclust_flash_combo_cluster_var --resolution_var $resolution --case_group $case_group --min_sample $min_sample
fi
```





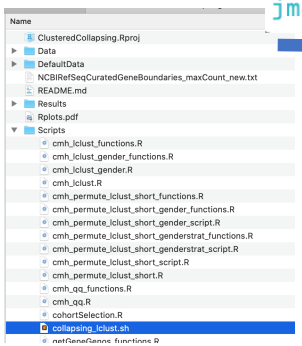
cluster	case	control	ratio
0	1242	1389	0.89
1	507	1584	0.32
2	288	1425	0.2
3	907	652	1.4
4	413	660	0.63
5	360	301	1.2
6	189	93	2
7	127	79	1.6
8	94	95	0.99


ClusteredCollapsing
Scripts
Data
doc
ClusteredCollapsing.Rproj
Rplots.pdf
Results
Coverage
Plots
KinshipFlashPCA
flashPCA_lclustering_res_0_2_cluster_9_min_sample_100_max_ratio_1_sample.txt
flashPCA_lclustering_res_0_2_cluster_0_sample_female.txt
flashPCA_lclustering_res_0_2_cluster_0_sample_male.txt

Clustering

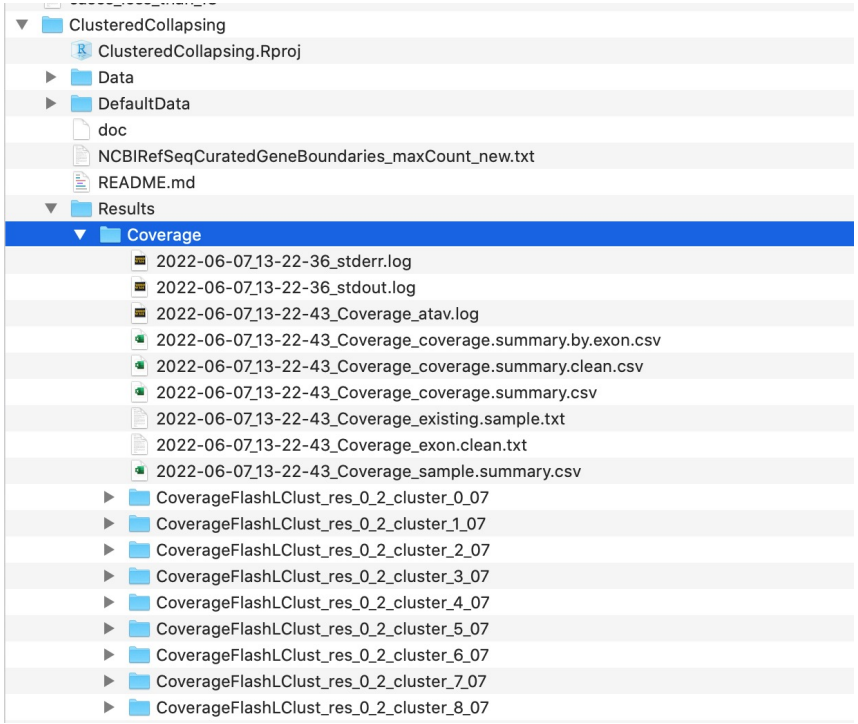
Multi-ancestry Cluster

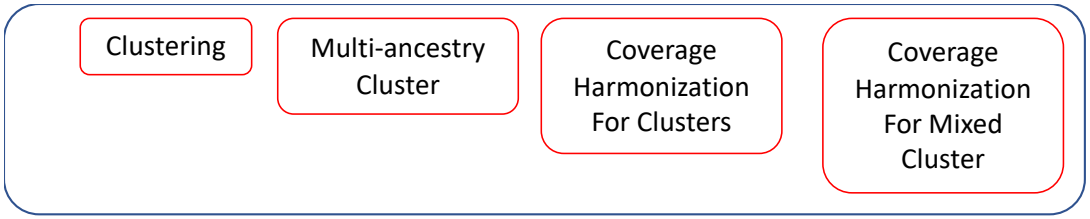
Coverage Harmonization For Clusters

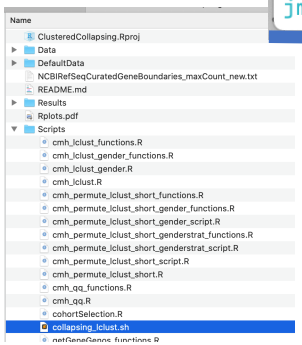



