MiX99

Solving Large Mixed Model Equations

MTT

# Genomic Models and MiX99

Ismo Strandén, MTT

# Talk outline

- Single-step in MiX99
  - CLIM
  - Directive file
- Making matrices for single-step
  - RelaX2
  - Hginv
- Bayesian models and abc_mix/abc_multi

# Genomic modeling

Commonly Genomic Modeling by two alternative paths:

## Multistep - evaluations

Three steps:

1. Genetic evaluations to get **EBV** or Daughter yield deviations (DYD)
   (or EBV are converted to deregressed proof **DRP**)
2. DYD or DRP are used to solve direct genomic values **DGV**
3. DGVs are combined with EBV to form Genomic Enhanced Breeding values **GEBV**

## Single step evaluations

Evaluations include directly all animals:

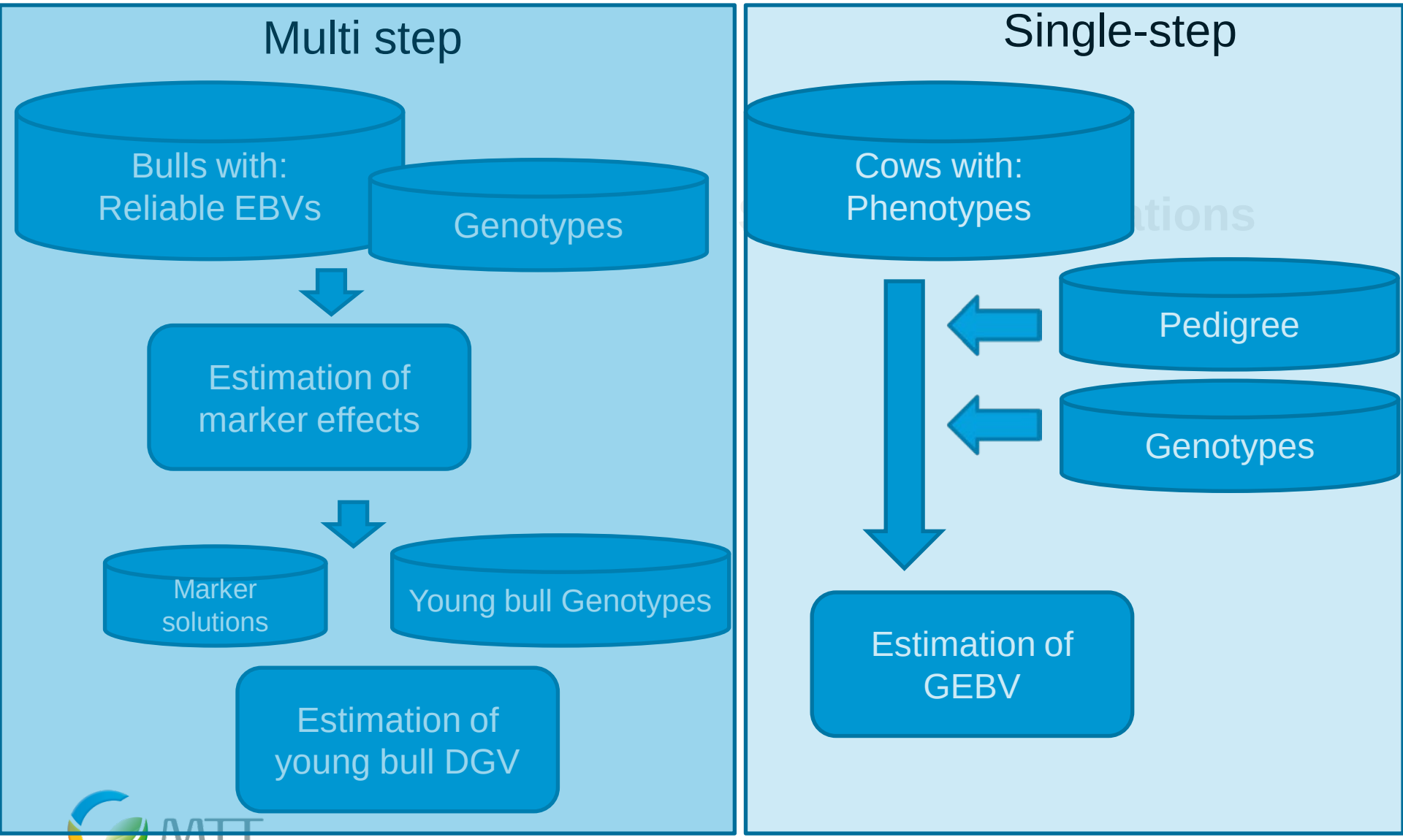
- Animals with genomic information
- Animals without genomic information

