# Friday morning session (MTG2) preparation

When you login the server 'hong1.une.edu.au',

```
cd <your direcotry>
cp ../honglee/For_participants.zip For_participants.zip
unzip For_participants.zip
cp -r ../honglee/example0/ example0
cp -r ../honglee/example1/ example1
cp mtg2.0 example0/
cp mtg2.0 example1/
```

Now, it gets ready for the prac.

For the manual, example files and binary files, you can also download from https://sites.google.com/site/honglee0707/mtg2

Summary
In the first session on Friday (9 - 10.30 am), I will introduce 'mtg2.0' software that is an extended version of GCTA including multivariate LMM and random regression, speeding up > 1000 times in some circumstance. I will go through how it has been used for complex traits analyses and some practical with 'mtg2.0'.

# Practical

S. Hong Lee

(Feb/16)

# Preparation

- Individual laptop
  - Install R
  - Install library(MASS)
  - Excel

- Linux server (provided by the organiser)
  - MTG2

# Preparation

- Literature
  - Lee, SH and van der Werf, JHJ. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* accepted (2016)
  - Maier, R., et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder and major depression disorder. *The American Journal of Human Genetics* 96, 283-294 (2015)
  - Lee, S. H.; Van der Werf, J. H. J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genetics Selection Evolution* 38: 25-43

# Genetic variation (lecture page 3)

- vg_sim.r (simulating g and e)
  - What is phenotypic variance?
  - What is genetic variance?
  - What is the highest value for genetic value?
  - What is the minimum value for genetic value?
  - What is the expected $h^2$?
  - What is the observed $h^2$?

# Genetic correlation

- vg_sim2.r (simulating 2 traits)
  - What is phenotypic variance structure?
  - What is genetic variance structure?
  - What is the expected genetic correlation?
  - What is the observed genetic correlation?

# Heritability

In the previous practical, you used $h^2$=var(g)/var(y) or $r_g$=cov(g1,g2)/sqrt(var(g1)var(g2)

However, you cannot observe g in real life.

So, you need a family design.

$h^2$ = phenotypic correlation / sample correlation

# Heritability (lecture page 5, 6, 7)

sib_sim.r (simulating 10000 sib pairs)

What is mean(yv[,1])?

What is mean(yv[,2])?

What is var(yv[,1])?

What is var(yv[,2])?

Make histogram (hist(yv[,1]) or hist(yv[,2]))

What does it look like?

# Heritability

sib_sim.r (simulating 10000 sib pairs)

What is phenotypic correlation?

If the pair is MZ twin, what is $h^2$?

If the pair is full sib, what is $h^2$?

If the pair is half sib, what is $h^2$?

# Heritability

sib_sim.r (simulating 10000 sib pairs)
In the code,
Change $h^2$=0.8, and run it.

What is phenotypic correlation?
If the pair is MZ twin, what is $h^2$?
If the pair is full sib, what is $h^2$?
If the pair is half sib, what is $h^2$?

Change $h^2$=0.3.

# Heritability

sib_sim.r (simulating 10000 sib pairs)
In the code,
Change sco=0.25, and run it.

What is phenotypic correlation?
If the pair is MZ twin, what is $h^2$?
If the pair is full sib, what is $h^2$?
If the pair is half sib, what is $h^2$?

Change sco=0.5.

# Heritability

sib_sim.r (simulating 10000 sib pairs)
In the code,

a. Change tn=100, and run it.
b. Change tn=10000, and run it

Can you find what is different between estimates from a and b?
Repeat 10 times for a
What is mean and variance of the estimates from these 10 replicates?

Repeat 10 times for b
What is mean and variance of the estimates from these 10 replicates?

# Heritability

Also, compare the confidence intervals with sqrt (your variance)

What can you conclude about this experiment?

# Heritability

Plot(yv[,1],yv[,2])

Change h2, sco or tn, and then Plot(yv[,1],yv[,2])

How does the plot change?

# Heritability

sib_sim.r

In the code, can you guess what is sn?

What does ‘sn=5’ mean?

Can you estimate $h^2$ with sn=5?

# Heritability

What if each family has different number of members?

This is unbalanced design problem.

‘Linear mixed model’ can be used.

