# Module 15: Prediction Practical

Need help? doug.speed@ucl.ac.uk

# Check List

If all has gone to plan you should:

- Have `putty` installed and configured, which allows you to "ssh" into Hong's server, hong1.une.edu.au (username asc2016, pw feb01). This is where you will run `PLINK, GCTA, LDAK and IMPUTE2`
- Have `winSCP` installed and configured, which allows you to copy files from Hong's server to your Desktop (which you can then read into R), and files from your Desktop to Hong's server - NOT AS IMPORTANT IN THIS ONE
- Have a copy of the slides, either from Julius' website `http://jvanderw.une.edu.au/AGSCcourse.htm`, or from the module15 folder on Hong's server

If so, you may now begin :)

# Datafiles

Stored in `doug.bed`, `doug.bim`, `doug.fam`

980 individuals recorded for 10 000 markers

Vinzent has also constructed a mystery phenotype, `doug.phen`. We will construct different prediction models, TRAINED USING ONLY INDIVIDUALS IN `doug.keep`

... and compare how well they predict the phenotypes of individuals in `doug.test`

First copy the files to your home folder

```
cd $HOME/XXX
cp ../module15/doug.* ./
```

# 1 - Polygenic Risk Scores

Polygenic Risk Scores construct a prediction model using estimates from single-SNP analysis

Therefore, we start by performing an association study

NOTE FOR THIS, WE MUST USE ldak.mm (not ldak.out)

```
../ldak.mm --pheno doug.pheno --linear linall \
--keep doug.keep --bfile doug
```

```
$ head linall.assoc
Chr Predictor BP A1 A2 Effect SD Alt_Like Null_Like Wald_Stat Wald_P A1_Mean MAF
1 SNP1 1 A B 1.271039 1.803377 -2548.8184 -2550.5753 0.7048 4.8093e-01 1.883871 0.0
1 SNP2 2 A B 0.593224 3.029093 -2548.5293 -2550.5753 0.1958 8.4473e-01 1.959677 0.0
1 SNP3 3 A B 1.250591 3.648633 -2548.3036 -2550.5753 0.3428 7.3178e-01 0.027419 0.0
1 SNP4 4 A B -1.278914 0.997218 -2548.8369 -2550.5753 -1.2825 1.9967e-01 0.440323 0
1 SNP5 5 A B 0.854818 2.200901 -2548.7923 -2550.5753 0.3884 6.9772e-01 1.924194 0.0
1 SNP6 6 A B 0.270115 1.139572 -2549.4980 -2550.5753 0.2370 8.1263e-01 1.658065 0.1
1 SNP7 7 A B -0.224289 0.919168 -2549.7112 -2550.5753 -0.2440 8.0722e-01 1.437097 0
1 SNP8 8 A B -0.095341 0.896938 -2549.7599 -2550.5753 -0.1063 9.1535e-01 1.353226 0
1 SNP9 9 A B 3.169010 2.058409 -2547.7501 -2550.5753 1.5395 1.2367e-01 1.908065 0.0
```

# 1 - Polygenic Risk Scores

This is our prediction model (one effect size for each SNP). Except we must put it in the correct "scorefile" format. Can do this using R

```
> res=as.matrix(read.table("linall.assoc",head=T))
> score=cbind(res[,c(2,4,5)],-1,res[,6])

> colnames(score)[4:5]=c("Centre","Effect")
> write.table(score,"linall.score",row=F,quote=F)
```

Alternatively, you could make this file using `awk`

```
awk < linall.assoc '{print $2, $4, $5, 0, $6}' > linall.score
```

# 1 - Polygenic Risk Scores

So now we can use `--calc-scores` to use this model to predict the phenotype for the test individuals in `doug.test`

```
../ldak.out --calc-scores linall --scorefile linall.score \
--bfile doug --keep doug.test --pheno doug.pheno
```