- Genomic information is used:
  - directly in the model for phenotypic observations

  Or EBV of animals with records are converted to DRP and DRP are used as observations

- Yields directly the **GEBV**

# Genomic modeling



# Single trait single-step

Model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}$, assume: $\text{Var}(\mathbf{e}) = \mathbf{I}\sigma_e^2$

$\text{Var}(\mathbf{a}) = \mathbf{H}\sigma_a^2$

Single trait ssGBLUP mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \lambda\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Order matrices to groups: 1= non-genotyped, 2= genotyped

Relationship matrix $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ and inverse: $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}$

We have $\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - (\mathbf{A}_{22})^{-1} \end{bmatrix}$

# Single-step in MiX99

- MiX99 needs
  - Pedigree
  - Matrix $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$

In order to make

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \left(\mathbf{A}_{22}\right)^{-1} \end{bmatrix}$$

Hence,

there is NO NEED to make complete $\mathbf{H}^{-1}$ for MiX99.

One approach is to use program: `hginv`

which needs $\mathbf{A}_{22}$

which can be made `RelaX2`

# Prerequisites for single-step

- High quality genotypes

  Typical step before genomic evaluation is extensive editing of genotypes: MAF, call rates, clones, high linkage
  - MiX99 uses data as -- animals in rows, SNPs in columns –

- Phenotype data
  - MiX99 data form as usual:
  - Columns:

    Fixed effects [random effects] observations [weights]

- Model parameters
- (co)Variances for markers – (co)variances for residuals

- Pedigree for all animals
- $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$

Walk through example

# Single-step

Terminal

File Edit View Search Terminal Help

```
esa@njok109k ~/R/mix99wrkshop $ ls -lrt bulls345.ped snp_genot345.dat phenotypes345.dat
-rw-r--r-- 1 esa esa    95364 loka  13  2009 bulls345.ped
-rw-r--r-- 1 esa esa 24548130 marra 20 16:05 snp_genot345.dat
-rw-r--r-- 1 esa esa    38439 marra 20 17:02 phenotypes345.dat
esa@njok109k ~/R/mix99wrkshop $ awk '{ print NF } ' snp_genot345.dat | head -1
35574
esa@njok109k ~/R/mix99wrkshop $ awk '{ NF=45; print $0 } ' snp_genot345.dat | head
3428378 1 1 1 1 1 0 1 2 1 1 1 1 0 0 2 0 2 2 0 2 2 2 1 2 0 1 0 0 0 2 0 0 0 2 0 1 2 1 2 1 2 2 1 2
3465276 1 1 2 2 1 1 1 2 2 0 2 2 2 0 0 0 2 2 0 2 2 2 1 2 0 2 0 0 0 2 0 0 0 2 0 1 2 1 2 0 2 2 1 1
3705482 0 1 2 2 1 0 1 2 2 2 0 0 0 0 2 0 2 2 0 2 2 2 2 2 1 2 1 1 0 2 0 0 0 2 0 1 1 2 1 1 2 2 2 2
3708802 2 2 2 2 1 0 1 2 1 0 2 2 1 1 1 1 1 1 1 1 1 2 1 0 1 1 1 0 2 0 0 0 2 0 1 2 1 2 0 2 2 0 1
3708853 2 2 2 2 0 1 0 2 1 0 2 2 1 0 1 0 2 2 1 2 2 1 2 2 0 2 0 0 1 2 0 0 0 2 0 0 2 2 2 0 1 1 2 2
3708863 2 2 2 2 1 1 1 2 1 0 2 2 1 0 1 0 2 2 0 2 2 2 1 2 0 2 0 0 1 2 0 0 0 2 0 1 2 2 2 0 1 1 2 1
3708868 2 2 2 2 2 2 2 1 1 0 2 2 2 0 1 0 1 2 0 2 2 2 1 2 0 2 0 0 0 2 0 0 0 2 0 2 2 2 2 0 2 2 1 1
3708869 1 2 2 1 2 0 2 2 2 1 1 1 0 0 2 0 2 2 0 2 2 2 2 2 1 1 1 0 0 2 0 0 0 2 0 2 1 2 1 1 2 2 2 2
3708874 1 2 2 2 1 1 1 2 1 0 2 2 1 0 1 0 2 2 0 2 2 2 2 2 0 2 0 0 0 2 0 0 0 2 0 1 2 1 2 0 2 2 0 2
3708877 1 2 2 2 0 0 0 2 1 0 2 2 0 0 2 0 2 2 0 2 2 2 2 2 0 2 0 0 1 2 0 0 0 2 0 0 2 1 2 0 1 1 2 1
esa@njok109k ~/R/mix99wrkshop $ head pheno
3428378 39994 1994 1 101 98 99 99 99 99 101 98 99
3465276 39933 1994 1 86 90 94 39 43 43 86 90 94
3705482 40764 1997 1 103 96 101 7 7 7 103 95 101
3708802 40158 1995 1 94 86 81 29 32 32 94 86 81
3708853 39775 1993 1 86 85 83 29 32 32 86 85 83
3708863 39602 1992 1 95 81 82 39 43 43 95 81 82
3708868 39777
3708869 39600
3708874 39641
3708877 39462
```