- Probability density function
- Likelihood

# Probability density function (lecture page 11)

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Can you draw a normal curve and what is probability density?

Can you write a R code for the function above?

Can you find probability density function (pdf) in R?

# Probability density function

pdf.r

Compare your code with pdf.r

Did you get the same pdf?

What is dnorm in R?

What is pnorm in R?

What is qnorm in R?

Can you explain these with a normal curve?

# Probability density function

Can you explain pdf?

Can you explain likelihood?

What is difference between pdf and likelihood?

# Grid search for maximum likelihood estimates (lecture page 12, 13)

Now, we will obtain maximum likelihood estimate for var(x) or var(y)

pdf_grid.r

Looking at the code, can you work out how pdf is used in the grid search?

# Grid search for maximum likelihood estimates (lecture page 12, 13)

pdf_grid.r

Run it, and turn off the simulation part (so that y variable is always the same)

With guess_sd=2, what is log likelihood?
With guess_sd=2.2, what is log likelihood?
…
With guess_sd=4.8, what is log likelihood?
With guess_sd=5, what is log likelihood?

Can you make a plot of the likelihood values for 2.0, 2.2, 2.4, … 4.6, 4.8 and 5.0?

What is guess_sd that makes the maximum likelihood value?

# Grid search for maximum likelihood estimates

pdf_grid.r

Try with different tmu, tsd, tn in the simulation.

Can you still find the maximum likelihood estimate?

Is the maximum likelihood estimate similar to the true value, i.e. tsd?

# Newton-Raphson method for maximum likelihood estimates (lecture page 14)

pdf_NR.r

Looking at the code, can you work out how pdf is used in the NR algorithm?

Can you find the first derivatives?

Can you find the second derivatives?

How is the updated value obtained?

# Newton-Raphson method for maximum likelihood estimates

pdf_NR.r

Turn off the simulation part (so that y variable is the same as in the grid search)

Run it.

Is the estimate from NR algorithm the same as that from the grid search?

Compare the likelihood too.

If different, which one has higher likelihood value?

If NR is higher, why?

- can you do pdf_grid.r with the estimated value from NR?

If grid search is higher, why?

# Newton-Raphson method for maximum likelihood estimates

What is your conclusion about the grid search and NR algorithm?

# Algorithm for REML (lecture page 19)

- uni_aireml.xlsm
- Can you see the difference between V-based and MME-based algorithm?
  - In the process?
  - In estimating solutions?
- When # variance components increases, which one is more efficient?

# Algorithm for REML

- aireml.r
- Have a look at the code.

- What kinds of input files are needed to run this code?
- Can you write model including variance covariance structure?
  (hint slide 23 in the presentation)
- Can you compare the code with the equations in the papers?

- Run it.
- How many iterations do you have?
- What is the maximum likelihood?
- What is the maximum likelihood estimate for h2?

# Algorithm for REML

- mtg2.0

  ./mtg2.0 -p {plink fam file name} -d {phenotype file name} –g {grm file name} -cc {class covariate file name} –qc {continuous covariate file name} -out {output file name} -sv {starting value file name} –mod {number of traits}

  Go to exapmple0/

  e.g. ./mtg2.0 -p toy.fam -d toy.phen -g toy.rtmx -mod 1

  Or ./mtg2.0 -fam toy.fam -pheno toy.phen -grm toy.rtmx -mod 1

Compare the result with that from aireml.r

Is the maximum likelihood the same?

Is the maximum likelihood estimate the same?

Is the the number of iteration the same?

Is the speed the same?

# Input files for MTG2

<fam file for -p>

The PLINK fam file is your *.fam file that used in estimating the grm.

<grm file for -g>

For the grm file, you should unzip the .gz file from GCTA, delete the third column.

Then, it looks like

```
1 1 0.999
2 1 0.011
2 2 1.031
3 1 0.02
…..
```

# Constructing GRM

- GCTA
- Plink2

- For *.gz file for grm, a slight modification is needed for the input file for MTG2

```
zcat test.grm.gz | awk {print $1,$2,$4}' > test.grm
```

- For binary GRM file (from GCTA), MTG2 can be used with –bg command

```
./mtg2.0 -p test.fam -d test.dat -bg test.grm.bin -cc test.cov -qc
test.pc -out test.out -mod 5
```

# Phenotype file for MTG2

<phenotype file for -d>

With 5 traits model, the columns have FID, IID, t1, t2, t3, t4 and t5 (phenotypes for trait 1 ~ 5). It looks like

```
1 1 0.02 0.71 -0.02 0.04 -0.62
1 2 0.12 0.31 -0.27 NA -0.35
2 1 0.22 0.25 -0.28 0.63 -0.15
……
```

Missing values should be coded as NA.

