MVP Terminal Commands and Files

Last accessed on 9/20/2024

#command_list.txt

```
bsub -q medium -e error%J -o output%J pullcode -G ros009 -R 64000 -M
93000 < pullphecodes_byloci.sh
bsub -q medium -e error%J -o output%J process -G ros009 -R 64000 -M
93000 -w "done(pullcode)" < dataprocessing.sh
```

#pullphecodes_byloci.sh

```
while IFS=$'\t' read -r fname
do
base=$(basename "$fname")
base=${base%.HARE.KDI.txt.gz}
while IFS=$'\t' read -r gene chr lowbp highbp
do
echo $gene
zcat $fname | awk -v g=$gene -v l=$lowbp -v h=$highbp -v f=$fname
'BEGIN{FS=OFS="\t"} {if ($3==c && $4>l && $4<h) {print $0}}' > /group/
research/ros009/Loci/"${base}_${gene}bylocus.txt"
done < MVP_genes_loci_expanded.txt
echo $base
sed -i "1s/^/$(zcat "$fname" | head -n1)\n/" /group/research/ros009/
Loci/"${base}_${gene}bylocus.txt"
done < Phecodes.txt
```

#dataprocessing.sh

```
./insertheaders_all.sh;
./AFR_conversion.sh;
./EUR_conversion.sh;
./HIS_conversion.sh;
./META_conversion.sh;
./rename.sh;
./combine_new.sh
```

#insertheaders_all.sh

```
for fname in *META*.txt
do
sed -i "1s/^/$(head -n1 /group/research/ros009/CFH/
Phe_480_1.META_CFHbySNP.txt)\n/" /group/research/ros009/Loci/"$fname"
echo $fname
done

for fname in *AFR*.txt
do
sed -i "1s/^/$(head -n1 /group/research/ros009/CFH/
Phe_480_1.AFR_CFHbySNP.txt)\n/" /group/research/ros009/Loci/"$fname"
echo $fname
done
```

```
for fname in *EUR*.txt
do
sed -i "1s/^/$(head -n1 /group/research/ros009/CFH/
Phe_480_1.EUR_CFHbySNP.txt)\n/" /group/research/ros009/Loci/"$fname"
echo $fname
done

for fname in *HIS*.txt
do
sed -i "1s/^/$(head -n1 /group/research/ros009/CFH/
Phe_480_1.HIS_CFHbySNP.txt)\n/" /group/research/ros009/Loci/"$fname"
echo $fname
done

#AFR_conversion.sh
for fname in *AFR*.txt
do
base=${fname%.txt}
echo $base
awk ' BEGIN{FS=OFS="\t"} {print $0, (NR==1?
"num_samples_actual\tq_pval\ti2\tdirection" : "99\t99\t99\t99")}'
$fname > ${base}_new.txt
done

#EUR_conversion.sh
for fname in *EUR*.txt
do
base=${fname%.txt}
echo $base
awk ' BEGIN{FS=OFS="\t"} {print $0, (NR==1?
"num_samples_actual\tq_pval\ti2\tdirection" : "99\t99\t99\t99")}'
$fname > ${base}_new.txt
done

#HIS_conversion.sh
for fname in *HIS*.txt
do
base=${fname%.txt}
echo $base
awk ' BEGIN{FS=OFS="\t"} {print $0, (NR==1?
"num_samples_actual\tq_pval\ti2\tdirection" : "99\t99\t99\t99")}'
$fname > ${base}_new.txt
done

#META_conversion.sh
for fname in *META*.txt
do
base=${fname%.txt}
echo $base
```

```
awk ' BEGIN{FS=OFS="\t"} {$9 = $9 FS
(NR==1?"case_af\tcontrol_af":"99\t99"); $13 = $13 FS
(NR==1?"r2":"99"); $14 = $14 FS
(NR==1?"num_controls\tnum_cases":"99\t99")}1' $fname > ${base}_new.txt
done

#rename.sh
for fname in *bylocus_new.txt
do
phecode=${fname%*.*.txt}
tmp=${fname#$phecode.}
race=${tmp%_*_*}
gene=${tmp#*_}
gene=${gene%*bylocus_new.txt}
echo $phecode
echo $race
echo $gene
mv "$fname" "${phecode}.${race}.${gene}bylocus_new.txt"
done

#combine_new.sh
awk 'BEGIN{FS=OFS="\t"} NR==1&&FNR==1 {print "Phecode", "Race",
"Gene", $0} {if (FNR!=1) {file=FILENAME; sub(/
bylocus_new.txt$/,"",file); split(file,id,"."); print id[1], id[2],
id[3], $0}}' *_new.txt > SNPsbylocus_combined_expanded.txt



#Phecodes.txt
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038.AFR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038.HIS.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038.META.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_1.AFR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_1.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_1.META.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_2.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_2.META.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_3.AFR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_3.EUR.HARE.KDI.txt.gz
```

Phe_038_3.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_3.HIS.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/Infectious_Diseases/
Phe_038_3.META.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994.AFR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994.HIS.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994.META.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_2.AFR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_2.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_2.HIS.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_2.META.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_21.AFR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_21.EUR.HARE.KDI.txt.gz
/data/data1/ros009/Other_Data/gwPheWas/Phecodes/
Injuries_and_Poisonings/Phe_994_21.META.HARE.KDI.txt.gz