# 1 - Polygenic Risk Scores

Now we can see how well it performed, which we will measure as correlation between true phenotype (column 3, below) and predicted phenotype (column 5, below)

```
> profile=as.matrix(read.table("linall.profile",head=T))

 head(profile)
     ID1 ID2 Phenotype Covariates   Profile1 Count1
[1,]  60  60  5.160747          0 -147.84518  10000
[2,]  62  62 30.020193          0  221.05192  10000
[3,]  65  65 18.120861          0 -624.75084  10000
[4,]  67  67 23.505274          0  916.48729  10000
[5,]  74  74 20.401616          0  241.84536  10000
[6,]  80  80 16.918050          0  -56.60087  10000

> cor(profile[,3],profile[,5])
```

# 1 - Polygenic Risk Scores

With Polygenic Risk Scores, it is common to only use those with $p$-value below a certain threshold. When association testing, can use (say) `--pfilter 0.1` to print out only those SNPs with $P < 0.1$

```
 ../ldak.mm --pheno doug.pheno --linear lin.1 \
--keep doug.keep --bfile doug --pfilter 0.1

 ../ldak.mm --pheno doug.pheno --linear lin.01 \
--keep doug.keep --bfile doug --pfilter 0.01

 ../ldak.mm --pheno doug.pheno --linear lin.001 \
--keep doug.keep --bfile doug --pfilter 0.001

 ../ldak.mm --pheno doug.pheno --linear lin.0001 \
--keep doug.keep --bfile doug --pfilter 0.0001
```

# 1 - Polygenic Risk Scores

Now make the score file for each of these

```
> res=as.matrix(read.table("lin.1.assoc",head=T))
> score=cbind(res[,c(2,4,5)],-1,res[,6])
> colnames(score)[4:5]=c("Centre","Effect")
> write.table(score,"lin.1.score",row=F,quote=F)

> resave=function(file)
{
res=as.matrix(read.table(paste(file,".assoc",sep=""),head=T))
score=cbind(res[,c(2,4,5)],-1,res[,6])
colnames(score)[4:5]=c("Centre","Effect")
write.table(score,paste(file,".score",sep=""),row=F,quote=F)
}

> resave("lin.1");resave("lin.01")
> resave("lin.001");resave("lin.0001")
```

# 1 - Polygenic Risk Scores

So now we can use `--calc-scores` to use this model to predict the phenotype for the test individuals in `doug.test`

```
../ldak.out --calc-scores lin.1 --scorefile lin.1.score \
--bfile doug --keep doug.test --pheno doug.pheno

../ldak.out --calc-scores lin.01 --scorefile lin.01.score \
--bfile doug --keep doug.test --pheno doug.pheno

../ldak.out --calc-scores lin.001 --scorefile lin.001.score \
--bfile doug --keep doug.test --pheno doug.pheno

../ldak.out --calc-scores lin.0001 --scorefile lin.0001.score \
--bfile doug --keep doug.test --pheno doug.pheno
```

# 1 - Polygenic Risk Scores

Now we can see how well it performed, which we will measure as correlation between true phenotype (column 3, below) and predicted phenotype (column 5, below)

```
> profile=as.matrix(read.table("linall.profile",head=T))
> cor(profile[,3],profile[,5])

> profile1=as.matrix(read.table("lin.1.profile",head=T))
> cor(profile1[,3],profile1[,5])

> profile01=as.matrix(read.table("lin.01.profile",head=T))
> cor(profile01[,3],profile01[,5])

> profile001=as.matrix(read.table("lin.001.profile",head=T))
> cor(profile001[,3],profile001[,5])

> profile0001=as.matrix(read.table("lin.0001.profile",head=T))
> cor(profile0001[,3],profile0001[,5])
```

# 2 - BLUP

BLUP (best linear unbiased prediction) is a direct extension of heritability analysis; heritability analysis stops once the variance components are computed, whereas BLUP continues one step further to estimate SNP effect sizes

We start by computing a kinship matrix and estimating heritability

```
../ldak.out --calc-kins-direct kins --bfile doug \
--ignore-weights YES --keep doug.keep
```

```
../ldak.out --reml blup --grm kins --pheno doug.pheno
```