# Single step needs $\mathbf{G^{-1}} - \mathbf{A_{22}^{-1}}$

ID Herdbook_Nbr Birth_year One

milkEBV protEBV fatEBV

milkEDC protEDC fatEDC milkDRP protDRP fatDRP

# Making $A_{22}$ using RelaX2: the 345 data

```
input pedigree                  # Pedigree input
  file prunenol.ped             # Use this file as pedigree file
  record id sire dam            # Th

input animals                   # Relationship for this animals
  file snp_geno_ids.dat         # in this file
  record id                     # id variable locations

output overwrite amatrix    genot.amat  # output pedigree
```

$A_{22}$ matrix between genotyped animals

```
herakles:~/genomics/one_step/345genotyped> head prunenol.ped
      2330 4895618       0
      2341 4899165    2330
      3988 4893380       0
      7603 4897202       0
      7610 4895613    7603
      7621 4899518    7610
      7635 4899683    7621
                  herakles:~/genomics/one_step/345genotyped> head snp_geno_ids.dat
                  3428378
                  3465276
                  3705482
                  3708802
                  3708853
                  3708863
```

# Single step by MiX99: hginv

Making the matrix: $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$

```
hginv_lapack_seq -m ost -w .1 -A genot.amat snp_geno345.dat iH_345.dat
```

-m ost= single-step matrix

-w .1 = 10% polygenic or $\mathbf{G}_w=(1-w)\mathbf{G} + w\,\mathbf{A}_{22}$

**CLIM**

```
TITLE     345 data

DATAFILE  phenotypes345.dat
INTEGER   animal hrdbookcode birthyr muu
REAL      milkEBV protEBV fatENB EDCm EDCp EDCf milkDRP proDRP fatDRP

MISSING  -9999.0
DATASORT PEDIGREECODE=animal

iGFILE    iH_345.dat
PEDFILE   prunenol.ped
PEDIGREE  animal am
PARFILE   gblup.var

MODEL
  milkDRP = muu animal ! WEIGHT= EDCm
```

GBLUP:

```
PEDFILE  /home/esa/R/mix99wrkshop/Gmat/bulls345.ginv
PEDIGREE  animal FILE     # Genetics associated with animal
```

Single-step: Normal pedigree file & file for $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$

# Solving single-step by MiX99

```
herakles:~/genomics/one_step/345_single_step> time mix99i single_step_345.clm > mix99i.log

real    0m0.160s
user    0m0.136s
sys     0m0.024s
herakles:~/genomics/one_step/345_single_step> time mix99s -s >mix99s.log

real    0m0.225s
user    0m0.198s
sys     0m0.027s
herakles:~/genomics/one_step/345_single_step> head Solani
       2330       1       0   0.13605
       2341       1       0   -2.3266
       3988       2       0   0.74216
       7603       1       0    1.0302
       7610       1       0   0.25884
       7621       1       0   -3.2088
       7635       1       0   -1.7215
       9733       1       0   -2.0168
    5067934       1       0    4.2366
      12528       1       0   0.19988E-01
herakles:~/genomics/one_step/345_single_step> head Solfix
 Fact. Trt       Level     N-Obs   Solution               Factor Trait
   1   1           1         345    97.249                    muu milkDRP
```