#MVP_genes_loci_expanded.txt

CFD	19	0	1863641
CFB	6	30946095	32952084
CFP	X	46623282	48630305
CFI	4	108730982	110801999
CFH	1	195652043	197747504
C3	19	5677704 7720650	
CD46	1	206752038	208795516
CD55	1	206321678	208360966
CD59	11	32703010	34736479
CR1	1	206496157	208641765
CR2	1	206454328	208489892
C3AR1	12	7056844 9066359	
C5AR1	19	46307477	48322066
C5AR2	19	46332175	48347329

###The CFHR genes are contained within the bounds of the expanded CFH loci and were not independently pulled as a result

Commands and files used on collection of SNPs obtained from the MVP database

```
#Processing SNPsbylocus_combined_expanded.txt to only include sepsis/
bacteremia Phecodes and AP genes
awk '{FS=OFS="\t"} FNR==1 {print $0} (FNR>1 && ($1 == "Phe_038" || $1
== "Phe_038_1" || $1 == "Phe_038_2" || $1 == "Phe_038_3" || $1 ==
"Phe_994" || $1 == "Phe_994_2" || $1 == "Phe_994_21") && ($3 == "CFD"
|| $3 == "CFB" || $3 == "CFP" || $3 == "CFI" || $3 == "CFH" || $3 ==
"C3" || $3 == "CR1" || $3 == "CR2" || $3 == "C3AR1" || $3 == "C5AR1"
|| $3 == "C5AR2" || $3 == "CD46" || $3 == "CD55" || $3 == "CD59" || $3
== "CFHR1" || $3 == "CFHR2" || $3 == "CFHR3" || $3 == "CFHR4")) {print
$0}' SNPsbylocus_combined_expanded.txt > snps_altsepsis.txt


#Obtaining 1000Genomes SNP data for use in Li and Ji calculations
#modified_loopinglocus_search.sh
#!/bin/zsh

# Check if the file argument is provided
if [ $# -eq 0 ]; then
  echo "Usage: $0 <loci_of_interest_file>"
  exit 1
fi

loci_file=$1

# Check if the loci file exists and is readable
if [ ! -f "$loci_file" ]; then
  echo "Error: File $loci_file does not exist or is not readable."
  exit 1
fi

set -x

while IFS=$'\t' read -r line
do

IFS=$'\t' read -r chr start end <<< "$line"

echo "Processing region: ${chr}:${start}-${end}"

url=ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/
1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/ALL.chr$
{chr}.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz

echo "URL: ${url}"

bcftools view -r ${chr}:${start}-${end} ${url} -o ${chr}_${start}_$
{end}.vcf.gz
```

```
# Check if the VCF file is empty (no variants)
if [ ! -s ${chr}_${start}_${end}.vcf.gz ]; then
    echo "No variants found for ${chr}:${start}-${end}. Skipping..."
    # Skip the current iteration if the file is empty
    continue
fi

/Users/kinman/Desktop/MVP_Data/plink_mac_20231211/plink --vcf ${chr}_$
{start}_${end}.vcf.gz --make-bed --out 1000G_MVP_SNPslist_${chr}_$
{start}_${end}

done < "$loci_file"
```

#sorted_loci_of_interest_AP.txt

```
1       196652043       196747504
1       196774840       196795407
1       196819731       196832189
1       196888052       196918633
1       196943738       196959622
1       207321678       207360966
1       207454328       207489892
1       207496157       207641765
1       207752038       207795516
4       109730982       109801999
6       31946095        31952084
11      33703010        33736479
12      8056844 8066359
19      47307477        47322066
19      47332175        47347329
19      6677704 6720650
19      859664  863641
```

#MVP_genes_loci_AP.txt

Gene	Chromosome	Start	End
CFHR1	1	196819731	196832189
CFHR2	1	196943738	196959622
CFHR3	1	196774840	196795407
CFHR4	1	196888052	196918633
CFD	19	859664 863641	
CFB	6	31946095	31952084
CFI	4	109730982	109801999
CFH	1	196652043	196747504
C3	19	6677704 6720650	
CD46	1	207752038	207795516
CD55	1	207321678	207360966
CD59	11	33703010	33736479
CR1	1	207496157	207641765
CR2	1	207454328	207489892

```
C3AR1   12      8056844 8066359
C5AR1   19      47307477        47322066
C5AR2   19      47332175        47347329
```

#Using VEP to annotate missing gene names (when able)
List of rsIDs manually pasted into VEP web interface at https://useast.ensembl.org/Tools/VEP and results downloaded in .txt format (VEP_missing_gene_results.txt)

#FUMA annotation
Text file uploaded at https://fuma.ctglab.nl/snp2gene and results downloaded in .txt format (multiple files, downloaded into a FUMA folder)

#Finding allele frequencies using downloaded 1000Genomes data to correct major/minor allele assignments (freq.vcf.gz contains the allele frequencies)

```
vcftools --gzvcf freq.vcf.gz --snps rsid_1000G_list.txt --stdout --recode > sepsisv2.txt && awk '{FS=OFS="\t"} FNR>9 {print $0}' sepsisv2.txt > freq_data_genereg.txt
```

#Missing RDB score collection
List of rsIDs manually pasted into search field of https://www.regulomedb.org/regulome-search/ and results downloaded in .tsv format (missing_rdb_scores.tsv)