```
jm4279@qs1:~$ /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/Scripts/collapsing_lclust.sh clusterCoverage
```

```
##############################################################################
#####                        COVERAGE2: Run on qs1                        ######
##############################################################################
cov_clust_exec_str_start="$atav --site-coverage-comparison --gene-boundaries /nfs/goldstein/software/atav_home/data/ccds/
  addjusted.CCDS.genes.index.r20.hg19.txt --site-max-percent-cov-difference 0.07 --min-coverage 10"
max_cluster=8 #this is defined by user
min_cluster=0 #this is defined by user
if [[ $1 = "clusterCoverage" ]]
then
  echo "Running coverage per cluster..."

  for (( i=$min_cluster; i<=$max_cluster; i++ ))
  do

  samples="$PROJECT/ClusteredCollapsing/Results/KinshipFlashPCA"${dir}"/flashPCA_lclustering_res_"${resolution}"_cluster_"$
    {i}"_sample.txt"
  out="$PROJECT/ClusteredCollapsing/Results/Coverage/CoverageFlashLClust_res_"${resolution}"_cluster_"${i}_${dir2}"07/"

  echo $i

  cov_clust_exec_str_end=" --sample $samples --out $out"
  cov_clust_exec_str="$cov_clust_exec_str_start$cov_clust_exec_str_end"
  echo $cov_clust_exec_str
  eval $cov_clust_exec_str