# Covariate files for MTG2

<files for –cc (class covariate) and –qc (continuous covariate)>

The FID and IID order for phenotype file, covariate files (cc, qc) should be the same.

Missing values should be coded as NA.

# Principal components using MTG2

In order to get eigenvalues and eigenvectors with the prefix “test.grm”, i.e. test.grm.eval and test.grm.evec, the following command can be used

```
./mtg2.0 -p test.fam -g test.grm -pca n (n is the number of individual).
```

# MTG2 extra options

-cove 1: parameterising residual covariance

-thread k: k paralleled computation

-sv {file name}

-mg {file name} instead of -g {file name}: multiple random effects model

-bv {file name}: BLUP estimation

-inv 1: inverting matrix

-bend 1: bending NPD matrix making it PD

-nit {value}: maximum number of iterations (default is 200)

-conv {value}: convergence criteria for log likelihood (default is 0.001)

-frq 1: estimating allele frequency given plink files

# Multivariate linear mixed model (5 traits model)

- Go to example1/

```
./mtg2.0 -p example1.fam -d example1.dat -g example1.grm -mod 5 (without residual covariance)
```

```
./mtg2.0 -p example1.fam -d example1.dat -g example1.grm -mod 5 -cove 1 (with residual covariance)
```

this will give estimated parameters in ascm.out (with -out <file_name>, the outputs will be in <file_name>)

# Multivariate linear mixed model (5 traits model) (lecture page 75)

- To speed up, using eigen-decomposition technique

```
./mtg2.0 -p example1.fam -g example1.grm -pca 1908 -thread 10
```

this will give example1.grm.eval and example1.grm.evec

Then, using -eig substantially reduce the computing time as,

```
./mtg2.0 -p example1.fam -d example1.dat -eig example1.grm -mod 5 -cove 1 -out ascm.out_eig
```

✓ Compare the speed and results with and without eigen-decomposition (i.e. ascm.out vs. ascm.out_eig)

# Multivariate linear mixed model (5 traits model)

- ✓ Can you also estimate breeding values (genetic profile scores)?

```
 ./mtg2.0 -p example1.fam -d example1.dat -eig
example1.grm -mod 5 -cove 1 -bv ascm.bv
```

# Random regression model (lecture page 66)

It needs a parameter file having # order for each random effect and value for each environment (see rrm.par3)

```
3 3            ! # order for the first and second random effects
0 15 30 75     ! environments variable (time at the measurements)
```

# Random regression model (lecture page 66)

```
./mtg2.0 -p example1.fam -d example1.dat -mg joint.rtmx -rrm
rrm.par3 -mod 4 -out rrm.out
```

Joint.rtmx is the file specifying genetic covariance matrices for the random effects, e.g.

example1.idm

example1.grm

if one gives example1.idm as n x n identity matrix, residual covariances can be modelled.

With –eig, this can be done using with –rrme 1.

```
./mtg2.0 -p example1.fam -d example1.dat -eig example1.rtmx -rrm
rrm.par3 -rrme 1 -mod 4 -out rrm.out_eig
```

✓ Compare the speed and results with and without eigen-decomposition (i.e. rrm.out vs. rrm.out_eig)