# MiX99: Directive file and single-step

```
# RELATIONSHIPS
 fg  # external file (lower triangle) and relationship
iH_345.dat
# Standard pedigree based relationship matrix
    1   1
# REGRESS
   2 cl cl
# COMBINE
n
# PEDIGREE
am
# DATAFILE  phenotypes345.dat
phenotypes345.dat
# VAR
    4    9 f
# MISSVA
 -9999
# SCALE
n
# PEDFILE    prunenol.ped
prunenol.ped
```

Standard single-step

Single step

i= inverse matrices given

Giving separately $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$

```
# RELATIONSHIPS
 ss1  # external file (lower triangle) and relationship matrix
../data_make/iG.dat
../data_make/iA.dat
# Standard pedigree based relationship matrix
    1   1
```

```
hginv_lapack_seq -m PvR1 -w .1 -A genot.amat snp_geno345.dat iG.dat
hginv_lapack_seq -A genot.amat - iA.dat
```

# Directive file and single-step

```
# RELATIONSHIPS
 fg  # external file
iH_345.dat
# Standard pedigree
    1   1
```

Standard single-step

```
# RELATIONSHIPS
 ssi  # external file (lower triangle) and relationship matrix
../data_make/iG.dat
../data_make/iA.dat
# Standard pedigree based relationship matrix
    1   1
```

Single step

i= inverse matrices given

Giving separately **G** and **A;** no inverses

```
# RELATIONSHIPS
ssM  # external file
G.dat
genot.amat
```

M= mixed format (both upper & lower triangle possible)

```
hginv_lapack_seq -G G.dat -m PvR1 -w .1 -A genot.amat snp_geno345.dat -
```

Correlations 1.00000

```
paste Solani_ss Solani_iG_iA Solani_G_A
2330      1      0  0.13605          2330      1      0  0.13566          2330      1      0  0.13493
2341      1      0  -2.3266          2341      1      0  -2.3243          2341      1      0  -2.3233
3988      2      0  0.74216          3988      2      0  0.74173          3988      2      0  0.74245
7603      1      0   1.0302          7603      1      0   1.0262          7603      1      0   1.0292
7610      1      0  0.25884          7610      1      0  0.25785          7610      1      0  0.25897
7621      1      0  -3.2088          7621      1      0  -3.2070          7621      1      0  -3.2025
7635      1      0  -1.7215          7635      1      0  -1.7219          7635      1      0  -1.7185
9733      1      0  -2.0168          9733      1      0  -2.0228          9733      1      0  -2.0247
67934     1      0   4.2366       5067934      1      0   4.2191       5067934      1      0   4.2185
```

# hginv program:

## 1 processor & multiple core versions

- Uses LAPACK subroutines in making **G** and inversion
- Methods:

```
-m method : Genomic matrix, method after -m is
    raw : use genotype data as such.
    101 : 101 coding (-1,0,1), assuming original is 012 coding.
 center : center coding, i.e. PvR1 without scaling by 2*sum(p*q), singular.
    avg : center around average allele frequency.
   PvR1 : P.VanRaden (JDS 2008) method 1, singular if allele frequencies from data (default).
  PvR1m : PvR1 with diag(G) multiplied by  1.0010
  PvR1a : PvR1 with diag(G) increased by 0.10000E-01
   PvR2 : P.VanRaden method 2, singular if allele frequencies from data.
  PvR2m : PvR2 with diag(G) multiplied by  1.0010
   poly : blending G(PvR1) and A: G_w = (1-w)*G+w*A, see option -w
    edm : Euclidean distance norm matrix, see option -theta.
    ost : inv(H)=inv(G)-inv(A_22) in single-step, G like poly but make inv(H) matrix, or -m poly -ss.
   ost2 : like ost but use PvR2 instead of PvR1, or -m PvR2 -ss.


 -ss          : inv(H)=inv(G)-inv(A_22) in single-step using given G matrix by option -m.
 -theta value : Euclidean distance norm divider in exp(-C/theta)
 -singularity method : singularity prevention by changing diagonal of G matrix
    mult : multiply diagonal of G by   1.0010
    add  : increase diagonal of G by  0.10000E-01
 -mult value   : change the diagonal multiplier in the singularity method mult
```