  sleep 5
  done
fi
```

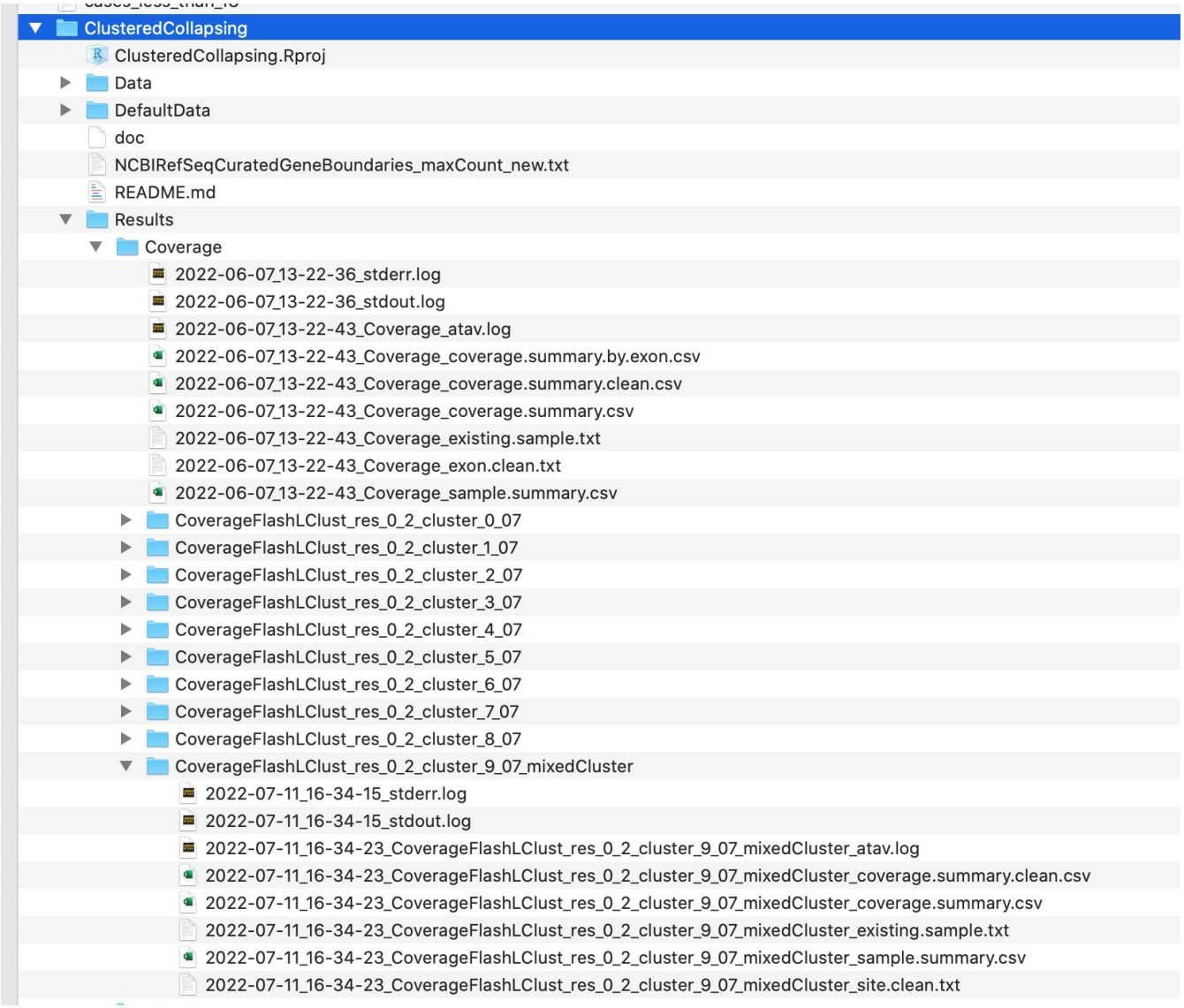

ClusteredCollapsing
ClusteredCollapsing.Rproj
Data
DefaultData
doc
NCBIRefSeqCuratedGeneBoundaries_maxCount_new.txt
README.md
Results
Coverage
2022-06-07_13-22-36_stderr.log
2022-06-07_13-22-36_stdout.log
2022-06-07_13-22-43_Coverage_atav.log
2022-06-07_13-22-43_Coverage_coverage.summary.by.exon.csv
2022-06-07_13-22-43_Coverage_coverage.summary.clean.csv
2022-06-07_13-22-43_Coverage_coverage.summary.csv
2022-06-07_13-22-43_Coverage_existing.sample.txt
2022-06-07_13-22-43_Coverage_exon.clean.txt
2022-06-07_13-22-43_Coverage_sample.summary.csv
CoverageFlashLClust_res_0_2_cluster_0_07
CoverageFlashLClust_res_0_2_cluster_1_07
CoverageFlashLClust_res_0_2_cluster_2_07
CoverageFlashLClust_res_0_2_cluster_3_07
CoverageFlashLClust_res_0_2_cluster_4_07
CoverageFlashLClust_res_0_2_cluster_5_07
CoverageFlashLClust_res_0_2_cluster_6_07
CoverageFlashLClust_res_0_2_cluster_7_07
CoverageFlashLClust_res_0_2_cluster_8_07






```
jm4279@qs1:~$ /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/Scripts/collapsing_lclust.sh clusterCoverageMixedAncestry
```

```
if [[ $1 = "clusterCoverageMixedAncestry" ]]
then
  echo "Running coverage for mixed ancestry_cluster..."

