Exercises – FImpute3

Summer Course – UNE – Armidale (Day2 – Feb 21, 2023)

Use the data in "`/home/guest011/imp_data/`" folder.

1) Write a control file to impute the 3 low density panels to high density.

2) Compute imputation accuracy (concordance rate with true genotypes in "`genotypes_true.txt`") for imputed genotypes. Do not include original genotypes in the accuracy calculation.

3) Do a parentage verification and discovery. Use 1% error rate.

4) Is there any duplicate record in the genotype file?

5) Panels 3 and 4 have high overlap. Merge these two panel with "`merge_chip`" command.

6) There are 25 sires with more than 8 progeny. Use "`add_ungen`" command to reconstruct their genotypes from their progeny genotypes.

7) Perform population imputation only.

8) Perform family imputation only; trace back the pedigree one generation.

Solutions:

1)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
ped_file="/home/guest011/imp_data/ped.txt";
output_folder="output_e1";
```

2)

```
##############################################################################
# Script to compute imputation accuracy
# Toni Reverter
##############################################################################

# Map the location of both genotype files

geno_imp= output_e1/genotypes_imp.txt
geno_true=/home/guest011/imp_data/genotypes_true.txt

# First collect those animals genotyped with the low density
# This is known from the second column in the imputed genotype file
# where it has to be other than one
# Also, the first line is a header so we skip that one

awk 'NR>1 && $2>1 {print $0}' $geno_imp > anim1.j

# Now we count how many we have and create a loop for each one of them

N=1
MN=`wc anim1.j | awk '{print $1}'`
while [ $N -le $MN ]
do
   # We collect the id of the animal we are currently processing
   id=`awk -v N=$N 'NR==N {print $1}' anim1.j`

   # We collect the imputed genotype of the animal currently processing
   # With a "printf" statement we transpose it so it turns into
   # a single column with 20000 rows (for as many SNPs)

   awk -v id=$id '$1==id {print $3}' anim1.j | \
   awk '{for(i=1;i<=length($1);i++) printf substr($1,i,1) " " "\n"}' > imp.j

   # Similarly, collect the imputed genotype of the same animal
   awk -v id=$id '$1==id {print $3}' $geno_true | \
   awk '{for(i=1;i<=length($1);i++) printf substr($1,i,1) " " "\n"}' > true.j

   # We paste the two files and after replacing codes "3" and "4" by "1"
   # in the imputed file we count how many coincidences
```

```
   # and divide the coincidences by 20000 to get the accuracy rate
   # ...if multiply by 100 you'll get the accuracy percentage
   paste imp.j true.j | awk '$1==3 || $1==4 {$1=1}; {print $0}' | \
   awk '$1==$2 {print $0}' | wc | \
   awk -v N=$N -v id=$id '{print N, id, $1/20000}'

   rm imp.j true.j

   N=`expr $N + 1`
done

# NOTE: This overestimates accuracy because this should be done only
#  for those SNPs that have been imputed, not all the 20000.
#  The correct approach would require keeping track of which LD chip
#  was each animal initially genotyped and then perform the above operation
#  only for the SNPs not included in that LD chip.
```

3)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
ped_file="/home/guest011/imp_data/ped.txt";
parentage_test /find_match_cnflt /find_match_mp
               /remove_conflict /find_identical
               /ert_mm=0.01 /ert_m=0.01;
output_folder="output_e3";
```

4)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
ped_file="/home/guest011/imp_data/ped.txt";
parentage_test /find_identical;
output_folder="output_e4";
```

5)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
```

```
ped_file="/home/guest011/imp_data/ped.txt";
merge_chip /min_overlap=0.95;
output_folder="output_e5";
```

6)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
ped_file="/home/guest011/imp_data/ped.txt";
add_ungen /min_fsize=8 /save_sep;
output_folder="output_e6";
```

7)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
ped_file="/home/guest011/imp_data/ped.txt";
turnoff_fam;
output_folder="output_e7";
```

8)

```
title="Summer Course UNE Feb 2023";
genotype_file = "/home/guest011/imp_data/genotypes_hd.txt"
                "/home/guest011/imp_data/genotypes_ld1.txt"
                "/home/guest011/imp_data/genotypes_ld2.txt"
                "/home/guest011/imp_data/genotypes_ld3.txt";
snp_info_file="/home/guest011/imp_data/snp_info.txt";
ped_file="/home/guest011/imp_data/ped.txt";
turnoff_pop;
ped_depth=1;
output_folder="output_e8";
```