# More options in hginv

```
-b mthd  : balancing G matrix with A matrix, mthd after option -b is
     ole : O.Christensen method of making G: G_n= b*G+a*J, where a&b by Ole
    PvR3 : P.VanRaden method 3: G_n= a*J + b*A where a&b by LS: G= a*J+b*A+E
      ls : LS method: G_n= b*G +(1-b)*I +a*J where a&b by LS G=f(A,a,b)
     ls3 : P.VanRaden method 3 twisted: (G-I)= a*J + b*(A-I) + E
     ls4 : P.VanRaden method 3 modified: G= a*I + b*(J-I) + c*A + E
     vit : Method by Vitezica et al.

-dG mthd : diagonal of G replacement where mthd is
      dA : diagonal of given A matrix:  a_ii * trace(G)/trace(A)
      PV : P.Visscher diag(G).

-D file  : external file for D in G = ZDZ'; one diagonal value per line
-ignacy  : I.Misztal method for inv(H): inv(H)= 1.5*inv(G)-0.6*inv(A)
-icoeff  : -ignacy using given coefficients t&o: inv(H)= t*inv(G)-o*inv(A)

-i ival  : majority imputation of missing value ival in genotype file.
```

The program allows:

- making **G** and $\mathbf{G}^{-1}$ when given marker information
- making $\mathbf{A}_{22}^{-1}$ when given $\mathbf{A}_{22}$
- making $\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ when given marker data and $\mathbf{A}_{22}$

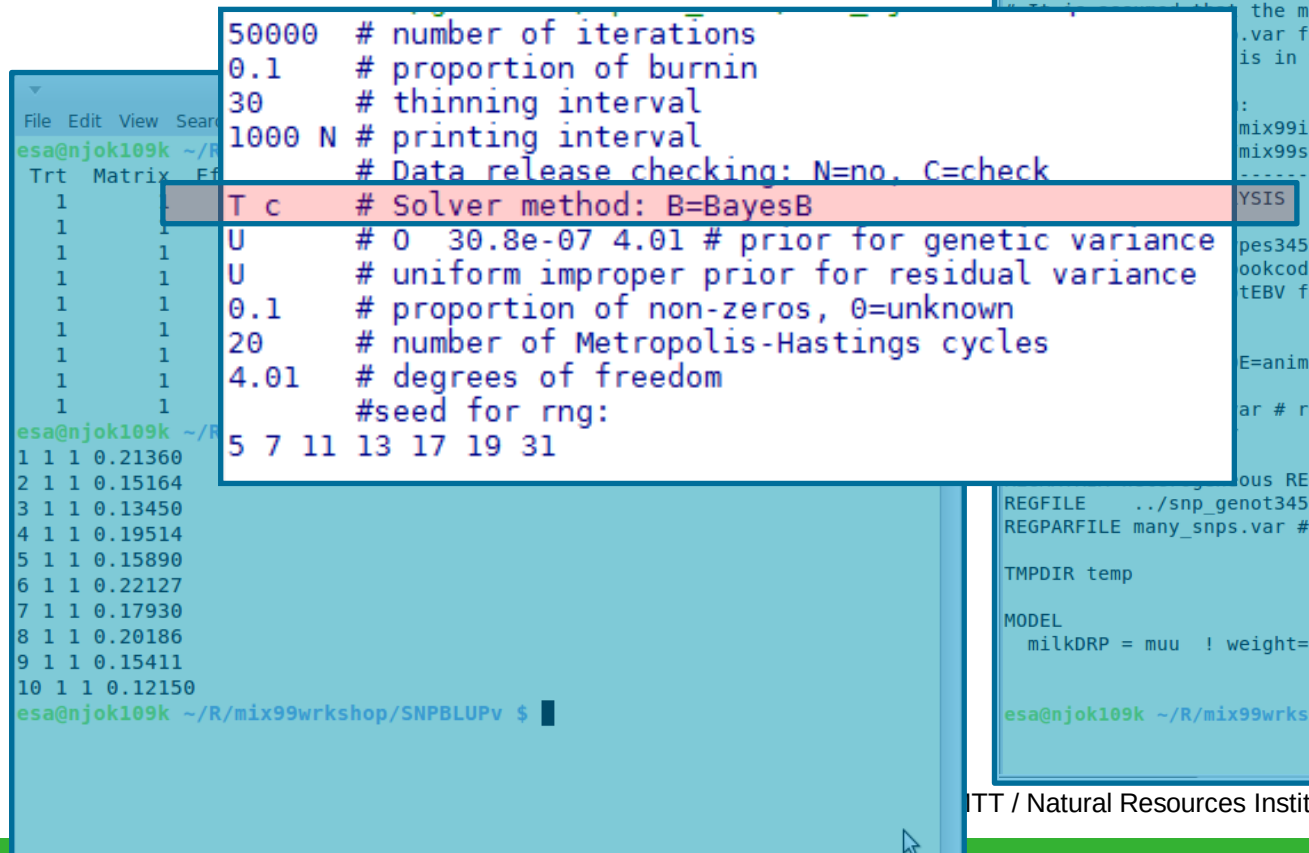
Flexibility in chosing method for genomic relationship matrix.