  mixed_ancestry_cluster=9


  samples="$PROJECT/ClusteredCollapsing/Results/KinshipFlashPCA"${dir}"/*flashPCA_lclustering_res_"${resolution}"_cluster_"
    ${mixed_ancestry_cluster}"_min_sample_"${min_sample}"_max_ratio_"${max_ratio}"_sample.txt"
  out="$PROJECT/ClusteredCollapsing/Results/Coverage/CoverageFlashLClust_res_"${resolution}"_cluster_"${
    mixed_ancestry_cluster}_${dir2}"07_mixedCluster/"

  cov_clust_exec_str_end=" --sample $samples --out $out"
  cov_clust_exec_str="$cov_clust_exec_str_start$cov_clust_exec_str_end"
  # echo $out

  echo $cov_clust_exec_str
  eval $cov_clust_exec_str
fi
```


ClusteredCollapsing
ClusteredCollapsing.Rproj
Data
DefaultData
doc
NCBIRefSeqCuratedGeneBoundaries_maxCount_new.txt
README.md
Results
Coverage
2022-06-07_13-22-36_stderr.log
2022-06-07_13-22-36_stdout.log
2022-06-07_13-22-43_Coverage_atav.log
2022-06-07_13-22-43_Coverage_coverage.summary.by.exon.csv
2022-06-07_13-22-43_Coverage_coverage.summary.clean.csv
2022-06-07_13-22-43_Coverage_coverage.summary.csv
2022-06-07_13-22-43_Coverage_existing.sample.txt
2022-06-07_13-22-43_Coverage_exon.clean.txt
2022-06-07_13-22-43_Coverage_sample.summary.csv
CoverageFlashLClust_res_0_2_cluster_0_07
CoverageFlashLClust_res_0_2_cluster_1_07
CoverageFlashLClust_res_0_2_cluster_2_07
CoverageFlashLClust_res_0_2_cluster_3_07
CoverageFlashLClust_res_0_2_cluster_4_07
CoverageFlashLClust_res_0_2_cluster_5_07
CoverageFlashLClust_res_0_2_cluster_6_07
CoverageFlashLClust_res_0_2_cluster_7_07
CoverageFlashLClust_res_0_2_cluster_8_07
CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster
2022-07-11_16-34-15_stderr.log
2022-07-11_16-34-15_stdout.log
2022-07-11_16-34-23_CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster_atav.log
2022-07-11_16-34-23_CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster_coverage.summary.clean.csv
2022-07-11_16-34-23_CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster_coverage.summary.csv
2022-07-11_16-34-23_CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster_existing.sample.txt
2022-07-11_16-34-23_CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster_sample.summary.csv
2022-07-11_16-34-23_CoverageFlashLClust_res_0_2_cluster_9_07_mixedCluster_site.clean.txt

Clustering

Multi-ancestry Cluster

Coverage Harmonization For Clusters

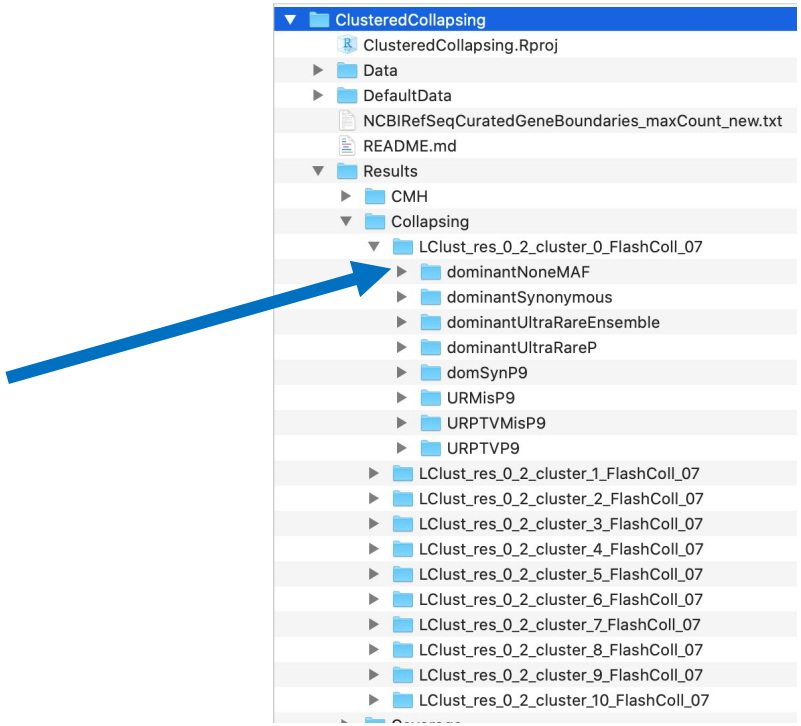
Coverage Harmonization For Mixed Cluster

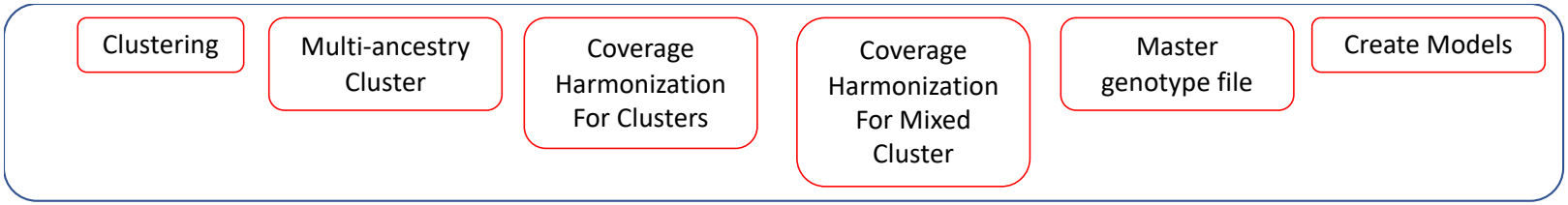
Master genotype file