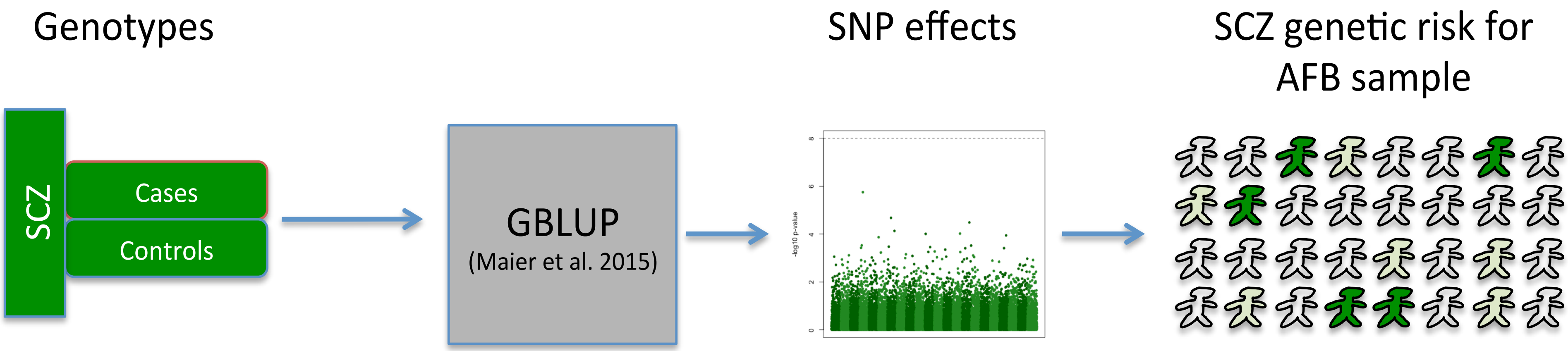
# Random regression model (lecture page 69)

To find the best model, change # order in the model (i.e. rrm.par3) and run it.

```
1 1              ! # order for the first and second random effects
0 15 30 75       ! environments variable (time at the measurements)

1 2              ! # order for the first and second random effects
0 15 30 75       ! environments variable (time at the measurements)

1 3              ! # order for the first and second random effects
0 15 30 75       ! environments variable (time at the measurements)
…

3 4              ! # order for the first and second random effects
0 15 30 75       ! environments variable (time at the measurements)
```

- ✓ Can you make a table like that in lecture note page 69 with the maximum likelihood values?
- ✓ Can you make a table like that in lecture note page 69 with BIC values?

# Converting individual BLUP to SNP BLUP or the other way around (lecture page 58)



Genotypes

SCZ

Cases

Controls

GBLUP

(Maier et al. 2015)

SNP effects

SCZ genetic risk for AFB sample

Diagram from Robert Maier

- Individual GBLUP in discovery -> SNP BLUP in discovery -> GBLUP in validation

# Equivalent model

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{Z}_i\mathbf{g}_i + \mathbf{e}_i$$

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i + \mathbf{W}_i\mathbf{u}_i + \mathbf{e}_i$$

- Equivalence between individual GBLUP and SNP BLUP (Hayes et al. 2011; Maier et al. 2015)
  - The relationship between **g** and **u**

# Equivalent model (Maier et al. 2015)

$$\begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} = \begin{bmatrix} \sigma^2_{g_1} & \cdots & \sigma_{g_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma_{g_{n1}} & \cdots & \sigma^2_{g_n} \end{bmatrix} \otimes \mathbf{A} \cdot \begin{bmatrix} \mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_n \end{bmatrix}' \cdot \mathbf{V}^{-1} \begin{bmatrix} \mathbf{y}_1 - \mathbf{X}_1 b_1 \\ \vdots \\ \mathbf{y}_n - \mathbf{X}_n b_n \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} \mathbf{ZA}\sigma^2_{g_1}\mathbf{Z}'+\mathbf{I}\sigma^2_{e_1} & \dots & \mathbf{ZA}\sigma_{g_{1n}}\mathbf{Z}'+\mathbf{I}\sigma_{e_{1n}} \\ \vdots & \ddots & \vdots \\ \mathbf{ZA}\sigma_{g_{n1}}\mathbf{Z}'+\mathbf{I}\sigma_{e_{n1}} & \cdots & \mathbf{ZA}\sigma^2_{g_n}\mathbf{Z}'+\mathbf{I}\sigma^2_{e_n} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \sigma^2_{u_1} & \cdots & \sigma_{u_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma_{u_{n1}} & \cdots & \sigma^2_{u_n} \end{bmatrix} \otimes \mathbf{I} \cdot \begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_n \end{bmatrix}' \cdot \Omega^{-1} \begin{bmatrix} \mathbf{y}_1 - \mathbf{X}_1 b_1 \\ \vdots \\ \mathbf{y}_n - \mathbf{X}_n b_n \end{bmatrix} \qquad \Omega = \begin{bmatrix} \mathbf{WI}\sigma^2_{u_1}\mathbf{W}'+\mathbf{I}\sigma^2_{e_1} & \dots & \mathbf{WI}\sigma_{u_{1n}}\mathbf{W}'+\mathbf{I}\sigma_{e_{1n}} \\ \vdots & \ddots & \vdots \\ \mathbf{WI}\sigma_{u_{n1}}\mathbf{W}'+\mathbf{I}\sigma_{e_{n1}} & \cdots & \mathbf{WI}\sigma^2_{g_n}\mathbf{W}'+\mathbf{I}\sigma^2_{e_n} \end{bmatrix}$$