Genomic data and

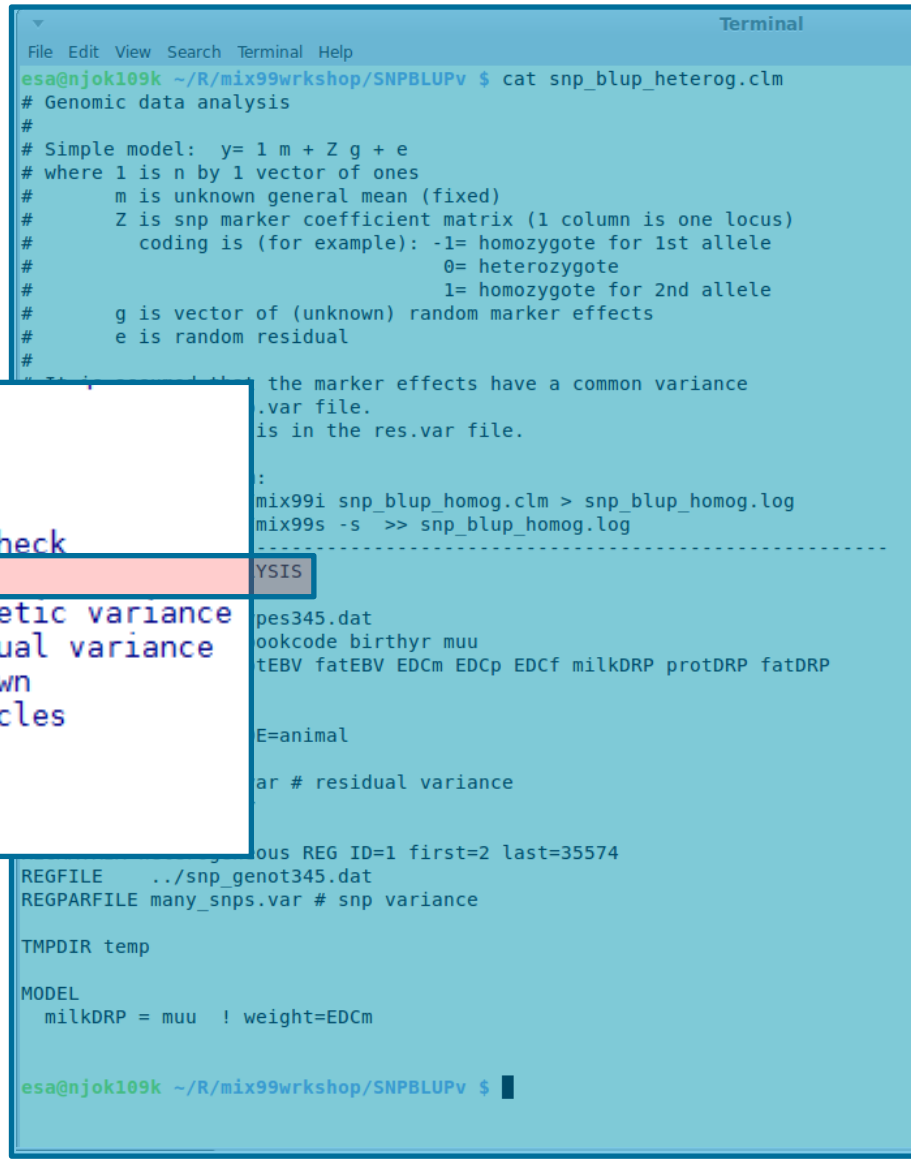
# Bayesian models

# Markov chain Monte Carlo for Bayesian genomic models

- Estimate the individual marker variances
  - eg. **abc_mix** <options.abc




```
50000  # number of iterations
0.1    # proportion of burnin
30     # thinning interval
1000 N # printing interval
       # Data release checking: N=no, C=check
T c    # Solver method: B=BayesB
U      # 0  30.8e-07 4.01 # prior for genetic variance
U      # uniform improper prior for residual variance
0.1    # proportion of non-zeros, 0=unknown
20     # number of Metropolis-Hastings cycles
4.01   # degrees of freedom
       #seed for rng:
5 7 11 13 17 19 31
```



# Program abc_mix/abc_multi

- Bayesian analysis using Markov chain Monte Carlo

- Model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Wg} + \mathbf{e}$,

where

often $\mathbf{X}=\mathbf{1} \rightarrow \mathbf{b}$ is general mean

vector $\mathbf{g}$ has marker effects.

SNP-BLUP when variance components known,

use mix99i/mix99s

# Models in abc_mix/abc_multi

**y** = **Xb** + **Wg** + **e**, where in general $g_j \sim N\left(0, \sigma^2_{g_j}\right), j = 1, \cdots, m$

SNP-BLUP: $g_j \sim N\left(0, \sigma^2_g\right)$ where all variances known: mix99s

BayesA: $\sigma^2_{g_j} \sim \chi^{-2}(\nu, S)$, i.e., $\sigma^2_{g_j} \sim S/\chi^2_\nu$ *S* and ν are known, residual variance estimated

OR: $g_j \sim t_\nu\left(0, \sigma^2_t\right)$

= *S*/ν is known

BayesB:

$$\sigma^2_{g_j} = \begin{cases} 0 & \text{with probability } \pi \\ \sim S/\chi^2_\nu & \text{with probability } 1-\pi \end{cases}$$

known: ν,π,*S*

OR:

$$g_j = \begin{cases} 0 & \text{with probability } \pi \\ \sim t_\nu\left(0, \sigma^2_t\right) & \text{with probability } 1-\pi \end{cases}$$

known: $\nu, \pi, \sigma^2_t$

# Some models in abc_mix/abc_multi

**y** = **Xb** + **Wg** + **e**, where in general $g_j \sim N\left(0, \sigma^2_{g_j}\right), j = 1, \cdots, m$

BayesA: $\sigma^2_{g_j} \sim \chi^{-2}(\nu, S)$, i.e., $\sigma^2_{g_j} \sim S/\chi^2_\nu$

S and ν are known, residual variance estimated

OR: $g_j \sim t_\nu\left(0, \sigma^2_t\right)$ known: $\nu, \sigma^2_t$

GtA: $g_j \sim t_\nu\left(0, \sigma^2_t\right)$ known: $\nu$; estimated $\sigma^2_t$

abc_mix/abc_multi allows estimation of π

BayesB:

$$\sigma^2_{g_j} = \begin{cases} 0 & \text{with probability } \pi \\ \sim S/\chi^2 & \text{with probability } 1-\pi \end{cases}$$

known: ν,π,S

Other models include:

Gaussian model for estimating marker and residual variance for SNP-BLUP

Stochastic variable selection (SSVS)

GtB:

$$g_j = \begin{cases} 0 & \text{with probability } \pi \\ \sim t_\nu\left(0, \sigma^2_t\right) & \text{with probability } 1-\pi \end{cases}$$

known: ν,π

estimated: $\sigma^2_t$

# First step: MiX99 preprocessor as usual

```
DATAFILE  /home/ejo13/genomics/April_2013/data/bull_FN_sel_train_ordered.dat

INTEGER   id mean
REAL      milk_ind Edc_milk prot_ind Edc_prot fat_ind Edc_fat drp_milk drp_prot drp_fat birth_year
MISSING   -99999
DATASORT  PEDIGREECODE=id

PARFILE   -

REGMATRIX  Fixed SNP id=1 first=2 last=36380
REGFILE   /share/data/genomic/june_2011/geno_impd_noeq.dat

MODEL
  drp_milk = mean
#  drp_milk = mean ! weight=Edc_milk
```

No PARFILE needed because given as fixed effect model!