```
jm4279@qs1:~$ /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/Scripts/collapsing_lclust.sh createMasterGeno
```

```
##########################################################################################
############ Run Master genotypes to be used for collapsing lite later: Run on qs1
##########################################################################################
if [[ $1 = "createMasterGeno" ]]
then
  echo "Creating master genotype.csv ..."


  for (( i=$min_cluster; i<=$max_cluster; i++ ))
  do
  samples="$PROJECT/ClusteredCollapsing/Results/KinshipFlashPCA"${dir}"/flashPCA_lclustering_res_"${resolution}"_cluster_"${
    {i}"_sample.txt"
  geneBoundaries="$PROJECT/ClusteredCollapsing/Results/Coverage/CoverageFlashLClust_res_"${resolution}"_cluster_"${i}_${
    dir2}"07/*_site.clean.txt"
  coverage2="$PROJECT/ClusteredCollapsing/Results/Coverage/CoverageFlashLClust_res_"${resolution}"_cluster_"${i}_${dir2}"
    07/*_coverage.summary.clean.csv"
  outputFolder="$PROJECT/ClusteredCollapsing/Results/Collapsing"${dir}"/LClust_res_"${resolution}"_cluster_"${i}_${dir2}"
    FlashColl_07"
  echo $samples


  # gundi version: max-qc-fail-sample missing, lenient effect filter, loose AF filter
  # $atav --collapsing-dom --mann-whitney-test --gene-boundaries $geneBoundaries --read-coverage-summary $coverage2
  $INCLUDES $QC --effect $CODING_SPLICE --gnomad-genome-pop global --gnomad-genome-maf 0.01 --gnomad-exome-pop global
  --gnomad-exome-maf 0.01 --exac-pop global --exac-maf 0.01 --loo-maf 0.01 --sample $samples --out $outputFolder/
  dominantNoneMAF/
  # sleep 2