Replacing **y** with **g**

$$\Omega = \begin{bmatrix} \sigma^2_{u_1} & \cdots & \sigma_{u_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma_{u_{n1}} & \cdots & \sigma^2_{u_n} \end{bmatrix} \otimes \mathbf{WW}' = \begin{bmatrix} \sigma^2_{u_1} & \cdots & \sigma_{u_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma_{u_{n1}} & \cdots & \sigma^2_{u_n} \end{bmatrix} \otimes \mathbf{A} \cdot M$$

$$\begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_n \end{bmatrix}' \otimes \mathbf{A}^{-1} \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} \cdot M^{-1} = \begin{bmatrix} \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_n \end{bmatrix}' \cdot \boxed{\begin{bmatrix} \sigma^2_{g_1} & \cdots & \sigma_{g_{1n}} \\ \vdots & \ddots & \vdots \\ \sigma_{g_{n1}} & \cdots & \sigma^2_{g_n} \end{bmatrix} \otimes \mathbf{I} \cdot \begin{bmatrix} \mathbf{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{Z}_n \end{bmatrix}' \cdot \mathbf{V}^{-1} \begin{bmatrix} \mathbf{y}_1 - \mathbf{X}_1 b_1 \\ \vdots \\ \mathbf{y}_n - \mathbf{X}_n b_n \end{bmatrix} M^{-1}}$$

This part is *.py output file with -bv

# Converting individual BLUP to SNP BLUP or the other way around (lecture page 58)

- Preparation of input (allele frequency file)

To estimate mean(x) and var(x)
(x is 0,1,2 SNP coefficient)

```
./mtg2.0 -plink toy -frq 1
```

NOTE: when scaling x, we use var(x) rather than 2p(1-p)
(p is RAF)

# Converting individual BLUP to SNP BLUP or the other way around

Go to example0/

<2 traits model>

- To get variance components (-out) and breeding values (-bv),

```
./mtg2.0 -p toy.fam -d toy.phen -g toy.rtmx -out toy.2t.ascm -mod 2 -bv toy.2t.bv -cove 1
```

- To get SNP BLUP for the first trait,

```
awk '$1==1 {print $2}' toy.2t.bv.py > tmp1
./mtg2.0 -plink toy -sbv a -vgpy tmp1 -out toy.2t.snpv_for_t1
```

- To get individual BLUP for the first trait in the data 'toy' (looking for toy.bed, toy.bim and toy.fam),

```
./mtg2.0 -plink toy -vgpy toy.2t.snpv_for_t1 -sbv b -out toy.2t.gbv_for_t1
```

✓ Can you get SNP BLUP for the second trait?

✓ Can you get individual BLUP for the second trait in the data ‘toy’?

✓ Can you compare the output files with toy.2t.bv (to get correlation)?

# Converting individual BLUP to SNP BLUP or the other way around

Here, the toy dataset has been used for both conversion, so the BLUP for the first trait in toy.2t.bv and toy.2t.gbv_for_t1 are identical. It is usually recommended to use an independent validation dataset by replacing the second argument, e.g.

```
./mtg2.0 -plink your_plink_file -vgpy toy.2t.snpv_for_t1 -sbv b -out your_plink_file.gbv_for_t1
```

NOTE: your_plink_file (target data set) has the same set of SNPs as in the discovery set.

# Converting individual BLUP to SNP BLUP or the other way around

Can you do individual-SNP BLUP with three trait model?

Can you do this with independent validation data set?

- E.g. lecture page 60