# abc_mix/abc_multi: BayesA & GtA

BayesA:

```
300000                  # number of iterations
0.05                    # proportion of burnin
30                      # thinning interval
30000 N                  # printing interval
                         # Data release checking: N=no, C=check
A c p                    # Solver method: C=Common, A=BayesA, B=BayesB
/home/ejo13/genomics/April_2013/data/geno_impd_noeq_valid.dat # validation animals genotype file
/home/ejo13/genomics/April_2013/data/bull_FIN_valid_milk.drp  # validation animals phenotypes
F 0.0038687 4.01 # prior for genetic variance
U  # uniform improper prior for residual variance
# seed for rng
3 5 7 11 13 17 19 31
```

c= centering for computations

p= prediction (validation)

GtA:

```
300000                  # number of iterations
0.05                    # proportion of burnin
30                      # thinning interval
30000 N                  # printing interval
                         # Data release checking: N=no, C=check
G c p                    # Solver method: C=Common, A=BayesA, B=BayesB
/home/ejo13/genomics/April_2013/data/geno_impd_noeq_valid.dat # validation animals genotype file
/home/ejo13/genomics/April_2013/data/bull_FIN_valid_milk.drp  # validation animals phenotypes
0 0 -2 # uniform improper prior for genetic variance
U  # uniform improper prior for residual variance
4.01 # d.f. known
# seed for rng
3 5 7 11 13 17 19 31
```

# Some output for GtA

```
herakles:~/genomics/April_2013/abc_ajoa/milk_GtA4> cat ABC_MeanVarResidual_milk_c
         1      0   1462   13.761        5.5712
herakles:~/genomics/April_2013/abc_ajoa/milk_GtA4> cat ABC_MeanVarGenetic_milk_c
     Trait Reg-Mat N-markers Gen-number      Mean        Variance
         1       1     36379     36380   0.30453E-02   0.55272E-07 Genetic
herakles:~/genomics/April_2013/abc_ajoa/milk_GtA4> cat ABC_Meanf01_milk_c
         1      1462   156.31
herakles:~/genomics/April_2013/abc_ajoa/milk_GtA4> head ABC_MeanMarker_milk_c
 Trt  Matrix  Effect  Mean          Variance    Prob   Mat-Name
   1       1       1  0.72826E-04   0.5200E-02 0.49558 SNP
   1       1       2  0.13888E-01   0.5787E-02 0.56316 SNP
   1       1       3  0.12531E-01   0.5879E-02 0.56737 SNP
   1       1       4  0.19897E-02   0.5245E-02 0.51032 SNP
   1       1       5 -0.26464E-01   0.6782E-02 0.62421 SNP
   1       1       6 -0.14881E-01   0.5751E-02 0.57221 SNP
   1       1       7  0.19168E-02   0.4975E-02 0.51505 SNP
   1       1       8  0.12456E-01   0.5570E-02 0.55474 SNP
   1       1       9  0.38337E-01   0.8568E-02 0.67242 SNP
```

Residual variance

Genetic marker variance

General mean estimate

Marker effects

Program output:

```
Posterior means of some parameters:

mean Gt_gvar=                  0.0030458045 improper uniform prior
mean var(e) =                 13.7352176953
mean mu     =                156.3105877286
genetic df  =                  4.0100002289 fixed at this value

Number of samples in means:      285000
-----------------------------------
```

Gives also:
Mean squared error (MSE),
Validation reliability
Regression coefficient ($b_1$)

# Bottom line

- MiX99 allows several ways to calculate single-step BLUP
- RelaX2 and Hginv programs can be used to prepare data for single-step model
- abc_mix/abc_multi can be used in genomic Bayesian model analysis
  - Preprocessing by MiX99 preprocessor OR directly (abc_multi)