  # Josh does same version but with list-var geno to save a little sapce
  $atav --list-var-geno --gene-boundaries $geneBoundaries $INCLUDES $QC --effect $CODING_SPLICE --gnomad-genome-pop global
    --gnomad-genome-maf 0.01 --gnomad-exome-pop global --gnomad-exome-maf 0.01 --exac-pop global --exac-maf 0.01 --loo-maf
    0.01 --sample $samples --out $outputFolder/dominantNoneMAF/
  sleep 2
  done
fi
```

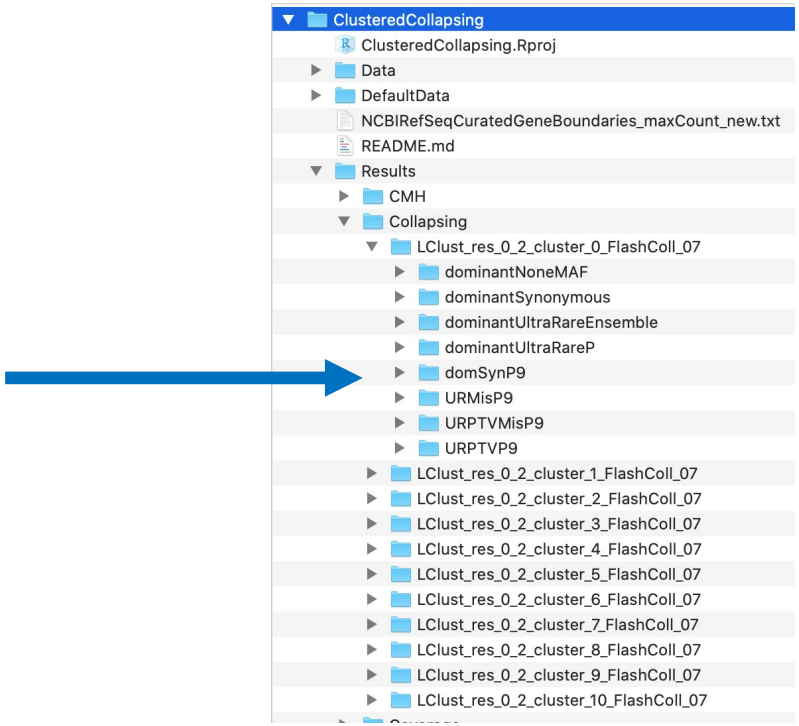
```
INCLUDES="--include-gnomad-genome --include-gnomad-exome --include-exac --include-mtr --include-gerp --include-rvis
  --include-sub-rvis --include-limbr --include-revel --include-trap --include-discovehr --include-known-var
  --include-primate-ai --include-ccr --include-loftee --include-gnomad-gene-metrics --include-pext --include-mpc
  --flag-repeat-region --include-syn-rvis --include-genome-asia --include-iranome --include-gme --include-top-med
  --include-igm-af"

QC="--exclude-artifacts --filter pass,likely,intermediate --exclude-evs-qc-failed --ccds-only --min-coverage 10
  --include-qc-missing --qd 5 --qual 50 --mq 40 --gq 20 --snv-fs 60 --indel-fs 200 --snv-sor 3 --indel-sor 10 --rprs -3
  --mqrs -10 --het-percent-alt-read 0.3-1 --min-exac-vqslod-snv -2.632 --min-exac-vqslod-indel 1.262
  --gnomad-genome-rf-tp-probability-snv 0.01 --gnomad-genome-rf-tp-probability-indel 0.02
  --gnomad-exome-rf-tp-probability-snv 0.01 --gnomad-exome-rf-tp-probability-indel 0.02"
```


ClusteredCollapsing
ClusteredCollapsing.Rproj
Data
DefaultData
NCBIRefSeqCuratedGeneBoundaries_maxCount_new.txt
README.md
Results
CMH
Collapsing
LClust_res_0_2_cluster_0_FlashColl_07
dominantNoneMAF
dominantSynonymous
dominantUltraRareEnsemble
dominantUltraRareP
domSynP9
URMisP9
URPTVMisP9
URPTVP9
LClust_res_0_2_cluster_1_FlashColl_07
LClust_res_0_2_cluster_2_FlashColl_07
LClust_res_0_2_cluster_3_FlashColl_07
LClust_res_0_2_cluster_4_FlashColl_07
LClust_res_0_2_cluster_5_FlashColl_07
LClust_res_0_2_cluster_6_FlashColl_07
LClust_res_0_2_cluster_7_FlashColl_07
LClust_res_0_2_cluster_8_FlashColl_07
LClust_res_0_2_cluster_9_FlashColl_07
LClust_res_0_2_cluster_10_FlashColl_07