# 1 - BLUP

```
$ head blup.reml
Num_Kinships 1
Num_Regions 0
Num_Covars 1
Blupfile blup.indi.blp
Regfile none
Fixfile blup.fixed
Total_Samples 620
With_Phenotypes 620
Null_Likelihood -2550.575277
Alt_Likelihood -2535.707299
LRT_Stat 29.7360
LRT_P 2.4754e-08
Component Heritability Her_SD Size Intensity Int_SD
Her_K1 0.337863 0.090975 10000.00 3.378626 0.909749
Her_All 0.337863 0.090975 10000.00 3.378626 0.909749
```

# 2 - BLUP

```
$ head blup.indi.blp
56 56 1.343310 -3.273761
57 57 -4.358053 -8.025177
58 58 3.995647 9.484571
59 59 -9.203814 -9.447628
61 61 9.054783 4.287477
63 63 1.923466 -3.302493
64 64 6.027758 -1.252290
66 66 7.850807 9.908565
68 68 -5.591179 2.481732
69 69 10.920491 11.717133
```

# 2 - BLUP

Next we compute the BLUP SNP loadings, for which we use `--calc-blups <output>` (where `<output>` is the stem we provide for the output files). Give this a try

```
../ldak.out --calc-blups blup
LDAK - Software for obtaining Linkage Disequilibrium Adjusted Kinship estimates and
Help pages at http://dougspeed.com/ldak
Arguments:
--calc-blups blup

Error, it is necessary to provide prefix for datafiles using one from "--bfile", "-

../ldak.out --calc-blups blup --bfile doug --remlfile blup.reml
LDAK - Software for obtaining Linkage Disequilibrium Adjusted Kinship estimates and
Help pages at http://dougspeed.com/ldak
Arguments:
--calc-blups blup
--bfile doug
--remlfile blup.reml

Error, to calculate blups it is necessary to provide the kinship
stems used in the REML using either "--grm" or "--mgrm"
```

# 2 - BLUP

Next we compute the BLUP SNP loadings

```
../ldak.out --calc-blups blup --bfile doug \
--remlfile blup.reml --grm kins

LDAK - Software for obtaining Linkage Disequilibrium Adjusted Kinship estimates and
Help pages at http://dougspeed.com/ldak
Arguments:
--calc-blups blup
--bfile doug
--remlfile blup.reml
--grm kins

Original number of samples: 980 --- Number being used: 980
Original number of predictors: 10000 --- Number being used: 10000
 ---  ---  ---  ---  ---  ---  ---  ---  ---  ---  ---

Calculating BLUP effects based on 1 kinship matrices and 0 regions
Found 10000 of the 10000 predictors in kins.grm.details
Calculating BLUPs for Chunk 1 of 2
Calculating BLUPs for Chunk 2 of 2
Effect sizes saved to blup.blup (and blup.blup.full) and predictions
to blup.pred (and blup.pred.full)
Mission completed. All your base are belong to us :)
```

# 2 - BLUP

Next we compute the BLUP SNP loadings

```
../ldak.out --calc-blups blup --bfile doug \
--remlfile blup.reml --grm kins
```

```
$ head blup.blup
Predictor A1 A2 Centre Effect
SNP1 A B 1.883871 -0.0106669510
SNP2 A B 1.959677 0.0260730230
SNP3 A B 0.027419 0.0153178848
SNP4 A B 0.440323 -0.0101744587
SNP5 A B 1.924194 0.0244202864
SNP6 A B 1.658065 0.0061845342
SNP7 A B 1.437097 0.0020977432
SNP8 A B 1.353226 -0.0097492864
SNP9 A B 1.908065 0.0397327332
```

# 2 - BLUP

So now we can use `--calc-scores` to use this model to predict the phenotype for the test individuals in `doug.test`

```
../ldak.out --calc-scores blup --scorefile blup.blup \
--bfile doug --keep doug.test --pheno doug.pheno
```

# 2 - BLUP

Now we can see how well it performed, which we will measure as correlation between true phenotype (column 3, below) and predicted phenotype (column 5, below)

```
> profile=as.matrix(read.table("blup.profile",head=T))
```

```
> cor(profile[,3],profile[,5])
```