```
jm4279@qs1:~$ /nfs/projects/ped_pli/ped_pli_20220607/ClusteredCollapsing/Scripts/collapsing_lclust.sh models
```

```
##########################################################################################
if [[ $1 = "models" ]]
then
  echo "Creating models..."

  for (( i=$min_cluster; i<=$max_cluster; i++ ))
  do

  samples="$PROJECT/ClusteredCollapsing/Results/KinshipFlashPCA"${dir}"/flashPCA_lclustering_res_"${resolution}"_cluster_"$
    {i}"_sample.txt"
  outputFolder="$PROJECT/ClusteredCollapsing/Results/Collapsing"${dir}"/LClust_res_"${resolution}"_cluster_"${i}_${dir2}"
    FlashColl_07"
  genos=$outputFolder"/dominantNoneMAF/*genotypes.csv"
  # echo $genos


  ### Dominant Syn
  # $atav --collapsing-lite --mann-whitney-test --genotype $genos --effect $SYN_EFFECTS --min-exac-vqslod-snv 5000
  --min-exac-vqslod-indel 5000  --gnomad-genome-pop global --gnomad-genome-maf 0 --gnomad-exome-pop global
  --gnomad-exome-maf 0 --exac-pop global --exac-maf 0 --loo-maf 0.0005 --max-qc-fail-sample 0 --sample $samples --out
  $outputFolder/dominantSynonymous/
  # sleep 2

  ### Josh: Dominant Syn
  $atav --collapsing-lite --mann-whitney-test --genotype $genos --effect $SYN_EFFECTS --min-exac-vqslod-snv 5000 --
    min-exac-vqslod-indel 5000  --gnomad-genome-pop global --gnomad-genome-maf 0 --gnomad-exome-pop global --
    gnomad-exome-maf 0 --exac-pop global --exac-maf 0 --loo-maf 0.0005 --min-pext-ratio 0.9 --max-qc-fail-sample 0 --
    exclude-repeat-region --sample $samples --out $outputFolder/domSynP9/
  sleep 2
```


ClusteredCollapsing
ClusteredCollapsing.Rproj
Data
DefaultData
NCBIRefSeqCuratedGeneBoundaries_maxCount_new.txt
README.md
Results
CMH
Collapsing
LClust_res_0_2_cluster_0_FlashColl_07
dominantNoneMAF
dominantSynonymous
dominantUltraRareEnsemble
dominantUltraRareP
domSynP9
URMisP9
URPTVMisP9
URPTVP9
LClust_res_0_2_cluster_1_FlashColl_07
LClust_res_0_2_cluster_2_FlashColl_07
LClust_res_0_2_cluster_3_FlashColl_07
LClust_res_0_2_cluster_4_FlashColl_07
LClust_res_0_2_cluster_5_FlashColl_07
LClust_res_0_2_cluster_6_FlashColl_07
LClust_res_0_2_cluster_7_FlashColl_07
LClust_res_0_2_cluster_8_FlashColl_07
LClust_res_0_2_cluster_9_FlashColl_07
LClust_res_0_2_cluster_10_FlashColl_07

CMH, permutations and QQ plots in process…